ECOLE SUPERIEURE DE STATISTIQUE

ET D'ANALYSE DE L'INFORMATION

(ESSAIT)

Cours de méthodes de simulation

Préparé par Hassen MATHLOUTHI

Année universitaire 2014-2015

AVANT PROPOS

Ce document propose un cours sur les méthodes de simulation aléatoire. Il est le résultat de l'enseignement de ce module durant ces dernières années à l'Ecole de Statistique et d'Analyse de l'Information.

Pour une bonne compréhension des méthodes proposées, ce cours doit être accompagné d'exercices aussi bien de travaux dirigés ainsi que de travaux pratiques. A cet effet, nous donnons à la fin du document les énoncés des épreuves des examens des années 2008 à 2014.

Ce cours reste néanmoins très incomplet. Il ne traite en effet que quelques une des méthodes usuelles de simulation aléatoire. En particulier, les méthodes de type MCMC (chaine de Markov- Monte Carlo) ne sont pas examinées. Ces autres méthodes devraient être aussi étudiées par tout lecteur cherchant à approfondir ses connaissances en la matière.

Il reste également assez théorique. En effet, les considérations d'ordre pratique liées notamment à la programmation informatique ne sont que partiellement ou pas du tout abordées.

D'autre part et quoique ayant fait l'objet de plusieurs lectures et de vérifications, le risque de présence d'erreurs mathématiques (et d'erreurs de langue aussi) n'est pas nul. Je serais très reconnaissant aux lecteurs me signalant les éventuelles erreurs ou incompréhensions.

Table de matières

Chapitre 1

INTRODUCTION GENERALE

Plusieurs applications scientifiques ont besoin de nombres choisis au hasard comme données nécessaires à leur fonctionnement. Ces nombres sont couramment appelés **nombres aléatoires**.

Comme exemples d'applications demandant ces nombres, nous citons en particulier

- La conception de programmes informatiques de jeux du hasard
- Les techniques de sélection probabiliste d'échantillons dans une population
- Les techniques de cryptologie (méthodes d'affectation de codes numériques cachés comme dans le cas des cartes de recharges téléphoniques).
- etc.

Ce cours a pour objet de présenter certaines des méthodes permettant de simuler la génération de tels nombres. Dans cette introduction générale, nous donnons une présentation sommaire de ces méthodes, de leur démarche méthodologique ainsi que de leur principale utilisation en statistique.

1. PRESENTATION

Un **nombre aléatoire** est à priori une des **valeurs prises par une certaine variable aléatoire réelle**.

En conséquence, pour avoir un nombre aléatoire, on n'a que **réaliser l'expérience aléatoire** qui convient, noter son résultat et déterminer son image par une variable aléatoire adéquate. C'est cette image qui constitue le nombre aléatoire demandé.

Ainsi par exemple, si une application nécessite, disons dix nombres aléatoires valant 0 ou 1 avec une probabilité égale à ½ chacun, on peut par exemple réaliser 10 fois l'expérience aléatoire consistant à jeter au hasard une pièce de monnaie et prendre le nombre 0 si le résultat est « Face » et le nombre 1 sinon.

Ce procédé de **génération** de nombres aléatoires qu'on peut qualifier de matériel est évidemment le plus naturel. Il présente cependant des limites importantes en pratique. Parmi ces limites citons notamment :

- Il est difficile de l'utiliser lorsque la **quantité** de nombres aléatoires demandés est **importante**. Or la plupart des applications concernées nécessitent souvent des milliers sinon des dizaines ou des centaines de milliers de nombres aléatoires.

- Ce procédé est même impossible à réaliser lorsqu'on **ne connait pas la nature de l'expérience aléatoire** sous jacente ou la définition de la variable aléatoire considérée. C'est le cas notamment de nombres aléatoires issus de variables aléatoires continues.

Ces limites ont conduit les utilisateurs, notamment avec l'avènement de l'ère informatique, à concevoir et mettre en œuvre d'autres procédés de génération de nombres aléatoires utilisables dans leurs applications. Parmi ces autres procédés, les procédés de type **algorithmique** qui constituent l'objet de ce cours sont d'une grande utilisation pratique.

Les méthodes algorithmiques de génération de nombres aléatoires se présentent comme des **formules mathématiques** permettant de disposer d'une suite de nombres qu'on peut considérer comme étant choisi au hasard selon une **loi de probabilité donnée**.

Ainsi, au lieu de réaliser matériellement une expérience aléatoire pour obtenir un nombre aléatoire, on procède à sa détermination en calculant une formule mathématique.

Du fait de leur caractère mathématique, les méthodes algorithmiques peuvent faire l'objet de programmation informatique, ce qui permet d'obtenir très rapidement autant qu'on veut de nombres aléatoires. Dans la pratique, on fait en effet tourner un certain programme informatique autant de fois qu'on veut de nombres aléatoires.

Il convient cependant de noter que de part leur construction, **il est impossible d'obtenir avec les méthodes algorithmiques de nombres aléatoires**. En effet, ces méthodes consistent en l'application d'une formule mathématique. La connaissance de cette formule permet ainsi de connaitre au préalable le nombre que ces méthodes donnent, ce qui est contraire à la définition même d'un nombre aléatoire. En effet, un nombre aléatoire est par définition non prévisible.

Néanmoins, malgré l'impossibilité d'obtenir de véritables nombres aléatoires avec les méthodes algorithmiques, les améliorations continues qu'ont connues ces méthodes ont conduit à l'obtention de nombres qui ressemblent dans plusieurs aspects à des vrais nombres aléatoires.

Aussi, appelle- t- on ces méthodes de méthodes de **simulation** aléatoire et les nombres qu'elles produisent des **nombres pseudo-aléatoires** ou des **nombres simulés**. On parle aussi d'**échantillon artificiel** pour un ensemble de nombres fournis par ces méthodes.

2. DEMARCHE GENERALE

Dans les applications, les générateurs algorithmiques permettent seulement de simuler des valeurs d'une variable X suivant la loi uniforme continue sur l'intervalle [0,1].

Pour la simulation d'une variable aléatoire Y suivant une autre loi de probabilité, on n'a pas besoin de générateurs particuliers. Il suffit de trouver la relation liant cette variable et la variable X. Or cette relation existe toujours.

En effet, selon un résultat mathématique, toute variable aléatoire Y peut s'écrire comme une certaine fonction de variables aléatoires réelles $X_1, X_2, \ldots, X_i, \ldots X_p$ suivant chacune une loi uniforme continue sur l'intervalle [0,1] :

$$Y = \varphi_Y(X_1, X_2, \ldots, X_i, \ldots, X_p)$$

En conséquence, pour avoir n valeurs simulées d'une variable Y étant donné sa loi de probabilité, on procède ainsi :

- Détermination de la fonction φ_Y : La théorie de probabilité propose à cet effet plusieurs résultats permettant d'aider à la détermination de cette fonction. Ce volet fera l'objet du troisième chapitre de ce cours.
- Génération de np valeurs simulées de la loi uniforme continue sur [0,1] : $(x_{11}, x_{21}, \ldots, x_{i1}, \ldots x_{p1})$, $(x_{12}, x_{22}, \ldots, x_{i2}, \ldots x_{p2})$, …,$(x_{1n}, x_{2n}, \ldots, x_{in}, \ldots x_{pn})$
- Calculer simplement : $y_i = \varphi_Y(x_{1i}, x_{2i}, \ldots, x_{ii}, \ldots x_{pi})$ pour i = 1 à n

3. APPLICATIONS

Outre les applications générales sus indiquées, la simulation de nombres aléatoires trouvent leur application dans un grand domaine des mathématiques et des statistiques qui est le calcul intégral approché. Cette application porte le nom de la méthode de **Monte Carlo**

En particulier, les caractéristiques d'une loi de probabilité comme entre autres l'espérance mathématique et la variance, ou la probabilité attachée à un intervalle donnée, se présentent comme des intégrales et peuvent ainsi être approchés en

appliquant la méthode de Monte Carlo sur un certain nombre de valeurs simulées issues de la loi de probabilité considérée.

En général, chaque fois qu'on peut tolérer une certaine marge d'erreur, on peut utiliser les techniques de simulation pour disposer de nombres aléatoires. C'est le cas en général du calcul de probabilités. Mais une telle utilisation serait à priori non permise pour les applications de cryptologie par exemple.

4. CONCEPTS ET OUTILS DE BASE

Les concepts et outils utilisés sont ceux du calcul de probabilité. Il s'agit notamment des concepts et outils suivants :

- Variable et vecteur aléatoire
- Loi de probabilité
- Fonction de répartition
- Densité de probabilité
- Moments théoriques d'une variable aléatoire : Espérance mathématique, variance, etc.
- Formule de changement de variables
- Loi des grands nombres
- Théorème central limite, etc.

Ces concepts et outils doivent être bien compris pour une meilleure maitrise des méthodes de simulation.

Nous donnons ci après à titre de rappel une présentation succincte, sous forme de tableaux, des lois usuelles à une seule dimension donnant leur définition, leurs principales caractéristiques et quelques unes de leurs propriétés.

4.1 Les lois usuelles discrètes.

Appellation et symbole	Expérience aléatoire	Expression	Moments	Propriétés
Loi de Bernoulli **X ~> B (p)**	X = 1 si un individu choisi au hasard dans une population E a un caractéristique c , X= 0 sinon	$P(X= x) = p^x q^{n-x}$ $\forall x \in \{0,1\}$ et p = 1-q est la proportion d'individus dans E ayant la caractéristique c	E(X) = p V(X) = p q $M_X(t) = q + pe^t$	X_1 ~> B (p_1) , X_2 ~> B (p_2) et E(X_1X_2) =E(X_1)E(X_2) alors X_1 et X_2 sont indépendantes
Loi binomiale **X ~> B (n,p)**	X= nombre d'individus ayant une caractéristique c dans un ensemble de n individus tirés **avec remise** dans une population E	**$P(X= x) = C_n^x\ p^x\ q^{n-x}$** **$\forall x \in \{0,1,2,\ldots,n\}$** et p = 1-q est la proportion d'individus dans E ayant la caractéristique c	E(X) = np V(X) =n p q $M_X(t) =(q + pe^t)^n$	• $X_1, X_2,\ldots$et X_n indépendantes et suivant chacune B(p) alors : S = Σ X_i **suit** B (n,p) • X_1 suit B (n_1 ,p) et X_2 suit B (n_2 ,p) indép alors, $X_1 + X_2$ **suit** B (n_1+ n_2 ,p)

Loi Hyper – géométrique **X ~>H (n,N_1,N)**	X= nombre d'individus ayant une caractéristique c dans un ensemble de n individus tirés **sans remise** dans une population E (on note N l'effectif de E et N_1 l'effectif dans E des individus ayant la caractéristique c)	$P(X = x) = \frac{C_{N1}^{x} C_{N-N1}^{n-x}}{C_N^n}$ $\forall$ x $\in$ {a, a+1,…,b-1,b} où a = max (0, n-N_2) et b=min (n,N_1) , N_2 = N- N_1	E (X) = np V(X) = npqk Avec p = N_1/N q= 1 – p et k= (N-n)/(N-1)	Quand N$\rightarrow +\infty$, n et p restant fixes X suit approximativement B(n,p) avec p= N_1/N
Loi de Poisson **X ~> P(λ)**	X= nombre de réalisations d'un certain évènement au cours d'une période de temps donnée.	**P(X= x) = $e^{-\lambda}$** λ^x / x ! $\forall$ x $\in$ N et $\lambda \geq 0$	E (X) = λ V(X) = λ M_X (t) = exp[λ(e^t -1)]	X_1 suit P(λ_1) et X_2 suit P(λ_2) indépendantes alors, $X_1 + X_2$ suit P(λ_1+ λ_2)
Loi géométrique **X ~> G(p)**	X= nombre de réalisations indépendantes d'une expérience jusqu'à l'obtention d'un évènement A donné.	**P(X= x) = pq^{x-1}** $\forall$ x $\in$ N* et q=1-p est la probabilité d'avoir A	E(X) = 1/p V(X) = q/ p^2 M_X (t)=pe^t / (1-qe^t)	

4.2 Les lois usuelles absolument continues

Appellation et symbole	Densité de probabilité	Fonction de répartition	Moments	Propriétés
Loi uniforme continue sur [a.b] **X ~> *U* (a,b)**	$f(x) = \frac{1}{b-a}$ $\forall x \in [a,b]$ $= 0$ sinon	*F(x) = 0 $\forall x < a$.* $F(x) = \frac{x-a}{b-a}$ $\forall x \in [a,b]$ *F(x) = 1 $\forall x > b$.*	$E(X) = \frac{b+a}{2}$ $V(X) = \frac{(b-a)^2}{12}$	Soit $Y = \frac{X-\alpha}{\beta}$ alors Y suit aussi la loi uniforme continue.
Loi exponentielle **X ~> e(λ)**	f(x)=λ *exp (- λ x) $\forall x \geq 0$* *= 0 sinon*	*F(x) = 0 $\forall x \leq 0$.* *F(x) = 1-$e^{-\lambda x}$ $\forall x > 0$.*	$M(t) = \frac{\lambda}{\lambda - t}$ $\forall t < \lambda$ *E(X) = 1/λ et V(X) = 1/ λ^2.*	
Loi gamma **X ~>γ(a,b)**	$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$ $\forall$ ***x*** ***$\geq$ 0*** ***= 0*** sinon	Non définie analytiquement	$M(t) = \left(\frac{b}{b-t}\right)^a$ $\forall t < b$ $E(X^s) = \frac{\Gamma(s+a)}{b^s \Gamma(a)}$ $\forall s >$ *0*	• Soit **X_1 ~>γ(a_1,b)** et **X_2 ~>γ(a_2,b)** indépendantes Alors(**X_1+ X_2) ~>γ(a_1+a_2,b)** • Soit **X_1 ~>γ(a,b)** et λ>0 alors λ**X_1 ~>γ(a,b/λ)** • **e(b)≡ γ(1,b)**
Loi Normale **X ~> N(m,σ^2)**	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x-m)^2]$ $\forall x \in R$ *m* $\in$ R et σ > 0	*F* est non définie analytiquement mais existe une table pour Φ la fonction de répartition de la loi normale centrée réduite N(0,1)	$M_X(t) = e^{(tm + t2\sigma2/2)}$ *E(X) = m* *V(X) = σ^2*	• Soit X ~> N(m,σ^2) et U = (X-m)/ σ alors **U~> N(0,1)** et donc F(x) = Φ (u=(x-m)/ σ). • Soit $Z = \sum_{i=1}^{n} \lambda_i X_i$ ou les λ_i sont des réels et les X_i sont normales alors Z est

				normale.
Loi de Khi deux $X \rightsquigarrow \chi^2(n)$	$f(x) = \frac{(\frac{1}{2})^{n/2}}{\Gamma(\frac{n}{2})} x^{n/2-1} e^{-y/2}$ $\forall x \geq 0 \quad = 0$ ailleurs	F est non définie analytiquement mais existe une table.	$E(Y) = n$, $V(Y) = 2n$. $M_X(t) =$ $\left(\frac{1}{1-2t}\right)^{\frac{n}{2}} \quad \forall\, t < \frac{1}{2}$	• *Soit* $X = \sum_{i=1}^{n} U_i^2$ où $U_i \rightarrow N(0,1)$ $\forall i$ et indépendantes alors $X \rightarrow \chi^2(n)$ • $\chi^2(n) \equiv \gamma(n/2, 1/2)$
Loi de Student $T \rightsquigarrow S(n)$.	$T = \frac{U}{\sqrt{Y/n}}$ où $U \rightsquigarrow \mathbf{N(0,1)}$, $Y \rightarrow \chi^2(n)$ indépendantes	F est non définie analytiquement mais existe une table.	$E(T) = 0$ $V(T) = \frac{n}{n-2}$	
Loi de Fisher $R \rightsquigarrow F$	$R = \frac{Y_1/n_1}{Y_2/n_2}$ où $Y_1 \rightsquigarrow \chi^2(n_1)$ et $Y_2 \rightsquigarrow \chi^2(n_2)$ indépendantes .	F est non définie analytiquement mais existe une table.		

5. PLAN DU COURS ET BIBLIOGRAPHIE

En plus d'un chapitre introductif, ce cours comprend quatre autres chapitres:

- Chapitre 2 : Simulation de la loi uniforme standard
- Chapitre 3 : Simulation des autres lois de probabilité à une seule dimension
- Chapitre 4 : Simulation des lois de probabilités multidimensionnelles
- Chapitre 5 : La méthode de Monte Carlo

Comme éléments bibliographiques approfondissant ce cours, signalons l'existence de plusieurs livres traitant les méthodes de simulation de variables aléatoires. On trouve également sur Internet un grand nombre de sites présentant des cours et exercices concernant les dites méthodes. Nous nous contentons dans ce qui de signaler quelques unes de ces références qui nous semblent les plus intéressants.

- N. Bouleau. Probabilités de l'Ingénieur, variables aléatoires et simulation
- L. Devroye. Non-Uniform Random Variate Generation. Springer, 1986.
- KNUTH D.E.. (1968) "The art of computer programming", volumes 1 et 2, Addison-Wiley

Chapitre 2

SIMULATION DE LA LOI UNIFORME CONTINUE STANDARD

Ce chapitre a pour objet de présenter certaines des méthodes permettant de générer des nombres simulant les réalisations d'une variable aléatoire réelle suivant la loi uniforme continue sur l'intervalle [0.1]. Ces nombres qu'on appelle « nombres pseudo aléatoires » comme expliqué dans l'introduction générale sont à la base des méthodes de simulation des autres lois de probabilité. Certaines procédures de tests statistiques permettant de juger la performance de ces méthodes sont également présentées.

La littérature propose plusieurs procédés de génération de nombres pseudo aléatoire. Nous nous limitons dans ce chapitre aux générateurs de congruence qui sont les plus usuels.

Après une introduction présentant notamment les générateurs de nombres pseudo aléatoires dans leurs généralités, nous étudions en détail les deux grandes familles de générateurs de congruence : les générateurs de congruence linéaire et les générateurs de congruence multiplicative. Nous terminons le chapitre par la présentation de trois familles de tests statistiques pouvant être utilisés pour juger la performance des générateurs proposés.

Nous rappelons en annexe la définition de la loi uniforme standard et ses propriétés.

1. GENERALITES

Qu'est ce qu'un générateur de nombres pseudo aléatoires ? Quelles sont ses qualités souhaitées ? Comment le construire ? Ce sont les questions traitées dans cette introduction.

1.1 Définition

Un générateur de nombres pseudo aléatoires est un procédé mathématique (formule) permettant en l'appliquant de disposer d'une suite de nombres qu'on peut considérer comme des valeurs **indépendantes** d'une variable aléatoire suivant la loi uniforme continue sur [0,1].

Remarques

- Le générateur de nombres pseudo aléatoires est ainsi censé remplacer tout procédé matériel passant par la réalisation d'une certaine expérience aléatoire.
- Plusieurs termes utilisés sont presque des synonymes. Ainsi, on dit indifféremment, générateur, simulateur, méthode de simulation, algorithme de simulation etc.
- Toutes les machines à calculer scientifiques et tous les langages, logiciels et tableurs informatiques disposent de fonctionnalités permettant de générer des nombres pseudo aléatoires.
- Concrètement un générateur de nombres pseudo aléatoires est un programme informatique qui permet en l'appelant de donner un ou plusieurs nombres ressemblant à des réalisations indépendantes d'une variable aléatoire suivant la loi uniforme continue sur [0,1]. En général, ce programme n'est pas directement accessible à l'utilisateur.

1.2. Critères de qualité

Ce qu'on demande d'un générateur de nombres pseudo aléatoires est évidemment sa qualité à imiter le hasard en fournissant des nombres indépendants et uniformes mais aussi des qualités d'ordre :

- Mathématique : la formule se prête à une analyse mathématique permettant notamment de voir les avantages et les insuffisances du générateur en question
- Informatique : la formule donne lieu à des calculs simples et rapides. En effet, on cherche des générateurs algorithmiques pour exploiter les grandes capacités de calcul ainsi que la rapidité de mise en œuvre fournis par les ordinateurs. .

1.3 Principe général de construction

Comme il a été dit, la formule de génération nombres pseudo aléatoires se présente en fait comme un programme informatique. Pour obtenir un nombre pseudo aléatoire, il convient ainsi de tourner ce programme. Comme il s'agit d'une formule, cela nécessite qu'on fournisse au programme une donnée initiale.

Il n'est pas cependant commode de tourner le programme et lui fournir donc la donnée initiale à chaque fois qu'on a besoin d'un nombre pseudo aléatoire. En effet, dans les applications, on a besoin d'un nombre important de valeurs aléatoires.

La solution généralement adoptée est de concevoir le programme de telle sorte à considérer chaque nombre pseudo aléatoire sorti comme valeur initiale du prochain nombre aléatoire à sortir. Il suffit ainsi de disposer que d'une seule valeur initiale et de préciser au programme le nombre de valeurs pseudo aléatoires demandées.

D'une manière plus précise, soit S un ensemble fini d'entiers naturels appelé **espace d'états** et s_0 un élément de S appelé état initial ou **germe**.

