Feuille de TP n°10
Fonction de répartition empirique

FIXME: biblio précise !

1 Fonction de répartition empirique.

Définition 1.1. *Soit* $(X_1, \dots, X_n)$ *un échantillon de loi* μ *sur* $\mathbb{R}$*, de fonction de répartition* F*. On appelle* fonction de répartition empirique *associée à* $(X_1, \dots, X_n)$*, la fonction aléatoire* $F_n : \mathbb{R} \to [0,1]$ *définie pour tout* $x \in \mathbb{R}$ *par* $F_n(x) := \frac{1}{n}\#\{X_k \leqslant x; 1 \leqslant k \leqslant n\}$*. On peut également écrire de manière équivalente*

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{I}_{\{X_k \leqslant x\}}.$$

C'est la fonction de répartition de la *mesure empirique* $\mathbb{P}_n := \frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n})$ de l'échantillon : pour tout $x \in \mathbb{R}$, on a $\mathbb{P}_n(]-\infty, x]) = F_n(x)$. Pour tout $p \in]0,1[$, le *quantile empirique* d'ordre p de l'échantillon est par définition $X_{([np])}$ où $[np]$ est la partie entière de np et où $X_{(1)}, \dots, X_{(n)}$ sont les statistiques d'ordre de l'échantillon. On a $F_n(X_{([np])}) = \frac{1}{n}[np] \in [p, p + \frac{1}{n}[$. Ainsi, $X_{([np])}$ est bien le quantile d'ordre p de $\mathbb{P}_n$.

Théorème 1.2 (Convergence des quantiles empiriques). *Soit* $(X_1, \dots, X_n)$ *un échantillon de loi* μ *sur* $\mathbb{R}$*, de fonction de répartition* F*. Si* F *est continue, d'inverse généralisée* F^{-1}*, alors pour tout* $p \in]0,1[$*, en notant* $Q_n(p) := X_{([np])}$ *et* $Q(p) := F^{-1}(p)$*, on a*

$$Q_n(p) \xrightarrow[n\to+\infty]{p.s.} Q(p).$$

De plus, si F *est dérivable en* $Q(p)$ *de dérivée* $f(Q(p))$*, on a*

$$\sqrt{n}(Q_n(p) - Q(p)) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{p(1-p)}{f(Q(p))^2}\right).$$

Théorème 1.3 (Glivenko-Cantelli). *Soit* $(X_1, \dots, X_n)$ *un échantillon de loi* μ *sur* $\mathbb{R}$*, de fonction de répartition* F*. Pour tout* $x \in \mathbb{R}$*, on a* $F_n(x) \xrightarrow[n\to+\infty]{p.s.} F(x)$*, et cette convergence est uniforme sur* $\mathbb{R}$

$$\|F_n - F\|_\infty := \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to+\infty]{p.s.} 0.$$

De plus, si $X_{(1)}, \dots, X_{(n)}$ *désigne les statistiques d'ordre de* $X_1, \dots, X_n$*, on a*[1]

$$\|F_n - F\|_\infty = \max_{1\leqslant i\leqslant n}\left[\max\left(\left|\frac{i}{n} - F(X_{(i)})\right|, \left|\frac{i-1}{n} - F(X_{(i)})\right|\right)\right].$$

Théorème 1.4 (Kolmogorov-Smirnov). *Soit* $(X_1, \dots, X_n)$ *un échantillon de loi* μ *sur* $\mathbb{R}$*, de fonction de répartition* F*. Si* F *est continue, alors* $\sqrt{n}\,\|F_n - F\|_\infty \xrightarrow[n\to+\infty]{\mathcal{L}} \mu_{KS}$*, où la loi* μ_{KS} *est universelle et ne dépend pas de* F *en particulier. Elle est portée par* $\mathbb{R}_+$ *et a pour fonction de répartion pour* $t \geqslant 0$

$$F_{KS}(t) := \mu_{KS}(]-\infty, t]) = 1 + 2\sum_{k=1}^{\infty}(-1)^k \exp(-2k^2t^2).$$

[1] Cette formule permet de calculer $\|F_n - F\|_\infty$ et montre que sa loi ne dépend pas de la loi de l'échantillon (statistique libre). Comme F est croissante, $F(X_{(1)}), \dots, F(X_{(n)})$ est une statistique d'ordre de la loi uniforme.

