# Probabilités V

STEP, MINES ParisTech*

10 janvier 2020 (`#f420b20`)

## Table des matières



*Ce document est un des produits du projet `boisgera/CDIS`, initié par la collaboration de (S)ébastien Boisgérault (CAOR), (T)homas Romary et (E)milie Chautru (GEOSCIENCES), (P)auline Bernard (CAS), avec la contribution de Gabriel Stoltz (Ecole des Ponts ParisTech, CERMICS). 

# Intégrale de Monte-Carlo

Les méthodes de Monte-Carlo ont été développées initialement par des physiciens dans les années 1950 (notamment par les travaux de Metropolis, Ulam, Hastings, Rosenbluth) pour calculer des intégrales (déterministes) à partir de méthodes probabilistes numériquement assez économiques. Le nom a été donné en référence au célèbre casino du fait du caractère aléatoire de ces méthodes.

Les méthodes de simulation sont basées sur la production de nombres aléatoires, distribués selon une certaine loi de probabilité. Dans de nombreuses applications, pour une certaine fonction $h$, on souhaite calculer, pour une variable aléatoire $X$ de loi $\mathbb{P}_X$

$$\mathcal{I} = \mathbb{E}\left(h(X)\right) = \int_{\mathbb{R}^d} h(x)\mathbb{P}_X(dx),$$

En général, même si on sait évaluer $h$ en tout point, on ne peut pas calculer formellement l'intégrale $\mathcal{I}$. Le calcul d'intégrale par la méthode Monte-Carlo consiste dans sa version la plus simple à générer un *échantillon* $(X_1, \ldots, X_n) \sim_{i.i.d.} \mathbb{P}_X$, et à estimer $\mathcal{I}$ par la moyenne empirique

$$M_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i),$$

où i.i.d signifie indépendant et identiquement distribué. En effet, d'après la loi forte des grands nombres, si $h(x)$ est $\mathbb{P}_X$-intégrable, on a l'assurance que

$$M_n(h) \to \int h(x)\mathbb{P}_X(dx) \text{ p.s.}$$

Si de plus, $h(X)^4$ est intégrable la vitesse de convergence de $M_n(h)$ peut être évaluée, puisque la variance

$$\mathbb{V}(M_n(h)) = \frac{1}{n}\int \left(h(x) - \mathcal{I}\right)^2 \mathbb{P}_X(dx)$$

peut également être estimée à partir de l'échantillon $(X_1, \ldots, X_n)$ par la quantité

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} \left(h(X_i) - M_n(h)\right)^2.$$

Le théorème central limite assure alors que pour $n$ grand,

$$\frac{M_n(h) - \mathcal{I}}{\sigma_n}$$

suit approximativement une loi $\mathcal{N}(0,1)$ [1]. Cette propriété conduit à la construction de tests de convergence et de bornes de confiance asymptotiques pour $M_n(h)$. Par exemple, on aura

$$\mathbb{P}\left(\mathcal{I} \in [M_n(h) - 1.96\sigma_n, M_n(h) + 1.96\sigma_n]\right) \approx 0.95,$$

1. ce résultat sera démontré dans le cours de science des données au second semestre.

où $1.96 = \Phi^{-1}(0,975)$ avec $\Phi$ la fonction de répartition de la loi normale centrée réduite.

En outre, cette propriété indique que la vitesse de convergence de $M_n(h)$ est de l'ordre de $\sqrt{n}$ et ce indépendamment de la dimension du problème. Cela explique la supériorité de cette méthode par rapport aux méthodes d'intégration numérique déterministes dont les vitesses de convergence décroissent rapidement (exponentiellement) avec la dimension du problème.

# Génération de nombres pseudo-aléatoires

Les ordinateurs sont des machines déterministes. Il peut sembler paradoxal de leur demander de générer des nombres aléatoires [2]. En réalité, les algorithmes de génération de nombres aléatoires vont produire des séquences de nombres déterministes qui vont avoir l'aspect de l'aléatoire. On s'intéresse ici à la génération de nombres uniformes sur $]0,1[$ (on exclut les bornes par commodité), puisque, comme on le verra dans la suite, il s'agit de l'ingrédient de base de toutes les méthodes de simulation stochastique.

### Définition — Générateur de nombres uniformes pseudo-aléatoires

Un *générateur de nombres uniformes pseudo-aléatoires* est un algorithme qui étant donné une valeur initiale $u_0$ et une transformation $T$ produit une séquence $u_i = T^i(u_0),\ \ i \in \mathbb{N}$ de valeurs dans ]0,1[.

