TP 3 - Représentation graphique d'échantillons (Scilab) - Illustration de convergences en loi[1]

Le but de ce document est de présenter différentes illustrations graphiques possibles permettant de justifier qu'un échantillon simulé avec Scilab provient bien d'une loi donnée. On présente également les justifications théoriques sous-jacentes. Les mêmes méthodes pourront être utilisées pour illustrer des convergences en loi.

Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $(X_1, \dots, X_n)$ un $n-$échantillon de variables aléatoires réelles sur cet espace suivant une loi de probabilité μ de fonction de répartition F. On appellera *réalisation* de la suite de variables $(X_1, \dots, X_n)$ tout $n-$uplet $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$ pour un certain $\omega \in \Omega$.

1 Mesure empirique d'un échantillon

1.1 Mesure empirique

Définition 1 *On définit la* ***mesure empirique*** $\widehat{\mu}_n$ *de l'échantillon* $(X_1, \dots, X_n)$ *comme la mesure aléatoire*

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Ainsi, $\widehat{\mu}_n$ *est une application de* $\Omega \times \mathcal{F}$ *dans* $\mathbb{R}_+$*, telle que*
- *pour tout* $\omega \in \Omega$, $\widehat{\mu}_n(\omega, \cdot)$ *est une mesure de probabilité sur* $\mathbb{R}$,
- *pour tout* $A \in \mathcal{F}$, $\widehat{\mu}_n(\cdot, A) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{X_i \in A}$.

Propriétés de la mesure empirique. Pour tout $A \in \mathcal{F}$,

1. $(\widehat{\mu}_n(., A))_n$ est une suite de variables aléatoires qui converge presque sûrement (sous $\mathbb{P}$) vers $\mu(A)$ (Loi forte des grands nombres),
2. la variable $n\widehat{\mu}_n(., A)$ suit une loi binomiale $\mathcal{B}(n, \mu(A))$.

1.2 Fonction de répartition empirique

Définition 2 *On définit la* ***fonction de répartition empirique*** $\widehat{F}_n$ *de l'échantillon* $(X_1, \dots, X_n)$ *comme l'application aléatoire sur* $\Omega \times \mathbb{R}$ *telle que pour tout* $\omega \in \Omega$, $\widehat{F}_n(\omega, .)$ *est la fonction de répartition de la loi* $\widehat{\mu}_n(\omega, .)$*. Ainsi,*

$$\forall x \in \mathbb{R},\ \widehat{F}_n(., x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i(.) \leq x}.$$

Propriétés de la fonction de répartition empirique.

1. La fonction de répartition empirique est une fonction constante par morceaux. Soit $(x_1, \dots, x_n)$ une réalisation de la suite $(X_1, \dots, X_n)$, pour un $\omega \in \Omega$ fixé. Alors (faire un dessin !)

$$\widehat{F}_n(\omega, x) = \begin{cases} 0 \text{ si } x < x_{(1)} \\ \dfrac{i}{n} \text{ si } x_{(i)} \leq x < x_{(i+1)} \\ 1 \text{ si } x \geq x_{(n)}, \end{cases} \tag{1}$$

1. Enseignant : G. Chagny, bureau M.2.35. gaelle.chagny@univ-rouen.fr.

où $(x_{(1)}, \ldots, x_{(n)})$ est le $n-$uplet $(x_1, \ldots, x_n)$ trié par ordre croissant.

2. La Loi forte des grands nombres assure la convergence presque sûre de la suite de variables aléatoires $(\widehat{F}_n(., x))_n$ vers $F(x)$ (cas particulier de la Propriété 1 de la mesure empirique). Le Théorème de Glivencko-Cantelli renforce le résultat :

$$\sup_{x\in\mathbb{R}} |\widehat{F}_n(., x) - F(x)| \xrightarrow{p.s.} 0.$$

On peut se référer à [**?** , (7.4) p.59] ou [**?** , p.71] ou [**?** , p.116] ou [**?** , p.85] pour la démonstration.

1.3 Quantiles empiriques

On rappelle tout d'abord la définition d'un quantile de la loi d'une variable aléatoire X de loi μ.