Corollaire 1.5 (Test d'adéquation de Kolmogorov-Smirnov). *Soit $(X_1, \dots, X_n)$ un échantillon de loi inconnue μ sur $\mathbb{R}$, de fonction de répartition F continue. On note F_n la fonction de répartition empirique associée. Soit ν une loi connue de fonction de répartition G. Posons $\mathcal{H}_0$:=« $\mu = \nu$ » et $\mathcal{H}_1$:=« $\mu \neq \nu$ ». Soit $\alpha \in [0,1]$ et $k_{1-\alpha}$ le quantile $1-\alpha$ de la loi de Kolmogorov-Smirnov μ_{KS}. Le test qui consiste à rejeter $\mathcal{H}_0$ si $\sqrt{n}\,\|F_n - G\|_\infty > k_{1-\alpha}$ et à accepter $\mathcal{H}_0$ sinon est asymptotiquement de niveau α et sa puissance converge vers 1. Cf. [2, Chap. 15.4.2.B].*

Remarque 1.6 (Forme équivalente). Si F_{KS} est la fonction de répartition de la distribution de Kolmogorov-Smirnov μ_{KS}, alors le test $\sqrt{n}\,\|F_n - G\|_\infty > k_{1-\alpha}$ est équivalent au test $F_{\mathrm{KS}}(\sqrt{n}\,\|F_n - G\|_\infty) > 1-\alpha$. La fonction `pks` de Stixbox implémente F_{KS}.

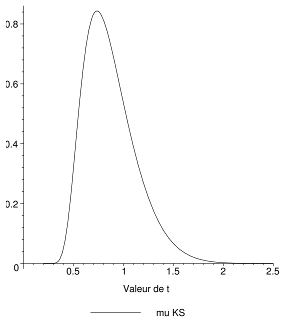

Fig. 1 – Densité de probabilité de la loi de Kolmogorov-Smirnov μ_{KS}.

Remarque 1.7 (Test d'adéquation de Cramer-von Mises). Soit $(X_1, \dots, X_n)$ un échantillon de loi μ sur $\mathbb{R}$, de fonction de répartition F. Soit F_n la fonction de répartition empirique. La statistique de Cramer-von Mises C_n est définie par

$$nC_n := n\int_{-\infty}^{+\infty} (F_n(x) - F(x))^2\, dF(x).$$

Cette statistique s'exprime simplement en terme des statistiques d'ordre de l'échantillon

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

Le thérème de Glivenko-Cantelli assure que $C_n \xrightarrow[n\to+\infty]{\text{p.s.}} 0$. D'autre part, lorsque F est continue, on peut montrer que $nC_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mu_{CM}$ où la loi μ_{CM} est universelle et ne dépend pas de F en particulier. On peut montrer que c'est la loi de $\sum_{k=1}^{+\infty} (\pi k)^{-2} Z_k$ où $(Z_k)_{k\in\mathbb{N}^*}$ est une suite i.i.d. de loi $\chi^2(1)$. On construit avec nC_n un test d'adéquation similaire au test de Kolmogorov-Smirnov. Cf. [2, Chap. 15.4.2.C].

2 Utilisation en modélisation.

Exercice 2.1 (Glivenko-Cantelli). Créer un code Matlab permettant d'illustrer le théorème de Glivenko-Cantelli sur un N-échantillon de loi binomiale $\mathcal{B}(n,p)$, de loi de Poisson $\mathcal{P}(\lambda)$, de loi exponentielle $\mathcal{E}(\lambda)$ et de loi normale $\mathcal{N}(m,\sigma^2)$ où les paramètres sont affectés par l'utilisateur.

Exercice 2.2 (Test de Kolmogorov-Smirnov pour l'adéquation à la loi normale). Créer un code Matlab permettant de générer, avec l'algorithme de Box-Muller ou l'algorithme polaire, un N-échantillon de loi normale $\mathcal{N}(m,\sigma^2)$ où N, m et σ^2 sont affectés par l'utilisateur. Effectuer ensuite un test de Kolmogorov-Smirnov d'adéquation à la loi normale $\mathcal{N}(m,\sigma^2)$ en utilisant la fonction Matlab `pks`. Essayer d'autres lois comme la loi uniforme $\mathcal{U}([0,1])$, la loi exponentielle $\mathcal{E}(\lambda)$ et la loi de Cauchy $\mathcal{C}(\lambda)$ avec $\lambda > 0$.

