UNIVERSITÉ DE NANTES

Master I : Ingénierie mathématique

Exercices de Probabilités

pour les séances de TP avec le logiciel R

Anne Philippe

12 octobre 2011

Bureau 118 bâtiment 10
Laboratoire de Mathématiques Jean Leray
anne.philippe@math.univ-nantes.fr

1. Simulation de variables aléatoires

Exercice 1.1. Méthode d'inversion.

Rappel

Soit F une fonction de répartition. On définit, pour tout $u \in]0,1[$, le pseudo inverse

$$F^{-}(u) = \inf\{x : \ F(x) \geq u\}.$$

Si U suit une loi uniforme sur $[0,1]$ alors la loi de $F^{-}(U)$ admet pour fonction de répartition F.

1) Écrire une fonction pour simuler suivant une loi exponentielle de paramètre $\ell > 0$ en utilisant ce résultat.

2) Appliquer cette fonction pour $\ell = 2.5$ et simuler un échantillon de taille $n = 100$.
   - Tracer l'histogramme de l'échantillon (voir l'option `proba` de la fonction `hist`)
   - Ajouter sur le graphique précédent la densité de la loi exponentielle de paramètre 2.5.
   - Calculer l'espérance de la loi et comparer avec la moyenne de l'échantillon simulé
   - Calculer la variance de la loi et comparer avec la variance de l'échantillon simulé.

   Recommencer avec $n = 5000$ puis 10000.

3) Reprendre les mêmes questions pour la loi de Cauchy qui admet pour densité

$$\frac{1}{\pi}\frac{1}{1+x^2}$$

Exercice 1.2. Algorithme d'acceptation rejet. Soit f la densité d'une loi de probabilité. On veut simuler des nombres pseudo aléatoires suivant la loi de densité f. On suppose qu'il existe une densité g telle que

H-1) on sait simuler des nombres pseudo aléatoires suivant la loi de densité g

H-2) Il existe une constant M telle que $f(x) \leq Mg(x)$.

L'algorithme pour simuler une variable aléatoire X suivant la loi de densité f s'écrit

1. `Générer` $Z \sim g$ `et` $U \sim \text{Uniforme}[0,1]$.
2. `Si` $U \leq \frac{f(Z)}{Mg(Z)}$, `on prend` $X = Z$;

   `Sinon on répète 1.`

Remarque : La loi de densité g est appelée la loi instrumentale de l'algorithme d'acceptation rejet.

Partie théorique.

On considère (Z_n) une suite de variables aléatoires iid suivant la loi de densité g et (U_n) une suite de variables aléatoires iid suivant la loi uniforme sur $[0,1]$. On suppose que les deux suites sont indépendantes. En posant $N = \inf\{n : \ U_n \leq \frac{f(Z_n)}{Mg(Z_n)}\}$, on a

$$X = Z_N$$

I - 1) Ecrire la densité du couple (U_1, Z_1).

I - 2) Exprimer $P(U_1 \leq \frac{f(Z_1)}{Mg(Z_1)})$ à partir de la densité du couple (U_1, Z_1), puis montrer que cette probabilité est égale à $\frac{1}{M}$.

I - 3) En déduire que N suit une loi géométrique de paramètre $\frac{1}{M}$.

I - 4) En utilisant la question 1), montrer que

$$P(U_1 \leq \frac{f(Z_1)}{Mg(Z_1)}\ ,\ Z_1 \leq x) = \frac{1}{M}\int_{-\infty}^{x} f(t)\ \mathrm{d}t$$

I - 5) Pour tout i on pose $A_i = [U_i \leq \frac{f(Z_i)}{Mg(Z_i)}]$ et $B_i = [Z_i \leq x]$.

Montrer

a)

$$[X \leq x] = \bigcup_{i=1}^{\infty}([N = i] \cap B_i),$$

b) $[N = 1] = A_1$,

c) pour $i > 1$

$$[N = i] = \left(\bigcap_{j=1}^{i-1} A_j^c\right) \cap A_i.$$

I - 6) En déduire que la loi de X admet pour densité f.