On appelle **fonction de transition** toute fonction f de S dans S : $s_n = f(s_{n-1})$. Comme son nom l'indique la fonction f permet de passer d'un élément à un autre au sein de l'ensemble S. En se donnant s_0 et en appliquant cette fonction un certain nombre N de fois, on obtient ainsi une suite $\{s_n\}$ de N+1 éléments de S : $s_0, s_1, s_2, \ldots, s_n, \ldots, s_N$.

Soit maintenant U un autre ensemble appelé **ensemble de sorties**. On appelle **fonction de sortie** une fonction g de S dans U : $u_n = g(s_n)$.

La fonction g permet ainsi de construire une deuxième suite $\{u_n\}$ a partir de $\{s_n\}$: $u_0, u_1, u_2, \ldots, u_n, \ldots, u_N$.

Dans la pratique, S est un ensemble fini d'entiers et U une partie de [0.1].

2. LES GENERATEURS DE CONGRUENCE

Les générateurs de congruence sont les générateurs les plus usuels compte tenu de la facilité de leur mise en œuvre. Ils sont introduits par Lehmer en 1948.

Comme leur nom l'indique, ces générateurs sont basés sur la relation de congruence.

Dans les applications, on distingue principalement entre deux types de générateurs de congruence :

- Les générateurs de congruence linéaires
- Les générateurs de congruence multiplicatifs

2.1 Générateurs de congruence linéaire

a. Définition

Un générateur de congruence linéaire (GCL) est défini par la fonction de transition suivante :

$$s_n = (as_{n-1}+c)(\bmod m)$$

où a, c et m sont des entiers positifs appelés respectivement multiplicateur, incrément et module.

Remarques

- s_n est en fait le reste de la division entière de $(as_{n-1}+c)$ sur m
- les termes de la suite qu'on obtient en tournant le générateur sont des nombres compris entre 0 et m-1. L'espace d'états est ainsi défini par S= {0, 1, 2,,,m-1}
- Pour démarrer le générateur, on se donne s_0 choisi au hasard entre 0 et m-1.
- Pour avoir des nombres compris entre 0 et 1, on divise les s_i par m par exemple.

Exemples

- Exemple 1 : Soit $s_n = (10s_{n-1} + 5) \pmod{12}$. On note que S = {0, 1, 2,,,11}. Si on fixe $s_0 = 5$. on trouve : $s_1 = 7$, $s_2 = 3$, $s_3 = 11$, $s_4 = 7$, $s_5 = 3$, $s_6 = 11$, $s_7 = 7$…

- Exemple 2 : Soit $s_n = (5s_{n-1} + 1) \pmod{8}$. On choisit $s_0 = 0$. En appliquant, on trouve : $s_1 = 1$, $s_2 = 6$, $s_3 = 7$, $s_4 = 4$, $s_5 = 5$, $s_6 = 2$, $s_7 = 3$, $s_8 = 0$, $s_9 = 1$,…

b. Propriétés

- Le nombre de valeurs possibles pouvant être fournies par un générateur de congruence linéaire est au plus égal à m.
- D'autre part, si un nombre apparait une deuxième fois, tous les nombres qui le suivent apparaissent aussi une deuxième fois et selon le même ordre. Ainsi, un générateur de congruence linéaire est nécessairement périodique. Sa période maximale vaut m.

Remarques :

- la période maximale n'est pas toujours atteinte (voir exemple 1 ci-dessus)
- Dans tous les cas, la qualité de l'indépendance n'est pas respectée. En effet, en notant ρ la période on a ainsi $s_{n+k\rho} = s_n \ \forall$ n et $k \in N$, ce qui est contraire au caractère aléatoire souhaité.

c. Optimisation

Les générateurs de congruence linéaire ont des bonnes propriétés mathématiques et informatiques. En effet, la fonction « modulo » est très facile à manipuler. Ils présentent cependant un grave inconvénient du fait qu'ils sont périodiques.

Aussi, a t- on chercher à les améliorer par un bon choix des paramètres m, a et c.

- **Choix de s_0** : ce paramètre est choisi au hasard (selon des procédés électroniques) parmi les éléments de S.

- **Choix de m** : Dans ce type de modèle, le nombre de termes possibles du générateur vaut m. On en déduit que la période maximale vaut m. Il convient en conséquence de choisir m le plus élevé possible afin que le générateur ne se répète pas. En général, on le choisit de la forme 2^k ou k est le nombre maximum de bits permis par la capacité de l'ordinateur. D'autre part ce choix permet de gagner du temps de calcul car il permet d'éviter la division.

  Exemple : Soit un processeur à 8 bits (2^3). Cherchons le reste de la division de 135 sur 2^3. Le nombre 135 s'écrit en binaire, la numérotation de base de l'ordinateur, 10000111. En effet, on vérifie que :
  $135 = 1.2^7+0.2^6+0.2^5+0.2^4+0.2^3+1.2^2+1.2^1+1.2^0$
  Le reste de la division par 2^3 est direct. Il est donné par le contenu des trois derniers bits 0000111 soit 7.

- **Choix de a et c :** Choisir m, la période maximale, la plus élevée possible ne suffit pas car cette période maximale peut ne pas être atteinte (voir exemple 1 ci-dess**us).** Néanmoins, il est possible d'atteindre cette période maximale grâce à un choix adéquat des paramètres a et c. On a à cet effet, la proposition suivante :

Proposition : le paramètre m étant donné, pour atteindre la période maximale soit m, il faut et il suffit que :

1. c et m soient premiers entre eux (leur pgcd = 1).
2. Pour chaque nombre premier p divisant m, (a-1) soit multiple de p
3. Si m est multiple de 4, (a-1) soit multiple de 4.

Exemple : soit le générateur de congruence linéaire : $s_n = (as_{n-1}+c)(\text{mod } 16)$
Trouvons a et c pour que ce GCL atteigne sa période maximale soit 16.

- Pour vérifier la condition 1, on peut prendre c = 3. En effet 3 et 16 sont premiers entre eux.
  NB : on aurait pu prendre c = 5 ou c = 7 mais pas 2 ni 4.
- Le seul nombre premier divisant m = 16 est 2. Le nombre (a-1) doit être donc multiple de 2. D'autre part, m =16 étant multiple de 4, (a-1) doit être aussi multiple de 4. On peut par conséquent prendre a = 5 ce qui permet de vérifier les deux conditions en même temps.

En supposant s_o à 0, ce générateur donne : : $s_1 = 3$, $s_2 = 2$, $s_3 = 13$, $s_4 = 4$, $s_5 = 7$, $s_6 = 6$, $s_7 = 7$, $s_8 = 8$, $s_9 = 11$, $s_{10} = 10$, $s_{11} = 5$, $s_{12} = 12$, $s_{13} = 15$, $s_{14} = 14$, $s_{15} = 9$, $s_{16} = 0$,etc.

Remarque : Avoir un générateur de période maximale ne suffit pas d'avoir un « bon » générateur.
Exemple : Prenons a = 1, c =1, m =1024 et $s_0 = 0$. Ce générateur donne $s_1 = 1$, $x_2 = 2$, $s_3 = 3,\ldots, s_{1023} = 1023, s_{1024} = 0$
Les nombres donnés par ce générateur n'ont pas manifestement le caractère aléatoire.

2.2 Générateurs de congruence multiplicative

a. Définition

Un générateur de congruence multiplicative est défini par la fonction de transition suivante :

$$s_n = (as_{n-1})(\text{ mod } m)$$

où a et m sont des entiers positifs appelés respectivement multiplicateur et module.

Remarques

- les termes de la suite qu'on obtient en tournant le générateur sont des nombres compris entre 0 et m-1. Il est à noter cependant que si le générateur donne 0 , il continue toujours à donner 0. Par conséquent, il convient d'éliminer le nombre 0 des résultats possibles d'un générateur de congruence multiplicative. L'espace d'états est ainsi défini par S= {1, 2,…,m-1}
- Comme le générateur de congruence linéaire, le générateur de congruence multiplicative est périodique. Sa période maximale vaut m-1 (nombre d'éléments de S).
- Ce type de générateur est plus avantageux que le générateur de congruence linéaire sur le plan de calcul informatique.

Exemples

- Exemple 1 : Soit $s_n = (10s_{n-1})(\text{mod } 11)$. On note que S = { 1, 2,,,10}. Si on fixe $s_0 = 1$. on trouve : $s_1 = 10$, $s_2 = 1$, $s_3 = 10$, $s_4 = 1$, $s_5 = 10$, $s_6 = 1$, $s_7 = 10\ldots$
- Exemple 2 : Soit $s_n = (5s_{n-1})$ (mod 8). Choisissons $s_0 = 3$. En appliquant, on trouve : $s_1 = 7$, $s_2 = 3$, $s_3 = 7$, …

Remarque :

- la période maximale n'est pas toujours atteinte (voir exemple 1 ci-dessus)
- Dans tous les cas, la qualité de l'indépendance n'est pas respectée. En effet, en notant ρ la période on a ainsi $s_{n+k\rho} = s_n \ \forall$ n et k $\in$ N.

c. Optimisation

Comme les générateurs de congruence linéaires, les générateurs de congruence multiplicative ont des bonnes propriétés mathématiques et informatiques. Ils présentent cependant un grave inconvénient du fait qu'ils sont périodiques.

Aussi, a t- on chercher à les améliorer par un bon choix des paramètres m et a.

Proposition 1 : Pour avoir la période maximale soit (m-1), il faut choisir m nombre premier et prendre a racine primitive de m, soit :

$$a^n \bmod m \neq 1 \ \forall \ n = 1,2, \ldots, (m-2).$$

Exemples :

- $s_n = 2s_{n-1} \bmod 11$

On note que 11 est bien un nombre premier. D'autre part on peut vérifier que les restes des divisions de 2, 4 , 8,… sur 11 sont différents de 1.

- $s_n = 7^5 s_{n-1} \bmod (2^{31}-1)$. On a vérifié que $(2^{31}-1)$ est premier et 7^5 est racine primitive de $(2^{31}-1)$.

Remarque : m étant premier ne peut pas être de la forme 2^k. Si on tient à cette forme pour des raisons de facilités de calcul informatiques, il convient de choisir a et x_0 selon la proposition suivante :

Proposition 2 : Soit $m = 2^k$ ($k \geq 3$). La période maximale sous cette contrainte vaut 2^{k-2} (donc inférieure à m-1 qui est la période maximale sans contrainte). Pour atteindre cette période maximale, il faut prendre x_0 impair et $a = \pm 3 \bmod 8$ $(a = 8t \pm 3)$

- Exemple : $s_n = 5s_{n-1} \bmod 32$. On note que $32 = 2^5$. Donc la période maximale vaut $2^{5-2} = 8$. Pour l'atteindre, on prend par exemple $x_0 = 1$ qui est impair et a = 5 qui vérifie 8x1 -3 . On aboutit en effet, à la suite suivante : 1,5,25,29,17,28,9,13,1, 5,…

2.3 Exemples réels de générateurs de congruence

Les générateurs de congruence ont été très utilisés en pratique par les logiciels et les langages de programmation notamment au commencement de l'ère informatique. Citons à titre d'exemples :

- Rand( ) du langage C ANSI avec $m = 2^{31}$, a= 1103515245 et c=12345.
- Randu d'IBM des années 60 avec $m = 2^{31}$, a= 65539 et c = 0.
- Générateur de Knuth et Lewis : $s_n = 69069\ s_{n-1} \bmod (2^{32})$
- la fonction drand48() (en C ANSI) qui utilise les paramètres :

m = 248 , a = 25214903917 , c = 11,

- – le générateur du logiciel MAPLE avec

m = 1012, a = 427419669081 , c = 0

Remarques :

- Les générateurs de congruence linéaires ont connu des développements importants ces dernières années notamment au niveau de la période maximale qui atteint dans certains cas des dimensions spectaculaires.
- Un grand nombre d'autres types de générateurs est proposé dans la littérature.
- Chaque nouvelle proposition vient corriger les problèmes d'anciens générateurs
- A noter aussi que beaucoup de générateurs considérés comme bons autrefois ne le sont plus maintenant
- On peut créer son propre générateur, mais il est bien plus prudent d'utiliser un générateur établi ayant été **testé** complètement (voir la suite du cours) que d'en inventer un nouveau.

3. LES TESTS STATISTIQUES

Posséder une très grande période et l'atteindre est, pour un générateur de nombres pseudo aléatoires, une condition nécessaire à remplir mais pas suffisante. En effet, un générateur peut satisfaire cette condition sans pour autant fournir des nombres présentant un caractère aléatoire.

La question qui se pose maintenant est comment savoir qu'une suite de nombres fournis par un générateur ont ou non un caractère aléatoire. D'une manière plus précise, il s'agit de « vérifier » si les nombres donnés par le générateur en question peuvent être considérés comme des réalisations **indépendantes** d'une variable aléatoire réelle suivant **la loi uniforme continue** sur l'intervalle **[0,1].**

S'agissant du domaine de l'aléatoire, la vérification ne peut être réalisée que par le biais de tests d'hypothèses. Les hypothèses à tester ici concernent deux aspects en même temps : l'uniformité et l'indépendance.

Plusieurs tests sont développés dans la littérature. Ils n'ont pas la même puissance et ne concernent souvent qu'un seul aspect du problème : l'uniformité ou l'indépendance. Aussi, convient il en général utiliser plusieurs tests pour accepter un générateur comme « un bon » générateur et l'utiliser par la suite pour produire des nombres aléatoires.

Dans ce qui suit, nous présentons en détail deux exemples de test. Il s'agit du test d'adéquation de Khi deux et d'un test d'indépendance appelé « Run Test ». Pour des raisons pédagogiques, nous considérons aussi et en premier lieu le test de la moyenne. Au préalable, nous rappelons les principes généraux de construction d'un test statistique.

3.1 Principes généraux

Etant donné un problème de test, c'est-à-dire une hypothèse nulle H_0 et une hypothèse alternative H_a, il s'agit au vu d'une suite de réalisations $x_1, x_{2,\dots,} x_{i,\dots,} x_n$ indépendantes d'une variable aléatoire réelle X de prendre une décision : Accepter ou de refuser H_0 au risque α de se tromper (α étant donné).

Un test se présente ainsi comme une règle de décision dans un contexte d'incertitude. Formellement, un test est défini par toute application de l'ensemble des échantillons possibles dans l'ensemble de décisions (Ce dernier est constitué par deux éléments « Accepter H_0 » et « Refuser H_0 »), ou d'une manière équivalente par la partie de l'ensemble des échantillons possibles conduisant à refuser H_0. Cette partie qu'on note W est appelée la région critique du test (ou sa région de rejet).

A chaque test est associé deux types de risque : le risque de première espèce noté couramment α et défini par la probabilité de refuser H_0 alors qu'elle est vraie et le risque de seconde espèce correspondant à la probabilité d'accepter H_0 alors qu'elle est fausse. Il n'existe pas de test optimal, celui minimisant à la fois les deux types de risque. A la place, on a développé des optiques générales permettant d'opérer des choix partiels dans l'ensemble des tests disponibles. Entre autres de ces optiques, citons notamment celle de Neyman, de Bayes, etc.

En pratique, pour construire un test, on passe en général par les étapes suivantes :

- Trouver une statistique S fondant le test, c'est-à-dire une fonction de l'échantillon : $S = \varphi(x_1, x_{2,\dots,} x_{i,\dots,} x_n)$, dont on connait la loi de probabilité (ou du moins la loi asymptotique) sous H_0. Concrètement, S est un indicateur tiré de l'échantillon et dont les valeurs renseignent sur la plausibilité de H_0.
- Se donner un niveau de risque de première espèce α =P(refuser H_0/ H_0 est vraie) = P(W/ H_0 est vraie)
- Déterminer la forme de la région critique W comme une partie « significative » de l'ensemble des valeurs prises par la statistique S
- Déterminer les frontières de W étant donné la forme retenue et le niveau de risque fixé.

3.2 Test de la moyenne

Soit $x_1, x_{2,\dots,} x_{i,\dots,} x_n$ n nombres donnés par un générateur. Il s'agit de tester l'hypothèse (H_0) selon la quelle ces nombres peuvent être considérés comme des réalisations indépendantes d'une variable aléatoire réelle suivant la loi uniforme continue sur [0,1].

Intuitivement, si le générateur utilisé est un « bon » générateur, la moyenne arithmétique des nombres qu'il donne doit se situer à proximité de la valeur 0.5 qui est l'espérance mathématique de la loi uniforme continue sur [0,1]. En conséquence, une moyenne très différente de 0.5 doit nous conduire à douter de la qualité de ce générateur.

Plus formellement, désignons par X_1, $X_{2,\dots,}$ $X_{i,\dots,}$ X_n les variables aléatoires dont sont issues les réalisations x_1, $x_{2,\dots,}$ $x_{i,\dots,}$ x_n . Lorsque H_0 est vraie, ces variables sont identiquement et indépendamment distribuées. Leur loi commune est la loi uniforme continue sur [0,1] et donc :

$$E(X_i) = \frac{1}{2} \ \forall\, i = 1 \text{ à } n \quad et$$
$$V(X_i) = \frac{1}{12} \ \forall\, i = 1 \text{ à } n$$

Considérons maintenant la statistique $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ (moyenne empirique). On calcule aisément :

$$E(\bar{X}_n) = \frac{1}{2} \quad et$$
$$V(\bar{X}_n) = \frac{1}{12n}$$

On sait alors d'après le théorème central limite que la suite $\{Z_n\}$ définie par :

$$Z_n = (\bar{X}_n - \tfrac{1}{2})\sqrt{12n}$$

converge en loi vers la loi normale centrée réduite.

Dés lors, on peut considérer Z_n comme la statistique fondant notre test puisque sa loi de probabilité sous H_0 est connue (asymptotiquement). Comme la loi normale est symétrique, il s'ensuit la région critique suivante :

$$W=\{ (x_1,x_2,\dots,x_i,\dots,x_n)/z_n < z_{\alpha/2} \text{ ou } z_n > z_{1-\alpha/2} \},$$

(Les nombres $z_{\alpha/2}$ et $z_{1-\alpha/2}$ sont respectivement les quantiles d'ordre ($\alpha/2$) et ($1-\alpha/2$) de la loi normale centrée réduite).

Exemple : On a utilisé le générateur de nombres aléatoires incorporé dans le tableur Excel pour générer les 48 nombres suivants :

0,382	0,101	0,596	0,899	0,885	0,958	0,014	0,407	0,863	0,139
0,245	0,045	0,032	0,164	0,220	0,017	0,285	0,343	0,554	0,357
0,372	0,356	0,910	0,466	0,426	0,304	0,976	0,807	0,991	0,256
0,952	0,053	0,705	0,817	0,973	0,466	0,300	0,750	0,351	0,776
0,074	0,1098	0,064	0,358	0,487	0,511	0,373	0,986		

Doit - on refuser ce générateur au risque de 5% de se tromper ?

Etant donné le niveau de risque fixé, la zone de rejet se définit ainsi :

$$W=\{ (x_1,x_2,\dots,x_i,\dots,x_{48})/z_{48} < z_{2.5\%}=-1.96 \text{ ou } z_{48} > z_{97.5\%}=1.96\},$$

Avec les données fournies, la statistique Z_{48} prend la valeur $z_{48} = 0.25$. Cette valeur n'appartient pas à la zone de rejet W. En conséquence, au risque de 5% de se tromper on ne peut pas considérer le générateur d'Excel comme un « mauvais » générateur de nombres aléatoires.

Remarque importante :

Le test de la moyenne est manifestement insuffisant pour tester un générateur. Comme son nom l'indique, il s'agit simplement d'un test de l'égalité de la moyenne à la valeur ½ . Or il n'ya pas que la loi uniforme continue sur [0.1] qui a une moyenne égale à ½. En conséquence, ce test peut accepter des données provenant d'autres lois de probabilité ayant une moyenne égale à ½. Le caractère uniforme n'est donc pas testé. A noter cependant, que ce test suffit pour refuser un générateur. En effet, une moyenne différente de ½ est incompatible avec la loi uniforme continue sur [0,1].

3.2 **Test d'adéquation de khi-deux**

Avec le test de la moyenne, il s'agissait de comparer la moyenne empirique avec la moyenne théorique. On a fait remarquer qu'une telle procédure est insuffisante pour juger la qualité d'un générateur. L'idée intuitive derrière le test d'adéquation de khi-deux est de comparer toute la distribution empirique, (et non pas seulement sa moyenne) avec son équivalent théorique sous l'hypothèse nulle. En effet, lorsque le générateur utilisé est « mauvais », ces deux distributions doivent se distinguer assez significativement.

D'une manière formelle, soit $x_1, x_{2, \dots,} x_{i,\dots,} x_n$ n nombres donnés par un générateur. Notons $n_1, n_2,\dots,n_j,\dots, n_k$ les effectifs correspondant à une répartition en k classes d'amplitude égale de ces n nombres. Ces effectifs sont des variables aléatoires. On les appelle les effectifs empiriques.

