Simulation des distributions de probabilité continues à une seule dimension.

1 Introduction

Soit X une variable aléatoire de distribution de probabilité quelconque donnée. On veut construir, à l'aide des ordinateurs, $(x_1, x_2, ..., x_n)$ un échantillon issue de X.
Une réalisation d'une suite de nombres au hasard (aléatoires) indépendants et uniformement distribués sur le ségment $[0, 1]$,

$$U_1, U_2, ... \tag{1}$$

est souvant utilisée pour obtenir une réalisation d'une suite de variables aléatoires indépendantes et arbitrairement distribuées.
L'algorithme le plus connu et le plus utilisé est de la méthode linéaire congruentielle avec $U_n = Z_n/m$ et $\{Z_n\}$ est la suite définie par la relation de récurrence:

$$Z_{n+1} = aZ_n + c \ \ [\text{mod}m]; \quad n = 0, 1, 2, ...$$

où Z_0 est la valeure initiale, a, c, m, sont des entiers positifs.
$\{U_n\}$ sont des nombres pseudo-aléatoires.
Un bon choix des entiers a, c, m et la valeur initiale Z_0, nous assure un comportement de ces $\{U_n\}$ très proche de celui de la suite (1).
Conditions pour un bon choix des valeurs qui définissent la méthode linéaire congruentielle:

1. c et m sont premiers entre eux;
2. $b = a - 1$ est dévisible par p pour n'importe quel nombre premier p qui est un diviseur de m;
3. b est divisible par 4 si m est divisible par 4.

2 Principes généraux

2.1 Méthode d'inversion

Dans sa version la plus simple, la méthode d'inversion se base sur le resultat, bien connu, suivant:

Proposition 1 *Supposons que la variable aléatoire X a une fonction de distribution F continue et strictement croissante, toujours que $0 < F(x) < 1$. Soit U une variable aléatoire de distribution uniforme sur $(0,1)$. Alors, la variable aléatoire $F^{-1}(U)$ a comme fonction de distribution F.*

Démonstration:
Soit F_X la fonction de distribution de $F^{-1}(U)$. Le résultat se déduit des égalités suivantes:

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

La deuxième égalité est justifiée par la monotonie de F, la troisième par $F(F^{-1}(U)) = U$ et la dernière parce que U a une distribution uniforme sur $(0,1)$. □

Cette proposition suggère que pour échantillonner à partir d'une variable aléatoire X que nous connaissons sa F^{-1}, nous pouvons générer des nombres U uniformes sur $(0,1)$ et faire $X = F^{-1}(U)$. Nous avons, alors, l' algorithme d'inversion suivant:
`Générer` $U \rightsquigarrow \mathcal{U}(0,1)$
`Faire` $X = F^{-1}(U)$
`Sortir` X.

Par conséquent, une condition minime pour pouvoir appliquer cette méthode est de savoir la forme explicite de F^{-1}. Cela est satisfait par plusieurs distributions, comme la uniforme, la exponentielle, la de Weibull, de Cauchy,... Remarquons que telle condition n'est pas suffisante; par exemple, pour la distribution beta, la simulation par inversion est théoriquement possible, mais elle peut être coûteuse. Parfois, nous disposons d'une bonne approximation de F^{-1}, avec laquelle nous pouvons utiliser la méthode par approximation.

Exemple 1 *(1) Génération d'une variable aléatoire $\mathcal{U}(a,b)$. Sa fonction de distribution est*

$$F(x) = \begin{cases} 0 & si \quad x < a \\ \frac{x-a}{b-a} & si \quad a \leq x < b \\ 1 & si \quad x \geq b \end{cases}$$

Dans le schéma général, il suffit de faire $X = F^{-1}(U) = a + (b-a)U$.
(2) Génération d'une variable de Weibull $\mathcal{W}(\alpha, 1)$. *Sa fonction de distribution est*

$$F(x) = \begin{cases} 0 & si \quad x < 0 \\ 1 - \exp(-x^{\alpha}) & si \quad x \geq 0 \end{cases}$$

Nous faisons $X = F^{-1}(U) = [-\ln(1-U)]^{\frac{1}{\alpha}}$, *ou bien* $X = (-\ln U)^{\frac{1}{\alpha}}$, *puisque* $1 - U \rightsquigarrow \mathcal{U}(0,1)$.