Implementation et application de l'algorithme

II - 1 Montrer que l'on peut simuler des nombres pseudo aléatoires suivant la loi gaussienne standard par la méthode d'acceptation rejet en prenant la loi de Cauchy comme loi instrumentale.

II - 2 Implémenter puis valider l'algorithme d'acceptation rejet.
Pour valider l'algorithme, on utilise les critères suivants :
- comparaison de l'histogramme de l'échantillon et de la densité de la loi que l'on souhaite simuler.
- comparaison de l'espérance de la loi et de la moyenne de l'échantillon simulé
- comparaison de la variance de la loi et de la variance de l'échantillon simulé.
- Retrouver par la simulation la valeur du taux d'acceptation.
- Vérifier que le nombre de variables simulées suivant la loi instrumentale suit une loi géométrique.

II - 3 Montrer que l'on peut simuler la loi de densité

$$f(x) = 2/\pi\sqrt{1 - x^2}\mathbb{I}_{[-1,1]}(x)$$

par la méthode d'acceptation rejet en prenant la loi uniforme sur $[-1, 1]$ comme loi instrumentale.

II - 4 Implémenter puis valider l'algorithme en suivant les étapes proposées à la question 2.

Exercice 1.3. Loi uniforme sur le disque. On veut simuler suivant la loi uniforme sur le disque unité en utilisant le changement de variables en coordonnées polaires notées (R, θ) Soit (U, V) un vecteur qui suit une loi uniforme sur le disque. On pose

$$U = R\cos(\theta)$$
$$V = R\sin(\theta)$$

1) Déterminer la loi du couple (R, θ)
2) Les variables aléatoires R et θ sont elles indépendantes ?
3) Donner les lois marginales de R et de θ.
4) Proposer des algorithmes pour simuler des nombres pseudo aléatoires suivant la loi de R, puis celle de θ.
5) Implémenter l'algorithme pour simuler (U, V) suivant la loi uniforme sur le disque unité
6) Valider l'algorithme à l'aide d'outils graphiques sur la loi du couple et les lois marginales.

2. Vecteurs Gaussiens

Soit $X = {}^t(X_1, ..., X_5)$ un vecteur aléatoire gaussien à valeurs dans $\mathbb{R}^5$ de moyenne μ et de variance Σ.

-I- Simulation de vecteurs gaussiens : la fonction `chol`(Σ) (décomposition de choleski) retourne une matrice triangulaire supérieure C telle que $\Sigma = {}^tCC$. Si $Y = {}^t(Y_1, ..., Y_p)$ est un vecteur dont les coordonnées sont iid suivant la loi normale $\mathcal{N}(0, 1)$, alors on peut exprimer X de la forme $X = {}^tCY + \mu$.

I-a) Justifier ce résultat.

I-b) Écrire une fonction qui retourne un n réalisations de vecteurs aléatoires gaussiens suivant la même loi que X. Les paramètres d'entrée sont (n, μ, Σ) et la sortie est une matrice constituée de n colonnes et 5 lignes.

-II- Pour tester votre générateur, on prend $\mu = {}^t(1, 2, 3, 4, 5)$ et $\Sigma = \left(\Sigma_{i,j}\right)_{1 \le i,j \le 5}$ avec

$$\Sigma_{i,j} = \frac{a^{|i-j|}}{1 - a^2}$$

où a est un réel, $a \in]-1, 1[$. On fixe $a = -0.8$.

II-a) Simuler un échantillon de vecteurs gaussiens $(X^{(1)}....X^{(n)})$ avec $n = 5000$

II-b) Représenter sur un même graphique l'évolution en fonction de n de la moyenne empirique de chacune des coordonnées. Commenter les résultats.

Soit $Z_1....Z_n$ un échantillon de variables aléatoires réelles iid suivant une loi qui appartient à L^2. Un estimateur de la variance est la variance empirique

$$S_n = \frac{1}{n} \sum_{i=1}^{n} Z_i^2 - (\bar{Z}_n)^2.$$

II-c) Représenter sur un même graphique l'évolution en fonction de n de la variance empirique de chacune des coordonnées. Commenter les résultats.