Pour tout $n$, les valeurs $(u_1, \ldots, u_n)$ reproduisent le comportement d'une suite de variables aléatoires $(V_1, \ldots, V_n)$ i.i.d de loi uniforme sur ]0,1[, lorsqu'on les compare au travers d'un ensemble de tests statistiques [3], par exemple que la corrélation entre deux nombres successifs soit suffisamment faible.

### Exemple : la méthode des congruences

Cet algorithme, dû à Lehmer (1951), est l'un des premiers à avoir été proposé et implémenté. Il repose sur 2 paramètres :

— le multiplicateur $a$
— le modulo $m$

Etant donné $u_0 \in ]0,1[$, la séquence de nombres est générée par la transformation suivante :

$$u_{n+1} = \frac{(m\, a\, u_n) \bmod m}{m}$$

2. Von Neumann (1951) résume ce problème très clairement : "Any one who considers arithmetical methods of reproducing random digits is, of course, in a state of sin. As has been pointed out several times, there is no such thing as a random number — there are only methods of producing random numbers, and a strict arithmetic procedure of course is not such a method."

3. Par exemple, la suite de tests Die Hard, due à Marsaglia.

On peut remarquer que l'algorithme va produire une séquence de valeurs qui sera périodique, c'est-à-dire qu'après un certain nombre d'itérations, la suite se répétera, et qu'il pourra fournir au plus $m-1$ valeurs différentes. Ce trait est commun à tous les générateurs de nombre aléatoires et est lié aux limitations matérielles des ordinateurs (on ne peut représenter qu'un nombre fini de nombres). Le choix des valeurs de $a$ et $m$ est par conséquent crucial. Il existe des critères qui permettent de s'assurer du bon comportement de cette suite :

— $m$ est un nombre premier (le plus grand possible)
— $p = m - 1/2$ est un nombre premier
— $a^p = -1 \bmod m$

Ce critère assure une période pleine et donc que tous les nombres $(1/m, \ldots, (m-1)/m)$ seront générés. Cependant, les nombres générés ne sont pas indépendants, pas même non corrélés. On peut montrer que la corrélation entre 2 nombres successifs vaut approximativement $1/a$. Il convient donc de choisir $a$ suffisamment grand pour que celle-ci devienne négligeable. Par exemple, en prenant $a = 1000$ et $m = 2001179$, on obtient une période de 2001178 et une corrélation de l'ordre de $10^{-3}$.

Ce générateur de nombres uniformes pseudo-aléatoires, bien que rudimentaire, a ouvert la voie à des générateurs plus sophistiqués, toujours basés sur des opérations arithmétiques. Le plus couramment utilisé à ce jour est l'algorithme de Mersenne-Twister. Il s'agit du générateur par défaut de la plupart des logiciels tels que Python, R, MATLAB, Julia, MS Excel,... Il passe notamment avec succès toute la batterie de tests Die Hard. En particulier, sa période vaut $2^{19937} - 1$.

Pour certains usages, cet algorithme n'est cependant pas recommandé du fait de sa prédictibilité. C'est notamment un défaut rédhibitoire pour les applications en cryptographie. Il existe des variantes mieux adaptées à ce cas de figure. On notera enfin qu'il existe des générateurs de nombres aléatoires basés sur des phénomènes physiques comme un bruit électrique ou des phénomènes quantiques et donc parfaitement imprévisibles.

## Méthodes de simulation de variables aléatoires réelles

On a vu au chapitre II du cours Probabilités que l'on pouvait transformer des variables aléatoires réelles suivant certaines lois pour obtenir de nouvelles. Par exemple, si $X_1, \ldots, X_n$ sont $n \in \mathbb{N}^*$ variables gaussiennes centrées réduites indépendantes, alors $X_1^2 + \ldots, X_n^2$ suit une loi du $\chi^2$ à $n$ degrés de liberté. Dans le même esprit, on va voir ici comment simuler des v.a.r. de lois diverses à partir de la simulation de variables uniformes sur $]0, 1[$. On introduit une notation qui sera utile dans la suite : pour spécifier que deux v.a.r. $X$ et $Y$ ont même loi, on écrira $X \overset{\mathcal{L}}{=} Y$.

## Méthode d'inversion

L'objectif de ce paragraphe est de définir quand et comment il est possible de simuler une variable aléatoire réelle $X$ de fonction de répartition (f.d.r.) $F_X$ en transformant la simulation d'une variable aléatoire $U$ de loi Uniforme sur $]0,1[$. En d'autres termes, on cherche à déterminer les conditions sous lesquelles il est possible d'identifier une fonction borélienne $\psi : ]0,1[ \to \mathbb{R}$ telle que $X \overset{\mathcal{L}}{=} \psi(U)$.

Commençons par un cadre simple, où $F_X$ est **bijective** d'un intervalle non vide de $\mathbb{R}$ sur $]0,1[$.

