TP 3 - Simulation de variables aléatoires [1]

Objectifs

Le problème de la *simulation* consiste à trouver des méthodes qui permettent de produire ou *générer* des échantillons finis de variables (ou vecteurs) aléatoires de lois données.

On supposera dans la majeure partie de ce document que l'on sait générer une suite *i.i.d.* de variables aléatoires de loi uniforme $\mathcal{U}_{[0;1]}$ (voir paragraphe suivant), à l'aide du générateur `rand` de Scilab. On se propose de bâtir des algorithmes permettant, à partir d'une telle suite, de simuler une variable aléatoire ou un vecteur aléatoire de loi donnée. La présentation est non exhaustive, les références principales sont [BC07] (chapitre 1), [Yca02] (chapitre 2) et [RS09] (chapitre 9 + paragraphe 8.1.2 du chapitre 8).

⚠ On n'utilisera bien sûr pas la commande `grand` pour ce T.P. ! Pour chaque exercice, penser à vérifier les fonctions implémentées, c'est-à-dire par exemple à illustrer graphiquement la qualité des échantillons simulés, en représentant les lois empiriques associées (voir le document complémentaire, **Représentation graphique d'échantillons**).

1 Obtention de variables de lois uniformes

1.1 Lois uniformes sur $[0;1]$ - Générateurs pseudo-aléatoires

Dans toute la suite de ce document, on partira du principe selon lequel la fonction `rand` de Scilab permet de générer une suite *i.i.d.* $(U_n)_n$ de loi $\mathcal{U}_{[0;1]}$. Les postulats sont donc les suivants (voir [Yca02]) :

1. Un appel à la fonction `rand` retourne une réalisation d'une variable aléatoire de loi $\mathcal{U}_{[0;1]}$.
2. Des appels successifs à la fonction `rand` retournent des variables aléatoires indépendantes.

En réalité, une suite générée par un algorithme connu est rarement aléatoire : connaissant la valeur initiale de l'algorithme, on peut en dérouler les commandes pour trouver la valeur retournée. La fonction `rand` de Scilab, de même que ses équivalents dans les autres logiciels, est donc seulement un *générateur pseudo-aléatoire*, c'est-à-dire un algorithme déterministe qui fournit une liste de nombres censés reproduire le comportement d'un échantillon $\mathcal{U}_{[0;1]}$.

La production d'une telle suite est généralement récursive à partir de valeurs initiales, appelées *graines* (*seed* en anglais). On parle de méthode prédictible (voir [BC07] pour plus de détails). Le générateur `rand` de Scilab est un générateur par *congruence linéaire* : une première suite $(X_n)_n$ est générée selon la relation de récurrence

$$X_{n+1} = aX_n + b \bmod m, \quad n \in \mathbb{N},$$

avec $m \in \mathbb{N}\backslash\{0\}$, $a, b \in \{0, \dots, m-1\}$, puis on en déduit la suite $(U_n)_n$ de $[0;1[$, avec $U_n = X_n/m$. Dans Scilab, $a = 843314861$, $b = 453816693$, et $m = 2^{31}$. La suite construite est donc périodique de période m. À partir de la graine initiale, qui est toujours 0 par défaut (rentrer la commande `rand('seed')` pour le voir), la séquence générée est toujours la même d'une session à une autre. On peut initialiser la graine du générateur différemment (voir l'aide de la fonction `rand`) dans le but de produire d'autres séquences, mais vu la grandeur de la période m, on peut considérer la suite produite comme aléatoire.

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

1.2 Autres lois uniformes

Exercice 1 1. Écrire une fonction qui simule un $n-$échantillon de loi uniforme $\mathcal{U}_{[a;b]}$ sur un intervalle $[a;b]$, pour $a < b$.

2. Comment simuler un vecteur aléatoire de loi uniforme sur un pavé $[a;b] \times [c;d]$?
3. Écrire une fonction qui simule un $n-$échantillon de loi uniforme $\mathcal{U}_{\{1,\dots,N\}}$ sur l'ensemble discret $\{1,\dots,N\}$.