On définit également des effectifs théoriques qu'on note $n_1^*, n_2^*,\dots,n_j^*,\dots, n_k^*$, ceux qui prévalent lorsque H_0 est vraie. Par définition, l'on a ainsi :

$$n_j^* = \frac{n}{k} \forall\, j = 1 \text{ à } k$$

A ce niveau, on démontre que la statistique Dn définie par :

$$D_n = \sum_{j=1}^{k} \frac{(n_j - n_j^*)^2}{n_j^*}$$

converge en loi sous H_0 vers la loi de Khi deux à (k-1) degrés de liberté.

C'est cette statistique qui fond le test d'adéquation de Khi deux. Intuitivement, cette statistique devrait prendre des valeurs assez proche de zéro lorsque H_0 est vraie, c'est-à-dire lorsque le générateur utilisé est de bonne qualité. En effet, sous cette hypothèse, il ne devrait pas y avoir de différences significatives entre les effectifs empiriques et leurs équivalents théoriques.

La région critique s'en déduit directement étant donné un niveau de risque de première espèce α :

$$W=\{ (x_1,x_2,\ldots,x_i,\ldots,x_n)/d_n > d_{1-\alpha}\,(k-1) \},$$

où la quantité $d_{1-\alpha}\,(k-1)$ représente la quantile d'ordre $(1-\alpha)$ de la loi de Khi deux à (k-1) degrés de liberté.

Pour appliquer en pratique un test d'adéquation de Khi deux sur un jeu de données issues d'un générateur $(x_1,x_2,\ldots,x_i,\ldots,x_n)$, on passe par les étapes suivantes :

- Construire la distribution empirique en se donnant d'abord le nombre k de ses classes (on choisit k entre 8 et 12 en pratique et des classes de même amplitude) et en comptant pour chacune de ces classes les nombres d'observations y appartenant. Ces nombres définissent les effectifs empiriques notés n_j .
- On calcule pour chaque classe j la distance de Khi deux correspondante définie par $d_j = (n_j-n_j^*)^2/n_j^*$ et leur somme d_n.
- On compare cette somme avec la quantité $d_{1-\alpha}\,(k-1)$ représentant la quantile d'ordre $(1-\alpha)$ de la loi de Khi deux à (k-1) degrés de liberté pour juger la qualité du générateur.

Exemple : Considérons les mêmes données issues du générateur du tableur Excel ci-dessus présentées. Au risque de 5% de se tromper refuse t- on ce générateur en utilisant le test d'adéquation de Khi deux ?

En fixant le nombre de classes à 8, l'on obtient la distribution suivante :

Classes	*Effectif réels*	Distance
0, 000 - 0,125	6	0
0,125 - 0,250	6	0
0,250 - 0,375	5	0,17
0,375 - 0,5	6	0,00
0,500 - 0,625	4	0,67
0,625 - 0,75	8	0,67
0,750 - 0,875	7	0,17
0,875 -1,000	6	0,00
Total	48	1,67

La dernière colonne du tableau donne la distance de Khi deux de chaque classe et leur somme d_{48} =1.67. Cette somme est inférieure au quantile d'ordre 95% de la de khi deux à 7 degrés de liberté qui vaut 14. En conséquence, on ne peut pas refuser ce générateur au risque de 5% de se tromper.

Remarque

Considérons le générateur de congruence linéaire défini par : $s_n =(s_{n-1}+1) \bmod 2^{10}$. En fixant s_0 à 0, ce générateur donne : 0,1,2, 3, 4,…,1023. Manifestement il s'agit

d'un très mauvais générateur. Cependant, en divisant ces nombres par leur maximum, on obtient une suite de nombres bien uniformément répartis entre 0 et 1. La statistique D_{1024} prend la valeur 0. On accepte ainsi avec le test d'adéquation de khi deux ce générateur comme étant un « bon » générateur de nombres aléatoires.

Ce contre exemple montre ainsi que le test d'adéquation de Khi deux est insuffisant pour juger la qualité d'un générateur. Il ne teste en fait qu'un seul aspect de la qualité d'un générateur à savoir l'uniformité. L'indépendance, l'autre aspect de la qualité d'un générateur, n'est pas prise en considération.

3.4 Test d'indépendance

Le paragraphe précédent a été achevé en faisant remarquer la nécessité de compléter le test d'adéquation de khi deux par un test d'indépendance. Il existe dans la littérature statistique plusieurs types de test d'indépendance. Nous présentons dans ce qui suit un test connu sous l'appellation « Run test » ou test des séquences croissantes et décroissantes.

Soit une suite de n valeurs données par un générateur de nombres pseudo aléatoires : $x_1, x_2, .., x_i, \ldots, x_n$. On symbolise par + une différence positive entre deux x_i successifs et par - une différence positive entre deux x_i successifs. On appelle séquence croissante (respectivement décroissante) une succession de symboles « + » (respectivement « -«)

Intuitivement un « bon » générateur ne doit pas donner une seule séquence croissante de valeurs, ni une seule séquence décroissante de valeurs, ni non plus une séquence alternée de valeurs car cela permettrait de prévoir les valeurs successives ce qui est contraire au caractère aléatoire demandé. Les valeurs fournies par un « bon » générateur devraient constituer des séquences croissantes et décroissantes en nombre « suffisant ».

Plus formellement soit R le nombre de séquences croissantes et décroissantes qu'on relève dans la suite des valeurs données par un générateur. Ce nombre est à priori une variable aléatoire. On démontre que sous H_0 elle suit asymptotiquement une loi normale de moyenne m et de variance σ^2 définis par :

$$m = (2n-1)/3 \text{ et } \sigma^2 = (3n- 5)/18$$

En conséquence, la statistique $Z = (R-m)/ \sigma$ suivant asymptotiquement N(0,1) peut fonder un test d'indépendance des valeurs successives fournies par un générateur. Plus précisément, pour un niveau de risque α donné, on refuse H_0 chaque fois que la valeur observée z dépasse $t_{1-\alpha/2}$ ou est inférieur à $t_{\alpha/2}$ ($t_{1-\alpha/2}$ et $t_{\alpha/2}$ étant les quantiles de rang $(1-\alpha/2)$ et $\alpha/2$ de la loi normale centrée réduite)

Exemple : Un générateur donne les valeurs suivantes : 0.604, 0.091, 0.297, 0.059, 0.776, 0.120, 0.48, 0.005, 0.075, 0.306, 0.392, 0.608, 0.382, 0.783, 0.717, 0.355, 0.815, 0.829, 0.493 ,0.061, 0.743, 0.358, 0.275, 0.149, 0.237.

Peut on, au risque de 5% de se tromper, refuser l'hypothèse nulle d'indépendance (et donc conclure que ce générateur est mauvais) ?

Trouvons r la réalisation de R. Pour cela, symbolisons par « + » et par « -« les différentes séquences croissantes et décroissantes :
-+-+---++++-+--++--+---+

On note la présence de 7 séquences décroissantes et 7 séquences croissantes. La valeur r vaut donc 14. Comme n vaut 25, m est égal à 49/3 soit 16.33 et σ^2 = (75- 5)/18 (=3.88). On en déduit la valeur z = -1.18 qui est bien comprise entre les seuils -1.96 et 1.96 correspondant aux quantiles de rang 2.5% et 97.5% de la loi normale centrée réduite. Les données observées ne permettent pas de refuser ce générateur.

Annexe

La loi uniforme continue standard

On dit qu'une variable aléatoire réelle absolument continue suit la loi uniforme continue standard ($X \rightarrow U(0,1)$) si sa densité de probabilité f est définie par :

$$f(x) = 1_{[0,1]}$$

On en déduit que F la fonction de répartition de X est donnée par

$$F(x) = 0 \text{ si } x \leq 0$$
$$= x \text{ si } x \in [0,1]$$
$$= 1 \text{ si } x \geq 1$$

Une propriété caractéristique de la loi uniforme continue sur [0,1] est que la probabilité d'un intervalle vaut sa longueur et ce indépendamment de la position qu'il occupe sur le support [0,1]:

$$P([a,b]) = (b-a) \ \forall \ [a,b] \subset [0,1].$$

On en déduit que tous les sous intervalles de [0,1] ayant la même longueur ont la même probabilité. Par exemple $P([0.1,0.2]) = P([0.2,0.3]) = P([0.3,0.4]) = \ldots$ $P([0.8,0.9]) = P([0.9,1]) = 0.1$, traduisant ainsi l'uniformité de la distribution sur l'intervalle [0,1].

Génération physique de la loi uniforme standard :

Soit l'expérience aléatoire ξ consistant à choisir au hasard un point du segment [0,1]. Notons l'ensemble Ω des résultats possibles de cette expérience. Soit maintenant la variable X de Ω dans Ω définie par $X(\omega) = \omega \ \forall \ \omega \in \Omega$. On peut montrer que X suit la loi uniforme continue sur [0,1].

En conséquence, pour générer des valeurs d'une variable aléatoire suivant la loi uniforme continue, on peut procéder à la réalisation de cette expérience aléatoire.

Chapitre 3

SIMULATION DE LOIS NON UNIFORMES A UNE SEULE DIMENSION

Le chapitre précédent a proposé des formules spécifiques permettant de simuler des réalisations d'une variable aléatoire suivant la loi uniforme standard. Il en est autrement en ce qui concerne les autres lois de probabilité. En effet, pour simuler une loi de probabilité non uniforme, on passe d'abord par la simulation de la loi uniforme standard. Ensuite, on transforme les valeurs obtenues, selon des techniques appropriées, pour obtenir les valeurs demandées.

Ce sont ces techniques de transformation des valeurs issues de la loi uniforme standard qui définissent les méthodes de simulation de lois non uniformes. De telles méthodes existent en grand nombre dans la littérature statistique. Elles se distinguent notamment sur le plan informatique en donnant plus ou moins rapidement la quantité de valeurs simulées demandée.

Dans ce qui suit, on se limite à certaines d'entre elles qui sont les plus utilisées.

1. GENERALITES

La simulation de valeurs d'une variable aléatoire réelle suivant une loi de probabilité autre que la loi uniforme standard repose sur une propriété mathématique connue. Nous rappelons dans ce qui suit cette propriété et nous montrons comment elle est utilisée pour simuler des lois de probabilité non uniformes.

1.1 Rappel d'une propriété mathématique

Toute variable aléatoire X à valeurs dans R^p peut être simulée sous la forme :

$$X = f(U)$$

où $U = (U_1, U_2, \ldots, U_q)$ est uniformément répartie sur $[0; 1]^q$

La fonction f de R^q dans R^p est borélienne et a ses points de discontinuité dans un ensemble Lebesgue-négligeable.

Remarques :

- Il s'agit d'une égalité en loi
- La fonction a une expression explicite et n'est pas nécessairement unique.

Nous ne donnons pas une démonstration à cette propriété générale. Nous nous contentons de la vérifier dans certains cas particuliers qui apparaitront dans la suite de ce chapitre.

1.2 Démarche de génération

Soit X une variable aléatoire réelle. On souhaite disposer de n valeurs simulées $x_1, x_2, \ldots, x_n$ de X .

Lorsque X suit la loi uniforme continue sur [0,1], on sait que les générateurs de nombres pseudo aléatoires étudiés dans le chapitre précédents permettent de fournir des nombres compris entre 0 et 1 qu'on peut assimiler à des valeurs simulées de X.

Dans le cas où X est une variable aléatoire réelle suivant une loi de probabilité d'un autre type, on peut adopter la démarche générale suivante qui est basée sur la propriété mathématique ci-dessus présentée.

- Générer en utilisant un « bon » générateur de nombres pseudo aléatoires, une suite indépendante de valeurs d'une variable aléatoire réelle U suivant la loi uniforme continue sur [0,1] : $u_1, u_2, \ldots$.
- Trouver une transformation *f* permettant de passer de *U* à *X* et consommant le moins possible du temps de calcul.
- Appliquer cette transformation pour trouver les valeurs $x_1, x_2, \ldots$. comme fonction des valeurs $u_1, u_2, \ldots$.

L'objet de ce cours est de présenter quelques unes de ces transformations qui sont les plus usuelles en ce qui concerne les lois unidimensionnelles.

Remarques :

- Il est possible qu'il existe plusieurs transformations permettant de passer de *U* à *X*. On choisira évidemment celle consommant le moins de temps de calcul.
- Des procédés permettant de simuler directement une loi non uniforme sans passer par la loi uniforme existent dans certains cas particuliers. Etant relativement récents, ces procédés ne sont pas étudiés dans ce cours.

2. LA METHODE D'INVERSION

Appelée également la méthode de la fonction réciproque, cette méthode est la plus directe des méthodes de transformation. Pour des raisons pédagogiques, nous distinguons le cas des variables absolument continues du cas des variables aléatoires discrètes. Au préalable nous rappelons la définition et les propriétés de la fonction de répartition d'une variable aléatoire réelle

2.1 Fonction de répartition

La fonction de répartition d'une variable aléatoire X est la fonction F de R dans R définie par

$$\mathbf{F(x) = P(X \leq x)}$$

La fonction de répartition d'une variable aléatoire réelle présente les propriétés suivantes :

- F(R) =[0,1]
- F est croissante au sens large.
- F est partout continue à droite (pouvant présenter des discontinuités à gauche)
- $F(+\infty) = 1$ et $F(-\infty)=0$

En outre, si X est une variable continue ($X(\Omega)$ est infini non dénombrable), alors F est partout continue aussi bien à droite qu'à gauche. Si en plus F est dérivable, on dit que X est absolument continue. Dans ce dernier cas, la loi de X est également caractérisée par sa densité de probabilité f définie par $f(x) = F'(x)$.

2.1. Cas des variables continues.

On sait que dans ce cas la fonction de répartition F est strictement croissante et partout continue. L'application réciproque existe donc et est également strictement croissante et partout continue.

Proposition : Soit X une variable aléatoire absolument continue et F sa fonction de répartition. Alors U = F(X) suit la loi uniforme continue sur [0,1].

En effet, notons G la fonction de répartition de U. Par définition,

$$G(u) = P(Y \leq u) = P(F(X) \leq u)$$

soit,

$$G(u) = P(X \leq F^{-1}(u))$$

d'où

$$G(u) = F(F^{-1}(u)) = u$$

ce qui correspond à la fonction de répartition d'une variable aléatoire réelle suivant la loi uniforme continue sur [0,1].

Ce qui précède montre que $X = F^{-1}(U)$ et autorise à utiliser l'algorithme suivant pour simuler des valeurs de X:

- Déterminer l'expression de F^{-1} à partir de F (programmer F^{-1})

- Générer, en utilisant un bon générateur de nombres pseudo aléatoires, n valeurs $u_1, u_2, \ldots, u_n$.
- Déterminer les valeurs simulées de X en utilisant l'expression $x_i = F^{-1}(u_i)$.

Exemple :

Soit X suivant la loi exponentielle e(θ). On sait alors que F prend la forme suivante :

$$F(x) = 1- \exp(-\theta x) \ \forall \ x > 0$$

On en déduit l'expression de F^{-1}

$$X = F^{-1}(u) = \left(\frac{\ln(1-U)}{-\theta}\right).$$

Ainsi si $u_1, u_2, \ldots, u_i, \ldots, u_n$ est une suite de n nombres pseudo aléatoires, on en déduit une suite de n valeurs simulées de X comme suit :

$$x_1 = \left(\frac{\ln(1-u_1)}{-\theta}\right), x_2 = \left(\frac{\ln(1-u_2)}{-\theta}\right) \ldots\ldots x_i = \left(\frac{\ln(1-u_i)}{-\theta}\right) \ldots\ldots x_n = \left(\frac{\ln(1-u_n)}{-\theta}\right).$$

Remarques :

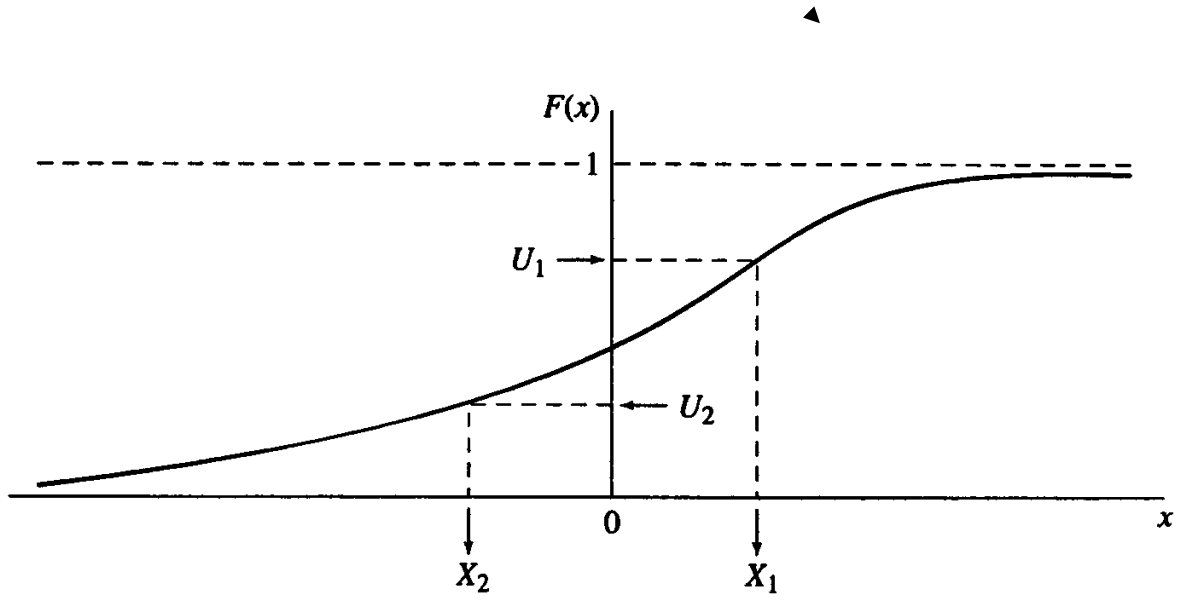
- On peut facilement démontrer que si U suit la loi uniforme continue sur [0,1], la variable V = 1-U suit aussi la loi uniforme continue sur [0,1]. On peut par conséquent calculer aussi x_i selon l'expression :

$$x_i = \left(\frac{\ln(u_i)}{-\theta}\right)$$

ce qui est plus intéressante sur le plan calcul informatique.

- En pratique, on commence par déterminer analytiquement l'expression de F^{-1}. Ensuite, on crée un programme informatique faisant appel au générateur de nombres pseudo aléatoire disponible et calculant ensuite les valeurs simulées x_i selon l'expression de F^{-1} avec les valeurs u_i données par le générateur.

Illustration graphique



2.2. Cas de variables discrètes.

Lorsque X est discrète, sa fonction de répartition F n'est pas partout continue et donc son inverse F^{-1} n'existe pas. Mais on peut définir sa pseudo inverse F^* par :

$$F^*(u) = \inf \{x \in E / F(x) \geq u\} \ \forall\ u \in [0,1]$$

où E est l'ensemble des valeurs possibles prises par X

Illustration soit X une variable discrète suivant la loi binomiale B(3,0.5). On en déduit directement la fonction de répartition de X :

- $F(x) = 0 \ \forall\ x < 0$
- $F(x) = 1/8 \ \forall\ x \in [0,1[$
- $F(x) = 1/2 \ \forall\ x \in [1,2[$
- $F(x) = 7/8 \ \forall\ x \in [2,3[$
- $F(x) = 1 \ \forall\ x \geq 3$

Déterminons par exemple $F^*(0.35)$. Par définition on a :
$F^*(0.35) = \inf \{ t \in E / F(t) \geq 0.35\} = \inf \{ 1,2,3\} = 1$

Remarque : $F^* = F^{-1}$ si F est bijective (cas de variable absolument continue)

Propositions

- Si $F^*(u) \leq a$ alors $u \leq F(a)$

En effet, puisque F est croissante (au sens large), on a $F(F^*(u)) \leq F(a)$. D'autre part, on sait que $F(F^*(u)) = F(\inf\{t \in E / F(t) \geq u\}) \geq u$ et donc $u \leq F(a)$.

• Soit U suivant la loi uniforme continue sur [0,1] et F* la pseudo inverse d'une fonction F croissante et continue à droite alors la variable $X = F^*(U)$ a pour fonction de répartition la fonction F.

En effet, trouvons la fonction de répartition de X :

$$P(X \leq x) = P(F^*(U) \leq x)$$

Soit compte tenu de la première proposition :

$$P(X \leq x) = P(U \leq F(x)) = F(x)$$

On déduit de ce qui précède l'algorithme suivant pour simuler des valeurs de X:

- Déterminer l'expression de F^* à partir de F (programmer F^*)
- Générer, en utilisant un bon générateur de nombres pseudo aléatoires, n valeurs $u_1, u_2, \ldots, u_n$.
- Déterminer les valeurs simulées de X en utilisant l'expression $x_i = F^*(u_i)$

Exemple : Simulation de X suivant la loi binomiale B(3,0.5)

Déterminons d'abord F* la pseudo inverse de F. L'on a par définition :

- $F^*(u) = 0 \ \forall u \leq 1/8$
- $F^*(u) = 1 \ \forall u \in]1/8, ½]$
- $F^*(u) = 2 \ \forall u \in]1/2, 7/8]$
- $F^*(u) = 3 \ \forall u \in]7/8, 1]$

Ainsi, si on a $u_1 = 0.21$; $u_2 = 0.005$; $u_3 = 0.95$, on prend $x_1 = 1$, $x_2 = 0$ et $x_3 = 3$.

c. Limites de la méthode

La principale insuffisance de la méthode est qu'elle nécessite la connaissance de l'expression explicite de la fonction de répartition. Or plusieurs lois de probabilité dont notamment la loi normale n'ont pas cette propriété.