Remarquons que cette stratégie peut-être précieuse si nous avons des tables de F^{-1} ou de F, ou si nous décidons d'employer la fonction de distribution empirique.

2.2 Méthode du rejet

Pour la méthode d'inversion, il convient de savoir la fonction de distribution. Parfois, on connaît la fonction de densité mais pas la fonction de distribution; comme il arrive, par exemple, dans le cas de la distribution normale. Dans quelques uns de ces cas, on peut appliquer la méthode du rejet, introduite par Von Neumann (1951).

Supposons que nous désirons échantillonner à partir d'une variable aléatoire X avec fonction de densité f. Nous ne savons pas le faire directement, mais nous disposons d'un prodédé pour échantillonner à partir d'une fonction de densité g telle que $f(x) \leq ag(x)$ pour tout x (avec $a < \infty$). La méthode du rejet suggère:
`Jusqu'où` $U \leq f(X)/ag(X)$
`Générer` $X \rightsquigarrow g$
`Générer` $U \rightsquigarrow \mathcal{U}(0,1)$
`Sortir` X .

La méthode du rejet est équivalente à générer des valeurs $Y \rightsquigarrow \mathcal{U}(0, ag(X))$ et les accepter si $Y \leq f(X)$. Prouvons que la procedure est correcte. Pour cela, nous devons voir que

$$P(X \leq x / X \text{acceptée}) = F(x),$$

où F est la fonction de distribution de X. Nous avons

$$P(X \leq x / X \text{acceptée}) = \frac{P(X \leq x, X \text{acceptée})}{P(X \text{acceptée})}$$

En plus,

$$P(X\text{acceptée}) = \int_{-\infty}^{+\infty} P(Y \leq f(X)/X = z)g(z)dz$$

$$P(X\text{acceptée}) = \int_{-\infty}^{+\infty} \frac{f(z)}{ag(z)}g(z)dz = \frac{1}{a} \qquad (2)$$

$$P(X \leq x, X\text{acceptée}) = \int_{-\infty}^{+\infty} P(X \leq x, Y \leq f(X)/X = z)g(z)dz$$

$$P(X \leq x, X\text{acceptée}) = \int_{-\infty}^{x} \frac{f(z)}{ag(z)}g(z)dz = \frac{F(x)}{a}$$

d'où le résultat.

L'équation (2) indique que dans chaque itération de l'algorithme, on accepte une valeur d'une manière indépendante avec probabilité $1/a$, ce qui assure l' efficacité de la procedure. Le nombre des itérations ou des "tests" indépendants avant l'acceptation d'une valeur suit une distribution géométrique de paramètre $1/a$; avec a le nombre esperé de "tests". Plus que ce a soit proche de 1 (toujours on aura $a \geq 1$) plus la méthode sera efficace, parce que les densités f et g seront plus proches.

Exemple 2 *Génération de la variable beta $\mathcal{B}(3,4)$ de fonction de densité*

$$f(x) = 60x^2(1-x)^3; \qquad 0 < x < 1$$

Prenons comme fonction g la densité de la distribution $\mathcal{U}(0,1)$. Déterminons une constante a telle que $f \leq ag$. Le maximum de la fonction $f(x)/g(x)$ est atteind en $x = 2/5$.

Ainsi,

$$\frac{f(x)}{g(x)} \leq 60\left(\frac{2}{5}\right)^2\left(1-\frac{2}{5}\right)^3 = \frac{1296}{625} \equiv a$$

Donc,

$$\frac{f(x)}{ag(x)} = \frac{3125}{108}x^2(1-x)^3.$$

L'algorithme du rejet devient

`Jusqu'où` $U_2 \leq \frac{3125}{108}U_1^2(1-U_1)^3$
`Générer` $U_1, U_2 \rightsquigarrow \mathcal{U}(0,1)$
`Sortir` U_1

Le nombre moyen d'itération jusqu'à l'acceptation est $a = 2.07$. L'efficacité est $1/a \simeq .48$.

Remarquons que, en fait, il n'est pas nécessaire de connaître f. Il suffit d' utiliser une fonction $f_1 \propto f$ et une borne supérieure a du quotient f_1/g.

