Université Paris Dauphine — Nom de l'étudiant(e) :
Année 2009-2010
U.F.R. Mathématiques de la Décision / MD3

Examen de rattrapage du 2 septembre 2010 (3 heures)

Chaque couleur de question est indépendante des autres couleurs. Les questions à choix multiple se verront attribuer une note positive uniquement en présence d'un code R valide recopié sur la feuille d'examen. L'ordre des réponses des QCM est aléatoire.
Les documents papier du cours et l'aide en ligne de R sont autorisés. L'utilisation de messagerie ou de mail est interdite et, en cas d'utilisation avérée, se verra sanctionnée par la commission de discipline de l'Université, comme toute autre tentative de fraude. **La feuille du texte de cet examen doit être rendue avec la copie à la sortie de la salle d'examen.**

Question magenta [3 points]

La loi de Gumbel a pour densité

$$f(x) = \exp\{x - \exp(x)\}$$

sur la droite réelle.

1. [**1**] Montrez que l'espérance de $\exp(X)$ est bien définie pour cette loi.
2. [**.5**] Créez une matrice `x` de simulations normales à 100 colonnes en utilisant `rnorm(100*10^4)` et déduisez les poids d'échantillonnage d'imporance `we`.
1. [**1.5**] Déduisez des versions régulières et auto-normalisées des estimateurs de $\mathbb{E}[\exp(X)]$ par un code similaire à

```
> nore=apply(we*exp(x),2,cumsum)/(1:10^4)
> reno=apply(we*exp(x),2,cumsum)/apply(we,2,cumsum)
```

et comparer leur variance.

Question bleue [7 points]

Soit l'intégrale

$$I = \int_0^4 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, \mathrm{d}y$$

1. [**.5**] Donner la valeur exacte de I (à 10^{-10} près) par une commande R.
2. [**1.5**] Proposer une méthode de Monte Carlo d'évaluation de I reposant sur la génération d'un n-échantillon de variables aléatoires gaussiennes. Fournir un intervalle de confiance à 95% sur I pour $n = 1000$.
3. [**1.5**] Proposer une autre méthode de Monte Carlo reposant sur un n-échantillon de variables aléatoires de loi uniforme. Fournir un intervalle de confiance à 95% sur I pour $n = 1000$. Parmi ces 2 méthodes, laquelle est la meilleure ?

Nous cherchons à améliorer la méthode précédente. Pour cela, *nous admettons que* I s'écrit aussi de la façon suivante :

$$I = \frac{1}{2}\int_0^4 \left[\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}} + \frac{1}{\sqrt{2\pi}}e^{-\frac{(4-y)^2}{2}}\right] \mathrm{d}y$$

4. **[1.5]** Déduire une nouvelle méthode d'estimation de I reposant sur des variables aléatoires de loi uniforme. Fournir un intervalle de confiance à 95% sur I pour $n = 1000$.

5. **[1]** Fournir les codes R pour illustrer sur un même graphique, la convergence des 3 méthodes précédentes en fonction de la taille n de l'échantillon simulé. On souhaite faire apparaître sur le graphique la valeur exacte de I.

Question ocre [6 points]

Soit une densité de probabilité f sur $\mathbb{R}$ telle que

$$f(x) \propto 0.7\exp(-x^2/2) + 0.3\exp(-x^2/2 + 7(x - 7/2))\,.$$

C est la constante de proportionalité définie par l'identité

$$C\int_{\mathbb{R}} \left\{0.7\exp(-x^2/2) + 0.3\exp(-x^2/2 + 7(x - 7/2))\right\} \mathrm{d}x = 1\,.$$

On génére suivant f par acceptation-rejet en partant d'une densité g, avec $f/g \leq MC$. *(On rappelle que la densité d'une loi de Student de paramètre d est donnée par* `dt(x,df=d)` *et la densité d'une loi de Cauchy par* `dcauchy(x)`*.)*

1. **[2]** Entourer les couples (g, M) valides. (On pourra utiliser la fonction `optimise`.)

1. $\mathcal{E}xp(3)$ et $M = 345$
2. $N(2.1, 3^2)$ et $M = 7$
3. $Student_3$ et $M = 300$
4. $U_{[-10,10]}$ et $M = 24$
5. $N(2.1, 3^2)$ et $M = 11$
6. Cauchy et $M = 40$

2. **[1]** Entourer le couple (g, M) le plus efficace en termes de nombres de rejets.

1. $N(2.1, 3^2)$ et $M = 11$
2. Cauchy et $M = 40$
3. $\mathcal{E}xp(3)$ et $M = 345$
4. $U_{[-10,10]}$ et $M = 24$
5. $N(2.1, 3^2)$ et $M = 7$
6. $Student_3$ et $M = 300$

On choisit finalement d'utiliser pour g une loi de Cauchy et $M = 50$.

3. **[1]** Simuler un échantillon de départ de taille $n = 10000$ réalisations de $Y \sim g$ en utilisant le couple (g, M) trouvé dans la question précédent. Parmi les propositions suivantes, entourer le taux d'acceptation le plus proche du résultat obtenu ?

1. 0.05
2. 0.89
3. 0.01
4. 0.34
5. 0.82

4. **[1]** Déduire la valeur de la constante de normalisation C du taux d'acceptation obtenu précédemment.