Soit $(Z_1, Y_1)....(Z_n, Y_n)$ un échantillon de couples de variables aléatoires réelles iid suivant une loi qui appartient à L^2. Un estimateur de la covariance de (Z_1, Y_1) est la covariance empirique $C_n = \frac{1}{n} \sum_{i=1}^{n} Z_i Y_i - \bar{Z}_n \bar{Y}_n$.

II-d) Representer sur un même graphique l'évolution en fonction de n de la covariance empirique pour chacun des couples de coordonnées (X_i, X_j) $(i < j)$. Commenter les résultats.

II-e) Estimer la densité de chacune des coordonnées et comparer graphiquement avec la densité théorique.

II-f) On pose $W = X_1 + \cdots + X_5$ la somme des coordonnées de X. Quelle est la loi de W. Simuler un échantillon suivant la loi de W à partir de l'échantillon simulé à la question II-a).

II-g) Estimer la densité de W et comparer graphiquement avec la densité théorique.

-III- Visualisation des corrélations. *Soit (Z, Y) un couple de variables aléatoires réelles qui appartient à L^2. La corrélation entre Z et Y est définie par*

$$\mathrm{Corr}(Y, Z) = \frac{\mathrm{Cov}(Y, Z)}{\sqrt{\mathrm{Var}(Y)\mathrm{Var}(Z)}}.$$

III-a) Montrer que $\mathrm{Corr}(Y, Z) \in]-1, 1[$.

III-b) Pour l'exemple précédent $(a = -0.8)$, calculer la matrice $R = (\mathrm{Corr}(\mathrm{X_i}, \mathrm{X_j}))_{1 \le i,j \le 5}$ appelée matrice de corrélation du vecteur X

III-c) Representer les nuages de points $\{(X_i(k), X_j(k)),\ k = 1..., n\}$ pour tout les couples $(i, j) \in \{1, ..., 5\}^2$, $i \neq j$. (Utiliser la fonction pairs).

III-d) Faire le lien entre la forme du nuage points et les valeurs de R.

3. Convergence presque sûre

Exercice 3.1. Estimation d'une densité. On suppose que f est une densité de probabilité définie sur $]-1,1[$. On suppose que f est une fonction continue dérivable

Comme f est continue , dérivable en tout x, on a la convergence ponctuelle de la série de Fourier pour tout $x \in]-1,1[$

$$f_N(x) = \frac{1}{2} + \sum_{k=1}^{+N} a_k(f) cos(\pi k x) + \sum_{k=1}^{+N} b_k(f) sin(\pi k x) \xrightarrow[N\to\infty]{} f(x)$$

avec pour $k > 0$,

$$a_k(f) = \int_{-1}^{1} f(t)\cos(kt\pi)\, dt \quad \text{et} \quad b_k(f) = \int_{-1}^{1} f(t)\sin(kt\pi)\, dt.$$

En prenant N assez grand, on peut approcher $f(x)$ par $f_N(x)$.

Dans cet exercice on cherche à construire un estimateur de $f(x)$ à partir de n observations $(X_1, \dots, X_n)$ indépendantes et identiquement distribuées suivant la loi de densité f. On pose

$$\alpha_k(n) = \frac{1}{n}\sum_{i=1}^{n} \cos(\pi k X_i)$$

$$\beta_k(n) = \frac{1}{n}\sum_{i=1}^{n} \sin(\pi k X_i)$$

et

$$f_N^n(x) = \frac{1}{2} + \sum_{k=1}^{+N} \alpha_k(n)\cos(\pi k x) + \sum_{k=1}^{+N} \beta_k(n)\sin(\pi k x)$$

1) Justifier le résultat suivant :

   pour tout $k \in \mathbb{N}$ fixé, on a

$$\alpha_k(n) \xrightarrow{p.s.} a_k \qquad n \to \infty$$
$$\beta_k(n) \xrightarrow{p.s.} b_k \qquad n \to \infty.$$