### Proposition

Soient $X$ une variable aléatoire réelle de fonction de répartition $F_X$ et $U$ une variable uniforme sur $]0,1[$. S'il existe un intervalle non vide $]a,b[ \subset \mathbb{R}$ tel que $F_X : ]a,b[ \to ]0,1[$ est bijective, de bijection réciproque $F_X^{-1} : ]0,1[ \to ]a,b[$, alors $F_X^{-1}(U) \overset{\mathcal{L}}{=} X$ et $F_X(X) \overset{\mathcal{L}}{=} U$.

**Démonstration** Le premier résultat est immédiat : pour tout $x \in \mathbb{R}$, par croissance de $F_X$ et donc de $F_X^{-1}$, on a

$$\mathbb{P}\left(F_X^{-1}(U) \leq x\right) = \mathbb{P}\left(U \leq F_X(x)\right) = F_X(x).$$

Concernant le second, notons $G$ la fonction de répartition de la variable aléatoire $F_X(X)$. Comme $F_X(X)$ est à valeurs dans $]0,1[$, pour tout $x \in \mathbb{R}$ on a bien

$$G(x) = \left| \begin{array}{ll} 1 & \text{si } x \geq 1, \\ \mathbb{P}\left(F_X(X) \leq x\right) = \mathbb{P}\left(X \leq F_X^{-1}(x)\right) = x & \text{si } 0 < x < 1, \\ 0 & \text{sinon.} \end{array} \right.$$

■

**Exercice – Exemples d'application** Donner un algorithme de simulation d'une v.a.r. $X$ suivant une loi

- Uniforme sur un intervalle $I \subset \mathbb{R}$,
- Exponentielle de paramètre $\lambda \in \mathbb{R}_+^*$,
- de Cauchy, de densité $x \in \mathbb{R} \mapsto \left(\pi\left(1+x^2\right)\right)^{-1}$,
- de Laplace de paramètres $\mu \in \mathbb{R}$ et $s \in \mathbb{R}_+^*$, de densité $x \in \mathbb{R} \mapsto \frac{1}{2s} \exp\left\{-\frac{|x-\mu|}{s}\right\}$,
- Logistique de paramètres $\mu \in \mathbb{R}$ et $s \in \mathbb{R}_+^*$, de fonction de répartition $x \in \mathbb{R} \mapsto \left(1+\exp\left\{-\frac{x-\mu}{s}\right\}\right)^{-1}$ ?

Dans cette situation idéale, $\psi = F_X^{-1}$ est une solution à notre problème. Que se passe-t-il en revanche si $F_X$ n'est pas bijective ?

**Exercice** Proposer une méthode pour simuler un tir à pile ou face à partir de la simulation d'une variable uniforme sur $]0,1[$.

On a déjà vu au chapitre II que les fonctions de répartition de v.a.r. possèdent un nombre au plus dénombrable de points de discontinuité. Sur chaque intervalle où elles sont continues, on peut alors considérer qu'elles sont bijectives, quitte à réduire les zones de palier à un point. Cela permet de généraliser la notion de bijection réciproque pour ces fonctions.

**Définition**

Soit $F$ une fonction de répartition. On définit sa *réciproque généralisée* (aussi appelée *inverse généralisée* ou *pseudo-inverse*) comme la fonction

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

**Remarques**

- Cette fonction est bien définie sur tout $]0,1[$, car quel que soit $u$ dans cet intervalle, l'ensemble $\inf \{x \in \mathbb{R} : F(x) \geq u\}$ n'est ni vide ni égal à $\mathbb{R}$ tout entier. S'il était vide ou égal à tout $\mathbb{R}$ pour un certain $u_0 \in ]0,1[$, pour tout $x \in \mathbb{R}$ on aurait dans le premier cas $F(x) < u_0 < 0$ et dans le second $F(x) \geq u_0 > 1$. L'une comme l'autre de ces inégalités est impossible pour une fonction de répartition.
- La réciproque généralisée de la f.d.r. $F_X$ d'une v.a.r. $X$ est aussi appelée *fonction quantile*. On pourra notamment remarquer que $F_X^- \left(\frac{1}{2}\right)$ n'est autre que la médiane de $X$.
- Lorsque $F$ réalise une bijection d'un intervalle non vide $I \subset \mathbb{R}$ sur $]0,1[$, sa réciproque généralisée coïncide avec sa bijection réciproque.

On a alors le résultat suivant, qui stipule que $\psi = F_X^-$ est une solution universelle à notre problème. La preuve détaillée est donnée en Annexe.

**Théorème – Méthode d'inversion**

Soient $U$ une variable uniforme sur $]0,1[$ ainsi que $X$ une variable aléatoire réelle de fonction de répartition $F_X$ et de réciproque généralisée $F_X^-$. Alors $F_X^-(U) \overset{\mathcal{L}}{=} X$.