2 Simulation par inversion

En plus des références ci-dessus (par exemple [RS09] pour cette section), on pourra aussi se reporter aux exercices 2.18 et 2.19 de [CGCDM11].

2.1 Principe de la méthode

Soit X une variable aléatoire réelle de fonction de répartition F. On rappelle que F est croissante (donc en particulier admet un nombre fini ou dénombrable de point de discontinuité). Elle est de plus continue à droite, limitée en tout point à gauche. En général, elle n'est cependant pas bijective. On peut toutefois introduire la notion suivante.

Définition 1 *On définit la* ***fonction inverse généralisée*** *ou* ***fonction quantile*** $F^{(-1)}$ *de* F *par*

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

Lemme 2 *Quels que soient* $x \in \mathbb{R}$ *et* $u \in [0;1]$, *on a l'équivalence*

$$u \leq F(x) \qquad \Longleftrightarrow \qquad F^{(-1)}(u) \leq x.$$

En particulier, $F \circ F^{(-1)}(u) \geq u$, *avec égalité si et seulement si* $u \in F(\mathbb{R})$.

Remarques.

- L'intervalle $]0;1[$ est inclus dans $F(\mathbb{R})$ si et seulement si F est continue, c'est-à-dire si et seulement si la loi de X ne charge aucun point. Attention, si F est continue, elle n'est pas nécessairement bijective (présence possible de plateaux).
- La fonction $F^{(-1)}$ coïncide avec l'inverse classique F^{-1} lorsque F est bijective.

On déduit du lemme la proposition suivante, sur laquelle est fondée la méthode de simulation par inversion (de la fonction de répartition).

Proposition 3 *Si* U *est une variable aléatoire de loi* $\mathcal{U}_{[0;1]}$, *alors la variable aléatoire* $F^{(-1)}(U)$ *suit la loi de fonction de répartition* F.

Remarque.

1. La fonction $F^{(-1)}$ est croissante, donc borélienne, donc $F^{(-1)}(U)$ est bien une variable aléatoire.
2. De plus, si la fonction de répartition F est continue, on peut aussi prouver que $F(X)$ suit la loi $\mathcal{U}_{[0;1]}$. C'est faux si F n'est pas continue : si la loi de X admet un atome, alors toute mesure image de sa loi en possède encore (et donc ne peut suivre la loi $\mathcal{U}_{[0;1]}$).

Méthode de simulation par inversion. Pour obtenir une variable X de loi de répartition F dont on sait calculer explicitement l'inverse généralisée $F^{(-1)}$, on peut simuler une variable U de loi $\mathcal{U}_{[0;1]}$ et poser $X = F^{(-1)}(U)$.

2.2 Application à la simulation de variables aléatoires de lois continues

Exercice 2 Écrire des fonctions permettant de simuler des $n-$échantillons de variables aléatoires dont les lois sont les suivantes :

1. **Loi exponentielle** $\mathcal{E}(\lambda)$, $\lambda > 0$: densité $x \mapsto \lambda \exp(-\lambda x)\mathbf{1}_{x>0}$ (penser à exploiter le fait que si U est de loi $\mathcal{U}_{[0;1]}$, alors $1-U$ a même loi que U).
2. **Loi de Cauchy** $\mathcal{C}(a)$, $a > 0$: densité $x \mapsto a/(\pi(a^2+x^2)$.
3. **Loi de Laplace** : densité $x \mapsto \exp(-|x|)/2$.
4. **Loi de Weibull** $\mathcal{W}(a,b,c)$, $a, c > 0$, $b \geq 0$: densité $x \mapsto (a/c)((x-b)/c)^{a-1}\exp(-((x-b)/c)^a)\mathbf{1}_{x>b}$.
5. **Loi de Pareto** de paramètre $a > 0$: densité $x \mapsto a/x^{a+1}\mathbf{1}_{x>1}$.

2.3 Application à la simulation de variables aléatoires suivant des lois discrètes

La méthode par inversion se réécrit pour des variables discrètes (variables à valeurs dans $\mathbb{N}$ ou $\mathbb{Z}$) de manière générale, en utilisant l'expression de l'inverse généralisée $F^{(-1)}$.