Définition 3 *On définit un* ***quantile d'ordre*** $p \in]0; 1[$ *de la loi de* X *comme étant un réel* q_p *tel que* $\mathbb{P}(X \leq q_p) \geq p$ *et* $\mathbb{P}(X \geq q_p) \geq 1 - p$. *Pour* $p = 1/2$, *on parle de* ***médiane*** *; pour* $p = 1/4, 1/2, 3/4$ *les quantiles correspondants sont appelés* ***quartiles****.*

Remarque. On note $F^{(-1)}$ l'inverse généralisée de la fonction de répartition F de X, définie par

$$F^{(-1)}(u) = \inf\{x \in \mathbb{R},\ F(x) \geq u\},\quad u \in [0; 1].$$

Alors $q_p = F^{(-1)}(p)$ est un quantile d'ordre p de la loi de X.

On suppose dans la suite de ce paragraphe que la fonction de répartition F est continue. Dans ce cas, on peut réordonner l'échantillon $(X_1, \ldots, X_n)$ en un échantillon $(X_{(1)}, \ldots, X_{(n)})$, vérifiant $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ (tri par ordre croissant). On parle des *statistiques d'ordre* associées à l'échantillon de départ. On peut montrer que les $X_{(i)}$ sont bien des variables aléatoires définies presque-sûrement, et que, si le $n-$uplet de départ admet une densité f, alors le $n-$échantillon réordonné aussi -cette densité pouvant alors s'exprimer en fonction de f (voir [**?** , Exercice 2.17 p.54]).

Définition 4 *Avec les notations précédentes, pour* $p \in]0; 1[$*, on définit le quantile empirique d'ordre* p *de l'échantillon* $(X_1, \ldots, X_n)$ *comme étant* $q_{n,p} = X_{([np]+1)}$ *([.] est la partie entière).*

Propriétés des quantiles empiriques.

1. Si la loi μ possède un unique quantile d'ordre $p \in]0; 1[$, $q_p = F^{(-1)}(p)$, alors

$$q_{n,p} \xrightarrow{p.s.} q_p.$$

En terme statistique, $q_{p,n}$ est un estimateur fortement consistant de q_p. C'est une conséquence du Théorème de Glivencko-Cantelli (voir aussi une autre preuve et d'autres hypothèses au Théorème 8.13 [**?** , p.93].

2. Sous des hypothèses plus fortes (existence d'une densité f par rapport à la mesure de Lebesgue pour la loi μ , et non-annulation de la densité au voisinage du quantile q_p, on a aussi la normalité asymptotique de $q_{p,n}$:

$$\sqrt{n}\,(q_{n,p} - q_p) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(q_p)}\right).$$

L'Exercice 16 du polycopié **Théorèmes limite en probabilités et statistiques** fournit un exemple de ce résultat (cas de la médiane d'une loi uniforme). La preuve générale (plus difficile, pouvant être omise) peut être trouvée dans [**?**].

Les différentes manières de tester si un $n-$échantillon simulé $(X_1, \ldots, X_n)$ provient bien d'une loi donnée μ présentées dans la suite sont fondées sur la loi empirique de l'échantillon.

2 Illustration graphique par comparaison des fonctions de répartition théorique et empirique

Méthode 1. Pour comparer la loi d'un échantillon avec une loi théorique, on peut illustrer le Théorème de Glivencko-Cantelli : on superpose sur un même graphique la fonction de répartition empirique du $n-$échantillon simulé (commande `plot2d2`), et la fonction de répartition théorique de la loi sous-jacente.

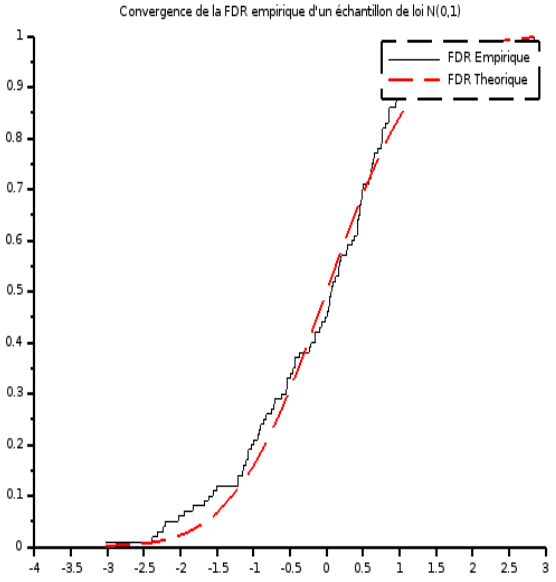
La fonction Scilab permettant de tracer une fonction en escalier comme la fonction de répartition empirique est `plot2d2`. On utilise l'expression donnée par (1), et il faut ainsi commencer par trier l'échantillon `X` dans l'ordre croissant à l'aide de la commande `gsort`.

