Feuille de TP n°10
Fonction de répartition empirique

1 Fonction de répartition empirique.

Définition 1.1. *Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de v.a.r i.i.d. de fonction de répartition F. On appelle* fonction de répartition empirique *associée à $(X_1,\dots,X_n)$, la fonction aléatoire F_n définie pour tout $x\in\mathbb{R}$ par*

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

Theorème 1.2 (Glivenko-Cantelli). *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 $\mathbb{R}$*

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

Définition 1.3 (Pont brownien). *On dit que $(P_t)_{0\leqslant t\leqslant 1}$ est un pont brownien si, pour tout $0\leqslant t\leqslant 1$, $P_t = B_t - tB_1$ où $(B_t)_{t\geqslant 0}$ est un mouvement brownien standard issu de zéro. Le pont brownien doit son nom au fait que chacune de ses trajectoires browniennes passe de 0 à l'instant 0 à 0 à l'instant 1 car $P_0 = P_1 = 0$.*

Theorème 1.4 (Kolmogorov-Smirnov). *Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de v.a.r. i.i.d. de fonction de répartition F est continue, alors*

$$\sqrt{n}\sup_{x\in\mathbb{R}}(F_n(x)-F(x)) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{L}(\sup_{0\leqslant t\leqslant 1} P_t) \quad et \quad \sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)| \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{L}(\sup_{0\leqslant t\leqslant 1}|P_t|),$$

où $(P_t)_{0\leqslant t\leqslant 1}$ est un pont brownien.

Les lois limites qui apparaissent dans le théorème 1.4 sont bien connues et sont appelées *lois de Kolmogorov-Smirnov*. Elle sont portées par $\mathbb{R}_+$ et leur fonction de répartition est donnée pour tout $u\geqslant 0$ par:

$$\begin{aligned}
\mathbb{P}\left(\sup_{0\leqslant t\leqslant 1} P_t \leqslant u\right) &= 1-\exp(-2u^2)\\
\mathbb{P}\left(\sup_{0\leqslant t\leqslant 1} |P_t| \leqslant u\right) &= 1+2\sum_{k=1}^{\infty}(-1)^k\exp(-2k^2u^2).
\end{aligned}$$

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 (Google!). Google! cherche à évaluer l'attirance des toulousains vers son moteur de recherches. Son service marketing a comptabilisé, sur cent journées choisies au hasard, le nombre de connexions sur Google! via Toulouse, dans le tableau suivant:

Milliers de connexions	[3.9 ; 6.0[	[6.0 ; 7.6[	[7.6 ; 8.4[	[8.4 ; 10.0[	[10.0 ; 12.0[
Effectis associés	4	35	37	21	3

Effectuer un test de Kolmogorov-Smirnov d'adéquation de ces observations à la loi $\mathcal{N}(8,1)$, avec un niveau de confiance de 95% puis de 99%, en utilisant la fonction `kstest` de Matlab. Effectuer également un test du χ^2 d'ajustement et comparer vos résultats.

Exercice 2.4 (Test d'homogénéité de Kolmogorov-Smirnov). Soit $(X_1,\dots,X_n)$ un n-échantillon de fonction de répartition F et soit $(Y_1,\dots,Y_m)$ 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 $(X_1,\dots,X_n)$ et $(Y_1,\dots,Y_m)$. Alors, sous H_0

$$\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}|F_n(x)-G_m(x)| \xrightarrow[n\to+\infty]{\mathcal{L}} \sup_{0\leqslant t\leqslant 1}|P_t|$$

où $(P_t)_{0\leqslant t\leqslant 1}$ est un pont brownien. 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 2.5 (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 2.6 (Estimation non paramétrique à noyau d'une densité). Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. indépendantes et de même 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.

