Travaux pratiques n° 3

Initiation à R

Une note tenant compte du déroulement de votre TP (implication, autonomie, ...) pourra être attribuée dans le cadre du contrôle continu. Une absence injustifiée ou un oubli de vos identifiants donne lieu à la note 0.

Le **problème de la simulation** d'une loi de probabilité μ consiste à trouver une méthode produisant des réalisations indépendantes de même loi μ.

Il est important de noter que générer une loi uniforme sur l'ensemble $[0, 1]$, dont chaque point a autant de chances d'être tiré que les autres, est déjà un problème difficile. Sur la plupart des logiciels (matlab, R, ...), il vaut mieux utiliser des algorithmes pré-implémentés produisant une suite par une méthode récursive et déterministe à partir d'une valeur initiale. Dans toute la suite de ce TP, nous utiliserons le générateur aléatoire de nombres uniformes sur $[0, 1]$. Nous allons voir comment générer des lois plus générales à partir de ce type de générateur.

A noter qu'il existe un certain nombre de méthodes permettant de tester si la loi d'un échantillon est bien conforme à ce qui est attendu (tests du χ^2, de Kolmogorov, ...). Ces méthodes ne seront pas abordées ici.

1 L'aiguille de Buffon

Proposée par G-L Leclerc de Buffon en 1733, cette expérience visait à proposer une approximation du nombre π par une méthode probabiliste.

Supposons que l'on ait à notre disposition un parquet composé de lattes parallèles de même largeur $L > 0$. On jette sur ce parquet une aiguille de longueur a (avec $0 < a < L$) et on regarde si l'aiguille coupe l'une des droites du réseau. On répète cette expérience N fois de manière indépendante en notant

$$D_i = \begin{cases} 1 & \text{si l'aiguille coupe une droite au lancer i} \\ 0 & \text{sinon.} \end{cases}$$

Questions : Quelle est la loi de D ? En particulier, en quoi l'observation d'une (ou plusieurs) réalisation(s) de D peut nous donner des informations sur π ?

On introduit les variables aléatoires suivantes :

- X : distance du centre de l'aiguille à la rainure à droite la plus proche dans la direction perpendiculaire aux lattes,
- Θ : l'angle que fait l'aiguille avec cette même droite.

On peut en particulier voir que ces variables sont respectivement à valeur dans $[0, L]$ et $[-\pi/2; \pi/2]$. De manière plus précise, à la vue de l'expérience décrite ci-dessus, on a

— $X \sim \mathcal{U}([0, L])$,
— $\Theta \sim \mathcal{U}([-\pi/2, \pi/2])$,
— X et Θ sont indépendantes.

Par ailleurs, on remarque que

$$D_i = 1 \Leftrightarrow X \in \left[0, \frac{a}{2}\cos(\Theta)\right] \cup \left[L - \frac{a}{2}\cos(\Theta), L\right].$$

En particulier,

$$\mathbb{P}(D = 1) = \mathbb{P}\left(X \leq \frac{a}{2}\cos(\Theta)\right) + \mathbb{P}\left(X > L - \frac{a}{2}\cos(\Theta)\right).$$

Exercice 1:

1. Si U et V sont deux variables aléatoires, on définit

$$\mathbb{P}(U \leq V) = \int_{\mathbb{R}} \int_{-\infty}^{v} f_{(U,V)}(u, v)\, dudv.$$

En déduire que

$$\mathbb{P}(D = 1) = \frac{2a}{\pi L}.$$

L'expérience imaginée par le comte de Buffon permet donc de générer des variables aléatoires de Bernoulli de paramètre $2a/\pi L$.

2. Nous allons nous intéresser à la simulation d'une loi de Bernoulli, de paramètre $p \in]0, 1[$ fixé. Soit U une variable aléatoire suivant une loi uniforme sur $[0, 1]$ (des réalisations de cette dernière pouvant être obtenues à l'aide du générateur dont nous disposons). On pose alors

$$V = \mathbf{1}_{\{U<p\}}.$$

La variable V étant à valeur dans $\{0, 1\}$, elle suit donc une loi de Bernoulli. Par ailleurs,

$$\mathbb{P}(V = 1) = \mathbb{P}(U < p) = \int_0^p dx = p.$$

Notons $(D_1, \ldots, D_n)$ le vecteur aléatoire i.i.d. de l'expérience de Buffon. La loi des grands nombres nous indique alors que

$$\frac{1}{n}\sum_{i=1}^{n} D_i \xrightarrow{\mathbb{P}} \mathbb{E}(D) = \frac{2a}{\pi L} \quad \text{quand } n \to +\infty.$$

Autrement dit, on dispose d'un algorithme aléatoire permettant d'obtenir une approximation de π.

