TP2 : SIMULATION DE VARIABLES ALÉATOIRES

Résumé. Ce tp présente les fondements de la simulation de variables aléatoires de loi donnée à partir d'une suite dénombrable de variables uniformes sur le segment unité $[0,1]$. Nous nous intéresserons aux variables aléatoires discrètes et continues et nous verrons plusieurs méthodes de simulations pour la loi normale. La simulation de variables aléatoires a de nombreuses applications pour observer des phénomènes de convergence, valider une méthode et même calculer des intégrales.

Toutes nos méthodes vont reposer sur le générateur de nombres aléatoires de Matlab. Un générateur de nombres aléatoires dans l'intervalle $[0,1]$ est une fonction `rand` qui vérifie les deux propriétés suivantes :

(i) un appel à la fonction `rand` donne une réalisation d'une variable aléatoire de loi uniforme sur $[0,1]$,

(ii) les appels successifs à la fonction `rand` fournissent une réalisation d'une suite de variables aléatoires indépendantes.

En pratique, les ordinateurs utilisent une suite de nombres dits "pseudo-aléatoires" pour simuler le tirage de nombres suivant la loi uniforme sur $[0,1]$. Une méthode pour générer une telle suite $(x_n)_{n\in\mathbb{N}}$ est de définir une suite d'entiers (y_n), par une relation de récurrence de la forme :

$$y_{n+1} = ay_n + b \bmod m,$$

où a, b et m sont des entiers bien choisis, et de poser $x_n = y_n/m$. Matlab dispose de deux générateurs de nombres aléatoires :

- `rand` pour la loi uniforme sur $[0,1]$,
- `randn` pour la loi normale $\mathcal{N}(0,1)$.

On supposera que la fonction `rand` de Matlab a les deux propriétés (i) et (ii) et que `randn` a les propriétés analogues pour la loi normale $\mathcal{N}(0,1)$.

NB : Au début d'un programme, on peut initialiser le générateur, en lui fournissant la première valeur x_0 appelée la graine du générateur. Sinon, à chaque fois que l'on démarre Matlab, c'est la même suite de valeurs qui sera donnée. (Voir `help rand`.) Avec la commande `rand('state',sum(100*clock))`, le générateur `rand` est initialisé avec une graine dont la valeur dépend de l'heure.

1. Variables discrètes

Dans le cas d'une variable aléatoire discrète, la méthode est canonique. Soit X une variable aléatoire à valeurs dans $\{x_1,\dots,x_r\}$ de loi $(p_1,\dots,p_r)$, avec $\mathbb{P}\{X = x_j\} = p_j$. Montrer que si U est une variable uniforme sur $[0,1]$ alors la variable aléatoire

$$\sum_{j=1}^{r} x_j \, \mathbb{1}_{]p_0+\ldots+p_{j-1},p_1+\ldots+p_j]}(U)$$

(où l'on note $p_0 = 0$) suit la même loi que X. Donc, pour simuler un échantillon $(y_1,\dots,y_k)$ d'une telle variable X, on simule un échantillon $(u_1,\dots,u_k)$ de variables de loi uniforme sur $[0,1]$ et on pose $y_i = x_j$ si $u_i \in]p_0+\ldots+p_{j-1}, p_1+\ldots+p_j]$.

Exercice 1. Nous avons déjà vu au TP1 comment simuler une loi de Bernoulli, montrer que nous avons en fait utilisé la méthode canonique décrite ci-dessus. Ecrire une fonction

`Bin(k1,k2,n,p)` qui prend en argument 3 entiers `k1`, `k2`, `n` et un réel `p` de $[0,1]$, et qui renvoie un `k1`×`k2`-échantillon de variables aléatoires de loi binomiale de paramètres `n` et `p`. Contrainte : il est interdit d'utiliser une quelconque instruction `if` ou une boucle `for`.

On a donc simulé une loi binomiale sans utiliser directement la méthode canonique. De la même façon, il existe de nombreuses méthodes spécifiques, comme le montrent les exercices suivants.

Exercice 2. Déterminer la loi de la variable aléatoire `X` en sortie des algorithmes :

(1) `X=floor(3*rand)`,

(2) `X=floor(3*rand)*floor(2*rand)`,

(3) `X=round(4*rand)`.