On considère X une variable aléatoire de loi $\sum_{k\in\mathbb{N}\setminus\{0\}} p_k\delta_k$. On note toujours F la fonction de répartition de X. Soit $s_k = \sum_{l=1}^{k} p_l$, $k \geq 1$, et $s_0 = 0$. Alors

$$F^{(-1)}(u) = k, \text{ avec } k \text{ tel que } s_{k-1} < u \leq s_k.$$

Par exemple, $F^{(-1)}(u) = 1$ si $u \leq p_1$, $F^{(-1)}(u) = 2$ si $p_1 < u \leq p_1 + p_2$...

Méthode de simulation d'une variable discrète. Soit X de loi $\sum_{k\in\mathbb{N}\setminus\{0\}} p_k\delta_{x_k}$ avec $\sum_{k\in\mathbb{N}\setminus\{0\}} p_k = 1$. On divise l'intervalle $[0;1]$ en intervalles de longueurs successives $p_1, p_2, \ldots, p_k, ...$ Les extrémités successives des intervalles sont donc $s_0, s_1, s_2 \ldots, s_k, ...$ On tire U de loi $\mathcal{U}_{[0;1]}$, et on détermine à quel intervalle de la partition U appartient : on choisit la valeur $X = x_k$ quand $s_{k-1} \leq U \leq s_k$.

La méthode est justifiée car en effet, on obtient alors

$$\mathbb{P}(X = x_k) = \mathbb{P}(s_{k-1} < U \leq s_k) = s_k - s_{k-1} = p_k.$$

Exercice 3

1. Comment simuler une variable X de **loi de Bernoulli** $\mathcal{B}(p)$ ($\mathbb{P}(X=1) = p = 1 - \mathbb{P}(X=0)$) ? Et une variable Y de **loi de Rademacher** $\mathcal{R}(p)$ ($\mathbb{P}(Y=1) = p = 1 - \mathbb{P}(Y=-1)$) ?
2. Écrire une fonction permettant de simuler un échantillon de variables aléatoires de **loi discrète sur un ensemble fini**. Tester la fonction pour simuler par exemple un échantillon de loi uniforme $\mathcal{U}_{\{1,\ldots,6\}}$.

Exercice 4 Écrire des fonctions permettant de simuler des $n-$échantillons de variables aléatoires dont les lois sont les suivantes :

1. **Loi de Poisson** $\mathcal{P}(\lambda)$, $\lambda > 0$: $\mathbb{P}(X = k) = \exp(-\lambda)\lambda^k/k!$, pour $k \in \mathbb{N}$.
2. **Loi géométrique** $\mathcal{G}(p)$, $p \in]0; 1[$: $\mathbb{P}(X = k) = p(1-p)^{k-1}$, pour $k \in \mathbb{N}\backslash\{0\}$.

3 Simulation par méthode d'acceptation-rejet

Les deux références principales pour cette section sont [BC07] et [Yca02]. On s'intéresse à la méthode de rejet pour générer des échantillons de lois uniformes et de loi à densité par rapport à la mesure de Lebesgue. La méthode permet également de simuler des lois discrètes (voir [Yca02]).

3.1 Simulation de lois uniformes par méthode de rejet

La méthode de rejet permet de simuler des variables de lois uniformes sur des boréliens bornés de $\mathbb{R}^d$. Elle est basée sur la proposition suivante.

Proposition 4 *Soit λ la mesure de Lebesgue sur $\mathbb{R}^d$. Soient A et B deux boréliens de $\mathbb{R}^d$, tels que*

$$B \subset A, \qquad 0 < \lambda(B) \leq \lambda(A) < \infty.$$

Soient $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme $\mathcal{U}_A$ uniforme sur A, et $T = \inf\{n \geq 1,\ X_n \in B\}$. Alors,

1. *La variable aléatoire T suit la loi géométrique $\mathcal{G}(p)$ de paramètre $p = \lambda(B)/\lambda(A)$.*
2. *Les variables T et X_T sont indépendantes.*
3. *La variable X_T suit la loi $\mathcal{U}_B$ uniforme sur B.*

Méthode de simulation de lois uniformes par rejet. Pour simuler une variable Y de loi $\mathcal{U}_B$, on commence donc par choisir un ensemble A tel que $B \subset A$ et pour lequel on va savoir simuler des variables de loi $\mathcal{U}_A$ (typiquement, A sera un pavé, voir Exercice 1). On tire ensuite des variables de loi $\mathcal{U}_A$ tant qu'elles n'appartiennent pas à B. La première appartenant à B convient pour Y.