En outre, la méthode est en comparaison avec d'autres, assez encombrante en place mémoire notamment dans le cas discret.

3. SIMULATION DE LA LOI NORMALE CENTREE REDUITE

Comme il a été déjà constaté la méthode d'inversion ne permet pas de simuler la loi normale. Or cette loi est d'une grande importance aussi bien théorique que pratique. Il convient donc de trouver d'autres méthodes. Celles-ci sont assez nombreuses. Nous nous contentons dans ce qui suit à deux d'entre elles : La première méthode conduit à des valeurs approchées, la seconde fournit quant à elle des valeurs exactes.

3.1 Méthode approchée

Cette méthode est basée sur le théorème central limite. Soit $\{X_j\}_{j=1 \text{ à } m}$ un échantillon IID(μ,σ^2). Rappelons que cela signifie que $\forall$ j = 1 à m, les X_j sont indépendantes et suivent la même loi de probabilité, la quelle loi possède une espérance mathématique μ et une variance σ^2. On sait alors d'après le théorème central limite que la variable aléatoire réelle définie par :

$$Y= \frac{(\sum_{j=1}^{m} X_j)-m\mu}{\sqrt{m\sigma^2}}$$

suit approximativement, pour m assez grand, la loi normale centrée réduite.

En particulier en prenant m = 12 et lorsque les X_j suivent chacune la loi uniforme continue standard $\mathcal{U}(0,1)$, μ vaut alors ½ et $\sigma^2 = {}^1/_{12}$, l'on a :

$$Y= (\sum_{j=1}^{12} X_j) - 6$$

Une procédure de simulation de n valeurs approchées de la loi normale centrée réduite s'en déduit directement :

- Générer, en utilisant un « bon » générateur de nombres pseudo aléatoires, 12.n valeurs simulées d'une variable aléatoire X suivant $\mathcal{U}(0,1)$:

$(x_{1,1},x_{2,1},\ldots,x_{12,1})$, $(x_{1,2},x_{2,2},\ldots,x_{12,2})$,…., $(x_{1,i},x_{2,i},\ldots,x_{12,i})$,…., $(x_{1,n},x_{2,n},\ldots,x_{12,n})$

- Calculer y_1, y_2,… y_i,… y_n en appliquant la définition précédente, soit :

$$y_i = (\sum_{j=1}^{12} x_{j,i}) - 6$$

3.2 Méthode exacte.

Cette méthode connue sous le nom d'algorithme de **Box Muller** est basée sur le théorème suivant :

Théorème de Box Muller : Soient *X* et *Y* deux variables aléatoires réelles indépendantes suivant chacune la loi normale centrée réduite N(0,1). On établit les égalités en loi suivantes :

$$X = \left(\sqrt{-2\ln U}\right)\cos(2\pi V)$$
$$Y = \left(\sqrt{-2\ln U}\right)\sin(2\pi V)$$

ou U et V sont deux variables aléatoires réelles indépendantes suivant la loi uniforme standard $\mathcal{U}(0,1)$

Démonstration :

En coordonnées polaires les variables *X* et *Y* s'écrivent :

$$X = R\cos A \quad (1)$$
$$Y = R\sin A \quad (2)$$

ou *R* (le module) est une variable aléatoire ayant pour support l'ensemble des réels positifs et *A* (l'angle) est une autre variable ayant pour support l'intervalle $[0,2\pi]$.

On note que $R^2 = X^2 + Y^2$. Comme *X* et *Y* sont indépendantes et suivant chacune N(0,1), il s'ensuit que R^2 suit la loi de khi deux à deux degrés de liberté ou encore la loi exponentielle de paramètre ½.

Notons *H* la fonction de répartition de *R*. On a donc :

$$H(r) = P(R \leq r) = P(R^2 \leq r^2)$$

Comme R^2 suit la loi exponentielle de paramètre 1/2; il vient que :

$$H(r) = 1 - e^{-r^2/2}$$

d'où :

$$H^{-1}(u) = \sqrt{-2\ln(1-u)}$$

On en déduit directement l'égalité en loi suivante (voir méthode d'inversion) :

$$\mathbf{R} = \sqrt{-2\ln U} \text{ ou U suit } \mathcal{U}(0,1) \quad (3)$$

Considérons maintenant la transformation suivante :

$$(X,Y) \longrightarrow (R, A)$$

et notons f la densité de probabilité de (X,Y), soit :

$$f(x,y) = \frac{1}{2\pi} e^{-(\frac{x^2+y^2}{2})} \forall\ (x,y) \in R^2$$

D'après la formule du changement de variables, la densité g du couple (R,A) est donnée par :

$$g(r,a) = f(r\cos a, r\sin a)|J|$$

ou |*J*| est le jacobien de la transformation précédente. En remplaçant, on trouve :

$$g(r,a) = \frac{1}{2\pi}[e^{-(\frac{r^2}{2})}r] \ \forall \ r \geq 0 \text{ et } a \in [0,2\pi]$$

Le terme entre crochets n'est autre que la densité h de La variable R. En effet, on vérifie que :

$$h(r) = H'(r) = r\, e^{-r^2/2} \ \forall \ r \geq 0$$

Déterminons maintenant la densité l de la variable A. Par définition :

$$l(a) = \int_0^{+\infty} g(r,a)dr = \frac{1}{2\pi}$$

Il s'ensuit que R et A sont indépendantes et que A suit la loi uniforme continue sur [0,2π]. On sait alors que L fonction de répartition de A s'écrit:

$$L(a) = a/2\pi$$

et donc :

$$a = L^{-1}(v) = 2\pi\, v$$

ce qui permet d'écrire l'égalité en loi suivante :

$$\mathbf{A = 2\pi\, V} \quad \text{ou V suit } \mathcal{U}(0,1) \quad (4)$$

En confrontant les égalités (1), (2), (3) et (4) on aboutit au résultat à démontrer.

<u>Utilisation pratique</u> :

Pour simuler 2n valeurs d'une variable aléatoire suivant N(0,1), on peut procéder comme suit :

- Générer, en utilisant un bon générateur de nombres pseudo aléatoires, 2 valeurs u et v d'une variable suivant $\mathcal{U}(0,1)$: .
- Calculer x et y ainsi :

$$x = \left(\sqrt{-2\ln u}\right)\cos(2\pi v)$$
$$y = \left(\sqrt{-2\ln u}\right)\sin(2\pi v)$$

- Répéter ces deux étapes n fois

4. **METHODE DE TRANSFORMATION**

4.1 Présentation

Soit Y et X deux variables aléatoires réelles telles que Y=T(X) où T est une transformation connue.

Supposons qu'on dispose déjà de n valeurs simulées $x_1, x_2,\ldots,x_i,..,x_n$ de X. Puisque T est connue, on peut en déduire directement n valeurs simulées $y_1, y_2,\ldots,y_i,..,y_n$ de Y en appliquant simplement cette transformation, soit

$$y_i = T(x_i) \ \forall \ i = 1 \text{ à } n$$

Une telle procédure de simulation est appelée méthode de transformation.

Remarques :

- La méthode d'inversion est un cas particulier de méthode de transformation où X suit $\mathcal{U}(0,1)$
- La méthode de transformation est utilisée chaque fois que la variable Y est difficile à simuler directement.

4.2 Quelques cas particuliers

- **Loi normale** (quelconque) : Soit Y suivant $N(m,\sigma^2)$ où m et σ^2 sont données. On sait alors qu'on peut écrire

$$Y = \sigma X + m$$

où X suit N(0,1).

En conséquence, pour simuler n valeurs $y_1, y_2,\ldots,y_i,..,y_n$ de Y on peut procéder comme suit :

- Utiliser l'algorithme de Box Muller pour simuler n valeurs $x_1, x_2,\ldots,x_i,..,x_n$ de X.
- En déduire n valeurs simulées $y_2,\ldots,y_i,..,y_n$ de Y en posant :

$$y_i = \sigma x_i + m \ \forall i = 1 \text{ à } n$$

- **Loi log-normale** : On dit que la variable Y suit la log-normale de paramètres m et σ^2 (donnés) si

$$Y = exp(X)$$

où X suit $N(m,\sigma^2)$.

Aussi, si on veut simuler *n* valeurs $y_1, y_2,\ldots,y_i,..,y_n$ de *Y* la procédure suivante peut être utilisée :

- Utiliser la procédure précédente pour simuler n valeurs $x_1, x_2,\ldots,x_i,..,x_n$ de X suivant N(m. σ^2).
- En déduire n valeurs simulées $y_2,\ldots,y_i,..,y_n$ de *Y* en posant :

$$y_i = exp(x_i) \quad \forall i = 1 \text{ à } n$$

- **Loi uniforme continue sur [a,b]**

Soit X suivant la loi uniforme continue sur [a,b]. On établit très facilement que l'on peut écrire l'égalité en loi suivante :

$$X = (b-a).U+a$$

où U suit la loi uniforme standard $\mathcal{U}(0,1)$

On en déduit directement une procédure de simulation de X :

- Générer, en utilisant un « bon » générateur de nombres pseudo aléatoires, n valeurs simulées d'une variable aléatoire U suivant $\mathcal{U}(0,1)$: $u_1,u_2,\ldots,u_n$,
- Calculer $x_1, x_2,\ldots x_i,\ldots x_n$ en appliquant la définition précédente, soit : $x_i = (b-a)u_i + a$
- Répéter les deux premières étapes pour i = 1 à n

5. LA METHODE DE REJET

Comme il va être précisé par la suite, la méthode de rejet est celle que l'on utilisera lorsque les méthodes précédentes ne sont pas applicables pour simuler une loi de probabilité donnée. C'est le cas notamment de la simulation de la loi Gamma.

La méthode de rejet est en rapport avec les procédures de simulation des lois conditionnelles. Aussi commençons-nous par présenter ces procédures.

5.1 Simulation des lois conditionnelles

Un exemple introductif permet de mieux comprendre les procédures de simulation des lois conditionnelles**.**

5.1.1 Exemple introductif

Soit à choisir au hasard un élève âgé de plus de 21 ans dans une classe de 50 élèves. On sait que dans cette classe il y'a trente élèves âgés de plus de 21 ans et donc vingt âgés de 21ans.

Deux procédures de choix sont à priori possibles. Une première manière de procéder consiste à effectuer d'abord un tri dans la liste des 50 élèves selon l'âge ce qui permet de créer deux sous listes : une pour les moins de 21 ans et une autre pour les plus de 21 ans. Ensuite, on choisit au hasard un élève dans la sous liste des élèves de plus de 21 ans. Il est aussi possible de procéder comme suit : On choisit au hasard un élève dans la liste des 50 élèves. S'il est âgé de plus de 21 ans on le retient. Sinon on **rejette** ce choix et on recommence l'expérience jusqu'à l'obtention d'un élève âgé de plus de 21 ans.

On peut noter que les deux procédures conduisent au même résultat à savoir un élève âgé de plus de 21 ans. Elles se distinguent cependant sur deux aspects :

- La première nécessite un tri préalable mettant en évidence la sous population concernée.
- La deuxième, en procédant au rejet du choix issu de la population non concernée, consomme en moyenne plus de temps pour sa réalisation.

5.1.2 Procédures de simulation

De l'exemple précédent, on peut faire les remarques suivantes :

- L'expérience aléatoire considérée consiste en un choix sous une condition donnée. La loi de probabilité correspondante est ainsi une loi de probabilité conditionnelle. En désignant par Ω l'ensemble de 50 nombres correspondant aux âges des 50 élèves, cette loi conditionnelle est en fait la loi uniforme discrète sur l'ensemble Ω conditionnellement à ce que l'âge tiré est supérieur à 21 ans.
- On note l'existence de deux procédures de simulation de cette loi conditionnelle : une procédure **directe** de simulation de cette loi (Choix dans la sous liste des élèves de plus de 21 ans) et une procédure **indirecte** simulant la loi marginale (la loi uniforme discrète sur Ω sans condition) suivie par une décision de rejet lorsque l'évènement conditionnant n'est pas réalisé.

Plus formellement, soit à simuler une variable aléatoire X selon une certaine loi de probabilité conditionnellement à la réalisation d'un événement A donné. Notons G la fonction de répartition de X et F sa fonction de répartition conditionnellement à A :

$$G(x) = P(X \leq x)$$
$$F(x) = P(X \leq x / A)$$

L'exemple introductif suggère deux procédures de simulation de la loi de X conditionnellement à la réalisation de l'évènement A :

- Première procédure : On simule X directement selon sa loi conditionnelle F , par exemple en utilisant la méthode d'inversion lorsque c'est possible.

• Deuxième procédure : On simule X selon sa loi marginale G et on vérifie la valeur simulée x. Si cette valeur satisfait la condition de réalisation de l'évènement A on l'accepte. Sinon on rejette cette valeur et on recommence la simulation selon G jusqu'à l'obtention d'une valeur vérifiant la condition.

Comme, il a été déjà remarqué, la deuxième procédure consomme plus de temps de calcul. On l'utilise cependant en pratique notamment lorsqu'il n'est pas possible d'utiliser la première procédure. C'est l'idée derrière la méthode de rejet qui sera développée dans la section suivante.

5.2 Présentation de la méthode de rejet

Soit f une densité difficilement ou carrément non simulable par les méthodes usuelles. Si on arrive à écrire cette densité comme une densité marginale g, facile à simuler, conditionnellement à un évènement A donné, on peut alors utiliser la procédure indirecte de simulation d'une loi conditionnelle telle que présentée ci-dessus. C'est la démarche suivie par la méthode de rejet.

5.2.1 Fondement mathématique

La méthode de rejet est basée sur la proposition suivante :
Soient f et g deux densités de probabilités sur R vérifiant la relation suivante :

$$\exists\, c > 0 \;\; telque \;\; \forall\, x \in R\,, f(x) \leq c.g(x)$$

Considérons une variable aléatoire réelle X ayant g pour densité de probabilité et Y une autre variable aléatoire réelle indépendante de X et suivant la loi uniforme standard $\mathcal{U}(0,1)$. On établit alors que la loi conditionnelle de X sachant $Y < f(X)/c.g(X)$ a pour densité f (en supposant évidement $g(X) \neq 0$ presque partout).

Démonstration

Notons $q(x) = f(x)/cg(x)$ et remarquons que $0 \leq q(x) \leq 1$. Calculons maintenant $P(Y < q(X))$ en nous plaçant dans le cas absolument continu . Par définition :

$$P\big(Y < q(X)\big) = \int_{x=-\infty}^{x=+\infty} \int_{y=0}^{y=q(x)} h(x,y)dxdy$$

où h est la densité de probabilité du couple (X,Y). Mais comme X et Y sont indépendantes $h(x,y) = 1.g(x)=g(x)$. Il s'en suit que :

$$P\big(Y < q(X)\big) = \int_{x=-\infty}^{x=+\infty} g(x)[\int_{y=0}^{y=q(x)} dy]dx$$

soit comme Y suit $\mathcal{U}(0,1)$,

$$P\big(Y < q(X)\big) = \int_{x=-\infty}^{x=+\infty} g(x)q(x)\,dx$$

D'où étant donné la définition de q(x) :

$$P\big(Y < q(X)\big) = \frac{1}{c}\int_{x=-\infty}^{x=+\infty} f(x)\,dx = \frac{1}{c}$$

Cherchons maintenant $P(X \in B \mid Y < q(X))$. Par définition, l'on a :

$$P\big(X \in B \mid Y < q(X)\big) = \frac{P(X \in B \text{ et } Y < q(X))}{P(Y < q(X))}$$

D'après ce qui précède, il vient que :

$$P\big(X \in B \mid Y < q(X)\big) = c\ P\big(X \in B \text{ et } Y < q(X)\big)$$

Ou encore :

$$P\big(X \in B \mid Y < q(X)\big) = c \int_{x\in B} g(x)[\int_{y=0}^{y=q(x)} dy]dx$$

Soit étant donné que Y suit $\mathcal{U}(0,1)$ et en remplaçant :

$$P\big(X \in B \mid Y < q(X)\big) = c \int_{x\in B} g(x)q(x)]dx = \int_{x\in B} f(x)dx$$

Ce qui prouve que f est la densité de probabilité de la loi conditionnelle de X sachant $Y < f(X)/c.g(X)$.

Remarques

- La proposition précédente est aussi vraie pour le cas où X est une variable discrète.
- Cette proposition est aussi valable pour le cas ou X est un vecteur aléatoire.

5.2.2 Procédure pratique de simulation

Soit une variable X telle que sa simulation selon une densité f ne peut pas être effectuée selon les méthodes précédentes. S'il existe une autre densité de probabilité g

de X facile à simuler et ayant le même support que f et un réel c vérifiant : $\forall$ x $\in$R, f(x)< c.g(x) , alors d'après ce qui précède on peut utiliser la procédure suivante, connue sous le nom de méthode de rejet, pour simuler une valeur de X selon f:

- Etape 1 : On simule X selon g. Soit x la valeur obtenue
- Etape 2 : On simule Y selon $\mathcal{U}$(0,1). d'une manière indépendante à X. Soit y la valeur trouvée.
- Etape 3 : On compare y au rapport q(x)= f(x)/cg(x). Si y < q(x) on accepte x comme une valeur simulée de X selon f, sinon on le rejette et on recommence à l'étape 1

Remarques

- La méthode de rejet est évidemment plus lente que les autres méthodes de simulation. On l'utilise seulement lorsque ces autres méthodes ne sont pas applicables.
- Par hypothèse, le nombre c est tel que $\forall\, x \in R\,, f(x) \leq c.g(x)$. En intégrant sur le support de f (et de g), l'on note que c $\geq$ 1. On note aussi qu'il n'est pas unique. En effet, soit c'$\geq$ c, l'on a aussi : $f(x) \leq c'.g(x)$
- On a établit dans ce qui précède que

$$P\big(Y < q(X)\big) = \frac{1}{c}$$

  Le nombre c est donc l'inverse de la probabilité d'acceptation d'un tirage.
- Soit N la variable aléatoire représentant le nombre de tirages de couples (x,y) rejetés avant l'obtention d'une valeur simulée de X selon f. On vérifie que la variable N suit la loi géométrique de paramètre p = 1/c. On sait alors que E(N) = c ce qui signifie que le nombre c représente le nombre moyen de rejet. En rapport avec la deuxième remarque ci dessus, il convient en conséquent de le choisir le plus petit possible. En pratique, on le choisit égal au maximum de h(x)=f(x)/g(x).
- La densité g doit être normalement simple à simuler. En pratique, on considère souvent la densité de la loi uniforme continue sur [a,b] ou la densité de la loi exponentielle comme exemples de densités g.

5.2.3 Exemple

Exemple : Soit X une variable aléatoire réelle ayant la densité de probabilité suivante : $f(x) = \frac{2}{\pi}\sqrt{1-x^2}\,1_{[-1,1]}$

On se propose de simuler X selon la méthode de rejet. Etant le support de f on peut prendre comme densité g celle de la loi uniforme continue sur [-1,1] , soit g(x) = ½ $.1_{[-1,1]}$. Pour trouver c, on cherche le max de h(x)=f(x)/g(x) sur [-1,], soit :

$$h(x) = \frac{4}{\pi}\sqrt{1-x^2}$$

Cette fonction atteint son maximum en x=0. Le nombre c vaut donc $4/\pi$.

En conséquence, pour simuler X selon la méthode de rejet, on procède comme suit :

- Etape 1 : On simule X selon g en utilisant par exemple la méthode d'inversion, soit en notant que G(x) = (x+1)/2 sur [-1,1] et donc $x = G^{-1}(u) = 2u-1$ sur [-1,1],
  - On génère u de U(0,1) en utilisant un « bon » générateur.
  - On calcule x = 2u -1
- Etape 2 : On génère, indépendamment de X, y de U(0,1) en utilisant un « bon » générateur.
- Etape 3 : On calcule q(x) =f(x) /c.g(x) = $\sqrt{1-x^2}$,
  - Si y < q(x) on accepte x comme valeur simulée de X selon f
  - Sinon on rejette x et on recommence à l'étape 1.

Avec u =0.125, soit x=-0.7 5 et avec y = 0.956 on a q(x) =0.661 et donc comme 0.956 > 0.661, on ne doit pas accepter x =-0.75 comme valeur simulée de X selon f.

En revanche avec u =0.754, soit x = 0.508 et avec y =0.254 on a q(x)= 0.861 et donc y < q(x), la valeur x = 0.508 est considérée comme valeur simulée de X selon f.