Vérifiez vos réponses numériquement, par simulation de n-échantillons et comparaison des fréquences empiriques aux fréquences que vous proposez

Application : écrivez une fonction qui simule élégamment un n-échantillon de loi uniforme sur $\{1,\dots,N\}$, où n et N seront les paramètres transmis en entrée.

Exercice 3. Soit $(X_n)_{n\geq 1}$ un échantillon de variables aléatoires de loi de Bernoulli de paramètre $p \in]0,1[$. Posons $Y = \inf\{k \in \mathbb{N}^* \,|\, X_k = 1\}$. (Remarquez que vu les hypothèses sur p, l'infimum précédent est atteint presque sûrement.)

(1) Quelle est la loi de Y ?

(2) Ecrire une fonction qui permette de simuler un n-échantillon de loi celle de Y. (Peut-on se passer d'instruction de boucle cette fois ?)

(3) Soit X une variable aléatoire de loi exponentielle de paramètre $\lambda = -\ln(1-p)$. Quelle est la loi de $\lceil X \rceil$, où $\lceil x \rceil$ désigne la partie entière supérieure de x ?

(4) Imaginons que vous disposiez d'un n-échantilllon de variables aléatoires de même loi que X (vous aurez à en construire un ci-dessous de vos propres mains, utilisez pour l'instant la fonction `rexpweib`). Comment en déduire un n-échantillon de loi celle de Y ? Ecrivez la fonction correspondante.

(5) Avec l'une ou l'autre de ces fonctions, tirez un n-échantillon, avec n petit puis grand, et déterminez graphiquement si votre n-échantillon est de qualité ou non. Remarquez que cela revient peu ou prou à tester le générateur de nombres uniformes de Matlab...

Exercice 4 [Loi de Poisson]. Pour simuler une loi de Poisson, on utilise généralement le résultat suivant :
Si $(U_i)_{i\in\mathbb{N}^*}$ est un échantillon de variables aléatoires de loi uniforme sur $[0,1]$, la variable

$$X = \inf\{n \geq 0 \,|\, U_1 \times \ldots \times U_{n+1} < e^{-\lambda}\}$$

suit une loi de Poisson de paramètre λ.

Ecrivez une fonction simulant un n-échantillon de loi de Poisson de paramètre λ en utilisant cette méthode. Quel avantage voyez-vous par rapport à la méthode canonique ? (Remarque : la loi de Poisson peut en fait être simulée avec la fonction `rpoiss` de *Stixbox*.)

2. Variables à densité

Il existe deux grandes méthodes de simulation de variables aléatoires continues, utilisant respectivement la fonction de répartition et la densité.

• Simulation par inversion

La méthode de simulation par inversion repose sur le résultat suivant.

Proposition 1. *Soit X une variable aléatoire réelle de fonction de répartition $F_X(t) = \mathbb{P}\{X \leq t\}$. On définit l'inverse généralisée F_X^{-1} de F_X sur $]0,1[$ par*

$$F_X^{-1}(x) = \inf\{t \in \mathbb{R} \,|\, F_X(t) \geq x\}.$$

Alors, si U est une variable aléatoire de loi uniforme sur $]0,1[$, $F_X^{-1}(U)$ a même loi que X.

Ainsi, si on tire n nombres au hasard uniformément répartis entre 0 et 1, $(u_1, u_2, \ldots, u_n)$, l'échantillon recherché, $(x_1, x_2, \ldots, x_n)$, de loi celle de X, sera déterminé par $x_i = F_X^{-1}(u_i)$.

• Simulation par méthode de rejet

La méthode d'inversion nécessitait la connaissance explicite de F_X^{-1}, ce qui n'est pas toujours le cas (notamment pour les lois normales). Une méthode alternative est la méthode de rejet. Celle-ci sert d'abord à simuler des lois conditionelles ou des loi uniformes sur des domaines de $\mathbb{R}^d$ en utilisant le résultat suivant.