**Exercice – Exemples d'application** Donner un algorithme de simulation d'une v.a.r $X$ suivant une loi

- Binomiale de paramètres $n \in \mathbb{N}^*$ et $p \in ]0,1[$,
- de Poisson de paramètre $\lambda \in \mathbb{R}_+^*$,
- Uniforme sur l'union de deux segments non vides et disjoints $[a,b], [c,d] \subset \mathbb{R}$, de densité $x \in \mathbb{R} \mapsto (b-a+d-c)^{-1} 1_{[a,b]\cup[c,d]}(x)$ ?

**Limitations**

La méthode d'inversion peut sembler universelle pour simuler toute v.a.r. $X$ à partir de $U \sim \mathcal{U}_{]0,1[}$. Cependant, elle nécessite en pratique de disposer d'une expression analytique de $F_X$ pour pouvoir en déduire sa réciproque généralisée. Or ce n'est typiquement pas le cas de nombreuses lois usuelles comme la loi Normale ! On va donc déterminer d'autres procédures pour simuler des variables suivant de telles lois.

## Méthode du rejet

## Simulation de variables aléatoires gaussiennes : Box-Muller

Nous avons vu que la méthode d'inversion est inappropriée pour simuler une variable gaussienne, puisqu'elle requiert une expression analytique de la fonction de répartition cible. Il existe des méthodes basées sur une intégration numérique de la densité gaussienne puis une inversion de cette approximation de la f.d.r. mais elle ne sont pas optimales en temps de calcul. La méthode du rejet est quant à elle sous-optimale, dans le sens où toutes les variables uniformes générées ne sont pas directement utilisées (une partie, potentiellement grande, est rejetée). La loi normale étant fondamentale en probabilité, il est plus que souhaitable de pouvoir en trouver une méthode de simulation exacte et efficace.

George E. P. Box et Mervin E. Muller ont proposé en 1958 une telle méthode. Elle exploite la propriété d'invariance par rotation de la densité d'un couple de variables gaussiennes indépendantes centrées réduites.

**Proposition**

Soient $U$ et $V$ deux variables indépendantes, de loi Uniforme sur $]0,1[$. Alors les variables aléatoires $X = \sqrt{-2\ln(U)}\cos(2\pi V)$ et $Y = \sqrt{-2\ln(U)}\sin(2\pi V)$ sont indépendantes et suivent toutes deux une loi normale centrée réduite.

**Démonstration** On considère $(\widetilde{X}, \widetilde{Y})$ un vecteur aléatoire dont les deux composantes sont indépendantes, de loi normale centrée réduite. On note ses coordonnées polaires aléatoires $\widetilde{R}$ et $\widetilde{\Theta}$. On a vu dans le cours de Probabilités II que dans ce cas, $\widetilde{R}$ et $\widetilde{\Theta}$ sont indépendantes, la première de densité $f_{\widetilde{R}} : r \in \mathbb{R} \mapsto r\, e^{-\frac{r^2}{2}} 1_{\mathbb{R}_+^*}(r)$ et la seconde de loi uniforme sur $]0, 2\pi]$.

Or on remarque que $X$ et $Y$ ont la forme de coordonnées cartésiennes obtenues à partir d'un rayon et d'un angle : en posant $R = \sqrt{-2\ln(U)}$ et $\Theta = 2\pi V$ on obtient $X = R\cos(\Theta)$ et $Y = R\sin(\Theta)$. Par indépendance de $U$ et $V$, on sait déjà que $R$ et $\Theta$ sont indépendantes. Pour que le vecteur $(X, Y)$ ait la même distribution que $(\widetilde{X}, \widetilde{Y}) = \left(\widetilde{R}\cos(\widetilde{\Theta}), \widetilde{R}\sin(\widetilde{\Theta})\right)$, il suffit donc de montrer que $R \overset{\mathcal{L}}{=} \widetilde{R}$ et $\Theta \overset{\mathcal{L}}{=} \widetilde{\Theta}$.