Chapitre 4

SIMULATION DES LOIS A PLUSIEURS DIMENSIONS

Le chapitre précédent a présenté un certain nombre de méthodes permettant de simuler des lois de probabilités usuelles à une seule dimension. Il peut être noté que ces méthodes peuvent également servir à simuler des lois de probabilités de vecteurs aléatoires dont les composantes sont indépendantes. En effet, si $X= (X_1, \cdots, X_k)$ est formé de variables aléatoires réelles indépendantes, il suffit de simuler indépendamment l'une de l'autre chacune des lois marginales correspondantes aux variables X_i pour obtenir une simulation du vecteur X.

Il convient également de rappeler, que les lois de probabilité usuelles sont utilisées en pratique pour modéliser et ajuster les distributions empiriques observées. En présence d'une seule variable aléatoire réelle, il existe plusieurs choix possibles de loi de probabilité usuelle pour ajuster sa distribution empirique. En revanche, dans le cas de vecteurs aléatoires, les choix sont assez limités. En effet, à part la loi normale multidimensionnelle dans le cas continu ou la loi multinomiale dans le cas discret, la littérature statistique offre peu d'alternatives pour modéliser le comportement aléatoire des grandeurs à plusieurs dimensions. En particulier, il n'existe pas ou peu de loi jointe usuelle dont les lois marginales sont de types différents. Aussi, en présence d'un vecteur aléatoire, ne cherche t- on pas en pratique à déterminer sa loi jointe en la modélisant par une distribution usuelle. On se contente souvent à spécifier les lois marginales de ce vecteur et l'intensité de la liaison probabiliste entre ses composantes.

Dans ce chapitre, nous présentons les méthodes de simulation d'un vecteur aléatoire,

- Lorsque sa loi jointe est connue (cas rare en pratique)
- Lorsque seules ses lois marginales sont données ainsi qu'éventuellement certaines de ses caractéristiques de liaison probabiliste.

Nous examinons également une méthode de simulation spécifique pour la loi normale multidimensionnelle qui se présente comme un cas particulier important.

1. SIMULATION D'UN VECTEUR ALEATOIRE A LOI JOINTE CONNUE.

Pour simplifier la présentation, considérons le cas d'un vecteur aléatoire à deux dimensions qu'on note Z = (X,Y).

Si les variables X et Y sont indépendantes, cela signifie que les chances de réalisation de chacune des valeurs prises par Y ne sont pas affectées par la valeur

réalisée de la variable X. En conséquence, pour simuler un couple (x,y) de Z = (X,Y), on peut simuler séparément X et Y selon leur loi marginale respective.

En revanche, si X et Y sont liés, la réalisation d'une valeur particulière x de X influence les chances de réalisation de chacune des valeurs possibles de Y. Pour bien comprendre cette remarque et celle qui l'a précédée, considérons l'exemple suivant d'une loi discrète à deux dimensions :

Y / X	0	1	P(X=x)
0	$^1/_4$	$^1/_4$	$^1/_2$
1	0	$^1/_2$	$^1/_2$
P(Y=y)	$^1/_4$	$^3/_4$	1

On note que lorsque X prend la valeur 0, Y peut prendre chacune des valeurs 0 et 1 avec la même probabilité. En revanche lorsque X prend la valeur 1, Y ne peut prendre la valeur 0. Seule la valeur 1 est possible.

En conséquence, une procédure de simulation du couple (X,Y) doit tenir compte de la liaison probabiliste entre X et Y. Ainsi, si X peut être simulé « librement », la simulation de Y différera selon la valeur simulée de X. En effet, la loi de probabilité de Y change selon la valeur prise par X. C'est en fait la loi conditionnelle de Y à la valeur simulée de X qu'il convient de considérer.

Plus précisément, notons $F_X(x)$ la fonction de répartition marginale de X et $F_Y(y/x)$ la fonction de répartition de Y conditionnellement à X= x :

$$F_X(x) = P(X \leq x)$$
$$F_Y(y/x) = P(Y \leq y / X=x)$$

Une procédure de simulation d'un couple (x,y) de (X,Y) peut être comme suit :

Etape 1 : Simuler X selon sa loi marginale $F_X(.)$ en utilisant une des procédures présentées dans le chapitre précédent.

Etape 2 : En utilisant également une des procédures présentées dans le chapitre précédent, simuler Y selon sa loi conditionnelle $F_Y(./x)$ où x est précisément la valeur simulée de X trouvée à l'étape 1.

La procédure précédente se généralise assez directement. Ainsi par exemple si on veut simuler un vecteur aléatoire à 4 dimensions Z=(X,Y,S,T), on procède comme suit :

- On simule X selon sa loi marginale.
- On simule ensuite Y selon sa loi conditionnelle à X=x, la valeur x étant celle obtenue dans l'étape précédente.
- Puis on simule S selon sa loi conditionnelle à X=x et Y=y, les valeurs x et y étant celles obtenues dans les étapes précédentes.
- Enfin on simule T selon sa loi conditionnelle à X=x, Y=y et S=s, les valeurs x et y et s étant celles obtenues dans les étapes précédentes.

Remarques :

- La procédure précédente s'applique chaque fois qu'on connait les lois marginales et conditionnelles d'une loi jointe.
- La procédure précédente est évidemment de plus en plus lourde au fur et à mesure que le nombre de dimensions augmente.

Exemple :

Soit un couple (X,Y) de variables aléatoires réelles absolument continues dont la densité de probabilité f est définie par :

$$f(x,y) = \frac{1}{x}\, e^{-x} 1_{\Delta}$$

où Δ est le domaine de R^2 défini par $\Delta = \{(x, y) \in R^2, 0 < y < x\}$.

Nous cherchons à simuler le couple de variables (X,Y) selon la méthode précédente. Déterminons d'abord la loi marginale de X. Par définition, en notant $f_X(.)$ la densité marginale de X, on a $\forall\ x > 0$

$$f_X(x) = \int_0^x f(x,y)dy = \frac{1}{x} e^{-x} \int_0^x dy = e^{-x}$$

La variable X suit ainsi la loi exponentielle de paramètre 1.

Cherchons maintenant la densité conditionnelle de Y sachant X= x. En notant cette densité $f_Y(.\backslash x)$, on a par définition $\forall\ y \in [0, x]$:

$$f_Y(y\backslash x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{x}$$

Ainsi, conditionnellement à X= x, la variable Y suit la loi uniforme continue sur [0,x]. En conséquence, une procédure de simulation de la loi de (X,Y) peut être comme suit :

- Etape 1 : Simuler X selon sa loi marginale. En utilisant par exemple la méthode d'inversion, cela peut se réaliser ainsi
  - Simuler u de $\mathcal{U}(0,1)$ en utilisant un « bon » générateur de nombres pseudo aléatoires
  - Calculer x = -Ln(u)
- Etape 2 : Simuler Y selon sa loi conditionnelle sachant X=x en procédant par exemple comme suit :
  - Simuler v de $\mathcal{U}(0,1)$ en utilisant un « bon » générateur de nombres pseudo aléatoires
  - Calculer y = xv

2. SIMULATION D'UN VECTEUR NORMAL

La simulation d'un vecteur normal peut s'opérer selon la méthode sus présentée. Cependant, l'analyse des propriétés de la loi normale multidimensionnelle conduit à adopter une autre méthode plus facile à mettre en œuvre.

A cet effet commençons par rappeler ces propriétés remarquables de la loi normale multidimensionnelle :

- Les lois marginales et conditionnelles sont également normales
- L'indépendance équivaut à l'absence de corrélation.
- Toute transformation linéaire d'un vecteur normal est également normale
- La connaissance des lois marginales et de la matrice de corrélation équivaut à la connaissance de toute la loi multidimensionnelle.

La simulation d'une loi normale multidimensionnelle quelconque repose sur la proposition suivante :

Soit μ un vecteur de R^k et Σ une matrice carrée de taille k symétrique positive et soit $X = (X_1, \cdots, X_k)$ un vecteur aléatoire de loi N (0, I) et A une matrice carrée d'ordre k telle que $AA' = \Sigma$. Alors le vecteur $Y = AX + \mu$ est un vecteur normal de moyenne μ et de matrice de variances et covariances Σ.

La démonstration de cette proposition est directe. On note en effet que Y est une transformation linéaire de X. Le vecteur Y est donc normal puisque X est normal (voir plus haut). On note aussi que :

$$E(Y) = E(AX+ \mu) = AE(X)+ \mu = \mu, \text{ et}$$
$$V(Y)=V(AX+ \mu)=V(AX)=AV(X)A'=AA'= \Sigma.$$

En conséquence, pour simuler Y suivant N(μ, Σ), on procède comme suit :

- Etape 1 : Simuler X suivant N(0,I). Comme X est un vecteur indépendant, cela revient à simuler d'une manière indépendante ses différentes composantes X_1, $X_2 \cdot \cdot$ et X_k qui suivent chacune la loi normale centrée réduite N(0,1) (en utilisant par exemple l'algorithme de Box Muller).
- Calculer simplement Y = AX + μ

Au préalable, il faut trouver la matrice A vérifiant AA' = Σ. La réponse est donnée par la décomposition de Cholesky

Décomposition de Cholesky

Une matrice réelle B est symétrique définie positive si, et seulement si, il existe une matrice C triangulaire inférieure et inversible telle que B = CC'.

La matrice Σ étant une matrice de variance covariance est symétrique et définie positive, il existe donc une matrice A triangulaire inférieure et inversible telle que Σ = AA'.

En pratique, pour calculer A, il convient de la poser triangulaire inférieure et résoudre colonne par colonne l'équation précédente.

Exemple :

Considérons le cas particulier où k = 2. Soit donc Y = (Y_1, Y_2)' un vecteur gaussien de R^2. Notons μ = (μ_1, μ_2) son espérance mathématique

et $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

On obtient directement la matrice A en posant A = $\begin{pmatrix} a & 0 \\ b & c \end{pmatrix}$

soit AA'= $\begin{pmatrix} a^2 & ab \\ ab & b^2c^2 \end{pmatrix}$ d'où par identification avec Σ :

- a = σ_1
- b = $\rho\sigma_2$
- c = $\sigma_2\sqrt{1-\rho^2}$

En conséquence, l'on peut écrire :

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Soit en développant :

$$\begin{cases} Y_1 = \sigma_1 X_1 + \mu_1 \\ Y_2 = \rho\sigma_2 X_1 + \sigma_2\sqrt{1-\rho^2}X_2 + \mu_2. \end{cases}$$

Une procédure de simulation de $Y = (Y_1, Y_2)'$ se déduit ainsi directement :

- Simuler, par exemple par Box Muller, deux valeurs indépendantes (x_1,x_2) de la loi normale centrée réduite
- Calculer (y_1,y_2) en utilisant l'expression précédente.

3. METHODE DE LA COPULE

Dans certaines applications, on ne cherche pas à simuler un vecteur aléatoire selon sa loi multidimensionnelle, car elle est inconnue, mais à simuler seulement ses composantes selon leur loi marginale respective tout en respectant un niveau donné de liaison probabiliste entre ces composantes. Par exemple, pour une application donnée on souhaite simuler un couple **corrélé** de variables aléatoires (X,Y) suivant respectivement la loi exponentielle de paramètre $\mu = 1$ et la loi de khi –deux à 5 degrés de liberté.

3.1 Présentation du problème

Soit $X = (X_1, X_2, \dots X_j,\dots, X_k)$ un vecteur aléatoire non indépendant. Notons F sa fonction de répartition et $F_1, F_2, \dots F_j,\dots, F_k$ les fonctions de répartitions marginales correspondantes :

$$F(x_1,..., x_j,\dots, x_k) = P(X_1 \le x_1 ..,X_j \le x_j,.. X_k \le x_k)$$

$$F_j(x_j) = P(X_j \le x_j) \ \forall \ j = 1 \text{ à } k$$

Dans le chapitre précédent, on a établi les égalités en loi suivantes :

$$X_j = F_j^{-1}(U_j) \ \forall \ j = 1 \text{ à } k$$

où les variables $U_1, U_2, \dots U_j,\dots, U_k$ suivent chacune la loi uniforme standard $\mathcal{U}(0,1)$.

Il est important de noter que les variables U_j ne sont pas indépendantes car les X_j ne le sont pas par hypothèse.

La fonction de répartition du vecteur $U= (U_1, U_2, \dots U_j,\dots, U_k)$ est couramment appelée **Copule**. On peut vérifier qu'elle est directement liée à la fonction de répartition de X. En effet, en la notant C, l'on a par définition :

$$C(u_1,.., u_j,.., u_k) = P(U_1 \le u_1 ,\dots U_j \le u_j \dots, U_k \le u_k)$$

Soit en remplaçant,

$$C(u_1, u_2, ..,u_j,.., u_k) = P[\ F_1(X_1) \leq\ u_1 ,\dots ,F_j(X_j) \leq\ u_j\dots, F_k(X_k) \leq\ u_k]$$

ou encore,

$$C(u_1,..,u_j,.., u_k) = P[X_1 \leq F_1^{-1}(u_1), \dots , X_j \leq F_j^{-1}(u_j)\dots, X_k \leq F_k^{-1}(u_k)\ ]$$

D'où enfin :

$$C(u_1,..,u_j,.., u_k) = F[F_1^{-1}(u_1), \dots , F_j^{-1}(u_j)\dots, F_k^{-1}(u_k)]$$

De même, on peut noter que l'on peut écrire :

$$F(x_1,..,x_j,.., x_k) = C(F_1\ (x_1), \dots , F_j\ (x_j)\dots, F_k\ (x_k))$$

Ce résultat est connu sous le nom du **théorème de Sklar**. Il montre que la donnée d'une fonction de répartition jointe équivaut à la donnée de ses fonctions de répartition marginales et de la copule correspondante. La copule d'un vecteur aléatoire se présente ainsi comme une représentation de la **dépendance probabiliste** entre les composantes de ce vecteur.

Il existe plusieurs mesures synthétiques de la dépendance entre composantes d'un vecteur aléatoire. On cite en particulier :

- Le coefficient de corrélation linéaire (de Pearson): $\rho(X_i, X_j) = \frac{Cov(X_i,X_j)}{\sigma(X_i)\sigma(X_j)}$
- Le coefficient de corrélation de Spearman : $\rho_S(X_i, X_j) = \rho(F_i(X_i), F_j(X_j))$

Il est important de noter que le coefficient de corrélation de Spearman reste inchangé en passant de U à X : $\rho_S\ (X_i, X_j) = \rho_S\ (U_i, U_j)$. En revanche, le coefficient de corrélation linéaire (de Pearson) $\rho\ (X_i, X_j)$ est en général différent de $\rho\ (U_i, U_j)$.

3.2 Simulation

Ce qui précède montre que si on dispose de valeurs simulées du k-uplet $(u_1, \dots u_j,\dots, u_k)$, on en déduit une simulation $(x_1, \dots x_j,\dots, x_k)$ de X.

Le problème se ramène ainsi à la simulation du vecteur $U= (U_1,. .U_j,..,U_k)$ Cependant, la loi jointe de U n'est pas connue. En effet, celle-ci est directement liée à celle de X qui est souvent inconnue.

En pratique, on remplace la copule inconnue par l'un des modèles de copules présentées dans la littérature statistique qui capture le mieux la structure de dépendance entre les composantes de X.

Plusieurs modèles de copule existent dans la littérature. Nous présentons dans ce qui suit deux familles de modèles qui sont les plus utilisées en pratique.

3.2.1 Les copules gaussiennes

Ce modèle de copules est définie par :

$$G(u_1,..,u_j,.., u_k) = F(\Phi^{-1}(u_1), \dots , \Phi^{-1}(u_j)\dots, \Phi^{-1}(u_j))$$

où F est la fonction de répartition d'une loi normale multidimensionnelle centrée réduite et de matrice de corrélation Σ donnée et Φ est la fonction de répartition de la loi normale centrée réduite à une seule dimension N(0,1).

Pour simuler $X= (X_1,. .X_j,..,X_k)$ selon une copule gaussienne, on peut procéder ainsi :

- Choisir Σ en rapport avec le niveau de corrélation demandé
- Simuler $Z = (Z_1,. .Z_j,..,Z_k)$ centré réduit ayant Σ pour matrice de corrélation en utilisant la procédure de simulation d'un vecteur normal sus présentée.
- Poser $U_j = \Phi(Z_j) \forall j =1$ à k.
- Calculer $X_j = F_j^{-1}(U_j) \forall j =1$ à k.

3.2.2Les copules archimédiennes

Cette famille de copule est définie ainsi :

$$G(u_1,..,u_j,.., u_k)= \varphi^{-1}(\varphi(u_1)+\dots \varphi(u_2) + \dots + \varphi(u_k))$$

avec φ strictement décroissante et convexe de [0, 1] dans R^+

Comme cas particuliers importants de copules archimédiennes citons notamment :

- La copule de Gumbel définie ainsi :

$C_\theta(u_1,..,u_j,\dots u_k) = \exp(-[\sum_{j=1}^{k}(-Lnu_j)^\theta]^{\frac{1}{\theta}})$ où $\theta \geq 1$

- La copule de Clayton définie par :

$C_\theta(u_1,..,u_j,...u_k) = (\sum_{j=1}^{k} u_j^{-\theta} - 1)^{-\frac{1}{\theta}}$ où $\theta \in [-1,0[\cup]0,+\infty[$.

On peut démontrer que le coefficient de Spearman dépend du paramètre θ. On peut en conséquence choisir ce paramètre en fonction du degré de corrélation souhaité entre les composantes X_j.

Pour simuler $X= (X_1,. .X_j,..,X_k)$ selon une copule archimédienne, on peut procéder ainsi :

- Choisir le modèle particulier de copule archimédienne et fixer le ou les paramètres correspondant en rapport avec le niveau de corrélation souhaité.
- Simuler $U = (U_1,. .U_j,..,U_k)$ en utilisant la procédure de simulation d'une loi jointe comme indiqué ci-dessus
- Calculer $X_j = F_j^{-1}(U_j) \forall$ j =1 à k.

Remarques

- La simulation ainsi obtenue n'est qu'une approximation du vecteur X. Celui-ci ne peut être exactement simulée qu'en connaissant sa fonction de répartition jointe.
- Les X_j simulées ont les lois marginales spécifiées. D'autre part elles sont corrélées du fait que les U_j sont corrélées.
- Les coefficients de corrélation linéaires des X_j peuvent être très différents de ceux des U_j. En revanche les coefficients de corrélation de rang de Spearman et de Kendall sont les mêmes.

En pratique, le choix du modèle de copule G n'est pas arbitraire. Il doit être modelé, à travers le choix de Σ ou de φ par exemple, en fonction du niveau de corrélation demandé des composante

Chapitre 5

LA METHODE DE MONTE CARLO

La méthode de Monte Carlo est l'une des principales applications mathématiques et statistiques utilisant les nombres aléatoires.

Cette méthode a pour objet de fournir une approximation numérique de la valeur d'une intégrale (quelque soit sa dimension) d'une fonction donnée pourvu qu'elle soit intégrable. Elle donne aussi, comme on va le préciser dans la suite, une mesure de la précision de l'approximation fournie.

En notant qu'une probabilité, une espérance mathématique, une variance ainsi que tout autre moment s'expriment comme des intégrales, on comprend l'importance pratique de la méthode de Monte Carlo dans les applications statistiques et probabilistes.

Dans ce qui suit, nous examinons son objet, son fondement mathématique, ses propriétés et sa mise en œuvre pratique. Nous terminons le chapitre par une présentation des techniques utilisées pour améliorer sa précision.

1. FONDEMENT

La méthode de Monte Carlo trouve son fondement mathématique dans la loi des grands nombres.

1.1 Loi des grands nombres

Etant donné un échantillon de n variables aléatoires réelles $\{X_i\}_{i=1 \text{ à } n}$ IID(m, σ^2)) c'est-à-dire identiquement , indépendamment et uniformément distribuées et possédant en outre une espérance mathématique m et une variance σ^2 . Désignons par $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ la variable aléatoire réelle définie par la moyenne arithmétique des X_i.

La loi des grands nombres stipule alors que sous les conditions précédentes la suite des $\bar{X}_n$ converge presque surement, lorsque n tend vers l'infini, vers l'espérance mathématique commune des X_i soit m.

Remarques :

- On démontre que l'on a aussi la convergence en moyenne quadratique ainsi que la convergence en probabilité. Le théorème de **Slutsky** s'applique en conséquence : toute fonction g ($\bar{X}_n$) converge en probabilité, lorsque n tend vers l'infini, vers g(m). Par exemple $\bar{X}_n^2$ converge en probabilité vers m^2.

- La loi des grands nombres s'applique également sur les autres moments empiriques (non centrés) $M_{k,n} = \frac{\sum_{i=1}^{n} X_i^k}{n}$ qui convergent ainsi presque surement, lorsque n tends vers l'infini, vers les moments théoriques correspondants $m_k = E(X^k)$ (à condition évidemment que ces moments théoriques existent). Par exemple $\frac{\sum_{i=1}^{n} X_i^2}{n}$ converge vers $E(X^2)$ et donc compte tenu de la première remarque, la variance empirique $V_n = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}_n^2$ converge en probabilité vers la variance théorique $\sigma^2 = E(X^2) - m^2$.
- La probabilité P(A) d'un évènement A donné peut être exprimée comme une espérance mathématique. Elle définit ainsi une limite d'une certaine moyenne empirique. En effet, soit X la variable aléatoire réelle suivant la loi de Bernoulli de paramètre p = P(A). On sait que E(X)= p. La loi des grands nombres montre alors que la fréquence $f_n = \frac{\sum_{i=1}^{n} X_i}{n}$ ou X_i vaut 1 si A est réalisé 0 sinon, converge vers p.