Donner une approximation de π par cette méthode avec `R`.

2 Simulation par la méthode d'inversion

Nous venons de voir comment générer une variable aléatoire suivant une loi de Bernoulli à partir d'un générateur de loi uniforme sur $[0,1]$.

De manière plus générale, il est possible de générer n'importe quel type de variable discrète en s'inspirant de ce principe. Nous allons voir par la suite que cette démarche s'inscrit dans une construction plus générale : la méthode d'inversion.

Soit F une fonction de répartition sur $\mathbb{R}$. La fonction F étant croissante, continue à droite, il est possible de définir son inverse généralisé $F^\dagger$ de la manière suivante

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

La fonction $F^\dagger$ n'est pas à proprement parler l'inverse (ou la fonction réciproque) de F , puisque celle-ci n'est pas nécessairement bijective de $\mathbb{R}$ dans $]0,1[$. Cependant, si F est inversible alors $F^\dagger = F^{-1}$.

Théorème 1

Si $U \sim \mathcal{U}([0,1])$, et F une fonction de répartition sur $\mathbb{R}$, alors la variable aléatoire $F^\dagger(U)$ a pour fonction de répartition F.

Preuve. On commence par montrer que pour tout $x \in \mathbb{R}$, $u \in [0,1]$,

$$u \leq F(x) \Leftrightarrow F^\dagger(u) \leq x. \tag{1}$$

Dans un premier temps, on remarque que si $u \leq F(x)$, alors

$$x \in \{t \in \mathbb{R} : F(t) \geq u\}.$$

Par définition de $F^\dagger$, on a donc $x \geq F^\dagger(u)$. Pour la réciproque, F étant continue à droite, on a

$$F^\dagger(u) \leq x \Rightarrow F\left(F^\dagger(u)\right) \leq F(x).$$

Dans la mesure où $u \leq F\left(F^\dagger(u)\right)$, on obtient finalement $u \leq F(x)$ ce qui prouve (1). En utilisant donc cette équivalence, on obtient, pour tout $x \in \mathbb{R}$

$$\mathbb{P}(F^\dagger(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

□

Dès lors que l'on dispose d'une forme explicite de la fonction de répartition (ce qui n'est pas toujours le cas ! Exemple : la loi normale), il est donc possible de générer la loi associée en utilisant ce principe d'inversion. Quelques exemples sont présentés ci-dessous.

1. **Simulation d'une loi exponentielle.**
   Si $X \sim \mathcal{E}(\lambda)$, alors
   $$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \forall x \geq 0, \\ 0 \quad \text{sinon.} \end{cases}$$
   La fonction F_X étant inversible dans ce cas précis, on a
   $$F^{-1}(u) = -\frac{1}{\lambda} \ln(1-u), \quad \forall u \in [0,1].$$
   Si $U \sim \mathcal{U}([0,1])$, le Théorème 1 indique alors que
   $$-\frac{1}{\lambda} \ln(1-U) \sim \mathcal{E}(\lambda).$$
   Effectuer des simulations en `R` pour illustrer cette méthode et tracer la fonction de répartition associée. Que constatez-vous ?

2. **Simulation d'une loi de Bernoulli.**
   Si maintenant $X \sim \text{Ber}(p)$,
   $$F_X(x) = \begin{cases} 0 & \text{si} \quad x < 0, \\ 1-p & \text{si} \quad x \in [0,1[, \\ 1 & \text{si} \quad x \geq 1, \end{cases}$$
   Dans ce cas, F_X n'est clairement pas inversible. En revanche, son inverse généralisé est donnée par la formule
   $$F_X^{\dagger}(u) = \begin{cases} 0 & \text{si} \quad u \in]0, 1-p[, \\ 1 & \text{si} \quad u \in [1-p, 1]. \end{cases}$$
   En utilisant à nouveau le Théorème 1, on obtient que si $U \sim \mathcal{U}([0,1])$, alors $F_X^{\dagger}(U) \sim \text{Ber}(p)$. On remarque en particulier que $F_X^{\dagger}(U) = 1$ si $U \in [1-p, 1]$. La variable $1-U$ suivant également une loi uniforme sur $[0,1]$, on retrouve le même algorithme que celui présenté à la section précédente.