2.3 Méthode du quotient des uniformes

Supposons que (U, V) suit une distribution uniforme sur le disque unité. On démontre que le quotient V/U suit une distribution de Cauchy. Nous pouvons, alors, nous demander s'il est possible d'échantillonner à partir d'autres distributions comme quotient des variables uniformes sur certain sous-ensemble. Nous avons le résultat suivant:

Proposition 2 *Soit h une fonction non-négative avec $0 < \int h < \infty$. Soit*

$$C_h = \{(u, v) : 0 \leq u \leq \sqrt{h(v/u)}\}$$

C_h a une aire finie. Si (U, V) suit une loi uniforme sur C_h, alors $X = V/U$ a une fonction de densité $h/(\int h)$.

Démonstration:
Faisons le changement de variable $u = u$, $x = v/u$, l'aire de C_h est

$$\int\int_{C_h} dudv = \int_{-\infty}^{\infty}\int_{0}^{\sqrt{h(x)}} ududx = \int_{-\infty}^{\infty} \frac{1}{2}h(x)dx$$

qui est finie par hypothèse. En plus, la densité de (U, V) est $1/aire(C_h)$ sur son support, par suite (U, X) a la densité $u/aire(C_h)$ sur son support et X a une distribution marginale

$$\int_0^{\sqrt{h(x)}} \frac{u}{aire(C_h)}du = \frac{h(x)}{2aire(C_h)} = \frac{h(x)}{\int h(x)dx} \qquad \square$$

Le résultat est spécialement utile lorsque C_h est contenu dans un rectangle $[0, m] \times [p^i, p^s]$, alors on peut utiliser l'échantillonnage par rejet. On peut donner l'algorithme suivant:
`Jusqu'où` $(U, V) \in C_h$
`Générer` $U_1, U_2 \rightsquigarrow \mathcal{U}(0, 1)$
`Faire` $U = mU_1, V = p^i + (p^s - p^i)U_2$
`Sortir` $X = \frac{V}{U}$

La proposition suivante donne des conditions pour que C_h soit contenu dans un rectangle, elle exige seulement savoir une fonction $h \propto f$, comme dans l'échantillonnage par rejet.

Proposition 3 *Supposons que les fonctions $h(x)$ et $x^2h(x)$ sont bornées (sur le domaine de f). Alors $C_h \subset [0, m] \times [p^i, p^s]$, avec $m = (\sup h)^{1/2}$, $p^s = (sup\{x^2h(x) : x \geq 0\})^{1/2}$ et $p^i = -(sup\{x^2h(x) : x \leq 0\})^{1/2}$.*

Démonstration:
Il est évident que la première est bornée, par les inégalités

$$0 \leq u \leq \sqrt{h(v/u)} \leq \sqrt{\sup h(x)}.$$

Voyons pour la majoration par p^s. Pour que $v \geq 0$, il doit exister $u > 0$ qui vérifie $0 < u^2 \leq h(v/u)$ ou, d'une manière équivalente, $x = v/u > 0$ telle que $v^2 \leq x^2 h(x)$. Ainsi, $(u, v) \in C_h$ implique que $v^2 \leq (p^s)^2$ c'est-à-dire, $v \leq p^s$.
Le cas de p^i se démontre d'une façon analogue. □

Exemple 3 *Génération de la distribution de Cauchy* $\mathcal{C}(1,1)$. *Prenons* $h(x) = \frac{1}{1+x^2}$ *sur* $(-\infty, \infty)$. *Puisque* $h(x)$ *et* $x^2 h(x)$ *sont bornées, de la proposition (3), on obtient* $m = 1$, $p^s = 1$ *et* $p^i = -1$. *On a les équivalences*

$$(u, v) \in C_h \Longleftrightarrow 0 \leq u \leq \sqrt{\frac{1}{1 + (\frac{v}{u})^2}} \Longleftrightarrow 0 \leq u \leq \sqrt{1 - v^2}.$$

Par conséquent, l'algorithme devient
`Jusqu'où` $U_1 \leq \sqrt{1 - V^2}$
`Générer` $U_1, U_2 \rightsquigarrow \mathcal{U}(0,1)$
`Faire` $V = 2U_2 - 1$
`Sortir` $X = V/U_1$

Parfois, il est possible d'inclure C_h dans des ensembles polygonaux plus efficaces qu'un rectangle. Mais, le majeur coût computationnel (calcul informatique) sera de faire des essais ("tests") pour voir si $(u, v) \in C_h$, c'est pourquoi la méthode s'utilise souvent avec des pré-essais ou des essais préalables.