Exemple. Le code suivant permet de tirer un échantillon de loi $\mathcal{N}(0,1)$ et d'illustrer la convergence annoncée.

```
//Simulation
n=100;
X=grand(1,n,'nor',0,1);

//Calcul des répartitions
   // pour la FDR empirique
XX=gsort(X,'g','i');
   // pour la FDR Theorique
FX=cdfnor('PQ',XX,zeros(XX),ones(XX))

//Tracé
scf(1)
clf
plot2d2(XX,(1:n)/n)
xset("line style",2)
xset("thickness",2)
plot2d(XX,FX,style=5)
xtitle('Convergence de la FDR empirique...
 d''un échantillon de loi N(0,1)')
legend(['FDR Empirique';'FDR Theorique'])
```



3 Illustration graphique par histogramme

On suppose que la loi μ est absolument continue par rapport à la mesure de comptage sur $\mathbb{Z}$ (cas d'une loi discrète) ou par rapport à la mesure de Lebesgue sur $\mathbb{R}$ (cas d'une loi continue). L'histogramme associé à un échantillon de données est un graphique constitué de barres verticales juxtaposées : chaque barre représente le nombre d'éléments de l'échantillon appartenant à une classe donnée.

Précisément, soit $(x_1, \dots, x_n)$ une réalisation de la suite $(X_1, \dots, X_n)$, et C l'ensemble des valeurs de l'échantillon. Pour bâtir un histogramme de l'échantillon,

1. on se donne une partition $(C_j)_{j\in J}$ de l'ensemble C : $C = \bigcup_{j\in J} C_j$, $C_j \neq \emptyset$ pour tout j, et $C_j \cap C_l = \emptyset$ pour $j \neq l$ (on note $|C_j|$ la mesure de Lebesgue ou de comptage de $|C_j|$, selon le cas considéré) ;
2. on compte le nombre N_j d'éléments de l'échantillon appartenant à la classe C_j pour tout j :
$$N_j = \sum_{i=1}^{n} \mathbf{1}_{x_i \in C_j}.$$
Remarquons que $N_j = n\widehat{\mu}_n(\omega, C_j)$ (avec $\omega \in \Omega$ tel que $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$).

L'*histogramme* associé est la fonction constante sur chaque élément de la partition, qui à $x \in C_j$ associe N_j pour tout j. On représente donc des barres de hauteur proportionnelle à l'effectif de la classe.

On s'intéresse plus souvent à l'*histogramme renormalisé* $\widehat{H}_{n,C}$, fonction toujours constante sur chaque élément de la partition, qui à $x \in C_j$ associe $N_j/(n|C_j|)$ pour tout j. On représente cette fois des barres d'aire proportionnelle à l'effectif de la classe. On a donc

$$\widehat{H}_{n,C}(x) = \frac{1}{n}\sum_{j\in J} \frac{N_j}{|C_j|}\mathbf{1}_{C_j}(x) = \sum_{j\in J} \frac{\widehat{\mu}_n(., C_j)}{|C_j|}\mathbf{1}_{C_j}(x) = \sum_{j\in J}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{x_i\in C_j}\right)\mathbf{1}_{C_j}(x), \qquad x \in \mathbb{R}.$$

L'aire totale de l'histogramme vaut 1, autrement dit, l'histogramme est une densité de probabilité.