2) En déduire que pour tout couple $(N, x) \in \mathbb{N}\times]-1,1[$ fixé, on a

$$f_N^n(x) \xrightarrow{p.s.} f_N(x) \qquad n \to \infty.$$

3) Construire une fonction R qui calcule f_N^n sur une grille de l'intervalle $]-1,1[$ (par exemple `seq(-1,1,.05)`)
4) Soit Z une variable aléatoire qui suit une loi beta de paramètres (3,10). On pose $X = 2Z - 1$. Construire une fonction qui retourne un échantillon $(X_1, \dots, X_n)$ de variables aléatoires indépendantes et de même loi que X.
5) Justifier que la densité f de X s'écrit

$$f(x) = \frac{1}{2} b_{(3,10)}\Big(\frac{x+1}{2}\Big)\, \mathbb{I}_{]-1,1[}(x)$$

   où $b_{(3,10)}$ est la densité de la loi beta de paramètres (3,10).
6) On fixe $n = 100$. Superposer f et f_N^n avec $N = 1, 5, 10, 15$
7) Reprendre la question précédente avec $n = 500, 5000, 10000$
8) Commenter les résultats obtenus.

Exercice 3.2. Processus empirique. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuée suivant une loi continue. On rappelle que le processus empirique est défini par

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{]-\infty,t]}(X_i).$$

1) Illustrer la convergence presque sûre de $F_n(x)$ vers $F(x) = P(X_1 \leq x)$ et le théorème de Glivenko Cantelli sur les exemples suivants
   - X_1 suit une loi uniforme sur $[0,1]$
   - X_1 suit une loi gaussienne standard
   - X_1 suit une loi beta de paramètre $(1/2, 1/4)$,
   - X_1 suit une loi de student à 1 degré de liberté.
2) Illustrer la propriété suivante :

   la loi de la variable aléatoire $\sup_t |F_n(t) - F(t)|$ ne dépend pas de la loi de X_1

   Reprendre les exemples de la question 1).
3) Par la simulation, trouver une suite (u_n) telle que

$$u_n \sup_t |F_n(t) - F(t)|$$

   converge en loi vers une variable aléatoire non dégénérée.

4. Convergence en loi

Exercice 4.1.

1) Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant la loi exponentielle de paramètre 1. On rappelle que $\mathbb{E}(X_1) = \mathrm{Var}(X_1) = 1$
2) Déterminer les suites a_n et b_n telles que

$$S_n^* = a_n S_n - b_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

3) A l'aide de la fonction `pgamma`, construire une fonction qui calcule la fonction de répartition $F_{S_n^*}$ de S_n^* .
4) Illustrer la convergence en loi en superposant la fonction de répartition Φ de la loi gaussienne standard et la fonction de répartition $F_{S_n^*}$ pour différentes valeurs de n.
5) Comparer numériquement la vitesse de convergence de $F_{S_n^*}(x) - \Phi(x)$ en fonction de n. Comparer le cas $|x| = 1$ et $|x| \neq 1$

   Indications : Pour différentes valeurs de x
   - représenter $|F_{S_n^*}(x) - \Phi(x)|$ en fonction de n,
   - representer $log(|F_{S_n^*}(x) - \Phi(x)|)$ en fonction de $log(n)$.

   Remarque On peut démontrer le résultat suivant

$$F_{S_n^*}(x) - \Phi(x) = C(1-x^2)\Phi'(x)\frac{1}{\sqrt{n}} + DQ(x)\Phi'(x)\frac{1}{n} + O(n^{-3/2})$$

   où Q est un polynôme qui ne s'annule pas en ± 1 et C, D sont des constantes non nulles.

Exercice 4.2. Illustration du TCL en dimension 2.

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de vecteurs aléatoires indépendants et de même loi uniforme sur le disque de centre $(0,0)$ et de rayon 1. On simule un échantillon $(X_1,\ldots,X_n)$ de taille n en utilisant un changement de variable en coordonnées polaires (voir l'algorithme obtenu à l'exercice 1.3)

On pose

$$S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

1) Calculer m_i la moyenne et v_i la variance de la i ème coordonnées des vecteurs S_n^*.