L'interprétation de la loi des grands nombres part de la constatation que la convergence est directement liée à la notion de distance. Ainsi dire que la suite des $\bar{X}_n$ converge vers la constante m c'est dire que la distance séparant cette constante à chacune des réalisations de $\bar{X}_n$ est de plus en plus petite au fur et à mesure que n augmente. On peut par conséquent approximer, lorsque n est grand, m par $\bar{x}_n$ ($\bar{x}_n$ étant n'importe quelle réalisation de $\bar{X}_n$):

$$m \approx \bar{x}_n$$

En conséquence, à supposer que m = E(X) est inconnu, une méthode de l'approximer est de se donner n réalisations indépendantes de la variable aléatoire réelle X et de calculer leur moyenne arithmétique.

1.2 Principe général de la méthode de Monte Carlo

A la base de la méthode de Monte Carlo est que d'après la loi des grands nombres, on peut approximer une espérance mathématique d'une variable X par la moyenne arithmétique de n réalisations indépendantes de cette variable.

Or dans le cas où cette variable est absolument continue, cette espérance mathématique s'exprime comme une intégrale d'une certaine fonction. Ainsi, en notant f la densité de probabilité de X et Δ le support correspondant, on a :

$$m = \int_{\Delta} xf(x)dx \approx \frac{\sum_{i}^{n} x_i}{n}.$$

(les x_i étant des réalisations indépendantes de X)

Plus généralement, considérons une fonction h (d'une ou plusieurs variables) intégrable et I la valeur de son intégrale sur un domaine Δ donné :

$$I = \int_{\Delta} h(x)d(x) .$$

Supposons qu'il n'est pas possible d'évaluer analytiquement cette intégrale et qu'on cherche en conséquence à l'approximer numériquement. Soit f une densité de probabilité ayant pour support Δ. On peut alors écrire :

$$I = \int_{\Delta} g(x)f(x)dx$$

$$o\grave{u}\ g(x) = \frac{h(x)}{f(x)}$$

soit,

$$I = E(g(X))$$

où X est une variable aléatoire ayant f pour densité de probabilité.

Ce développement montre que toute intégrale *I* peut s'écrire comme une espérance mathématique d'une certaine fonction g d'une variable aléatoire X.

Il s'ensuit en vertu de la loi des grands nombres qu'une approximation de l'intégrale *I* est donnée par la moyenne arithmétique de n réalisations indépendantes de la variable aléatoire g(X),

$$I \approx \frac{\sum_{i=1}^{n} g(x_i)}{n}$$

où les x_i sont n réalisations d'une variable aléatoire X ayant f pour densité de probabilité.

Il convient cependant de noter que dans les applications, on dispose rarement d'un échantillon de n réalisations de la variable X. A la place et par application des méthodes de génération de variables aléatoires présentées dans les chapitres précédents, on peut disposer de n valeurs simulées de X. Ces valeurs peuvent alors remplacer les vraies réalisations dans la formule d'approximation. L'erreur commise par ce remplacement ne devrait pas être à priori importante surtout lorsque le générateur utilisé est de bonne qualité.

Aussi en pratique cherche t on à approximer l'intégrale *I* par la quantité suivante :

$$\hat{I} = \frac{\sum_{i=1}^{n} g(x_i)}{n}$$

où les x_i sont n valeurs simulées d'une variable aléatoire X ayant f pour densité de probabilité.

La quantité $\hat{I}$ est appelée l'approximation de I par la méthode de Monte Carlo.

2. CALCUL PRATIQUE

On peut calculer l'approximation de Monte Carlo en utilisant l'algorithme suivant :

- Générer en utilisant un générateur de bonne qualité n nombres aléatoires (valeurs simulées d'une variable aléatoire suivant la loi uniforme continue sur [0,1]) : $u_1, u_2, ..., u_i, \ldots, u_n$.
- Choisir une densité f définie sur le support Δ (et facile à simuler par exemple par la méthode d'inversion) pour simuler à partir des u_i n valeurs d'une variable X : $x_1, x_2, ..., x_i, \ldots, x_n$.
- Déterminer les quantités : $g(x_1), g(x_2), ..., g(x_i), \ldots, g(x_n)$
- Calculer la moyenne arithmétique de ces quantités donnant $\widehat{I}$.

Remarque : L'approximation Monte Carlo de I n'est pas unique. En effet, on peut aussi écrire : $I = E(g^*(X))$ où $g^*(x) = \frac{h(x)}{f^*(x)}$ et f^* une autre densité de probabilité de X ayant pour support Δ et donc $\widehat{I^*} = \frac{\sum_{i=1}^{n} g^*(x_i)}{n}$ définit une autre approximation Monte Carlo de I. En fait, il y'a autant d'approximations que de densité de probabilité sur le supportΔ. En pratique, on choisit la densité la plus facile à simuler.

Exemples :

Soit à approximer par la méthode de Monte Carlo l'intégrale suivante :

$$I = \int_0^2 e^{-x^2} dx$$

Il convient au préalable d'écrire I comme une espérance mathématique en se donnant une densité de probabilité f sur [0,2] facile à simuler. On choisit pour cet exemple la densité de la loi uniforme continue sur [0,2] qui est la plus simple, soit :

$$f(x) = \frac{1}{2}\ 1_{[0,2]}$$

d'où, on déduit directement :

$$I = \int_0^2 2e^{-x^2}\ \frac{1}{2} dx$$

Soit ,

$$I = E\left(2e^{-X^2}\right)\ \text{où}\ X\ \rightarrow U(0,2)$$

et donc :

$$\hat{I} = \frac{\sum_{i=1}^{n} 2e^{-x_i^2}}{n}$$

(les x_i étant des valeurs simulées de $X\ \rightarrow U(0,2)$)

définit l'approximation de I par la méthode de Monte Carlo. En conséquence pour trouver cette approximation, on peut procéder comme suit (n étant donné):

- Générer n valeurs indépendantes $u_1, u_2, ..., u_i, \ldots, u_n$ en utilisant un bon générateur
- Calculer $x_1, x_2, ..., x_i, \ldots, x_n$ par la méthode d'inversion, soit $x_i = 2\ u_i$
- Calculer $g(x_i) = 2e^{-x_i^2}$ pour i = 1 à n et leur moyenne arithmétique simple pour trouver $\hat{I}$

3. PROPRIETES

On peut considérer $\hat{I}$ comme un estimateur calculée sur l'échantillon de variables $X_1, X_2, ..., X_i, \ldots, X_n$ qui est un échantillon IID qu'on suppose posséder une espérance mathématique m et une variance σ^2. On peut alors chercher si $\hat{I}$ possède les propriétés classiques d'un estimateur. Au préalable on donne l'expression de la variance de $\hat{I}$ dont l'estimation fournit une mesure de la précision de cette approximation.

3.1 Précision de $\hat{I}$

Par définition, l'on a :

$$V(\hat{I}) = V(\frac{\sum_{i=1}^{n} g(X_i)}{n}).$$

Comme les $g(X_i)$ sont IID, il s'ensuit que :

$$V(\hat{I}) = \frac{V(g(X))}{n}$$

ou encore :

$$V(\hat{I}) = \frac{E(g^2(X)) - E^2(g(X))}{n}$$

soit enfin :

$$V(\hat{I}) = \frac{E(g^2(X)) - I^2}{n}$$

Cette quantité dépendant de I ne peut pas être donc déterminée. On peut cependant l'approximer. En effet, on peut approximer I^2 par $\hat{I}^2$. Quant à $J = E(g^2(X))$ on peut , comme on l'a fait pour $I=E(g(X))$, utiliser la méthode de Monte Carlo pour l'approximer, soit

$$\hat{J} = \frac{\sum_i^n g^2(x_i)}{n}$$

où les x_i sont des valeurs simulées de X selon f.

Une approximation de $V(\hat{I})$ peut être ainsi donnée par ;

$$\widehat{V(\hat{I})} = \frac{1}{n}(\hat{J} - \hat{I}^2)$$

qui fournit une estimation de la précision de l'approximation de la méthode de Monte Carlo.

3.2 Absence de biais

L'estimateur $\hat{I}$ est sans biais. En effet,

$$E(\hat{I}) = E(\frac{\sum_{i=1}^{n} g(X_i)}{n})$$

soit compte tenu des propriétés de l'opérateur « espérance »,

$$E(\hat{I}) = \frac{\sum_{i=1}^{n} E(g(X_i))}{n}$$

Comme les X_i sont indépendantes et de même loi, les $g(X_i)$ le sont aussi, d'où

$$E(\hat{I}) = E(g(X)) = I$$

ou X est une variable de même loi que les X_i.

prouvant ainsi que $\hat{I}$ est un estimateur sans biais de I.

3.3 Convergence

L'estimateur $\hat{I}$ est par construction **convergent en probabilité**. En effet étant construit à partir de la loi des grands nombres, l'on a ainsi :

$$Plim\ \hat{I} = I$$

On note aussi, qu'il est convergent en moyenne quadratique. L'on a en effet à la fois $E(\hat{I}) = I$ et $V(\hat{I}) \to 0$

3.4 Normalité asymptotique

L'estimateur $\hat{I}$ se présente comme une moyenne arithmétique. On sait alors d'après le théorème central limite, que :

$$\frac{\hat{I}-I}{\sigma/\sqrt{n}} \quad \text{où} \quad \sigma^2 = V(g(X))$$

converge en loi vers la loi normale centrée réduite. A noter en outre que cette proposition reste vraie en changeant σ^2 par $\hat{\sigma}^2 = \hat{V}\left(g(X)\right)$. On peut en conséquence énoncer que :

$$\frac{\hat{I}-I}{\hat{\sigma}_{/\sqrt{n}}} \sim N(0,1)$$

ce qui permet de définir un intervalle de confiance (asymptotique) de I de niveau égal à (1-α)% :

$$[\hat{I} - t_{(\frac{\alpha}{2})}\hat{\sigma}_{/\sqrt{n}} \ , \ \hat{I} + t_{1-(\frac{\alpha}{2})}\hat{\sigma}_{/\sqrt{n}}].$$

En effet, l'on a :

$$P\left(\hat{I} - t_{(\frac{\alpha}{2})}\hat{\sigma}_{/\sqrt{n}} < I < \hat{I} + t_{1-(\frac{\alpha}{2})}\hat{\sigma}_{/\sqrt{n}}\right) \approx 1 - \alpha$$

où t_a désigne le quantile d'ordre a de la loi normale centrée réduite.

3.5 Vitesse de convergence

La méthode de Monte Carlo présente une vitesse de convergence relativement lente. En effet, du fait de la normalité asymptotique, l'on a :

$$\left|\hat{I} - I\right| < t_{1-(\frac{\alpha}{2})}\sigma_{/\sqrt{n}}$$

avec une probabilité égale à (1-α)

La quantité $t_{1-(\frac{\alpha}{2})}\sigma_{/\sqrt{n}}$ définit ainsi l'erreur d'approximation maximale de la méthode de Monte Carlo. Cette erreur converge vers 0 quant n tend vers l'infini. Cette convergence est cependant assez lente, de l'ordre de $n^{-1/2}$, puisque par exemple en multipliant le nombre d'observations par 100, l'erreur maximale d'approximation se trouve seulement divisée par 10.

Conclusion

Comme démontré ci-dessus, l'approximation de Monte Carlo présente des «bonnes propriétés ». On démontre également que la méthode de Monte Carlo est avantageuse pour l'approximation des intégrales multiples. En effet, cette méthode est insensible à la dimension de l'intégrale.

En revanche, la méthode de Monte Carlo est très concurrencée par les autres méthodes de calcul numérique d'intégrales notamment en ce qui concerne les intégrales simples à cause de la lenteur de sa vitesse de convergence.

4. TECHNIQUES DE REDUCTION DE LA VARIANCE

Il a été noté à la fin du paragraphe précédent que l'erreur d'approximation de la méthode de Monte Carlo se réduit au fur et à mesure que le nombre d'observations n augmente mais cette réduction est assez lente. Cependant, comme cette erreur dépend aussi positivement de la variance σ^2, une idée est de chercher à réduire cette dernière pour améliorer l'approximation de la méthode de Monte Carlo.

Plusieurs techniques de réduction de la variance ont été alors proposées. En fait, il s'agit de trouver une autre approximation $\tilde{I}$ de I ayant une variance plus petite que celle de $\hat{I}$. C'est l'objet des méthodes de réduction de la variance dont quelques unes, les plus usuelles, sont présentées dans ce qui suit.

Dans le reste de cette section, nous appelons $\hat{I}$ l'approximation initiale et $\tilde{I}$ l'approximation améliorée de I en utilisant une technique de réduction de la variance.

4.1 Méthode de la variable antithétique

Illustrons cette méthode d'abord sur un cas particulier. Soit

$$I = \int_0^1 g(x)dx.$$

On note que I = E(g(X)) ou X suit la loi uniforme continue sur [0.1].

L'approximation initiale de I est alors donnée par :

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

où les x_i sont des valeurs simulées d'une variable X suivant la loi uniforme continue sur [0.1].

On note qu'on peut aussi écrire

$$I = \int_0^1 g(1-x)dx$$

ou encore :

$$I = \int_0^1 \frac{(g(x)+g(1-x))}{2}dx$$

D'où une autre approximation de I par la méthode de Monte Carlo :

$$\tilde{I} = \frac{1}{2n}\sum_{i=1}^{n}(g(x_i) + g(1 - x_i))$$

où les x_i sont des valeurs simulées d'une variable X suivant la loi uniforme continue sur [0.1].

On peut alors montrer que $V(\tilde{I}) \leq V(\hat{I})$. En effet ,

$$V(\tilde{I}) = \frac{1}{4n}\,(\,V(g(X) + V\big(g(1 - X)\big) + 2\,cov\,(g(X), g(1 - X))$$

Puisque X suit la loi uniforme continue sur [0,1] la variable (1-X) suit également loi uniforme continue sur [0,1]. Les variables g(X) et g(1-X) suivent alors la même loi et donc V(g(X)) = V(g(1-X)).

D'autre part, on sait que :

$Cov\big(g(X), g(1 - X)\big) \leq \sqrt{V\big(g(X)\big)V\big(g(1 - X)\big)}$ = V(g(X)) (inégalité de Hodler)

En conséquence,

$$V(\tilde{I}) \leq \frac{1}{4n}\,(\,4V\big(g(X)\big) = V(\hat{I})$$

On peut généraliser cette méthode pour des dimensions quelconques de l'intégrale. Soit $I = \int_{\Delta} g(x)f(x)dx$ où f est une densité de probabilité d'une variable X ayant pour support Δ , donc I = E (g(X)). On sait qu'une approximation initiale de I par la méthode de Monte Carlo est donnée par $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ où les x_i sont des valeurs simulées d'une variable X selon la densité f. Soit Y = t(X) une transformation bijective et continue préservant la loi de X. Autrement dit Y et X ont la même loi de probabilité. On peut alors écrire :

$$I = \int_{\Delta} g(t^{-1}(y)f\big(t^{-1}(y)\big)(\frac{dt^{-1}(y)}{dy})dy$$

soit ,

$$I = \int_{\Delta} g(t^{-1}(y)l(y)dy$$

où *l (.)* est la densité de probabilité de Y.

Mais comme la transformation *t* préserve la loi de probabilité *l(.) = f(.)* sur Δ et donc :

$$I = \int_{\Delta} g(t^{-1}(y)f(y)dy$$

ou encore :

$$I = \frac{1}{2}\int_{\Delta} (g(t^{-1}(y) + g(y))f(y)dy = \frac{1}{2}E(g(X) + g(t^{-1}(X)))$$

où X est une variable aléatoire ayant f pour densité de probabilité.

A la place de $\tilde{I}$ on peut alors proposer :

$$\tilde{I} = \frac{1}{2n}\sum_{i=1}^{n}(g(x_i) + g(t^{-1}(x_i))$$

où les x_i sont des valeurs simulées d'une variable X selon f .

L'approximation $\tilde{I}$ a une variance plus faible que celle de $\hat{I}$. La démonstration est exactement la même que ci-dessus.

4.2 Méthode de l'échantillonnage préférentiel

On l'appelle aussi la méthode de la fonction d'importance. Soit à évaluer $I = \int_{\Delta} g(x)f(x)dx$ où f est une densité de probabilité d'une variable X ayant pour support Δ , soit I = E (g(X)). On sait qu'une approximation initiale de I par la méthode de Monte Carlo est donnée par $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ où les x_i sont des valeurs simulées d'une variable X selon la densité f.

On peut remarquer que *I* peut s'écrire aussi $I = \int_{\Delta} \frac{g(x)f(x)}{l(x)} l(x)dx$ où *l* (.) est une autre densité de probabilité de la variable X sur Δ .

Ce qui permet d'écrire :

I = E(k(X))

où *X* est une variable aléatoire de densité *l(.)* et

k(.) = g(.)f(.)/ l(.) sur Δ.

A la place de $\tilde{I}$ on peut alors proposer $\tilde{I} = \frac{1}{n}\sum_{i=1}^{n} k(x_i)$ où les x_i sont des valeurs simulées d'une variable *X* selon la densité *l* .

Comment choisir *l(.)* pour que la variance de $\tilde{I}$ soit plus petite que celle de $\hat{I}$ ou ce qui est équivalent que la variance de k(X) soit plus petite que celle g(X). ?

Posons $l(x) = \frac{g(x)f(x)}{E(g(X))}$ et calculons *V(k(X)) :*

$$V(k(X)) = \int \frac{g^2(x)f^2(x)}{g(x)f(x)} E\big(g(X)\big)dx - E^2\big(g(X)\big)$$

soit après simplification :

$$V(k(X)) = 0 < V(g(X)).$$

En conséquence, la variance s'est fait réduite au maximum. Cependant ce résultat n'a pas de portée pratique. En effet, on ne connait pas E(g(X)). Mais il a conduit à l'idée d'approcher *l(x)* par :

$$l^*(x) = \frac{Approx(g(x)f(x))}{\int approx(g(x)f(x))dx}$$

Exemple : soit à évaluer $I = \int_0^1 cos\frac{\pi x}{2} dx$

On note que $I = E(cos(\pi X/2))$ ou X suit U(0,1). L'approximation initiale de I par la méthode de Monte Carlo est donc définie par :

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \cos(\frac{\pi x_i}{2})$$

où les x_i sont des valeurs simulées d'une variable X suivant la loi uniforme continue sur [0.1].

Afin d'améliorer cette approximation avec la méthode de l'échantillonnage préférentiel, on peut approcher $\cos(\pi x/2)$ par son développement limité au voisinage de 0, soit

$$\cos\left(\frac{\pi x}{2}\right) \approx 1 - \frac{\pi^2 x^2}{8} \approx 1 - x^2$$

et prendre

$$l(x) = \frac{1-x^2}{\int_0^1 (1-x^2)dx} = \frac{3}{2}(1 - x^2)$$

Une approximation améliorée de I est alors donnée par

$$\tilde{I} = \frac{1}{n}\sum_{i=1}^{n} k(x_i)$$

où les x_i sont des valeurs simulées d'une variable *X* selon la densité *l* et $k(x) = \frac{\cos(\frac{\pi x}{2})}{\frac{3}{2}(1-x^2)}$.

4.3 Méthode de la variable de contrôle

On cherche à évaluer (approximativement) $I = \int_{\Delta} g(x)f(x)dx$ où f est une densité de probabilité d'une variable X ayant pour support Δ. On sait qu'une approximation initiale de I par la méthode de Monte Carlo est donnée par $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ où les x_i sont des valeurs simulées d'une variable X selon la densité f.

Remarquons que l'on peut écrire :

$$I= E(g(X)) = E(g(X) – h(X)+h(X)) = E(g(X)-h(X)) + E(h(X)).$$

Supposons que E(h(X)) peut se calculer analytiquement. Soit θ sa valeur et notons $I_1 = E(g(X)-h(X))$. L'approximation initiale de I_1 par la méthode de Monte Carlo est alors donnée par : $\widehat{I_1} = \frac{1}{n}\sum_{i=1}^{n}(g(x_i) - h(x_i))$ les x_i étant des valeurs simulées de X selon f.