Le nombre de tirages à générer pour obtenir une réalisation de Y suit la loi géométrique $\mathcal{G}(\lambda(B)/\lambda(A))$: il faut donc en moyenne $\lambda(A)/\lambda(B)$ tirages (on a donc tout intérêt à choisir A pas trop "grand" par rapport à l'ensemble B de départ).

Exercice 5

1. Implémenter le tirage d'un échantillon de **loi uniforme sur une ellipse**. On rappelle qu'une équation cartésienne d'une ellipse est donnée par $(x/a)^2 + (y/b)^2 = 1$, où $a > 0$ et $b > 0$ sont les longueurs des demi-axes.
2. Afficher sur une même figure un échantillon obtenu et le tracé de l'ellipse.
3. Vérifier la loi des grands nombres en calculant la moyenne empirique du nombre de points tirés appartenant à un rectangle inclu dans l'ellipse. On rappelle que l'aire à l'intérieur de l'ellipse dont l'équation est donnée ci-dessus est πab.

3.2 Simulation de lois continues par méthode de rejet

3.2.1 Cas des lois à densité bornée et à support compact

On a tout d'abord le résultat suivant.

Proposition 5 *Soient f une densité de probabilité sur $\mathbb{R}$, et D_f son hypographe :*

$$D_f = \{(x, u) \in \mathbb{R} \times \mathbb{R}_+,\ 0 < u < f(x)\}.$$

Si un vecteur aléatoire (X, U) suit la loi $\mathcal{U}_{D_f}$ uniforme sur D_f, alors X suit la loi de densité f (et la loi conditionnelle de U sachant "$X = x$" est la loi uniforme $\mathcal{U}_{[0;f(x)]}$).

Ainsi, pour tirer une variable de loi de densité f, il suffit de tirer un point au hasard sous la courbe représentative de f, et de prendre son abscisse. En combinant ce résultat avec la Proposition 4, on obtient le corollaire suivant.

Corollaire 1 *Soit f une densité de probabilité à support compact $[a; b]$ et bornée par une constante M. Soient $(X_n, U_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi $\mathcal{U}_{[a;b]\times[0;M]}$ et $T = \inf\{n \geq 1,\ U_n \leq f(X_n)\}$. Alors,*

1. *La variable aléatoire T suit la loi géométrique $\mathcal{G}(p)$ de paramètre $p = 1/(M(b - a))$.*
2. *Les variables T et X_T sont indépendantes.*
3. *La variable X_T suit la loi de densité f.*

Méthode de simulation de lois à densités bornées à support compact par rejet.
Pour générer une variable de densité f bornée par M à support compact $[a; b]$, on simule des variables uniformes sur $[a; b] \times [0; M]$ (qui est le pavé contenant le graphe de f), jusqu'à ce que l'un des points tirés appartiennent à l'hypographe de f. On prend alors pour réalisation de la loi f l'abscisse de ce point.

Exercice 6 Utiliser la méthode de rejet pour simuler un échantillon de loi de densité

$$f : x \mapsto \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - x^2}\,\mathbf{1}_{[-2\sigma;2\sigma]}(x), \qquad \sigma > 0.$$

Cette loi, dite **loi du demi-cercle de Wigner**, est la loi limite universelle pour le spectre de matrices aléatoires symétriques (théorème de Wigner, voir [BC07, Section 2.10]).

3.2.2 Généralisation

Le résultat suivant peut-être considéré comme une réciproque à la Proposition 5.

