Université Paris Dauphine
U.F.R. Mathématiques de la Décision / MD3
MAT-6-3 / NOISE
Année 2007/2008

Projet de Statistique Exploratoire :
Quatrième partie à rendre la semaine du 12 janvier [à la scolarité]

Ce projet est à rédiger en binôme de manière *autonome* [un binôme = un projet]. Quatre parties seront rendues séparément à l'enseignant(e) en charge du binôme. Les parties faisant appel à la programmation doivent être programmées en **R** et le code **R** correspondant doit être fourni en annexe. Chaque projet fera l'objet d'une soutenance orale devant un enseignant. La notation portera sur (a) la validité des réponses, (b) l'originalité des textes et programmes fournis, (c) la clarté de la rédaction et de la mise en page des sorties graphiques, (d) la cohérence de l'exposé (oral). *Attention, toute duplication avérée de programme ou de texte se verra sanctionnée par une note nulle !*

On cherche à déterminer la probabilité d'un événement "rare", à savoir la probabilité de dépasser une valeur a donnée,

$$\varrho = P(X > \alpha)$$

par simulation plutôt que par calcul exact.

La loi de la variable aléatoire X est dans la famille de densité

$$f_{a,b,c,d}(x) \propto \left\{1 + ax^2 + bx^4\right\} \exp\left\{-c\, x^{1/d}\right\}\, \mathbb{I}_{\mathbb{R}^+}(x) \tag{1}$$

où $a \geq 0$, $b \geq 0$, $c > 0$, $d > 0$. Pour la réalisation du projet, le binôme prendra `dcba` égal à la somme des dates de naissance du binôme, en remplaçant d par 3 si $d = 0$ et de même pour c.

1. On s'intéresse à la production d'un échantillon aléatoire tiré suivant la densité $f_{a,b,c,d}$ (pour les valeurs numériques retenues ci-dessus).
   1.1 Montrer que les densités $f_{a,b,c,d}$ sont bien intégrables quels que soient les paramètres $a \geq 0$, $b \geq 0$, $c > 0$, $d > 0$ et indiquer pourquoi les contraintes $d \neq 0$ et $c \neq 0$ sont nécessaires. En déduire que les constantes de normalisation

   $$C_{a,b,c,d} = 1 \Big/ \int_0^\infty \left\{1 + ax^2 + bx^4\right\} \exp\left\{-c\, x^{1/d}\right\}\, \mathrm{d}x$$

   existent quels que soient $a \geq 0$, $b \geq 0$, $c > 0$, $d > 0$.
   1.2 Représenter en R les courbes des densités $f_{a,b,c,d}$ pour 5 valeurs différentes de a, b, c, d, en utilisant une approximation *numérique* des constantes $C_{a,b,c,d}$.
   1.3 Construire et implémenter en R un algorithme de génération par *acceptation-rejet* suivant la loi de densité $f_{a,b,c,d}$ en effectuant tout d'abord le changement de variable $y^d = x$.
   1.4 Générer par ordinateur un échantillon de taille 500 d'une loi de densité $f_{a,b,c,d}$ et comparer l'histogramme de l'échantillon ainsi produit avec la courbe de la densité $f_{a,b,c,d}$.
2. Dans cette partie, on cherche à approcher la constante de normalisation $C_{a,b,c,d}$ par simulation.
   2.1 En utilisant un échantillon généré à partir d'une loi exponentielle $\mathcal{E}(c/2)$, de densité

   $$g_c(x) = (c/2) \exp\{-cx/2\}\, \mathbb{I}_{\mathbb{R}^+}(x)\,,$$

   produire une approximation de la constante $C_{a,b,c,d}$, ainsi qu'un intervalle de précision sur cette approximation. En déduire un nombre de simulations nécessaire pour obtenir une précision de 10^{-2}.

2.2 Comparer la précision précédente à celle associée à un échantillon (de même taille) provenant d'une loi $\mathcal{E}(c)$.