Question lilas [3.5 points]
On s'intéresse à une v.a. X de densité

$$f(x) = 5\,x^4 \mathbf{I}_{[0,1]}(x)$$

1. **[1.5]** On simule un échantillon de taille $n = 1000$ de f en utilisant un échantillon $(U_1, \cdots, U_n) \sim U_{[0,1]}$. Entourer la transformation à utiliser :

1. $X = U^{1/5}$
2. $X = (6U/5)^{1/6}$
3. $X = 5\,U^4$
4. $X = U/5$
5. $X = 4 * U^5$

2. **[.5]** Proposer une vérification graphique du résultat obtenu.

On rappelle que la fonction de répartition empirique, estimateur de la fonction de répartition F, est définie par $\widehat{F_n}(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{X_i<t}$.

3. **[1]** Ecrire une fonction R calculant $\widehat{F_n}(t)$ et donner une estimation de $F(t)$ au point $t = 0.3$ en fonction de cet échantillon.

4. **[.5]** Donner le code R permettant d'afficher le graphe de $\widehat{F_n}$.

Question pourpre [6 points]
Etant donnée la densité $f(x) \propto \exp\{-x^2\sqrt{x}\}[\sin(x)]^6$, $0 < x < \infty$, de la variable aléatoire X, on considère les lois de densité

$$g_1(x) = \frac{1}{2}e^{-|x|}, \quad g_2(x) = \frac{1}{2\pi}\frac{1}{1+x^2/4}, \quad g_3(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,.$$

Pour chaque question ci-dessous, fournir le code R utilisé.

1. **[4]** Pour des échantillons iid produits suivant chacune de ces lois g_i, estimer par une expérience de Monte Carlo le nombre M de simulations nécessaire pour obtenir une précision de trois décimales sur $\mathbb{E}_f[X]$. Fournir les trois valeurs estimées de $\mathbb{E}_f[X]$.

2. **[1]** En utilisant g_1 comme proposition, générer un échantillon suivant f par acceptation-rejet et en déduire une estimation de la constante de normalisation de f.

3. **[1]** Comparez aux estimations obtenues en utilisant g_2 et g_3.

Question sépia [6 points]
On travaille sur des données extraites de la base `faithful` par la commande

`erup=faithful[91:180,1].`

On suppose que la loi de densité f qui a généré les données appartient à la famille paramétrique de lois $\{f_{3,\lambda}\}_\lambda$ où $f_{3,\lambda}$ est la densité d'une loi Gamma$(3,\lambda)$. La densité $f_{p,\lambda}$ est donnée par

$$f_{p,\lambda}(x) = \frac{\lambda^p}{\Gamma(p)} e^{-\lambda x} x^{p-1} \mathbb{I}_{x>0}\,.$$

1. [.5] On précise que l'espérance d'une loi Gamma(p,λ) vaut $\frac{p}{\lambda}$. Calculer un estimateur $\widehat{\lambda}$ de λ en utilisant la moyenne empirique des données et entourer le résultat le plus proche dans la liste.

(a) 0.578
(b) 0.858
(c) 1.730
(d) 0.087
(e) 1.326

3. [1] Le mode d'une distribution de densité f est le point m où $f(x)$ est maximale. Entourer l'expression du mode de $f_{3,\lambda}$ (on pourra s'aider d'une représentation graphique) ci-dessous :

(a) λ
(b) $\frac{3}{\lambda}$
(c) $\frac{2}{\lambda}$
(d) $\frac{1}{\lambda}$
(e) $\frac{\lambda}{2}$

4. [.5] Déduire un estimateur $\widehat{m}$ du mode de f et entourer la valeur ci-dessous la plus proche :

(a) 0.754
(b) 1.730
(c) 2.331
(d) 2.924
(e) 0.578

On souhaite calculer un intervalle de confiance pour m par bootstrap non-paramétrique.

5. [1.5] Simuler $B = 1000$ échantillons bootstrap non-paramétriques X^{*l}, $l = 1,\ldots,B$. Calculer l'estimateur $\widehat{m}^{*l}$ du mode de f pour chaque échantillon bootstrap. En déduire une estimation de l'intervalle de confiance pour m à 95%. Donner les lignes de code R correspondant.

6. [.5] Donner l'intervalle ci-dessous le plus proche de l'estimation obtenue
(a) $[2.34; 3.62]$

(b) $[0.72; 0.79]$
(c) $[2.17; 2.49]$

(d) $[1.61; 1.85]$
(e) $[2.91; 2.97]$

On utilise dans cette question un estimateur de f par noyau gaussien K. L'estimateur suivant (qu'on ne demande pas de calculer)

$$\widehat{f}_{NP}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

est implémenté dans R pour une largeur de fenêtre h calculée automatiquement. On utilise pour cela la commande `d=density(erup)`. Le résultat `d` est une liste. On a accés à un vecteur d'antécédents et d'images par $\widehat{f}_{NP}$ avec les commandes `d$x` et `d$y`.

7. [2] On propose comme nouvel estimateur du mode de f le mode de $\widehat{f}_{NP}$, noté $\widehat{m}_{NP}$. Approcher le mode de $\widehat{f}_{NP}$ (indice : utiliser la fonction `which.max`). Quelle valeur s'en-rapproche le plus ?

(a) 4.49
(b) 4.38
(c) 4.29
(d) 4.34
(e) 4.42