   Effectuer des simulations en `R` pour illustrer cette méthode et tracer la fonction de répartition associée. Que constatez-vous ?

3 Méthode de Box Muller pour la simulation de variables gaussiennes

On présente ici une méthode dédiée tout particulièrement aux variables aléatoires gaussiennes, appelée méthode de **Box Muller**.

Son principe repose sur le changement de variable que nous rappelons :

Proposition 1

Soient $d \geq 2$ et ψ un $\mathscr{C}^1$-difféomorphisme d'un ouvert D de R^d sur $\psi(D) \subset \mathbb{R}^d$. Soit φ une application bornée de $\psi(D)$ dans $\mathbb{R}$ et f une application intégrable sur D. Alors,

$$\int_D \varphi(\psi(x)) f(x) dx = \int_{\psi(D)} \varphi(y) g(y) dy$$

où

$$\forall y \in \psi(D), \quad g(y) = \frac{f(\psi^{-1}(y))}{|\det(J_\psi(\psi^{-1}(y)))|}.$$

Notons que J_ψ est la matrice jacobienne de ψ et g peut aussi s'écrire

$$\forall y \in \psi(D), \quad g(y) = f(\psi^{-1}(y)) \; |\det(J_{\psi^{-1}}(\psi(y)))|.$$

Procéder à un changement de variable revient donc à déterminer le domaine $\psi(D)$ et la fonction g. C'est valable aussi pour les changements de variables dans les vecteurs aléatoires :

Proposition 2

Soit X un vecteur aléatoire à valeurs dans un ouvert $D \subset \mathbb{R}^d$ et de densité f. On pose $Y = \psi(X)$ où ψ est un $\mathscr{C}^1$-difféomorphisme de D sur $\psi(D) \subset \mathbb{R}^d$. Alors Y est un vecteur aléatoire qui admet pour densité la fonction g définie par

$$\forall y \in \mathbb{R}^d, \quad g(y) = \frac{f(\psi^{-1}(y))}{|\det(J_\psi(\psi^{-1}(y)))|} \mathbf{1}_{\psi(D)}(y)$$

ou

$$\forall y \in \mathbb{R}^d, \quad g(y) = f(\psi^{-1}(y)) \; |\det(J_{\psi^{-1}}(\psi(y)))| \mathbf{1}_{\psi(D)}(y).$$

Exercice 2: Soient $U \sim \mathcal{U}_{[0,1]}$ et $V \sim \mathcal{E}(1)$, deux variables indépendantes, et (X_1, X_2) le vecteur défini comme

$$X_1 = \sqrt{2V}\cos(2\pi U), \qquad X_2 = \sqrt{2V}\sin(2\pi U).$$

1. Montrer que $(X_1, X_2) \sim \mathcal{N}(0, I_2)$.
2. Illustrer et valider ce résultat avec des simulations en `R`.

Voici un exercice supplémentaire pour déterminer la densité d'un vecteur après changement de variable.

Exercice 3: Soit $X = (U, V)$ un vecteur aléatoire de $\mathbb{R}^2$ avec $U \sim \Gamma(\alpha, \theta)$ et $V \sim \Gamma(\beta, \theta)$ indépendantes. On pose $Y = \dfrac{U}{U+V}$ et $Z = U + V$.

1. Montrer que
$$\forall y \in \mathbb{R}, \quad f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}\mathbf{1}_{]0,1[}(y).$$
On appelle **loi bêta de paramètres α et β** la loi admettant cette densité.
2. Quelle est la loi de Z ?
3. Y et Z sont-elles indépendantes ?

4 Méthode de Monte-Carlo

Le principe de base des méthodes de Monte-Carlo est l'approximation numérique par estimation empirique. Plus précisemment, à partir de n simulations $x_1, \cdots, x_n$ de la loi d'une variable aléatoire X, on approche

— l'espérance d'une loi par la moyenne de l'échantillon simulé,
— un quantile par un quantile empirique de l'échantillon simulé,
— la probabilité d'un ensemble par la fréquence de cet ensemble dans l'échantillon simulé,
— ...

Dans tous ces cadres, on utilise la propriété fondamentale suivante : si la variable aléatoire X admet pour densité $f : \mathbb{R}^n \to \mathbb{R}$ et si $h : \mathbb{R}^n \to \mathbb{R}$ est une fonction, alors

$$\int_{\mathbb{R}^n} h(x)f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i).$$

Cette propriété peut-être formalisée dans la proposition suivante, conséquence directe de la LGN et du TCL.