Pour les valeurs de n suivantes $n = 2,\ 3,\ 4,\ 5,\ 25,\ 50,\ 100$

2) Générer N (prendre par exemple N=500) échantillons $(X_1,\ldots,X_n)$ de taille n. Puis calculer pour chacun des échantillons

$$S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

On note les N valeurs obtenues $S_n^{*1}, \ldots., S_n^{*N}$.

3) Pour $i = 1,2$ tracer l'histogramme de la i ème coordonnée des vecteurs $S_n^{*1}, \ldots., S_n^{*N}$, puis superposer la densité de la loi gaussienne $\mathcal{N}(m_i, v_i)$. Commenter
4) Tracer le nuage de points $S_n^{*1}, \ldots., S_n^{*N}$.
5) Tracer l'histogramme de cet échantillon de vecteurs aléatoires. Voir l'aide de la fonction `smoothScatter` pour les histogrammes en 2D.
6) Simuler un échantillon de taille N de vecteurs aléatoires iid suivant la loi $\mathcal{N}_2\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix}\right)$
7) Tracer l'histogramme de l'échantillon et comparer avec le graphique obtenu à la question 5.

5. Conditionnement

Soit (X, Z) un couple de variables aléatoires dont la loi est donnée par
- Z est une variable aléatoire à valeurs dans $\{0, 1\}$ et $p = 1/3 = P(Z = 1)$
- X est une variable aléatoire dont la loi conditionnelle sachant Z est la loi de Poisson de paramètre $5Z + 1$. Autrement dit pour tout $k \in \mathbb{N}$ et tout $z \in \{0, 1\}$ on a

$$P(X = k|Z = z) = \frac{1}{k!}e^{-5z-1}(5z + 1)^k$$

1) Quelle est la loi du couple (X, Z) ? Quelle est la loi de X ?
2) Quelle est la loi de Z sachant X
3) Calculer $E(X|Z)$ et $E(Z|X)$.
4) Ecrire une fonction `R` pour simuler une suite de couples aléatoires suivant cet algorithme

    `pour tout` $i = 1...n$

    a. `on simule` Z_i `suivant la loi de Bernoulli de paramètre` p

    b. `on simule` X_i `suivant la loi de Poisson de paramètre` $5Z_i + 1$
5) Simuler un échantillon de taille $n = 5000$.

   On veut verifier que cet algorithme retourne un échantillon simulé suivant la loi de (X, Z) calculée à la question 1).

Lois marginales

6) A partir de l'échantillon simulé, donner une estimation de la loi de X.
   Indications : calculer les fréquences empiriques à l'aide de la fonction `table`.
   Comparer graphiquement la loi théorique et la loi estimée.
7) Même question pour la loi de la variable aléatoire Z.

Loi du couple

8) Mettre l'échantillon simulé au format `data.frame`
   Indication : `D= data.frame(x,z)`.
9) Utiliser la fonction `table` pour calculer $N_{k,z}$ (pour tout $k \in \mathbb{N}$ et $z \in \{0, 1\}$) le nombre fois où le couple (k, z) est observé dans l'échantillon.
10) En déduire une estimation de la loi du couple (X, Z), puis comparer avec la loi théorique obtenue à la question 1).

Lois conditionnelles

11) Pour $z = 1$. Représenter graphiquement $\dfrac{N_{k,z}}{\sum_{i=1}^{n} Z_i}$ en fonction de k. C'est l'estimateur d'une loi conditionnelle, laquelle ? Ajouter la loi théorique sur le graphique.
12) Même question pour $z = 0$ et l'estimateur $\dfrac{N_{k,z}}{n - \sum_{i=1}^{n} Z_i}$
13) Représenter graphiquement

$$E_k = \frac{\sum_{i=1}^{k} X_i Z_i}{\sum_{i=1}^{k} Z_i}$$

   pour $k = 1...n$. Ajouter la droite d'équation $y = E(X|Z = 1)$. Expliquer théoriquement la convergence observée.
14) Même question pour

$$F_k = \frac{\sum_{i=1}^{k} X_i(1 - Z_i)}{\sum_{i=1}^{k}(1 - Z_i)}.$$

   Ajouter la droite d'équation $y = E(X|Z = 0)$. Expliquer théoriquement la convergence observée.