Exercice 2.3 (Test d'homogénéité de Kolmogorov-Smirnov). Soit $(X_1, \ldots, X_n)$ un n-échantillon de fonction de répartition F et soit $(Y_1, \ldots, Y_m)$ un m-échantillon de fonction de répartition G. On suppose que ces deux échantillons sont indépendants et que F et G sont continues. On veut tester H_0: « $F = G$ » contre H_1: « $F \neq G$ ». Soient F_n et G_m les fonctions de répartition empirique associées à $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$. Alors, sous H_0

$$\sqrt{\frac{nm}{n+m}} \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)| \xrightarrow[n \to +\infty]{\mathcal{L}} \mu_{\text{KS}}.$$

Effectuer un test d'homogénéité de Kolmogorov-Smirnov sur deux échantillons indépendants de loi uniforme $\mathcal{U}([0,1])$ et de tailles respectives $n = 100$ et $m = 1000$. Essayer d'autres lois.

Exercice 2.4 (Grosses boites). Les deux tableaux suivant représentent le revenu net en milliards d'Euros pour l'année 2002 de vingt groupes français et de vingt-quatre groupes allemands de l'industrie et des services.

Groupes Français

0.2	3.8	7.6	4.0	4.1	-2.8	4.7	3.6	5.4	-0.2
1.6	5.6	-0.6	0.8	-5.0	0.1	2.9	3.7	3.9	1.1

Groupes Allemands

1.8	4.0	1.4	1.9	1.9	1.8	1.4	1.9	1.4	4.5	2.2	2.4
3.1	0.3	-1.4	0.4	2.3	0.2	1.5	4.8	0.6	1.0	1.5	5.5

Effectuer un test d'homogénéité de Kolmogorov-Smirnov sur ces observations en utilisant la fonction `kstest2` de Matlab.

Exercice 2.5 (Estimation non paramétrique à noyau d'une densité). Soit $(X_1, \ldots, X_n)$ un échantillon de loi de densité de probabilité f. On suppose que $f \in \mathcal{C}^1(\mathbb{R})$ et que f' est bornée. Soit $K : \mathbb{R} \to \mathbb{R}_+$ une fonction bornée appelée noyau, telle que

$$\int_{\mathbb{R}} K(x)\, dx = 1 \quad \text{et} \quad \int_{\mathbb{R}} K^2(x)\, dx = \sigma^2.$$

On peut par exemple choisir le noyau uniforme $K(x) = (2a)^{-1}\,\mathrm{I}_{[-a,a]}(x)$ avec $a > 0$ ou encore le noyau gaussien $K(x) = (2\pi)^{-n/2} \exp(-x^2/2)$. On estime f par l'estimateur à noyau $\widehat{f}_n$ défini $\forall x \in \mathbb{R}$ par

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{X_i - x}{h_i}\right)$$

où $h_n := n^{-\alpha}$ avec $0 < \alpha < 1$. Montrer que $\widehat{f}_n(x) \xrightarrow[n \to +\infty]{\text{p.s.}} f(x)$ et que si $1/3 < \alpha < 1$,

$$\sqrt{nh_n}\left(\widehat{f}_n(x) - f(x)\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2 f(x)}{1+\alpha}\right).$$

Créer un code Matlab permettant d'illustrer cette méthode d'estimation de la densité par noyaux sur la loi normale $\mathcal{N}(m, \sigma^2)$ et sur la loi exponentielle $\mathcal{E}(\lambda)$, où les paramètres m, σ^2 et $\lambda > 0$ sont affectés par l'utilisateur.

Références

[1] D. Dacunha-Castelle et M. Duflo – *Probabilités et statistiques. Tome 1*, Masson, Paris, 1982, Problèmes à temps fixe.

[2] G. Saporta – *Probabilités, analyse des données et statistique*, Technip, 1989.

[3] P. Toulouse – *Thèmes de probabilités et statistique*, Dunod, 1999, INTERDIT À L'ORAL.