— Commençons par étudier la loi de $R$, de fonction de répartition notée $F_R$. On remarque que la fonction $u \in ]0,1[ \mapsto \sqrt{-2\ln(u)} \in \mathbb{R}_+^*$ est bijective,

strictement décroissante. Ainsi, pour tout $r \in \mathbb{R}_-$ on a $\mathbb{P}(R \leq r) = 0$ et pour tout $r \in \mathbb{R}_+^*$ on a

$$\mathbb{P}(R \leq r) = \mathbb{P}\left(\sqrt{-2\ln(U)} \leq r\right) = \mathbb{P}\left(U \geq e^{-\frac{r^2}{2}}\right) = 1 - e^{-\frac{r^2}{2}}.$$

En d'autres termes, pour tout $r \in \mathbb{R}$,

$$F_R(r) = \left| \begin{array}{ll} 1 - e^{-\frac{r^2}{2}} & \text{si } r > 0, \\ 0 & \text{sinon,} \end{array} \right.$$

qui correspond exactement à la fonction de répartition de $\widetilde{R}$ : quel que soit $r \in \mathbb{R}$

$$\int_{-\infty}^{r} f_{\widetilde{R}}(x)\,dx = \left| \begin{array}{ll} \displaystyle\int_0^r x\, e^{-\frac{x^2}{2}}\,dx = \left[-e^{-\frac{x^2}{2}}\right]_0^r = 1 - e^{-\frac{r^2}{2}} & \text{si } r > 0, \\ 0 & \text{sinon.} \end{array} \right.$$

— Regardons maintenant la loi de $\Theta$, de fonction de répartition $F_\Theta$. Puisque la fonction $v \in ]0,1[ \mapsto 2\pi v \in ]0, 2\pi[$ est bijective strictement croissante, on a directement que pour tout $\theta \in \mathbb{R}$

$$F_\Theta(\theta) = \left| \begin{array}{ll} 1 & \text{si } \theta \geq 2\pi, \\ \mathbb{P}\left(V \leq \frac{\theta}{2\pi}\right) = \dfrac{\theta}{2\pi} & \text{si } \theta \in ]0, 2\pi[, \\ 0 & \text{si } \theta \leq 0, \end{array} \right.$$

qui n'est autre que la fonction de répartition d'une loi uniforme sur $]0, 2\pi[$.

■

Cette méthode permet de simuler directement deux variables gaussiennes centrées réduites indépendantes à partir de deux variables uniformes indépendantes. Pour simuler une variable gaussienne d'espérance $m \in \mathbb{R}$ et de variance $\sigma^2 \in \mathbb{R}_+^*$ quelconques, il suffit de se rappeler le résultat préliminaire de l'exercice *Combinaisons linéaires de variables aléatoires Gaussiennes indépendantes* du cours Probabilités II : si $X$ suit une loi normale centrée réduite, alors $\sigma X + m$ suit une loi normale d'espérance $m$ et de variance $\sigma^2$.

## Simulation d'un vecteur gaussien à densité

La simulation d'un vecteur gaussien dont la matrice de covariance est inversible est extrêmement aisée. En effet, on souhaite simuler un vecteur gaussien $X = (X_1, \ldots, X_d)$ à valeurs dans $\mathbb{R}^d$ d'espérance $m$ et de matrice de covariance $C$ définie positive donnés.

Puisque la matrice $C$ est inversible, elle admet une racine carrée, c'est-à-dire qu'il existe une matrice $N$ telle que $C = N\,N^t$. En effet, on peut par exemple décomposer $C$ de la manière suivante :

$$C = V\,D\,V^t$$

où $V$ est une matrice orthogonale et $D$ est la matrice diagonale dont les termes diagonaux sont les valeurs propres (toutes strictement positives) de $C$. Il suffit alors de prendre $N = V\,D^{1/2}$, où $D^{1/2}$ est la matrice diagonale dont les termes diagonaux sont les racines carrées des valeurs propres.

En pratique, il est coûteux numériquement d'effectuer le calcul des valeurs propres et des vecteurs propres de $C$. On va plutôt calculer sa *décomposition ou factorisation de Cholesky* qui permet d'écrire

$$C = L\,L^t$$

avec $L$ une matrice triangulaire inférieure. [4]

Soit maintenant un autre vecteur gaussien $Y = (Y_1, \ldots, Y_d)$ à valeurs dans $\mathbb{R}^d$ et de matrice de covariance l'identité, notée $I_d$. Autrement dit, les $Y_i$ sont des variables aléatoires gaussiennes centrées, réduites et indépendantes.

Alors, le vecteur $Z = m + L\,Y$ est gaussien, d'espérance $m$ et de matrice de covariance $C$. En effet, $Z$ est gaussien comme combinaison linéaire de variables aléatoires gaussiennes, $\mathbb{E}(Z) = \mathbb{E}(m + L\,Y) = m$ et $\mathbb{V}(Z) = \mathbb{E}((L\,Y)^2) = L I_d L^t = C$.

# Echantillonnage d'importance

On introduit dans cette section la méthode d'échantillonnage d'importance (importance sampling en anglais), que l'on appelle aussi, de manière plus intuitive, échantillonnage préférentiel, pour les lois à densité. Pour ce faire, nous allons commencer par un exemple qui montre qu'il peut être plus efficace de simuler des valeurs selon une loi différente de celle d'intérêt, autrement dit de modifier la représentation de l'intégrale $\mathcal{I}$ sous la forme d'une espérance calculée selon une autre densité.