Proposition 6 *Soient g une densité de probabilité sur $\mathbb{R}$, et D_g son hypographe. Soient X une variable aléatoire de densité g, et U de loi uniforme $\mathcal{U}_{[0;1]}$, indépendante de X. Alors $(X, g(X)U)$ suit la loi $\mathcal{U}_{D_g}$ uniforme sur D_g.*

Ainsi, pour tirer un vecteur aléatoire de loi uniforme sur l'hypographe d'une densité g, il suffit de tirer une variable de loi de densité g et une variable indépendante de loi $\mathcal{U}_{[0;1]}$. En tenant compte des propositions 4, 5, et 6, on en déduit le résultat plus général suivant.

Corollaire 2 *Soit f une densité de probabilité sur $\mathbb{R}$. On suppose qu'il existe une autre densité g et une constante $c > 0$ telles que*

$$f \leq cg.$$

Soient $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d. de loi de densité g, $(U_n)_{n\geq 1}$ une autre suite i.i.d. de loi $\mathcal{U}_{[0;1]}$, indépendante de $(X_n)_n$ et $T = \inf\{n \geq 1,\ cU_n g(X_n) \leq f(X_n)\}$. Alors,

1. *La variable aléatoire T suit la loi géométrique $\mathcal{G}(1/c)$.*
2. *Les variables T et X_T sont indépendantes.*
3. *La variable X_T suit la loi de densité f.*

Méthode de simulation de lois à densité par rejet. Pour simuler une variable de densité f, on peut commencer par chercher une loi plus simple à simuler de densité g telle que $f \leq cg$. Puis on simule donc des variables U_n uniformes sur $[0;1]$ et des variables X_n de densité g (indépendantes) jusqu'à ce que $cU_n g(X_n) \leq f(X_n)$. On prend alors pour réalisation de la loi f la variable X_n correspondante.

Exercice 7

1. Soit $f : x \mapsto \exp(-x^2/2)/\sqrt{2\pi}$ la densité de la loi $\mathcal{N}(0,1)$, et $g : x \mapsto \exp(-|x|)/2$ la densité de la loi de Laplace. Justifier que $f \leq cg$ avec $c = \sqrt{2e/\pi}$. En déduire une fonction simulant un échantillon de **loi normale** $\mathcal{N}(0;1)$.
2. Soit $f : x \mapsto c_{a,b}x^{a-1}(1-x)^{b-1}\mathbf{1}_{[0;1]}(x)$ avec $c_{a,b} = \Gamma(a+b)/(\Gamma(a)\Gamma(b))$ la densité de la **loi beta** $\mathcal{B}(a,b)$ $a,b > 0$, et $g : x \mapsto ax^{a-1}\mathbf{1}_{[0;1]}(x)$. Justifier que $f \leq cg$ avec $c = c_{a,b}/a$, si $b > 1$. En déduire une fonction simulant un échantillon de loi $\mathcal{B}(a,b)$ $a > 0$, $b > 1$.

4 Méthodes particulières

En utilisant la définition ou la caractérisation de certaines lois, on obtient des procédés particuliers de simulations. On en a déjà vu à l'Exercice 1 par exemple, pour obtenir des lois uniformes discrètes ou continues sur des intervalles quelconques à partir de la loi uniforme sur $[0;1]$. Les exercices suivant donnent quelques autres exemples.

4.1 Quelques cas particuliers

Exercice 8 A partir de la loi de Bernoulli.

1. **Loi binomiale.** Montrer que si $(X_1,\dots,X_n)$ est un échantillon de loi $\mathcal{B}(p)$, $p \in]0;1[$, alors $S = X_1 + \cdots + X_n$ suit la loi binomiale $\mathcal{B}(n,p)$. En déduire une fonction ne contenant qu'une instruction permettant de simuler un échantillon de taille $n_1 \times n_2$ de variables de loi $\mathcal{B}(N,p)$.
2. **Loi géométrique.** Montrer que si $(X_i)_{i\geq 1}$ est une suite *i.i.d.* de variables de loi $\mathcal{B}(p)$, alors $G = \inf\{i \geq 1,\ X_i = 1\}$ suit la loi géométrique $\mathcal{G}(p)$.

Exercice 9 A partir de la loi exponentielle (1).