2.4 Utilisation des pré-essais

Les méthodes du rejet et du quotient des uniformes impliquent des essais des conditions de la forme $aU \leq f/g$ ou $(U, V) \in C_h$, respectivement, qui peuvent avoir une exécution lente. Marsaglia (1977) a introduit un procédé qui tente d'éviter de tels essais, en considérant des approximations inférieures et supérieures d'exécution plus rapide.

Pour la méthode du rejet, c'est simple d'obtenir des fonctions u, v satisfaisant les inégalités $u(x) \leq f(x)/g(x) \leq v(x)$ pour tout x. Nous pouvons établir la règle suivante:

Si $aU \leq u(X)$ accepter et sortir X
Si $aU > v(X)$ rejeter X.

Seulement dans le cas où $u(X) \leq aU \leq v(X)$, nous avons à vérifier la condition plus coûteuse $aU \leq f(X)/g(X)$.

Pour la méthode du quotient des uniformes, c'est simple d'obtenir des régions C_i, C_s telles que $C_i \subset C_h \subset C_s$ et telles qu'il est facile de déterminer si $(u,v) \in C_i$, et si $(u,v) \notin C_s$. Nous pouvons établir la règle:

Si $(U,V) \in C_i$ accepter et sortir $X = U/V$

Si $(U,V) \notin C_s$ rejeter X

Seulement dans le cas où $(U,V) \in C_s \setminus C_i$, nous avons à vérifier la condition plus coûteuse $(U,V) \in C_h$.

L'obtention des bornes peut-être compliquée, c'est pourquoi le procédé nécessite une importante "dose de génie" pour qu'il soit efficace et avantageux.

Exemple 4 *Génération de la distribution exponentielle tronquée sur* $(0,3)$. *Sa fonction de densité est*

$$f(x) = \frac{e^{-x}}{1-e^{-3}} \qquad 0 < x < 3$$

Appliquons la méthode du rejet. Pour cela, nous avons la fonction de densité $g \rightsquigarrow \mathcal{U}(0,3)$ *et considérons la fonction* $f_1(x) = e^{-x} \propto f(x)$. *Le maximum de* f_1/g *sur* $(0,3)$ *s'atteind en* $x = 0$. *Par conséquent, une constante a telle que* $f_1 \leq ag$ *est* $a = 3e^{-0} = 3$. *L'algorithme du rejet devient*

`Jusqu'où` $U \leq e^{-X}$
`Générer` $X \rightsquigarrow \mathcal{U}(0,3)$
`Générer` $U \rightsquigarrow \mathcal{U}(0,1)$
`sortir` X

Cherchons, maintenant, des pré-essais pour $U \leq e^{-X}$. *Nous avons les inégalités suivantes*

$$1 - x \leq e^{-x} \leq \frac{1}{1+x}.$$

Faisons $x = y - c$, *on obtient*

$$\frac{1-y+c}{e^c} \leq e^{-y}.$$

Par analogie, faisons $x = y - d$

$$e^{-y} \leq \frac{e^{-d}}{(1+y-d)}.$$

L'algorithme du rejet avec des pré-essais est:

1 Générer $X \rightsquigarrow \mathcal{U}(0,3), U \rightsquigarrow \mathcal{U}(0,1)$
2 Si $U \leq (1 - X + c)/e^c$, aller à 5
3 Si $U > e^{-d}/(1 + X - d)$, aller à 1
4 Si $U > e^{-X}$, aller à 1
5 Sortir X

Les opérations en 2 et 3 sont un peu moins coûteuses computationnellement que la 4. Reste à choisir c et d pour maximiser la probabilité d'accepter les conditions 2 et 3, ou pour que les pré-essais aient plus de possibilités en contenir l'essai d'origine.

2.5 Emploi des transformations

Parfois, il est possible d'utiliser des transformations entre les variables aléatoires, de façon que si nous savons générer à partir d'une d'entre elles, nous pouvons le faire à partir des autres.