### Exemple

Supposons que l'on s'intéresse à calculer la probabilité $p$ qu'une variable $X$ de loi de Cauchy standard soit plus grande que 2 (on peut le calculer directement et p=0.15)

$$p = \int_2^{+\infty} \frac{1}{\pi(1+x^2)}\,dx$$

Si on estime $p$ directement à partir d'un échantillon $(X_1, \ldots, X_n)$ simulé selon la loi de Cauchy standard soit

$$\widehat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i > 2},$$

4. Cette décomposition est très utile dans la résolution de systèmes linéaires de la forme $A\,x = b$, où $b$ est connu, $x$ inconnu et $A$ est définie positive. Cela revient à résoudre $L\,L^t\,x = b$. On pose alors $y = L^t\,x$ et on résoud d'abord $Ly = b$, ce qui est très rapide puisque $L$ est triangulaire inférieure (on commence par $y_1 = b_1/L_{11}$, puis $y_2 = (b_2 - L_{21}y_1)/L_{22}$, etc. en descendant). On résoud ensuite $L^t\,x = y$, ce qui est aussi très rapide pour la même raison (on commence par $x_n = y_n/L_{nn}$ puis on remonte).

la *variance de l'estimateur* $\mathbb{V}(\widehat{p}_1) = p(1-p)/n = 0.127/n$, puisque $\widehat{p}_1$ suit une loi binomiale de paramètre $(n, p)$. On peut réduire cette variance (et donc améliorer la qualité de l'estimateur) en tirant parti de la symétrie de la densité de la loi de Cauchy, en formant un second estimateur

$$\widehat{p}_2 = \frac{1}{n}\sum_{i=1}^{n} 1_{|X_i|>2},$$

dont la variance vaut $\mathbb{V}(\widehat{p}_2) = p(1-p/2)/2n = 0.052/n$. La relative inefficacité de ces méthodes est due au fait que la majeure partie des valeurs simulées seront en dehors de la zone d'intérêt $]2, +\infty[$. En passant par le complémentaire, on peut réécrire $p$ comme

$$p = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)} dx,$$

dont le second terme peut être vu comme l'espérance de $h(U) = \frac{2}{\pi(1+U^2)}$ avec $U \sim \mathcal{U}_{[0,2]}$. Tirant un échantillon $(U_1, \ldots, U_n)$ i.i.d. de loi uniforme sur $[0, 2]$, on obtient un troisième estimateur :

$$\widehat{p}_3 = \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n} \frac{2}{\pi(1+U_i^2)},$$

dont la variance vaut $\mathbb{V}(\widehat{p}_3) = (\mathbb{E}(h(X)^2) - \mathbb{E}(h(U))^2)/n = 0.0285/n$ (par intégration par parties). Enfin, on peut encore réécrire (voir Ripley (1987))

$$p = \int_0^{1/2} \frac{y^{-2}}{\pi(1+y^{-2})} dy,$$

qui peut être vue comme $\mathbb{E}\left(\frac{V^{-2}}{2\pi(1+V^{-2})}\right)$ avec $V \sim \mathcal{U}_{]0,1/2[}$. L'estimateur formé à partir de cette représentation et d'un échantillon $(V_1, \ldots, V_n)$ i.i.d. de loi uniforme sur $[0, 1/2]$ a une variance de $0.95\,10^{-4}/n$. Il est donc bien plus efficace que $\widehat{p}_1$ puisqu'il nécessite environ $\sqrt{10^3} = 32$ fois moins de simulations pour atteindre la même précision.

On a ainsi vu sur ce cas particulier que l'estimation d'une intégrale de la forme

$$\mathcal{I} = \mathbb{E}\left(h(X)\right) = \int_{\mathbb{R}^d} h(x) f(x) dx,$$

peut s'écrire de différentes manières, en faisant varier $h$ et $f$. Par conséquent, un estimateur "optimal" devrait tenir compte de l'ensemble de ces possibilités. C'est justement l'idée développée dans la méthode d'échantillonnage d'importance dont le principe est décrit dans la définition suivante :

**Définition**

La méthode d'*échantillonnage d'importance* est une évaluation de $\mathcal{I}$ basée sur la simulation d'un échantillon $X_1, \ldots, X_n$ de loi de densité $g$ et approximant :

$$\mathbb{E}_f\left(h(X)\right) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} h(X_i),$$

où la notation $\mathbb{E}_f$ signifie que l'espérance est calculée avec $X \sim f$. On appelle souvent les ratios $\frac{f(X_i)}{g(X_i)}$ les *poids d'importance* que l'on note $w_i$.