1. **Loi géométrique bis.** Montrer que si X suit une loi exponentielle $\mathcal{E}(\lambda)$, $\lambda > 0$, alors $\lfloor X+1 \rfloor$ suit la loi géométrique $\mathcal{G}(1-\exp(-\lambda))$ ($\lfloor \cdot \rfloor$ est la partie entière).
2. **Exemples de lois gamma.** La loi gamma $\gamma(a,\lambda)$, $a > 0$, $\lambda > 0$ a pour densité $x \mapsto (\lambda^a/\Gamma(a))x^{a-1}\exp(-\lambda x)\mathbf{1}_{x>0}$.
   (a) Justifier que la loi gamma $\gamma(1,\lambda)$ est la loi exponentielle $\mathcal{E}(\lambda)$.
   (b) Montrer que si X_1 et X_2 sont deux variables indépendantes de lois respectives $\gamma(a_1,\lambda)$ et $\gamma(a_2,\lambda)$, $a_1, a_2 > 0$, alors $X_1 + X_2$ suit la loi $\gamma(a_1+a_2,\lambda)$. En déduire une méthode de simulation pour les lois $\gamma(k,\lambda)$, $k > 0$ entier.

   Remarque. Il existe des méthodes de rejet pour simuler les autres lois gamma.

Exercice 10 A partir de la loi exponentielle (2). Simulation de la loi de Poisson.

1. Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires *i.i.d.* de loi exponentielle $\mathcal{E}(\lambda)$, $\lambda > 0$. Soient, pour $n \geq 1$, $S_n = X_1 + \cdots + X_n$, $S_0 = 0$, et pour $t \geq 0$,

$$N_t = \sum_{n=1}^{\infty} \mathbf{1}_{S_n \leq t} = \sup\{n \geq 0,\ S_n \leq t\} = \inf\{n \geq 0,\ S_{n+1} > t\}.$$

   Le processus $(N_t)_{t\geq 0}$ est appelé **processus de Poisson** homogène issu de 0 d'intensité λ. Montrer que N_t suit la loi de Poisson $\mathcal{P}(\lambda t)$. On pourra remarquer que $\{N_t = n\} = \{S_n \leq t < S_{n+1}\}$, puis utiliser la loi de S_n (voir Exercice 9).
2. En déduire que si $(U_i)_{i\geq 1}$ est une suite de variables aléatoires *i.i.d.* de loi $\mathcal{U}_{[0;1]}$, alors,

$$X = \inf\left\{n \geq 0,\ \prod_{i=1}^{n+1} U_i < \exp(-\lambda)\right\}$$

   suit une loi de Poisson $\mathcal{P}(\lambda)$.

Exercice 11 Simulation de la loi de Cauchy.

1. Montrer que si (U,V) suit la loi uniforme $\mathcal{U}_D$ sur le disque unité $D = \{(u,v) \in \mathbb{R}^2,\ u^2+v^2 \leq 1\}$ de $\mathbb{R}^2$, alors V/U suit la loi de Cauchy $\mathcal{C}(1)$ (penser au changement de variables en coordonnées polaires).
2. En déduire un algorithme de simulation de la loi $\mathcal{C}(1)$.

4.2 Simulation de lois normales

4.2.1 Simulation de lois normales réelles

L'exercice suivant donne deux méthodes de simulation de lois normales fondées sur les coordonnées polaires.

Exercice 12 1. **Méthode de Box-Müller.**

(a) Soit

$$\begin{cases} X = \sqrt{W}\cos(\Theta) \\ Y = \sqrt{W}\sin(\Theta) \end{cases} \quad \text{avec } W \sim \mathcal{E}(1/2) \text{ et } \Theta \sim \mathcal{U}_{[0;2\pi]}.$$

Montrer qu'alors les variables X et Y sont indépendantes et de même loi $\mathcal{N}(0,1)$.

   (b) En déduire un algorithme de simulation d'un $n-$échantillon de loi normale $\mathcal{N}(0;1)$ (n pair ou impair !).