Définissons une autre approximation de I par : $\tilde{I} = \hat{I}_1 + \theta$

Comme θ est une constante,

$$V(\tilde{I}) = V(\hat{I}_1) = (V(g(X)-h(X)))/n$$

En choisissant h(X) assez proche de g(X), cette variance deviendra petite et inférieure à $V(\hat{I}) = V(g(X))/n$

RECUEIL DE SUJETS D'EXAMEN 2009-2014

ESSAI **Epreuve de Techniques de Simulation**
2013-2014 **Session principale**
2ème Année **Durée 1 H30'**

Exercice 1 (10 points)

Soit X une variable aléatoire réelle suivant la loi triangulaire T(0,1,2). On note f la densité de probabilité de X :

$$\mathbf{f(x) = x.1_{[0,1]}+(2-x).1_{[1,2]}}$$

1. Vérifier graphiquement que f est bien une densité de probabilité
2. Trouver F(.) la fonction de répartition de X et son inverse.
3. Un générateur de nombres aléatoires donne u_1=0.254 et u_2=0.821. En déduire deux valeurs simulées de X en utilisant la méthode d'inversion.
4. On appelle fonction génératrice de X la fonction $M_X(.)$ définie par :
   $$M_X(t)=E(e^{tX})$$
   où E est l'opérateur « espérance mathématique »
   Calculer $M_X(t)$.
5. Soit U et V deux variables aléatoires indépendantes. On pose S = U+V
   a. Montrer que $M_S(t)=M_U(t).M_V(t)$
   b. On suppose que U et V suivent chacune la loi uniforme continue sur [0,1]. Calculer $M_U(t)$, $M_V(t)$ et $M_S(t)$. Que constate t on ?
6. En admettant que deux variables aléatoires ayant la même fonction génératrice ont la même loi de probabilité, déduire de ce qui précède une autre méthode de simulation de la loi triangulaire T(0,1,2).
7. Application numérique : Un générateur de nombres aléatoires donne u_1=0.254 et u_2=0.821. En déduire une valeur simulée de X.
8. Au vu de ce qui précède, quelles sont les différences essentielles entre les deux méthodes proposées pour la simulation de la loi triangulaire T(0,1,2).

Exercice 2 : (10 points)

Soient X, Y et Z trois variables aléatoires réelles indépendantes. On note σ_x^2, σ_y^2 et σ_z^2 leur variances respectives supposées finies. On pose par ailleurs :

$$S=X+Y \text{ et } T = Y+Z$$

1. Exprimer en fonction de σ_x^2, σ_y^2 et σ_z^2 :
   a. Les variances de S et T notées respectivement σ_S^2 et σ_T^2
   b. La covariance de S et T notée $\sigma_{S,T}$
   c. Le coefficient de corrélation linéaire de S et T noté $\rho_{S,T}$
2. En déduire que S et T sont positivement corrélées (linéairement)
3. On suppose en outre que X, Y et Z sont normales de moyennes nulles et de variances respectives $\sigma_x^2=\lambda$, $\sigma_y^2 = 1$ et $\sigma_z^2 = \lambda$ ($\lambda > 0$).

a. Montrer que S et T sont deux variables normales corrélées dont on précisera les moyennes, les variances et le coefficient de corrélation linéaire.

b. Exprimer λ en fonction de $\rho_{S,T}$.

4. On cherche à utiliser ce qui précède pour simuler un couple de variables normales S et T centrées réduites et ayant un coefficient de corrélation linéaire égal à 0.5.

Comment utiliser l'algorithme de Box Muller pour simuler des couple (s,t) du vecteur aléatoire (S,T). (Présenter d'une manière détaillée les étapes de calcul associées à cette procédure).

5. Par rapport aux procédures présentées dans le cours concernant la simulation de couple de variables normales corrélées, quels avantages et quels inconvénients présente la procédure objet de cet exercice.

ESSAI **Epreuve de Techniques de Simulation**
2013-2014 **Session de Contrôle**
2ème Année **Durée 1 H30'**

Problème : (15 points)

Ce problème a pour objet de donner une approximation du type Monte Carlo au nombre π. On considère à cet effet le carré unité $[0,1]^2$ noté C et le quart du disque centré en 0 et de rayon 1 noté D.

A. Préliminaires

1. Tracer les domaines C et D et calculer leur surface respective.
2. Donner l'expression de la fonction y = g(x) ayant pour courbe représentative la frontière supérieure de D.
3. Utiliser cette fonction pour caractériser les coordonnées des points de D.
4. On choisit au hasard un point du carré plein C. Quelle est la probabilité de tomber sur un point de D ?

B. Première méthode d'approximation

Soit (X,Y) un couple de variables aléatoires indépendantes suivant chacune la loi uniforme continue sur [0,1]. On pose $Z=\mathbf{1}_{[(X,Y)\in D]}$.

1. Quelle est la loi de Z ?
2. Calculer E(Z) et V(Z)
3. Donner l'approximation $\hat{I}$ de Monte Carlo de E(Z) et de sa variance V($\hat{I}$).
4. En déduire l'approximation $\hat{\pi}$ de Monte Carlo de π et de sa variance V($\hat{\pi}$).
5. Décrire en détail les différentes étapes de calcul de $\hat{\pi}$.

C. Deuxième méthode d'approximation

Soit Y = g(X) où X est une variable aléatoire suivant la loi uniforme continue sur [0,1] et g telle que définie ci-dessus.

1. Calculer E(Y) et V(Y)
2. Donner l'approximation $\tilde{I}$ de Monte Carlo de E(Y) et de sa variance V($\tilde{I}$).
3. En déduire l'approximation $\tilde{\pi}$ de Monte Carlo de π et de sa variance V($\tilde{\pi}$).
4. Décrire en détail les différentes étapes de calcul de $\tilde{\pi}$.

D. Comparaison entre les deux méthodes

1. Quelle est la méthode nécessitant le plus de calcul pour une même taille n de l'échantillon?
2. Quelle est la méthode la plus précise pour une même taille n de l'échantillon?
3. Conclusion

Exercice (5 points)

On considère un générateur de congruence linéaire : $s_n = (a.s_{n-1}+c) \bmod m$ où a,c et m sont des paramètres à choisir.

1. Quelle est la période maximale de ce générateur ?
2. Le paramètre m étant donné, rappeler les conditions sur a et c pour que ce générateur atteint sa période maximale.
3. Que deviennent ces conditions lorsque m est de la forme 2^k ?
4. On pose m =32.
   a. Quelles valeurs donner aux paramètres a et c pour atteindre la période maximale ?
   b. Soit $s_0 = 0$. En déduire x_1, x_2, x_3,x_4, x_5 les 5 premières valeurs simulées issues de la loi uniforme continue sur [0,1] données par ce générateur.

ESSAI **Epreuve de Simulation**
2012-2013 **Session Principale**
2ème Année **Durée 1h 30 '**

Problème (12 points)

Soit X une variable aléatoire réelle absolument continue admettant la densité de probabilité suivante :

$f(x) = k \sin(x) \mathbf{1}_{[0,\pi]}$

1. Trouver k pour que f soit effectivement une densité de probabilité.
2. Déterminer F la fonction de répartition de X.
3. Montrer que la fonction réciproque de F existe. Trouver cette fonction qu'on note F^{-1}.
4. Quelle est la loi de probabilité suivie par la variable aléatoire U définie par U= F (X) ? En déduire une méthode pour générer des valeurs de X.
5. Application numérique : Un générateur de nombres pseudo aléatoires donne : u_1= 0.021 , u_2= 0.452, u_3= 0.954. En déduire des valeurs simulées de X.
6. Soit $I = \int_0^{\pi} e^x \sin(x) dx$

a. Ecrire I comme une espérance mathématique d'une fonction de variable aléatoire dont on précisera la densité de probabilité.

b. Donner $\hat{I}$ l'approximation de Monte Carlo de I ainsi que sa variance approximée notée $\hat{\sigma}_{\hat{I}}^2$.

c. Indiquer les différentes étapes d'un algorithme de calcul de $\hat{I}$ et de sa variance approximée $\hat{\sigma}_{\hat{I}}^2$

7. Soit la variable aléatoire Y définie par Y= $\pi - X$

a. Montrer que Y et X ont la même loi de probabilité. (même ddp)

Nb : On rappelle que sin(π-x)=sin(x)

b. En déduire une autre approximation Monte Carlo plus précise de I basée sur la méthode de la variable antithétique.

Exercice : (8 points)

Soit X une variable aléatoire réelle absolument continue admettant la densité de probabilité suivante :

$f(x) = b\, x^k (1-x)^k \mathbf{1}_{[0,1]}$

avec k entier strictement positif et $b = \frac{\Gamma(2k+2)}{\Gamma(k+1)\Gamma(k+1)}$, Γ étant la fonction gamma.

Nb : On rappelle que Γ(n+1)= n ! ∀ n ∈ N.

1. Montrer que f(x) ≤ c où c est un réel à déterminer en fonction de b.
2. En notant que f(x) ≤ cg(x) où g est la densité de probabilité associée à la loi uniforme continue sur [0,1]

a. Montrer comment utiliser la méthode de rejet pour simuler des valeurs de X (indiquer les différentes étapes de l'algorithme).

b. A-t-on besoin de connaitre la valeur de b ?

3. Donner l'expression de b en fonction de k et en déduire celle de c en fonction de k.
4. Soit φ la fonction liant c à k (c = φ (k)).
a. Calculer φ (1), φ (2) et φ (3)
b. Montrer que φ est une fonction croissante de k en calculant
φ (k+1)- φ (k).
c. En se basant sur la signification de c, quelle conclusion tirer quant à l'efficacité de la méthode de rejet pour la simulation de loi de probabilité du type celle suivie par X ?

ESSAI **Epreuve de Simulation**
2012-2013 **Session de contrôle**
2ème Année **Durée 1h 30'**

Exercice (12 points)

Soit I= $\int_{-1}^{1} h(x)dx$ où h est une fonction intégrable et X une variable aléatoire réelle suivant la loi uniforme continue sur [-1,1]

1. Donner f la densité de probabilité de X
2. Exprimer I comme une espérance mathématique d'une certaine fonction g de X (expliciter g).
3. Soit $\hat{I}$ l'approximation initiale de Monte Carlo de I.

a. Soit $(x_1, x_2, \dots, x_i, \dots, x_n)$ un échantillon de valeurs simulées de X. Donner l'expression de $\hat{I}$

b. Comment calculer une estimation de la variance de $\hat{I}$?

4. On écrit I = $I_1 + I_2 = \int_{-1}^{0} h(x)dx + \int_{0}^{1} h(x)dx$

a. Soit X_1 (respectivement X_2) une variable aléatoire réelle suivant la loi uniforme continue sur [-1,0] (respectivement sur [0,1]). Rappeler f_1 et f_2 les densités respectives de X_1 et X_2.

b. Ecrire I_1 (respectivement I_2) comme une espérance mathématique d'une certaine fonction g_1 de X_1 (respectivement g_2 de X_2).

c. Soit $(x_1^1, x_2^1, \dots, x_i^1, \dots, x_{n1}^1)$ et $(x_1^2, x_2^2, \dots, x_i^2, \dots, x_{n2}^2)$ deux échantillons indépendants de valeurs simulées issus respectivement de X_1 et de X_2. Donner une nouvelle approximation de Monte Carlo de I. On note cette approximation $\tilde{I}$.

d. Donner une estimation de la variance de $\tilde{I}$.

5. On pose $n_1 = n_2 = n/2$ (n étant supposé pair)

Montrer que la variance de $\tilde{I}$ est inférieure à celle de $\hat{I}$.

6. Le remplacement de $\hat{I}$ par $\tilde{I}$ s'inscrit dans quels types de techniques ?

Exercice 2 (8points)

Soient R, S et T trois variables aléatoires indépendantes suivant des lois normales de moyennes égales à zéro et de variances égales respectivement à (1- |a|), (1-|a|) et (|a|) où |a| est compris entre 0 et 1.

1. On suppose a positif et on pose X =R+T et Y =S+T

a. Calculer les variances de X et Y notées respectivement V(X) et V(Y)

b. Calculer la covariance de X et Y notée Cov(X,Y) . En déduire $\rho(X,Y)$ le coefficient de corrélation linéaire de X et Y.

c. Trouver les lois de probabilité de X et Y.

2. On suppose a négatif et on pose X =R+T et Y =S-T

a. Calculer les variances de X et Y notées respectivement V(X) et V(Y)

b. Calculer la covariance de X et Y notée Cov(X,Y) . En déduire $\rho(X,Y)$ le coefficient de corrélation linéaire de X et Y.
c. Trouver les lois de probabilité de X et Y.
3. En se basant sur ce qui précède, on se propose de simuler un couple (X,Y) de variable aléatoires normales **corrélées** (non indépendantes)
a. Proposer un algorithme de simulation de (X,Y) dans le cas où X et Y sont corrélées négativement.
b. Proposer un algorithme de simulation de (X,Y) dans le cas où X et Y sont corrélées positivement.

ESSAI **Epreuve de Techniques de Simulation**
2011-2012 **Session principale**
2ème Année **Durée 1 H30'**

Exercice n°1 :

1. Soit X une variable aléatoire suivant la loi exponentielle de paramètre $\lambda > 0$.
a. Donner la fonction de répartition F (x) de X et son inverse $F^{-1}(u)$.
b. Décrire les étapes à suivre pour simuler des valeurs de X
2. Soit Z =ENT(X) où ENT( ) est la fonction partie entière et X une variable aléatoire réelle suivant la loi exponentielle de paramètre $\lambda > 0$.(L'évènement « Z=z » est donc équivalent à « $z \leq X < z+1$ »)
a. Quelles sont les valeurs possibles de Z ?
b. Trouver la loi de Z en calculant P(Z=z)
c . On pose Y = 1+Z. Montrer que Y suit la loi géométrique de paramètre p qu'on exprimera comme fonction de λ.

NB : On rappelle qu'une variable T suivant la loi géométrique de paramètre $p \in]0,1[$ est telle que $P(T=t) = p(1-p)^{t-1}$ $\forall t \in N^*$

3. En se basant sur ce qui précède, on se propose de simuler des valeurs d'une variable discrète Y suivant la loi géométrique de paramètre $p \in]0,1[$.
a. Décrire d'une manière détaillée les étapes à suivre pour obtenir des valeurs simulées de Y selon la loi géométrique de paramètre $p \in]0,1[$.
b. Un générateur de nombres pseudo aléatoires donne $u_1 = 0.055$ et $u_2 = 0.853$. En déduire des valeurs simulées d'une variable Y suivant la loi géométrique de paramètre p = 0.635. (On exprime d'abord λ en fonction de p).

Exercice n° 2 :

Soit (X,Y) un couple de variables aléatoires suivant la loi uniforme continue sur
$D = \{ (x,y) \in R^2 / 0<x<y<1 \}$ (on a donc $f(x,y) = k.1_D$ où f est la densité jointe)
1. Tracer D et trouver k.
2. Donner $f_x(x)$ la densité marginale de X et $F_X(x)$ sa fonction de répartition.
3. Déterminer $f_{Y/X}(y/x)$ la densité conditionnelle de Y sachant que X= x ainsi que sa fonction de répartition $F_{Y/X}(y/x)$.
4. En se basant sur ce qui précède, on se propose de simuler des couples (x,y) issus de la loi de (X,Y).
a. Décrire d'une manière détaillée les étapes à suivre pour simuler des couples (x,y) issus de la loi de (X,Y) en utilisant la méthode d'inversion.
b. Un générateur de nombres pseudo aléatoires donne u = 0.341 et $u_2 = 0.958$. En déduire un couple simulé de (X,Y)

Exercice 3 : Soit $I = \int_a^b g(x)dx$ où b > a sont donnés.

1. Ecrire I comme une espérance mathématique d'une fonction de variable aléatoire réelle X dont on explicitera la densité de probabilité f et la fonction de répartition F.
2. Soit $\hat{I}$ une première approximation de I par la méthode de Monte Carlo.
a. Donner l'expression de $\hat{I}$
b. Donner les expressions de la variance de $\hat{I}$ (notée $V(\hat{I})$) et de son estimation (notée $\widehat{V(\hat{I})}$).

c. Décrire d'une manière détaillée les étapes à suivre pour calculer $\hat{I}$ et $\widehat{V(\hat{I})}$ à partir de 1000 nombres pseudo aléatoires (a et b étant données)

3. Soit Y =(a+b-X) une nouvelle variable aléatoire dont on note G la fonction de répartition.

a. Montrer que I = $\frac{1}{2}\int_a^b \big(g(x) + g(a+b-x)\big)dx$

b. Montrer que X et Y ont la même loi de probabilité en prouvant que G(y)=F(y)

4. Soit une nouvelle approximation de I par la méthode de Monte Carlo définie par $\tilde{I} = \frac{(b-a)}{2n}\sum_{i=1}^{n}(g(x_i) + g(a+b-x_i))$ où les x_i sont des valeurs simulées de X selon f.

a. Montrer que $\tilde{I}$ est une approximation sans biais de I

b. Prouver que $\tilde{I}$ a une variance plus faible que celle de $\hat{I}$.

ESSAI **Epreuve de Méthodes de simulation**
2011-2012 **Session de contrôle**
2ème Année **Durée 1h30'**

Exercice 1 : (12 points)

Soit I l'intégrale suivante : $I = \int_0^{+\infty} \sqrt{t}\ e^{-t} \mathrm{dt}$

1. Exprimer I comme une espérance mathématique d'une variable aléatoire réelle dont on précisera la loi de probabilité.

NB : On considère évidemment la loi de probabilité la plus simple.

2. On note $\hat{I}_n$ l'approximation de I par la méthode de Monte Carlo.

a. Donner l'expression de $\hat{I}_n$

b. Quelle est son espérance mathématique

c. Montrer que la variance de $\hat{I}_n$ s'écrit comme une fonction de I.

d. En déduire l'expression de la variance estimée de $\hat{I}_n$.

3. On se propose de donner une approximation de I par un intervalle de confiance issu de la méthode de Monte Carlo.

a. Donner la loi limite de la variable $\frac{\hat{I}_n - \mathrm{I}}{\hat{\sigma}(\hat{I}_n)}$, $\hat{\sigma}(\hat{I}_n)$ étant l'écart type estimé de $\hat{I}_n$

b. En déduire un intervalle de confiance asymptotique de niveau 1-α = 95% de I

NB : Le quantile d'ordre 97.5% de la loi normale centrée réduite vaut 1.96.

4. Calcul pratique de $\hat{I}_n$

a. Indiquer les différentes étapes à suivre pour calculer $\hat{I}_n$ et sa variance estimée

b. Un générateur de nombres pseudo aléatoires a donné les valeurs suivantes : 0.792 ; 0.339 ; 0.828 ; 0.05 ; 0.642. En déduire la valeur de $\hat{I}_5$, approximation de I par la méthode de Monte Carlo et sa variance estimée.

c. En supposant n (=5) assez grand, en déduire une approximation de I par un intervalle de confiance de niveau 1-α = 95%.

Exercice 2 : (4 points)

On considère un couple de variables aléatoires (X,Y) dont la densité jointe est définie par : $f(x,y) = \lambda x y^{x-1} e^{-\lambda x} 1_{x>0} 1_{0\le y\le 1}$

1. Trouver la loi marginale de X (en donnant sa densité a(x) et sa fonction de répartition A(x))

2. Trouver la loi conditionnelle de Y sachant X=x (en donnant sa densité b(y/x) et sa fonction de répartition B(y/x))

3. Donner avec précision les différentes étapes à suivre pour simuler un couple de (X,Y) en utilisant la méthode d'inversion.

4. Application : Un générateur de nombres pseudo aléatoires donne : u = 0.098 et v =0.486. Avec λ = 2, en déduire une simulation d'un couple (x, y)

Exercice 3 : (4 points)

On a fait tourner un générateur de nombres pseudo aléatoires 1600 fois de suite. La distribution empirique des nombres obtenus se présente comme suit :

Classes	Effectif
0 - 0.125	180
0.125 – 0.25	195
0.25 – 0.375	225
0.375 – 0.5	170
0.5 – 0.625	230
0.625 – 0.75	190
0.75 – 0.875	220
0.875 - 1	190
Total	1600

On sait par ailleurs que la moyenne des nombres obtenus vaut 0.58.
On cherche à tester la qualité de ce générateur.
1. Test de la moyenne.
a. Donner Z la statistique fondant le test de la moyenne.
b. Selon ce test, peut- on refuser ce générateur au risque de 5% de se tromper ?
Nb : Le quantile d'ordre 97.5% de la loi normale centrée réduite vaut 1.96.
2. Test de Khi deux.
a. Donner K la statistique de Khi deux fondant le test d'adéquation de la distribution empirique obtenue à la loi uniforme continue.
b. Selon ce test, peut- on refuser ce générateur au risque de 5% de se tromper ?
NB : le quantile d'ordre 95% de la loi de khi deux à 7 degrés de liberté vaut 14.

ESSAI **Epreuve de Simulation**
2010-2011 **Session Principale**
2ème Année **Durée 1h 30 '**

Problème (12 points)

Soit X une variable aléatoire réelle absolument continue. On dit que X suit une loi triangulaire sur le support [0,1] et de mode 0,5 si sa densité de probabilité f est définie par :

$$f(x) = \begin{cases} 4x & si \quad 0 \le x \le 0.5 \\ 4(1-x) & si \quad 0.5 \le x \le 1 \\ 0 & ailleurs \end{cases}$$

On se propose de simuler la variable X.

A. Simulation de X selon la méthode d'inversion

1. Tracer la représentation graphique de f et vérifier à travers cette représentation que f est effectivement une densité de probabilité.