Proposition 3

Soient X_i des vecteurs aléatoires i.i.d. de densité f, et h une fonction telle que $\mathbb{E}[|h(X)|] < +\infty$. Alors
$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{p.s.} \int_{\mathbb{R}^n} h(x)f(x)dx = \mathbb{E}[h(X)].$$
Si de plus $Var(h(X)) < +\infty$, alors on a en outre
$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} h(X_i) - \int_{\mathbb{R}^n} h(x)f(x)dx}{\sqrt{Var(h(X))}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Exercice 4: *Simulation de π par une méthode de Monte-Carlo*
Soit M le point de coordonnées (x, y) dans le carré $\mathscr{C} = [0, 1]^2$.
M appartient au quart de disque $\mathscr{D} \subset \mathscr{C}$, de centre $(0, 0)$ et de rayon 1, si et seulement si $x^2 + y^2 \leq 1$.
La probabilité que M appartienne à $\mathscr{D}$ est $\dfrac{Aire(\mathscr{D})}{Aire(\mathscr{C})} = \dfrac{\pi}{4}$.
En remarquant que

$$\frac{\pi}{4} = \int_0^1 \int_0^1 \mathbf{1}_{\mathscr{D}}(x, y) dx dy$$

donner une approximation de π par la méthode de Monte-Carlo.
Indication : on pourra considérer les variables aléatoires indépendantes $X \sim \mathcal{U}([0, 1])$ et $Y \sim \mathcal{U}([0, 1])$.

5 Simulation par la méthode du rejet

Comme évoqué plus haut, il arrive qu'une forme explicite de $F^\dagger$ soit disponible afin de mettre en place une simulation par la méthode d'inversion. Dans de nombreuses applications, ce type de propriété n'est pas toujours garantie. Nous allons nous intéresser ici à la méthode du rejet, permettant de mettre en place des simulations quand seul un majorant de la densité cible est disponible.

Plus précisément, il s'agit d'un algorithme de simulation d'observations issues d'une v.a. de densité f utilisable si l'on dispose

— d'une densité g que l'on sait simuler,
— d'une constante $c \geq 1$ telle que $f(x) \leq cg(x)$ pour tout $x \in \mathbb{R}^d$.

Algorithme du rejet
Générer indépendamment $U \sim U([0, 1])$ et X de densité g.
Poser $T \leftarrow c \times \dfrac{g(X)}{f(X)}$.
Tant que $(UT \geq 1)$ faire :
- *Générer indépendamment $U \sim U([0, 1])$ et X de densité g,*
- *Poser $T \leftarrow c \times \dfrac{g(X)}{f(X)}$.*

Accepter X comme variable aléatoire de densité f.

Exercice 5:

1. Montrer que
$$\mathbb{P}\left(U < \frac{1}{T}\right) = \frac{1}{c}.$$
2. Montrer que, pour $B \subset \mathbb{R}^d$, on a
$$\mathbb{P}\left(X \in B \,\middle|\, U < \frac{1}{T}\right) = \int_B f(x) dx,$$
ce qui signifie que f est la densité de probabilité de la loi conditionnelle de X sachant $U < \dfrac{1}{T}$.

Remarques

- La méthode du rejet est plus lente que les méthodes précédentes. On ne l'utilise donc que lorsque les autres ne sont pas applicables.
- On peut voir que le nombre N d'itérations nécessaires avant d'avoir généré une réalisation de X (de densité f) suit une loi géométrique de paramètre $p = \frac{1}{c}$.
- Comme $\mathbb{E}(N) = c$, alors c représente le nombre moyen de rejets. La constante c joue donc un rôle de première importance dans la rapidité de l'algorithme. En particulier, plus c est grande (i.e. la majoration de f est mauvaise), plus l'algorithme mettra du temps à produire une réalisation de X. Il convient donc de le choisir le plus petit possible (en pratique, on peut prendre $c = \max_x \frac{f(x)}{g(x)}$).
- Si la densité f est bornée et à support compact, il est facile de travailler avec une loi uniforme pour la densité g. On utilise ensuite la méthode d'inversion afin de ne générer que des simulations de la loi $U(0,1)$.

Exercice 6: Soit X une variable aléatoire réelle ayant la densité de probabilité suivante

$$\forall x \in \mathbb{R}, \quad f(x) = \frac{2}{\pi}\sqrt{1-x^2}\mathbf{1}_{[-1,1]}.$$

A l'aide des remarques ci-dessus, simuler X selon la méthode du rejet et valider votre algorithme graphiquement.