Exemple 5 *Génération de la distribution Lognormale. Supposons que nous avons accès à un générateur de variables normales Y. Nous savons que si X est Lognormale, $\log X$ est normale. Par conséquent, il suffit de faire*

Générer Y normale
Sortir $X = \exp Y$

3 Méthodes spécifiques pour la distribution normale

Nous allons voir des méthodes pour générer à partir de la distribution normale centrée réduite $Y \rightsquigarrow \mathcal{N}(0,1)$. Si nous voulons le faire à partir de la distribution normale $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, il suffit de faire la transformation $X = \mu + \sigma Y$.

3.1 Somme de 12 uniformes

Ce procédé se base sur le théorème centrale limite (TCL) et peut-être considéré comme un exemple de transformation. Par le TCL, si les variables $U_i; i = 1, ..., n$ sont i.i.d. $\mathcal{U}(0,1)$, avec $E(U_i) = 1/2$, $Var(U_i) = 1/12$, la

variable

$$X = \frac{\left(\sum_{i=1}^{n} U_i - \frac{n}{2}\right)}{\sqrt{n/12}}$$

suit, approximativement, une loi normale, pour n suffisamment grande.

Une bonne approximation s'obtient, donc, pour $n = 12$, (En pratique sur le TCL, on voit que l'approximation dans le cas de la loi uniforme ne nécessite pas des valeurs très grandes de n), d'où $X = (\sum_{i=1}^{12} U_i) - 6$ et le procédé devient

Générer $U_1, U_2, ..., U_{12} \rightsquigarrow \mathcal{U}(0,1)$
Faire $X = (\sum_{i=1}^{12} U_i) - 6$
Sortir X

Remarquons que X satisfait $E(X) = 0$ et $Var(X) = 1$. L'approximation est suffisamment bonne.

3.2 Méthode de Box-Muller

La méthode exacte pour générer à partir de la loi normale, la plus connue, est celle de Box et Muller (1958), qui génére un couple de variables (X, Y) normales centrées réduites et indépendantes. La fonction de densité de (X, Y) est

$$f(x,y) = \frac{1}{2\pi} \exp\left[-\frac{(x^2+y^2)}{2}\right]$$

Soient (R, Θ) les coordonnées polaires de (X, Y); c'est-à-dire

$$R^2 = X^2 + Y^2; \qquad \tan\Theta = \frac{Y}{X}$$

qui ont la fonction de densité

$$g(r,\theta) = \left(\frac{1}{2\pi}\right)\left(r\exp\left(-\frac{r^2}{2}\right)\right) = g_1(\theta)g_2(r)$$

sur $(0,\infty) \times (0,2\pi)$; avec $g_1(\theta) = (\frac{1}{2\pi})\,\mathbb{I}_{(0,2\pi)}(\theta)$, c'est-à-dire une densité uniforme $\mathcal{U}(0,2\pi)$, et $g_2(r) = r\exp\left(-\frac{r^2}{2}\right)\mathbb{I}_{(0,\infty)}(r)$, c'est-à-dire une densité telle que R^2 est $\exp(1/2) \equiv \chi^2_2$. R et Θ sont indépendantes.

R se génére facilement par inversion, ona

$$F(r) = \int_0^r s\exp\left(-\frac{s^2}{2}\right)ds = 1 - \exp\left(-\frac{r^2}{2}\right)$$

Ainsi, si $U_1 \rightsquigarrow \mathcal{U}(0,1)$, on a

$$R = \sqrt{-2\ln(1-U_1)}$$

qui a la même distribution que

$$R = \sqrt{-2\ln U_1}$$

Alors on a l'algorithme:
Générer $U_1, U_2 \rightsquigarrow \mathcal{U}(0,1)$
Faire $R = \sqrt{-2\ln U_1}, \Theta = 2\pi U_2$
Faire $X = R\cos\Theta = \sqrt{-2\ln U_1}\cos 2\pi U_2$
Faire $Y = R\sin\Theta = \sqrt{-2\ln U_1}\sin 2\pi U_2$
Sortir X, Y

Les équations pour obtenir X, Y s'appellent transformations de Box-Muller.

3.3 Variante de Marsaglia

Marsaglia a introduit sa variante polaire de la méthode de Box-Muller, qui utilise la méthode du rejet pour éviter les opérations trigonométriques de sinus et cosinus, peu efficaces. Cette procédure est, pratiquement, plus rapide que celle de Box-Muller, aux dépens de complexité additionnelle dans la programmation près.