Cette méthode est basée sur la représentation suivante de $\mathcal{I}$ :

$$\mathbb{E}_f\left[h(X)\right] = \int h(x)\frac{f(x)}{g(x)}g(x)\,dx$$

que l'on appelle l'*identité fondamentale de l'échantillonnage d'importance* et l'estimateur converge du fait de la loi forte des grands nombres.

Cette identité indique qu'une intégrale du type $\mathcal{I}$ n'est pas intrinsèquement associée à une loi donnée. L'intérêt de l'échantillonnage d'importance repose sur le fait qu'il n'y a aucune restriction sur le choix de la densité $g$, dite *instrumentale*, que l'on peut donc choisir parmi les densités des lois que l'on sait simuler aisément. Il y a bien évidemment des choix qui sont meilleurs que d'autres. Remarquons tout d'abord que bien que l'estimateur proposé dans la définition ci-dessus converge presque sûrement, sa variance est finie si

$$\mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) = \mathbb{E}_f\left(h^2(X)\frac{f(X)}{g(X)}\right) = \int h^2(x)\frac{f^2(x)}{g(x)}\,dx < \infty.$$

On préconise alors l'usage de densités instrumentales $g$ dont la queue de distribution est plus épaisse que celle de $f$ pour éviter que cette variance puisse être infinie (on notera que cela dépend aussi de la fonction $h$ à intégrer). En pratique, on utilise généralement l'estimateur suivant, de variance finie et qui donne des résultats plus stables numériquement que celui de la définition :

$$\frac{\sum_{i=1}^n w_i h(X_i)}{\sum_{i=1}^n w_i}$$

où on a remplacé $n$ par la somme des poids d'importance. Puisque $\frac{1}{n}\sum_{i=1^n} w_i = \frac{1}{n}\sum_{i=1}^n \frac{f(x_i)}{g(x_i)}$ tend vers 1 quand $n \to \infty$, cet estimateur converge presque sûrement vers $\mathbb{E}_f\left(h(X)\right)$ par la loi forte des grands nombres (voir C. P. Robert and G. Casella (2004)).

Parmi les densités $g$ qui fournissent des estimateurs de variance finie, il est possible d'exhiber la densité optimale (au sens de la variance de l'estimateur associé) pour une fonction $h$ et une densité $f$ données.

**Théorème**

Le choix de $g$ qui minimise la variance de l'estimateur donné dans la définition ci-dessus est

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)dx}.$$

**Démonstration** Notons d'abord que

$$\mathbb{V}_g\left(\frac{h(X)f(X)}{g(X)}\right) = \mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) - \mathbb{E}_g\left(h(X)\frac{f(X)}{g(X)}\right)^2$$

et le second terme ne dépend pas de $g$. On minimise donc le premier terme. D'après l'inégalité de Jensen, on a

$$\mathbb{E}_g\left(h^2(X)\frac{f^2(X)}{g^2(X)}\right) \geq \mathbb{E}_g\left(|h(X)|\frac{f(X)}{g(X)}\right)^2 = \left(\int |h(x)|f(x)dx\right)^2,$$

qui nous donne une borne inférieure indépendante du choix de $g$. Elle est atteinte en prenant $g = g^*$. ■

Ce résultat formel n'a que peu d'intérêt pratique : le choix optimal de $g$ fait intervenir $\int |h(x)|f(x)dx$, qui est à une valeur absolue près la quantité que l'on souhaite estimer ! Il suggère néanmoins de considérer des densités $g$ telles que $|h|f/g$ est quasi constante et de variance finie. On se reportera au chapitre 3 de C. P. Robert and G. Casella (2004) pour des exemples où un bon choix de $g$ permet des améliorations considérables par rapport à des estimateurs de Monte-Carlo plus naïfs.

# Annexe

## Preuve de la méthode d'inversion

Pour pouvoir démontrer le théorème de la méthode d'inversion, il faut d'abord établir un certain nombre de propriétés de la réciproque généralisée d'une fonction de répartition. Elle peuvent être visualisées sur la figure REF ICI.

### Proposition

Soit $F$ une fonction de répartition. Alors sa réciproque généralisée $F^-$ satisfait les propriétés suivantes.

1. $F^-$ est croissante.
2. $\forall\, x \in \mathbb{R} : F^- \circ F(x) \leq x$.
3. $\forall\, u \in ]0,1[ : F \circ F^-(u) \geq u$ avec égalité si $u \in F(\mathbb{R})$.
4. $\forall\, (u,x) \in ]0,1[ \times \mathbb{R} : \{F(x) \geq u\} \Leftrightarrow \{x \geq F^-(u)\}$ et $\{F(x) < u\} \Rightarrow \{x \leq F^-(u)\}$.