2.3 La loi Gamma $\mathcal{G}(\alpha, \beta)$ est associée à la densité

$$h_{\alpha,\beta}(x) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} \exp\{-\beta x\}\, \mathbb{I}_{\mathbb{R}^+}(x)\,,$$

où $\Gamma(\alpha)$ est la fonction gamma codée gamma sous R. En utilisant le changement de variable $y^d = x$, montrer que la densité $f_{a,b,c,d}$ s'écrit comme somme pondérée de 3 densités de lois gamma et en déduire une formule exacte pour $C_{a,b,c,d}$.

2.4 Reprendre la comparaison de la précision des deux simulations par loi $\mathcal{E}(c/2)$ et par loi $\mathcal{E}(c)$ en effectuant une expérience de Monte Carlo approchant l'erreur moyenne

$$\mathbb{E}[\widehat{C}_{a,b,c,d} - C_{a,b,c,d}]$$

pour les deux approximations $\widehat{C}_{a,b,c,d}$ correspondantes.

3. On s'intéresse au quantile à 99.9% associé à $f_{a,b,c,1}$, noté $q_{.999}$. *(Attention, $d = 1$!)*

3.1 Expliquer pourquoi une approche par simulation directe suivant $f_{a,b,c,1}$ ne peut pas évaluer correctement la probabilité $\varrho = P(X > q_{.999})$, même si elle peut bien approcher le quantile à 90%, noté $q_{.9}$, par $\hat{q}_{.9}$.

3.2 On remplace la simulation suivant $f_{a,b,c,1}$ par une simulation suivant une loi exponentielle $\mathscr{E}(\lambda)$. Donner un estimateur sans biais de $\varrho = P(X > \alpha)$ fondé sur cet échantillon exponentiel quand $X \sim f_{a,b,c,1}$. En déduire un estimateur de $q_{.999}$.

En fixant $\lambda = 1/\hat{q}_{.9}$ et $n = 10000$, estimer le quantile à 90%, noté $q_{.99}$, par $\hat{q}_{.99}$.

En fixant $\lambda = 1/\hat{q}_{.99}$ et $n = 10000$, évaluer par une estimation bootstrap la variabilité de l'estimateur de $q_{.999}$.

3.3 Développer une expérience de Monte Carlo sélectionnant un paramètre λ conduisant à une petite variance pour l'estimateur de la question 3.2. (On vérifiera si la variance de cet estimateur existe pour tout $\lambda > 0$.)

3.4 On teste l'adéquation de l'échantillon produit ci-dessus à la famille de lois $\mathcal{E}(\beta)$ ($\beta > 0$).

Construire le seuil d'acceptation à 5% du test de Kolmogorov–Smirnov par une méthode de Monte Carlo (et non pas à partir de la fonction R ks.test() correspondante).

Calculer la statistique de test pour l'échantillon de la question 1.4 et conclure.

Donner une estimation bootstrap de de la probabilité d'acceptation du test.

4. On cherche à estimer non-paramétriquement la densité de la log-transformée de l'échantillon produit à la question 1.4 par la méthode du noyau.

4.1 Tester, de manière graphique, l'influence de la fenêtre sur la forme de l'estimateur du noyau, $\hat{f}_n$.

4.2 On utilise la mesure de divergence de Kullback–Leibler

$$\mathfrak{K}(f, \hat{f}_n) = \int \log\left(f(x)/\hat{f}_n(x)\right) f(x)dx$$

Donner une approximation de $\mathfrak{K}(f, \hat{f}_n)$ par simulation à partir d'un échantillon de f.

4.3 En utilisant cette approximation de la distance $\mathfrak{K}(f, \hat{f}_n)$, en déduire par simulation un choix optimal de la fenêtre dans l'estimateur du noyau. (On discrétisera le domaine de variation de cette fenêtre.)