3.1 Cas d'une loi discrète sur $\mathbb{Z}$

Si la mesure de probabilité μ est absolument continue par rapport à la mesure de comptage $\sum_{k\in\mathbb{Z}} \delta_{\{k\}}$ sur $\mathbb{Z}$, on représente souvent l'histogramme $\widehat{H}_{n,C}$ associé à la partition $C = (\{k\},\ k \in \mathbb{Z})$. On obtient alors un *diagramme en bâtons*. La hauteur du "bâton" d'abscisse k est la fréquence de k dans l'échantillon (proportion d'éléments de l'échantillon ayant pour valeur k) :

$$\widehat{H}_{n,C}(k) = \widehat{\mu}_n(\omega, \{k\}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i=k}.$$

Propriété du diagramme en bâtons. On a la convergence suivante :

$$\sup_{k\in\mathbb{Z}} |\widehat{\mu}_n(., \{k\}) - \mu(k)| \xrightarrow{p.s.} 0.$$

C'est une conséquence du Théorème de Glivencko-Cantelli.

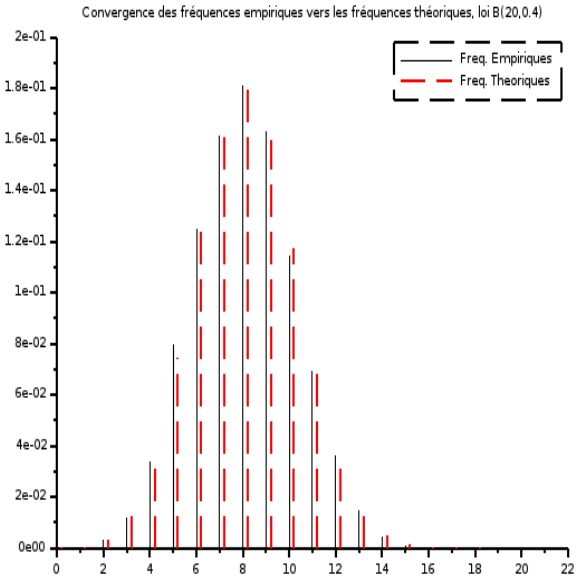
Méthode 2 - lois discrètes. Pour illustrer la qualité d'un $n-$échantillon de loi discrète, on peut donc représenter le diagramme en bâtons associé, et superposer les "bâtons" de la loi théorique sous-jacente à l'aide de la fonction `plot2d3` de Scilab : ceci revient à superposer fréquences empiriques dans l'échantillon et fréquences théoriques.

Exemple. Le code suivant permet de tirer un échantillon de loi binomiale $\mathcal{B}(20, 0.4)$ et de représenter les fréquences empiriques obtenues en les comparant aux fréquences théoriques de la loi $\mathcal{B}(20, 0.4)$.

```
//Simulation
n=200;
X=grand(1,n,'bin',20,0.4);

//Calcul des fréquences
FreqTheo=binomial(0.4,20);
[ind, occ]=dsearch(X,0:20,"d");
FreqEmp=occ/n;

//Tracé
scf(2)
clf
plot2d3(0:20,FreqEmp)
xset("thickness",2)
xset("line style",2)
plot2d3((0:20)+0.2,FreqTheo,style=5)
xtitle('Convergence des fréquences,...
    empiriques vers les fréquences,...
    théoriques loi B(20,0.4)')
legend(['Freq. Empiriques';...
    'Freq. Theoriques'])
```



3.2 Cas d'une loi absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}$

Supposons que μ ait une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$, et notons $[a;b]$ son support. On choisit alors souvent une partition de $[a;b]$ en intervalles de même longueur h_n : $C = (C_1, \dots, C_{r_n})$. On obtient cette fois un "vrai" histogramme aussi appelé éventuellement diagramme en barres,

$$\widehat{H}_{n,C}(x) = \frac{1}{nh_n} \sum_{j=1}^{r_n} N_j \mathbf{1}_{C_j}(x), \qquad x \in \mathbb{R}.$$

On choisit généralement une largeur h_n d'intervalle qui tend vers 0 quand n tend vers ∞.

Propriété de l'histogramme. En faisant des hypothèses sur la vitesse de convergence de $(h_n)_n$ vers 0, on pourra montrer que $(\mathbb{E}[(\widehat{H}_{n,C}(x) - f(x))^2])_n$ converge vers 0, ou encore, en ajoutant des hypothèses de régularité sur la densité f, que l'on peut choisir une largeur h_n qui minimise $\mathbb{E}[\int_a^b (\widehat{H}_{n,C}(x) - f(x))^2 dx]$.