2. Trouver F la fonction de répartition de X

3. Déterminer F^{-1} l'inverse de la fonction de répartition

NB : On pourra écrire $F(x) = a+b(\beta+\alpha x)^2$ avec $\beta+\alpha x \ge 0$

4. Décrire la procédure de simulation de X par la méthode d'inversion en précisant ses différentes étapes.

B. Autre méthode de simulation de X

Soit Y_1 et Y_2 deux variables aléatoires réelles suivant indépendamment la loi uniforme continue sur [0,1]. On considère la variable $X= (Y_1+Y_2)/2$ dont on note F la fonction de répartition

1. Trouver la densité jointe du couple (Y_1 , Y_2)

2 . Soit le domaine $\Delta = \{(y_1 , y_2) / 0 \le y_1 \le 1, 0 \le y_2 \le 1 \text{ et } ((y_1+y_2)/2) \le x\}$

   a. Tracer le domaine Δ pour x compris entre 0 et 0.5 et exprimer sa surface en fonction de x. A quoi correspond cette surface ?

   b. Tracer le domaine Δ pour x compris entre 0.5 et 1 et exprimer sa surface en fonction de x. A quoi correspond cette surface ?

   c. En déduire l'expression de F pour x compris entre 0 et 0.5 et pour x compris entre 0.5 et 1

3. Compte tenu de ce qui précède,

   a. Quelle est la loi suivie par X ?

b. En déduire une autre procédure de simulation de X. Est-elle préférable à la première procédure ? Justifier votre réponse.

Exercice : (8 points)

1. Soit $X = |Z|$ où Z est une variable aléatoire réelle suivant la loi normale centrée réduite. Donner la densité de probabilité de X notée f.

NB : On rappelle que si $X = \varphi_1(Z)$ sur $[z_0 , z_1[$, $X=\varphi_2(Z)$ sur $[z_1 , z_2[$, $X = \varphi_3(Z)$ sur $[z_2 , z_3[$,…, $X = \varphi_n(Z)$ sur $[z_{n-1} , z_n[$ où les φ_i sont strictement monotones , alors $f(x) = \sum_{i=1}^{n} l(\varphi_i^{-1}(x)) \left| \frac{d\varphi_i^{-1}(x)}{dx} \right|$ où *f* et *l* sont respectivement les densités de probabilité de X et de Z .

2. On veut simuler X en utilisant la méthode de rejet

a. Montrer que : $\exp(-\frac{x^2}{2}) \leq \exp(\frac{1}{2})\exp(-x)$

NB : On peut se baser sur la propriété : $(x-1)^2 \geq 0$.

b. En déduire que : $f(x) \leq c\, g(x)$ $\forall$ $x \geq 0$ où g est la densité de probabilité de la loi exponentielle de paramètre 1 et c est une constante à déterminer.

c. En se basant sur ce qui précède, donner les différentes étapes d'une simulation de X selon la méthode de rejet.

3. On veut maintenant simuler la loi normale centrée réduite en utilisant les résultats trouvés en 1. et 2.

a. Soit S une variable aléatoire réelle discrète prenant les valeurs -1 et 1 avec P(S=-1) = P(S=1) = ½. Montrer que Z et $T=|Z|S$ suivent la même loi de probabilité (même fonction de répartition).

b. En déduire une procédure de simulation de la loi normale centrée réduite (présenter les différentes étapes).

ESSAI **Epreuve de Techniques de Simulation**
2010-2011 **Session de contrôle**
2ème Année **Durée 1 H30'**

Exercice n°1 : (8 points)

On dit qu'une variable aléatoire réelle absolument continue suit la loi de Gumbel si sa densité de probabilité est définie par :

$$f(x) = e^{-x}e^{-e^{-x}} \; \forall \; x \in R$$

1. En notant que f(x) s'écrit sous la forme u'(x)exp(u(x)) ou u est une fonction de x et u' sa dérivée,
a. Montrer que f est effectivement une densité de probabilité.
b. Trouver la fonction de répartition F de X.

2. On veut simuler X par la méthode d'inversion.
a. Déterminer l'inverse F^{-1} de F
b. Un générateur de nombres pseudo aléatoires donne : u1 = 0.281, u2 = 0.735, u3= 0.0110. En déduire trois valeurs simulées de X selon la méthode d'inversion.

3. Soit $Y = e^{-X}$
a. Trouver la loi de probabilité de Y
b. Comment simuler Y ?
c. En déduire une méthode de simulation de X
d. Montrer que l'on aboutit à la même méthode trouvée en 2.

Exercice n° 2 : (12 points)

Soit X une variable aléatoire réelle absolument continue suivant la loi béta suivante :

$$f(x) = \frac{x^2(1-x)^2}{B(3,3)} 1_{0<x<1}$$

On se propose de simuler X par la méthode de rejet
1. En supposant qu'il existe $c \in R$ et g une autre densité de probabilité de X simulable par la méthode d'inversion et tels que $f(x) \leq c\, g(x) \; \forall \; 0 < x < 1$ présenter les étapes à suivre pour simuler X selon f en utilisant la méthode de rejet.

Est-il nécessaire de connaitre B (3,3) pour réaliser ces étapes ?

2. Soit g_1 la densité de probabilité de la loi uniforme continue sur [0,1]
a. Rappeler l'expression de g_1

b. Trouver en fonction de B(3,3) le plus petit réel c vérifiant ;

$$f(x) \leq c\, g_1(x) \qquad \forall\ 0 < x < 1$$

c. En déduire en fonction de B(3,3) le nombre moyen de rejet d'une méthode de rejet basée sur g_1.

3. Soit g_2 la densité de probabilité de la loi triangulaire sur [0,1] et de mode 0.5 :

$$g_2(x) = 4x \text{ si } 0< x<0.5$$
$$g_2(x) = 4(1-x) \text{ si } 0.5< x<1.$$
$$g_2(x) = 0 \text{ sinon}$$

a. Vérifier géométriquement que g_2 est effectivement une densité de probabilité

b. Trouver en fonction de B(3,3) le plus petit réel c vérifiant :

$$f(x) \leq c\, g_2(x) \ \forall\ 0 < x < 1$$

NB: On cherche c pour $0< x<0.5$ et pour $0.5< x<1$ et on note qu'il s'agit du même réel

c. En déduire en fonction de B(3,3) le nombre moyen de rejet d'une méthode de rejet basée sur g_2.

4. Compte tenu de ce qui précède la quelle des densités choisir pour baser la méthode de rejet : g_1 ou g_2 ? Justifier votre réponse.

5. On veut utiliser la méthode de Monte Carlo pour évaluer approximativement le nombre B(3,3).

a. Montrer que B(3,3) s'écrit comme une intégrale qu'on précisera

b. Exprimer cette intégrale comme une espérance mathématique d'une fonction de variable aléatoire suivant la loi uniforme continue sur [0,1]

c. Donner l'approximation de Monte Carlo de B(3,3)

ESSAI **Epreuve de Techniques de Simulation**
2009-2010 **Session Principale**
2ème Année **Durée 2 Heures**

Exercice n°1 (9 points)

Soit X une variable aléatoire réelle suivant la loi de Cauchy. On note g sa densité de probabilité :

$$g(x) = \frac{1}{\pi(1+x^2)} \; \forall \; x \in R$$

1. Trouver la fonction de répartition G de X
2. Montrer que G est bijective et trouver son inverse
3. On veut utiliser la méthode d'inversion pour simuler des valeurs de X
a. Rappeler le principe général de cette méthode.
b. Utiliser cette méthode pour simuler des valeurs de X basées sur les nombres pseudo aléatoires suivants : $u_1 = 0.313$, $u_2 = 0.03$, $u_3 = 0.872$.

4. On pose $h(x) = f(x)/g(x) \; \forall \; x \in R$ où f est la densité de probabilité de la loi normale centrée réduite.
a. Trouver les réels x maximisant h
b. En déduire que $f(x) \leq M \; g(x) \; \forall \; x \in R$ où M est un réel à déterminer.

5. On se propose de simuler la loi normale centrée réduite par une méthode de rejet en se basant sur 4.
a. Décrire les différentes étapes de mise en application de la méthode de rejet.
b. Déterminer la probabilité de rejet et le nombre moyen de rejet.
c. Un générateur de nombres pseudo aléatoires donne $v_1 = 0.975$, $v_2 = 0.55$, $v_3 = 0.328$, $v_4 = 0.045$. Utiliser ces nombres avec les valeurs simulées de X selon g en 3.b pour simuler des valeurs de X selon f.

Exercice n°2 : (11 points)

On rappelle qu'une variable aléatoire réelle suit une loi gamma de paramètres $a \in N^*$ et $b > 0$ a pour densité de probabilité :

$$f(x) = \frac{b^a}{(a-1)!} x^{a-1} e^{-bx} \; \forall \; x > 0 \; ; f(x) = 0 \; sinon$$

1. Soit Y une variable aléatoire réelle suivant la loi exponentielle de paramètre $b > 0$. Quelle est la méthode usuelle utilisée pour simuler Y ? la présenter assez brièvement.

2. Soit $X = \sum_{i=1}^{a} Y_i$ où les Y_i sont des variables indépendantes suivant la même que Y. Montrer que X suit une loi gamma dont on précisera les paramètres.

NB : on rappelle que la loi gamma est stable pour l'addition

3. On cherche à simuler des valeurs de X
a. En se basant sur ce qui précède, donner une méthode de simulation de X (préciser les différentes étapes de l'algorithme de simulation).
c. Un générateur de nombres pseudo aléatoires donne : $u_1 = 0.195$, $u_2 = 0.946$, $u_3 = 0.732$, $u_4 = 0.013$. Avec a =2 , utiliser ces nombres pour simuler des valeurs de X.

3. Soit l'intégrale $I = \int_0^1 2e^{-2x}\, dx$. On suppose que I ne peut pas être calculée analytiquement et on se propose de l'approximer par la méthode de Monte Carlo.
a. Ecrire I comme une espérance mathématique d'une fonction de variable aléatoire suivant la loi uniforme continue sur [0,1].
b. En déduire l'approximation de I par la méthode de Monte Carlo. On note $\hat{I}$ cette approximation.
c. Ecrire I comme une espérance mathématique d'une fonction de variable aléatoire suivant une loi gamma dont on précisera les paramètres.
d. En déduire l'approximation correspondante I par la méthode de Monte Carlo. On note $\tilde{I}$ cette approximation.
e. Calculer les variances respectives de $\hat{I}$ et de $\tilde{I}$. Commenter.

ESSAI **Epreuve de Techniques de Simulation**
2009-2010 **Session de contrôle**
2ème Année **Durée 2 Heures**

Exercice n°1

NB : On arrondi à 2 chiffres après la virgule

1. Soit le générateur de congruence linéaire mixte : $x_n = (ax_{n-1}+c) \bmod m$

a. quelle est la période maximale de ce générateur ?

b. Rappeler les conditions sur les paramètres a,c et m pour que ce générateur atteigne sa période maximale.

c. Comment se présente ces conditions lorsque $m = 2^k$ ($k \geq 2$) ?

2. On considère le générateur suivant : $x_n = (69x_{n-1}+1) \bmod 2^{10}$.

a. Ce générateur atteint il sa période maximale ? Justifier votre réponse.

b. Donner 3 nombres pseudo aléatoires issus de ce générateur à partir d'une germe $x_0 = 255$. On note ces nombres x_1, x_2, x_3.

c. En déduire une suite de 4 valeurs simulées de la loi uniforme continue sur [0,1].On note ces nombres u_0, u_1, u_2, u_3.

3. À partir des valeurs obtenues en 2.c, trouver des valeurs simulées de la loi normale centrée réduite.

a. En utilisant la méthode de Box Muller.(On note ces nombres y_0, y_1, y_2, y_3.)

b. En utilisant l'algorithme polaire.(On note ces nombres z_0, z_1, z_2, z_3.)

c. Quels sont les avantages et inconvénients de chacune de ces méthodes ?

Exercice n°2

1. Soit X une variable aléatoire réelle absolument continue définie sur [0,1]. On note f et F respectivement sa densité de probabilité et sa fonction de répartition. On suppose en outre que X possède une espérance mathématique notée E(X).

Montrer que $E(X) = \int_0^1 \big(1 - F(u)\big)du$

NB : On peut utiliser la technique d'intégration par partie.

2. On cherche à approximer E(X) par la méthode directe de Monte Carlo. On note $\widehat{m}$ cette approximation calculée à partir de n valeurs simulées de X.

a. Donner l'expression générale de $\widehat{m}$ et de sa variance.

b. Sachant que les valeurs simulées de X ont été obtenues par la méthode d'inversion, exprimer $\widehat{m}$ en fonction d'une suite de n valeurs simulées de la loi uniforme continue sur [0,1].

3. Soit $\widetilde{m}$ une deuxième approximation de E(X) déduite en se basant sur 1.
Donner l'expression de $\widehat{m}$ et de sa variance.

4. On suppose que f prend la forme suivante : $f(x) = 3x^2 . \mathbf{1}_{[0,1]}$

Calculer $V(\widehat{m})$ et $V(\widetilde{m})$. Commenter.

ESSAI **Epreuve d Techniques de Simulation**
2008-2009 **Session Principale**
2ème Année **Durée 2 Heures**

Exercice n°1 (4 points)

On considère l'intégrale suivante I= $\int_0^1 x^2\,dx$

1. Soit X une variable aléatoire réelle suivant la loi uniforme continue U(0,1)
a. Ecrire I comme une espérance mathématique d'une certaine fonction de X
b. En déduire un estimateur $\hat{I}_1$ de I par la méthode de Monte Carlo.
c. Calculer la variance de $\hat{I}_1$.

2. Soit X une variable aléatoire réelle absolument continue ayant une densité de probabilité définie par : f(x) = 2x $\forall$ x$\in$ [0,1] et f(x) = 0 ailleurs.
a. Ecrire I comme une espérance mathématique d'une certaine fonction de X
b. En déduire un estimateur $\hat{I}_2$ de I par la méthode de Monte Carlo.
c. Calculer la variance de $\hat{I}_2$.

3. Soit X une variable aléatoire réelle suivant la loi uniforme continue U(0,1)
a Ecrire I comme une espérance mathématique en se basant sur la méthode de la variable antithétique.
b. En déduire un estimateur $\hat{I}_3$ de I par la méthode de Monte Carlo.
c . Calculer la variance de $\hat{I}_3$.

4. La quelle des trois méthodes utiliser ? Justifier votre réponse.

Exercice n°2 : (7 points)

1. Soient X et Y deux variables aléatoires discrètes indépendantes suivant respectivement les lois uniformes discrètes : P(X = x) = ¼ $\forall$ x $\in${0,1,2,3} et P(Y = y) = 1/3 $\forall$ x $\in${0,1,2}.
On pose Z = (X+Y) mod 4. Donner les valeurs possibles de Z et sa loi de probabilité.
2. Soit le générateur de congruence linéaire suivant : $x_n = (ax_{n-1}+c) \bmod 4$.
a. Quelle est la période maximale ?
b. Trouver a et c permettant d'avoir la période maximale.
c. On pose $x_0 = 0$. Calculer les valeurs données par ce générateur jusqu'à l'ordre 4.
3. Soit le générateur de congruence linéaire suivant : $y_n = (\alpha x_{n-1}+\gamma) \bmod 3$.
a. Quelle est la période maximale ?
b. Trouver α et γ permettant d'avoir la période maximale.
c. On pose $y_0 = 0$. Calculer les valeurs données par ce générateur jusqu'à l'ordre 3
4. Soit la relation suivante : $z_n = (x_n+y_n) \bmod 4$

a. Expliquer pourquoi on peut considérer cette relation comme un générateur de nombres pseudo aléatoires.
b. En posant $x_0 = y_0 = 0$, Calculez $z_0, z_1, z_2, \ldots, z_{25}$.
c. Quelle est la période de ce générateur ? Commenter.

Exercice n°3 : (9 points)

1. Soit Z une variable aléatoire suivant la loi uniforme continue U(0,1) et Y une autre variable aléatoire absolument continue prenant ses valeurs dans [0,1]. On suppose Z et Y indépendantes.

a. Représenter graphiquement l'évènement « Z< Y » dans le plan.
b. Montrer que P(Z< Y) = E(Y) , E étant l'opérateur espérance mathématique.

2. Soit X une variable aléatoire absolument continue et F sa fonction de répartition (supposée admettre une réciproque F^{-1}).

a. Comment utiliser la méthode d'inversion pour simuler des valeurs de X ?
b. On suppose que F est inconnue mais peut être estimée sans biais par $\hat{F}$ en tout x∈ R : $0 \leq \hat{F}(x) \leq 1 \; et \; E\left(\hat{F}(x)\right) = F(x) \forall x \in R$

En supposant que $\hat{F}^{-1}$ existe, montrer que $P(\hat{F}^{-1}(Z) < x) = F(x) \; \forall \; x \in R$ ou Z est une variable aléatoire suivant U(0,1) et est indépendante de $\hat{F}(x)$.

NB : $\hat{F}$ (x) est bien une variable aléatoire.

c. En déduire une procédure pour simuler des valeurs de X lorsqu'on ne connait qu'une estimation sans biais de F(x).

3. Soient $x_1, x_2, \ldots, x_n$ n réalisations passées de X (supposées indépendantes).

a. Montrer que $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i \leq x]}$ est un estimateur sans biais de F(x).
b. Les n (= 360) réalisations passées de X ont été regroupées dans le tableau suivant :

$]e_{i-1}, e_i]$	]0, 0.5]	]0.5, 1]	]1, 1.5]	]1.5, 2]	]2, 2.5]	]2.5, 3]
Nombre D'observations	40	40	60	120	60	40

Calculer $\hat{F}$ (e_i) pour e_i = 0.5, 1, 1.5, 2, 2.5, 3 et représenter graphiquement $\hat{F}$ en supposant qu'elle linéaire dans chaque intervalle $]e_{i-1}, e_i]$.

c. Un générateur de nombres pseudo aléatoires donne $z_{1=}$ 0.468, z_2 = 0.754, z_3 =0.191. En déduire des valeurs simulées de X.

ESSAI **Epreuve de Techniques de Simulation**
2008-2009 **Session de contrôle**
2ème Année **Durée 1h30**

Exercice 1(10 points)

Soit X une variable aléatoire réelle absolument continue dont la densité de probabilité est définie par :$g(x) = \frac{1}{2}e^{-|x|} \; \forall \; x \in R$

1. Vérifier que f est effectivement une densité de probabilité.
2. Trouver la fonction de répartition correspondante.
3. Un générateur de nombres pseudo aléatoires a donné $u_1 = 0.046$, $u_2=0.945$, $u_3 = 0.5$. En utilisant la méthode d'inversion, en déduire trois valeurs simulées de X selon g.
4. On considère la fonction : $h(x) = \frac{f(x)}{g(x)} \forall \; x \in R$ où f la densité de probabilité de la loi normale centrée réduite :

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \; \forall \; x \in R$$

a. Déterminer les réels qui maximisent h et en déduire sa valeur maximale.
b. En déduire le plus petit réel c tel que $f(x) \leq cg(x) \; \forall \; x \in R$.

5. On considère l'évènement E défini par « cUg(X) < f(X) » où U est une variable aléatoire réelle indépendante de X et suivant la loi uniforme U(0.1). Calculer la probabilité de E.

6. Quelle est la densité de probabilité de la loi de X conditionnellement à la réalisation de l'événement E ?
7. En déduire une méthode de rejet pour simuler X selon la loi normale centrée réduite.
8. Quel est le nombre moyen de rejet ?
9. Un générateur de nombres pseudo aléatoires donne les nombres suivants : $v_1 = 0.988$; $v_2 = 0.113$; $v_3 = 0.311$. Utiliser ces nombres et les valeurs simulées de X selon g trouvés en 3. pour simuler des valeurs de X selon f en utilisant la méthode de rejet.

Exercice 2: (10 points)

Soit X une variable aléatoire réelle absolument continue dont la densité de probabilité est définie par :

$$g(x) = k \; \forall \; x \in \mathbb{I}_{[-1,1]}$$

1. Trouver k pour que g soit effectivement une densité de probabilité.
2. On pose Y = (1+X)/2.

a. Quelle est la loi de Y ?

b. En déduire une procédure de simulation de X.

3. Soient X_1 et X_2 deux variables indépendantes ayant la même loi que X.

a. Donner la densité $f(x_1, x_2)$ du couple (X_1, X_2).

b. Comment simuler le couple (X_1, X_2) ?

c. Calculer la probabilité de l'évènement :

$B=\{(x_1, x_2)/ -1\leq x_1 \leq 1, -1\leq x_2 \leq 1 \text{ et } x_1^2 + x_2^2 \leq 1)\}$

NB : Il n'est pas nécessaire de faire du calcul intégral.

4. On veut approcher le nombre π par la méthode de Monte Carlo.

a. Ecrire π comme une espérance mathématique d'une fonction $h(X_1, X_2)$ à préciser.

b. Comment obtenir des valeurs simulées de $h(X_1, X_2)$?

c. Quelle est l'approximation Monte Carlo $\hat{\pi}$ de π ?

d. Calculer la variance de cette approximation. Comment l'estimer ?