2. **Algorithme polaire-rejet.**

   (a) Soit (U,V) un couple de loi $\mathcal{U}_D$ uniforme sur le disque unité $D=\{(u,v)\in\mathbb{R}^2,\ u^2+v^2\le 1\}$. On considère $(U,V)=(R\cos(\Theta),R\sin(\Theta))$ l'écriture de ces variables en coordonnées polaires. Montrer que R et Θ sont deux variables indépendantes, que Θ suit la loi uniforme $\mathcal{U}_{[0;2\pi]}$, et R suit la loi beta $\mathcal{B}(2,1)$. En déduire que $-4\ln(R)$ suit la loi exponentielle $\mathcal{E}(1/2)$, puis que

$$\begin{cases} X = 2\sqrt{-\ln(R)}\cos(\Theta) \\ Y = 2\sqrt{-\ln(R)}\sin(\Theta) \end{cases}$$

   forme un couple de variables indépendantes et de même loi $\mathcal{N}(0,1)$.

   (b) En déduire un autre algorithme de simulation d'un $n-$échantillon de loi normale $\mathcal{N}(0,1)$, en mettant en oeuvre la méthode de rejet pour simuler la loi $\mathcal{U}_D$.

3. Comment obtenir une variable Z de loi normale $\mathcal{N}(\mu,\sigma^2)$ ($\mu\in\mathbb{R}$, $\sigma^2\in\mathbb{R}_+^*$) à partir d'une variable N de loi normale $\mathcal{N}(0,1)$?

4.2.2 Simulation de lois normales multidimensionnelles.

On note $\mathcal{N}_d(m,K)$ la loi normale multidimensionnelle de $\mathbb{R}^d$ de vecteur moyenne $m\in\mathbb{R}^d$ et de matrice de covariance $K\in\mathcal{M}_d(\mathbb{R})$ (matrice symétrique positive). Il s'agit de la loi d'un vecteur gaussien.

1. **Simulation d'un vecteur gaussien standard.** Un tel vecteur gaussien a pour moyenne $m=0$ et pour matrice de covariance $K=I_d$ (matrice identité de dimension d). Ses composantes sont des variables aléatoires *i.i.d.* de loi $\mathcal{N}(0,1)$, donc les méthodes de l'Exercice 12 permettent de simuler un tel vecteur.

2. **Simulation d'un vecteur gaussien quelconque.** On utilise le résultat suivant :

   Lemme 7 *Soient $m\in\mathbb{R}^d$, $K\in\mathcal{M}_d(\mathbb{R})$ symétrique positive, et X un vecteur de loi $\mathcal{N}_d(0,I_d)$. Si C est une matrice carrée telle que $K=C^tC$ alors $Y=CX+m$ est un vecteur de loi $\mathcal{N}_d(m,K)$.*

   Le problème de la simulation de la loi $\mathcal{N}_d(m,K)$ se ramène donc à celui de la construction d'une matrice C telle que $K=C^tC$. Si K est symétrique définie positive, la ***décomposition de Cholesky*** assure l'existence d'une telle matrice C, triangulaire inférieure et inversible (avec unicité si on impose la positivité des coefficients diagonaux de C). Ce résultat se démontre en utilisant le procédé d'orthonormalisation de Gram-Schmidt : si K est symétrique définie positive, c'est la matrice d'un certain produit scalaire sur $\mathbb{R}^d$, dans la base canonique :

$$(x,y)\mapsto\langle x,y\rangle={}^txKy.$$

   Le procédé de Gram-Schmidt permet d'obtenir une base orthonormée pour ce produit scalaire à partir de la base canonique de $\mathbb{R}^d$. La matrice du produit scalaire dans cette nouvelle base est I_n, et on passe de K à I_n à l'aide de la matrice de passage P de la base canonique à la nouvelle base, par la formule $I_n={}^tPKP$. Par construction (voir le détail du procédé de Gram-Schmidt), la matrice P est triangulaire supérieure. En posant $C={}^tP^{-1}$, on a le résultat cherché.

La fonction `chol` de Scilab appliquée à une matrice K symétrique définie positive retourne une matrice triangulaire supérieure L telle que $K = {}^tLL$. Il suffit donc de prendre $C = {}^tL$.