Proposition 2. *Soit (M_n) une suite de variables aléatoires indépendantes de loi μ et B un borélien tel que $\mu(B) > 0$. Soit T le plus petit entier $n \geq 1$ tel que $M_n \in B$. Alors*

(i) *T est une variable aléatoire de loi géométrique de paramètre $\mu(B)$,*

(ii) *M_T est une variable aléatoire ayant pour loi la loi conditionnelle $\mu(.|B)$. En particulier lorsque μ est la loi uniforme sur un borélien C contenant B et tel que $0 < \lambda(B) < \lambda(C)$, cette loi conditionnelle est la loi uniforme sur B.*

On a noté λ la mesure de Lebesgue. En pratique on peut donc simuler une variable X de loi uniforme sur B de la façon suivante : on tire M_1 uniforme sur C, si $M_1 \in B$, on pose $X = M_1$, sinon on retire une variable M_2 uniforme sur C et on recommence [1].

Cette méthode permet également de simuler des lois à densités en considérant le domaine sous le graphe de la densité. Dans le cas particulier où f est une densité continue à support compact inclus dans un intervalle $[a,b]$, majorée par un réel $K > 0$, on peut utiliser la proposition 2 avec μ la loi uniforme sur le rectangle $[a,b] \times [0,K]$ et $B = \{(x,y) \in \mathbb{R}^2, 0 \leq y \leq f(x)\}$. En notant $M_n = (U_n, V_n)$, on obtient que (U_T, V_T) suit la loi uniforme sur B ce qui entraîne que U_T a pour densité f. Pour simuler X selon la densité f, on utilise donc la procédure suivante. On tire un point $M = (U,V)$ selon une loi uniforme sur le rectangle $[a,b] \times [0,K]$. Si le point M se trouve dans la région située sous le graphe de f, on l'accepte et on pose $X = U$, sinon on le rejette et on tire un nouveau point uniformément, et ainsi de suite jusqu'à obtenir un point M qui se trouve dans la région située sous le graphe de f.

La méthode précédente peut être généralisée au cas où il existe une densité de probabilité g telle que

– on sait simuler une variable de densité g,
– il existe une constante K ($K \geq 1$!) telle que $f \leq Kg$.

On utilise alors la procédure suivante, qui termine presque sûrement et simule une variable X de densité f :

(1) On simule une variable U de densité g et une variable W indépendante de X de loi uniforme sur $[0,1]$. On pose $V = K\,W\,g(U)$.

(2) Si $V \leq f(U)$ on pose $X = U$, sinon on retourne en (1).

1. C'est en fait la même procédure (bien connue des rôlistes) que jeter un dé à 6 faces et ne retenir que les chiffres inférieurs à 4 pour obtenir une loi uniforme sur $\{1, \ldots, 4\}$.

Il existe bien sûr de nombreuses autres méthodes, spécifiques à certaines lois, comme nous allons le voir dans les derniers exercices.

Exercice 5. Prouver la proposition 1, en comparant la fonction de répartition de $F_X^{-1}(U)$ à celle de X.

Exercice 6. Montrer que la méthode canonique pour les variables discrètes est l'application exacte de la proposition 1.

Exercice 7. Ecrire des fonctions pour simuler un n-échantillon :

(1) de loi uniforme sur $[a, b]$, pour $a < b$ quelconque ;

(2) de loi de Laplace, de densité sur $\mathbb{R}$ donnée par $x \mapsto e^{-|x|}/2$.

Attention, ces fonctions ne doivent pas comporter de boucles `for`.

Exercice 8 [Loi de Cauchy]. Cette loi admet pour densité sur $\mathbb{R}$

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \ .$$

Ecrire une fonction qui simule un n-échantillon de loi de Cauchy. Nous allons maintenant faire plus ample connaissance avec cette loi. Admet-elle un moment d'ordre 1 ? Pourtant elle est centrée... Essayer d'estimer une hypothétique "moyenne" en faisant comme si la loi des grands nombres s'appliquait. Or, la loi de Cauchy est parfois notée $C(0, 1)$: à quoi correspond le 0 ? Et le 1 ? Vérifiez numériquement vos hypothèses sur un n-échantillon.