**Démonstration** Pour tout $u \in ]0,1[$ on note $\mathcal{X}_u = \{x \in \mathbb{R} : F(x) \geq u\}$ l'image réciproque de $[u,1[$ par $F$.

1. Soit $(u,v) \in ]0,1[^2$. Si $u < v$ alors $\mathcal{X}_v \subset \mathcal{X}_u$ d'où $F^-(u) = \inf \mathcal{X}_u \leq \inf \mathcal{X}_v = F^-(v)$. La fonction $F^-$ est donc bien croissante.
2. Soit $x \in \mathbb{R}$, alors $F^- \circ F(x) = \inf\big\{z \in \mathbb{R} : F(z) \geq F(x)\big\} = \inf \mathcal{X}_{F(x)}$. Comme $F$ est croissante sur $\mathbb{R}$ on a $[x, +\infty[ \subseteq \mathcal{X}_{F(x)}$ donc $F^- \circ F(x) = \inf \mathcal{X}_{F(x)} \leq \inf [x, +\infty[ = x$.
3. Soit $u \in ]0,1[$.

— Puisque $\mathcal{X}_u$ est nécessairement non vide, il existe une suite décroissante $(x_n)_{n\in\mathbb{N}} \subseteq \mathcal{X}_u$ convergeant vers $\inf \mathcal{X}_u = F^-(u)$. La croissance de $F$ implique que la suite $\big(F(x_n)\big)_{n\in\mathbb{N}}$ est elle aussi décroissante, minorée par $u$ car $(x_n)_{n\in\mathbb{N}} \subseteq \mathcal{X}_u$ donc convergente. Sa limite est de même supérieure ou égale à $u$. Comme $F$ est continue à droite, cette dernière n'est autre que

$$\lim_{n\to+\infty} F(x_n) = F\left(\lim_{n\to+\infty} x_n\right) = F \circ F^-(u).$$

Nous avons donc bien $F \circ F^-(u) \geq u$.

— Supposons maintenant que $u \in F(\mathbb{R})$. Alors $\mathcal{X}_u^* = \{x \in \mathbb{R} : F(x) = u\} \neq \varnothing$. Il existe donc une suite décroissante $(x_n)_{n\in\mathbb{N}} \subseteq \mathcal{X}_u^*$ convergeant vers $\inf \mathcal{X}_u^* = F^-(u)$ par croissance de $F$. Comme $F$ est continue à droite en tout point de $\mathbb{R}$ et $F(x_n) = u$ pour tout $n \in \mathbb{N}$, on a bien

$$u = \lim_{n\to+\infty} F(x_n) = F\left(\lim_{n\to+\infty} x_n\right) = F \circ F^-(u).$$

4. Soit $(u, x) \in ]0, 1[ \times \mathbb{R}$.

— **Equivalence.** Supposons $F(x) \geq u$. On a $F^- \circ F(x) \geq F^-(u)$ par croissance de $F^-$ (propriété 1.) et $F^- \circ F(x) \leq x$ d'après la propriété 2. Réciproquement, supposons que $x \geq F^-(u)$. Alors par croissance de $F$ on a $F(x) \geq F \circ F^-(u)$ puis $F \circ F^-(u) \geq u$ d'après la propriété 3.

— **Implication.** Supposons $F(x) < u$. Tout $z \in \mathcal{X}_u$ vérifie $F(z) \geq u > F(x)$. Comme $F$ est croissante sur $\mathbb{R}$ on a donc $z \geq x$, ce qui implique $x \leq \inf \mathcal{X}_u = F^-(y)$.

■

FIGURE ICI

Nous pouvons maintenant établir la preuve du théorème de la méthode d'inversion.

**Démonstration – Méthode d'inversion** Soient $x \in \mathbb{R}$ et $(\Omega, \mathcal{A}, \mathbb{P})$ l'espace probabilisé sur lequel sont définies $U$ et $X$. D'après la propriété 4 ci-dessus, on a $\big\{\omega \in \Omega : F_X^-(U(\omega)) \leq x\big\} = \{\omega \in \Omega : U(\omega) \leq F(x)\}$, d'où

$$\mathbb{P}\left(F_X^-(U) \leq x\right) = \mathbb{P}\left(U \leq F_X(x)\right) = F_X(x).$$

■

# Références


C. P. Robert, and G. Casella. 2004. *Monte-Carlo Statistical Methods.* 2nd edition, Berlin: Springer.

Lehmer, Derrick H. 1951. "Mathematical Methods in Large-Scale Computing Units." *Annu. Comput. Lab. Harvard Univ.* 26: 141–46.

Ripley, Brian D. 1987. *Stochastic Simulation.* New York: Wiley.