Remarque : en terme statistique, on dira que la fonction $\widehat{H}_{n,C}$ est un estimateur non paramétrique de la densité f. Son principal défaut est de ne pas être lui-même régulier quand parfois la densité à estimer l'est : $\widehat{H}_{n,C}$ n'est même pas continu. Une façon de résoudre le problème consiste à lisser les histogrammes en définissant des estimateurs plus généraux, les estimateurs à noyaux (voir exemple de texte de modélisation).

Méthode 2 - lois continues. Pour illustrer la qualité d'un $n-$échantillon de loi continue, on peut donc représenter un histogramme associé, et superposer la densité de la loi théorique

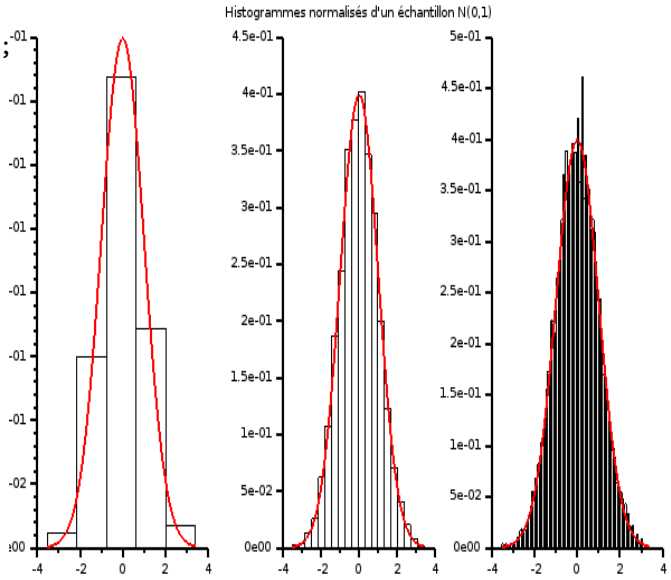
sous-jacente. On utilisera la commande `histplot`, qui fournit directement un histogramme renormalisé.

Exemple. Le code suivant permet de tirer un échantillon de loi $\mathcal{N}(0,1)$ et de représenter trois histogrammes (pour trois partitions différentes) ainsi que la densité théorique.

```
//Simulation
n=10000;
X=grand(1,n,'nor',0,1);

//Calcul de la densité théorique
vect_x=linspace(min(X),max(X),200);
densite_theo=exp(-vect_x.^2/2)/sqrt(2*%pi);

//Tracé
scf(3)
clf
subplot(131)
histplot(5,X)
xset("thickness",2)
plot2d(vect_x,densite_theo,style=5)
subplot(132)
histplot(20,X)
xset("thickness",2)
plot2d(vect_x,densite_theo,style=5)
xtitle('Histogrammes normalisés,...
   d''un échantillon N(0,1)')
subplot(133)
histplot(100,X)
xset("thickness",2)
plot2d(vect_x,densite_theo,style=5)
```



4 Illustration graphique par QQ-plot

Méthode 3 - lois de fonctions de répartition continues. On peut justifier graphiquement qu'un échantillon provient d'une loi donnée de fonction de répartition continue en vérifiant que les quantiles empiriques sont proches des quantiles théoriques. On peut par exemple tracer la courbe passant par les points d'abscisses les quantiles théoriques et d'ordonnées les quantiles empiriques. Cette courbe devrait être proche d'une droite.

Pour effectuer une telle représentation, on peut commencer par implémenter une fonction permettant de calculer le quantile empirique d'ordre `p` d'un échantillon `X`.

```
function y=Quantile_emp(X,p)
  // X=echantillon, p=vecteur à composantes dans ]0,1]
  //retourne les quantiles d'ordre p de X
    n=length(X);
    XX=gsort(X,'g','i');
    y=XX(n*p)
```

```
endfunction
```