Exercice 9 [Loi exponentielle]. Dans l'exercice 3, nous avions eu besoin d'un n-échantillon de loi exponentielle. Ecrivez une fonction qui réalise ceci, à partir d'un n-échantillon de loi uniforme... sans boucle `for` bien sûr.
Application : extrait du partiel 2009. La densité d'une loi Gamma $G(\alpha, \lambda)$ est donnée par

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbb{1}_{]0,\infty[}(x)$$

où Γ est la fonction Gamma définie pour tout $x > 0$ par $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$.

(1) Expliquer pourquoi la méthode de simulation par inversion est difficile à réaliser pour cette loi.

(2) A quoi correspond la loi gamma quand $\alpha = 1$? Faire N simulations d'une telle loi (fonction `gamma1` qui prend en argument λ et N).

(3) Soit X de loi $G(\alpha, \lambda)$ et Y de loi $G(\beta, \lambda)$ indépendante de X. Montrer qu'il existe une constante $C(\alpha, \beta, \lambda)$ telle que la densité de $X + Y$ s'écrive

$$f_{X+Y}(x) = C(\alpha, \beta, \lambda) x^{\alpha+\beta-1} e^{-\lambda x} \mathbb{1}_{]0,\infty[}(x).$$

(4) Si les X_i sont des variables indépendantes de loi Gamma $G(\alpha_i, \lambda)$ pour tout $i = 1, \ldots, n$, quelle est la loi de $\sum_{i=1}^n X_i$?

(5) En déduire une méthode de simulation d'une loi Gamma quand α est entier : écrire une fonction `gammaentier` qui prend en argument α (entier), λ et N et qui renvoie un N-échantillon de loi Gamma.

(6) Si $X \sim \mathcal{N}(0, 1)$, quelle est la loi de X^2 (sachant que $\Gamma(1/2) = \sqrt{\pi}$) ? En déduire une méthode de simulation de la loi $G(k/2, 1/2)$ pour $k \in \mathbb{N}$, et écrire la fonction `gammademi` correspondante.

Exercice 10. Démontrer la proposition 2. L'utiliser pour simuler des réalisations i.i.d. de la loi uniforme sur le disque unité de $\mathbb{R}^2$.

Exercice 11. Il existe plusieurs méthodes pour simuler une loi gaussienne, commençons par la méthode de rejet. Si f est la densité de la loi normale, on peut prendre pour g celle de la loi de Laplace :

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \leq Kg(x)$$

avec $K = \sqrt{\frac{2e}{\pi}}$ et $g(x) = \exp(-|x|)/2$. Ecrire une fonction permettant de simuler un n-échantillon de loi normale standard.

Exercice 12 [algorithme de Box-Muller]. Soient U et V deux variables aléatoires indépendantes de loi uniforme sur $[0, 1]$. Les variables X et Y définies par

$$\begin{aligned} X &= \sqrt{-2\ln(U)}\cos(2\pi V) \ , \\ Y &= \sqrt{-2\ln(U)}\sin(2\pi V) \end{aligned}$$

sont indépendantes et de loi $\mathcal{N}(0, 1)$. (Cela se prouve par changement de variables.) En déduire une fonction permettant de simuler un n-échantillon de loi normale standard. Comparez ensuite la vitesse d'éxécution de cette fonction et de la précédente à celle obtenue par `randn`.

Exercice 13 [Simulation de la loi $\mathcal{N}(m, \Gamma)$]. Ici, $m \in \mathbb{R}^d$ et Γ est une matrice de $\mathcal{M}_d(\mathbb{R})$, symétrique et postive. Soit C une matrice telle que $C^t C = \Gamma$ et Z un vecteur aléatoire gaussien de loi $\mathcal{N}(0, I_d)$. Alors $X = CZ + m$ a pour loi $\mathcal{N}(m, \Gamma)$. En pratique pour déterminer C, on utilise la décomposition de Cholesky lorsque Γ est définie positive (commande `chol`). Cette décomposition calcule une matrice triangulaire supérieure A telle que ${}^tAA = \Gamma$. Il suffit alors de prendre $C = {}^tA$.

Simuler un n-échantillon de vecteurs gaussiens de dimension 2, de moyenne $m = (7, 9)$ et de matrice de covariance $\Gamma = [4\ 3; 3\ 4]$. Calculer la moyenne et la covariance empirique de votre échantillon pour vérifier, et représenter le nuage de points.