5 Simulation de lois mélange

5.1 Méthode générale

Définition 8 *Soit $(\mu_\theta)_{\theta\in\Theta}$ une famille quelconque de probabilités sur un espace mesurable $(\Omega, \mathcal{F})$. Soit π une probabilité sur Θ. La* ***loi mélange*** *de la famille $(\mu_\theta)_{\theta\in\Theta}$ selon π est la mesure μ définie par*

$$\forall F \in \mathcal{F}, \;\; \mu(F) = \int_\Theta \mu_\theta(F) d\pi(\theta). \tag{1}$$

Remarque. On vérifie que μ est bien une probabilité.

Méthode de simulation d'une loi mélange. Si μ est une probabilité s'exprimant comme le mélange de probabilités $(\mu_\theta)_{\theta\in\Theta}$ selon une loi π sur Θ, pour obtenir une variable X de loi μ, on tire une réalisation d'une variable Y suivant la loi π, puis sachant que $Y = \theta$, on tire X selon la loi $\mu_Y = \mu_\theta$.

La méthode consiste à décomposer la loi d'une variable aléatoire en lois conditionnelles plus faciles à simuler. On peut donc la reformuler ainsi.

Méthode de simulation d'une loi mélange - autre formulation. On souhaite générer une variable aléatoire réelle X de loi $\mathbb{P}_X$. Si on sait tirer y une réalisation d'une variable Y de loi $\mathbb{P}_Y$, ainsi qu'une réalisation x de la loi conditionnelle $\mathbb{P}_{X|Y=y}$ de X sachant $Y = y$, alors on obtient (x, y) réalisation du couple (X, Y). En particulier, x est une réalisation de X.

On a ainsi décomposer $\mathbb{P}_X$ comme mélange des lois conditionnelles $\mathbb{P}_{X|Y=y}$: pour tout borélien A,

$$\mathbb{P}_X(A) = \int_{y\in\mathbb{R}} \mathbb{P}_{X|Y=y}(A) d\mathbb{P}_Y(y).$$

5.2 Cas des mélanges finis

Avec les notations ci-dessus, on s'intéresse au cas où π est une loi discrète à support fini, donc de la forme $\pi = \sum_{i=1}^r p_i\delta_i$. La loi μ s'exprime alors comme combinaison linéaire convexe des p_i. En effet, la réécriture de (1) est dans ce cas :

$$\forall F \in \mathcal{F}, \;\; \mu(F) = \sum_{i=1}^r p_i\mu_i(F).$$

La procédure est alors la suivante.

Méthode de simulation d'un mélange fini. Pour générer une variable aléatoire réelle X de loi $\sum_{i=1}^r p_i\mu_i$ (avec $0 \le p_i \le 1$ pour tout i, $\sum_{i=1}^r p_i = 1$, et $(\mu_i)_i$ une famille de probabilités). On choisit l avec probabilité p_l (simulation d'une loi discrète finie). Puis on tire X selon la loi μ_l.

Remarques.

1. Si $F_1, \ldots, F_r$ sont les fonctions de répartition respectives de $\mu_1, \ldots, \mu_r$, alors μ a pour fonction de répartition $\sum_{i=1}^r p_i F_i$.
2. Si les lois $\mu_1, \ldots, \mu_r$ admettent des densités $f_1, \ldots, f_r$, alors le mélange aussi, et sa densité est $\sum_{i=1}^r p_i f_i$.

Exercice 13 Implémenter une fonction permettant de simuler le mélange de deux lois gaussiennes.
Remarque. Les modèles de mélange gaussiens sont très utilisés en statistique, par exemple pour modéliser une population constituée de sous-populations se répartissant selon des lois normales différentes.

Références

[BC07] Bernard Bercu et Djalil Chafaï : *Modélisation stochastique et simulation - Cours et applications*. 2007.

[CGCDM11] Marie Cottrell, Valentine Genon-Catalot, Christian Duhamel et Thierry Meyre : *Exercices de probabilités : licence, master, écoles d'ingénieurs*. Cassini, 2011.

[RS09] Vincent Rivoirard et Gilles Stoltz : *Statistique en action*. Cassini, 2009.

[Yca02] Bernard Ycart : *Modèles et algorithmes markoviens*, volume 39. Springer, 2002.