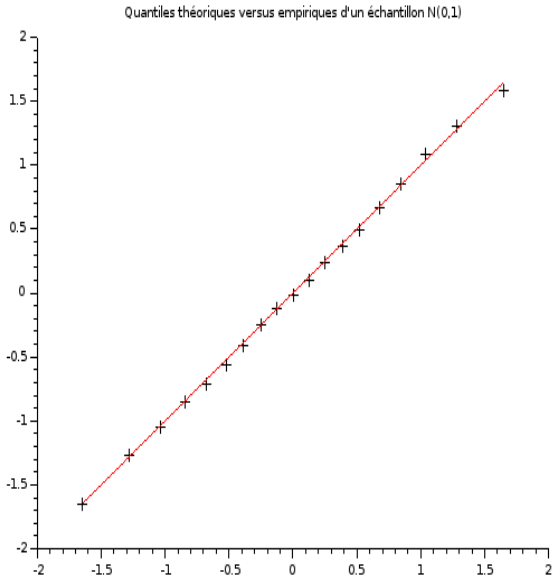
Exemple. Le code suivant permet de tirer un échantillon de loi $\mathcal{N}(0,1)$ et de comparer les quantiles empiriques et théoriques.

```
exec('Quantile_emp')

//Simulation
n=1000;
X=grand(1,n,'nor',0,1);

//Calcul des quantiles
p=0.05:0.05:0.95;
q_emp=Quantile_emp(X,p)
q_theo=cdfnor('X',zeros(p),ones(p),p,1-p);

//Tracé
scf(4)
clf
plot2d(q_theo, q_emp, style=-1)
plot2d(q_theo,q_theo,style=5)
xtitle('Quantiles théoriques versus,...
 empiriques d''un échantillon N(0,1)')
```



5 Application : illustrations de convergences en loi - Exercices

Exercice 1 *Approximation de la loi binomiale.* Pour de grandes valeurs de n, les calculs pratiques des probabilités d'une loi binomiale $\mathcal{B}(n,p)$ ($p \in]0;1[$) deviennent quasiment impossible, à cause du calcul des coefficients binomiaux $\binom{n}{k}$, $k \in \{0,\dots,n\}$ (voir Exercice 1, TP1). On utilise en pratique deux types d'approximation.

1. **Premier cas.** Si n tend vers l'infini, et $p = p_n$ dépend de n de telle sorte que $\lim_{n\to\infty} np_n = \lambda > 0$, alors la loi $\mathcal{B}(n,p_n)$ converge étroitement vers une loi de Poisson de paramètre λ (voir Exercice 7 du polycopié de cours). En pratique, on remplace la loi binomiale par une loi de Poisson dès que $n > 30$ et $np_n < 5$ ou dès que $n > 50$ et $p < 0.1$. Illustrer cette convergence à l'aide de diagrammes en bâtons.
2. **Second cas.** Si n tend vers l'infini, et si p est fixé, l'approximation est donnée par le Théorème de Moivre-Laplace (cas particulier historique du Théorème Central Limite pour les lois binomiales), utilisée en pratique dès que $n > 30$, $np > 5$ et $n(1-p) > 5$. En rappeler l'énoncé, et illustrer la convergence étroite associée, via les fonctions de répartition par exemple.

Exercice 2 *Autour du Théorème Central Limite.*

1. Illustrer le Théorème Central Limite dans le cas où la suite de variables de départ est de loi uniforme $\mathcal{U}_{[0;1]}$.
2. Soit $(X_n)_{n\geq 1}$ une suite de variables indépendantes de loi de Pareto de paramètre $a \in]1;2[$, c'est-à-dire de densité $x \mapsto a/x^{a+1}\mathbf{1}_{x>1}$. Justifier que les X_n admettent une

espérance (et la calculer), mais pas de moment d'ordre 2. Illustrer le fait que dans ce cas il n'y a pas convergence en loi de la suite $(\sqrt{n}(\sum_{i=1}^{n} X_i/n - \mathbb{E}[X_1]))_n$ vers une variable de loi gaussienne. On admettra pour l'instant (jusqu'au TP suivant !), que si U suit la loi $\mathcal{U}_{[0;1]}$, alors $U^{-1/a}$ suit la loi de Pareto de paramètre a.

Exercice 3 *Une autre convergence en loi.* Soit $(U_n)_{n\geq 1}$ une suite de variables aléatoires *i.i.d* de loi $\mathcal{U}_{[0;1]}$. Soit $V_n = n \min_{i=1,\dots,n} U_i$, pour $n \geq 1$.

1. Montrer que la suite $(V_n)_{n\geq 1}$ converge en loi vers une variable de loi exponentielle de paramètre 1.
2. Illustrer numériquement cette convergence.