Université Paris Dauphine — Année 2008-2009
U.F.R. Mathématiques de la Décision / MD3

Examen final du 5 janvier 2009
Séance de 8 heures à 10h45

Préliminaires

Vous devez enregistrer la page Web contenant vos réponses périodiquement pour les valider. La note finale sera attribuée au vu du document informatique seul et l'absence de document enregistré donnera lieu à une note nulle sans possibilité de contestation.
Les questions à choix multiple comportant une plage d'entrée du code informatique se verront attribuer une note positive uniquement en présence de codes R valides. Chaque question est indépendante des autres questions. L'ordre des questions et des réponses est aléatoire.
Les documents disponibles sur votre compte informatique sont autorisés, ainsi que les documents papier du cours et l'aide en ligne de R. L'utilisation de tout service de messagerie ou de mail est interdite et, en cas d'utilisation avérée, se verra sanctionnée par une note nulle pour les deux parties. La copie papier de l'examen doit être rendue à la sortie de la salle informatique.

Question bleue [4 points]
Le code suivant vise à résoudre une grille de Sudoku simple : on rappelle qu'un Sudoku est une grille 9x9 partiellement remplie par des chiffres pour laquelle il existe une seule solution telle qu'un chiffre entre 1 et 9 n'apparaisse qu'une seule fois dans une ligne, dans une colonne ou dans un bloc 3x3. Pour donner la grille de Sudoku, on définit

```
s=matrix(0,ncol=9,nrow=9)
s[1,c(6,8)]=c(6,4)
s[2,c(1:3,8)]=c(2,7,9,5)
s[3,c(2,4,9]=c(5,8,2)
s[4,3:4]=c(2,6)
s[6,c(3,5,7:9)]=c(1,9,6,7,3)
s[7,c(1,3:4,7)]=c(8,5,2,4)
s[8,c(1,8:9)]=c(3,8,5)
s[9,c(1,7,9)]=c(6,9,1)
```

où les cases vides sont indiquées par des zéros.

1. [**1**] Corriger la faute de frappe présente dans cette définition de la grille, fournir la ligne de code corrigée et imprimer la grille dans la case ci-dessous.

On définit une table des valeurs possibles de chaque case par `pool=array(TRUE,dim=c(9,9,9))`, `pool[i,j,]` représentant les valeurs possibles de `s[i,j]` entre 1 et 9 par des `TRUE` et les impossibles par des `FALSE`.

2.[**.5**] Déterminer si l'élimination des cases `s[i,j]` déjà remplies se fait par la boucle commençant par
`for (i in 1:9) for (j in 1:9){`
et continuant par

1. `pool[(s[i,j]>0),-s[i,j]]=FALSE}`
2. `if (s[i,j]>0) pool[i,j,s[i,j]]=FALSE}`
3. `if (s[i,j]>0) pool[i,j,-s[i,j]]=FALSE}`
4. `if (s[i,j]>0) pool[i,j,1:9-s[i,j]]=FALSE}`

Ayant ainsi éliminé les cases remplies de `s`, on parcourt les cases non remplies de `s` par

```
for (t in 1:100){
for (i in sample(1:81)){
    if (s[i]==0){
       a=((i-1)%%9)+1
       b=trunc((i-1)/9)+1
       boxa=3*trunc((a-1)/3)+1
       boxa=boxa:(boxa+2)
       boxb=3*trunc((b-1)/3)+1
       boxb=boxb:(boxb+2)
       for (u in (1:9)[pool[a,b,]]){
         pool[a,b,u]=(sum(u==s[a,])+sum(u==s[,b])+sum(u==s[boxa,boxb]))==0
         }
       if (sum(pool[a,b,])==1){
         s[i]=(1:9)[pool[a,b,]]
  }}}}
```

3. [**.5**] A quoi est égal `s[i]` dans le code ci-dessus ?

1. `s[i,]`
2. `s[a,b]`
3. `s[,i]`
4. `s[b,a]`
5. `s[i,i]`

4. [**1**] La boucle extérieure `for (t in 1:100)` est employée ci-dessus par défaut pour garantir la complétion de toutes les entrées de la grille de Sudoku. Remplacer cette boucle par une contrainte plus efficace utilisant `while` et portant sur la présence ou non de cases de `s` égales à zero.

5. [**1**] Produire la solution de cette grille de Sudoku dans la case ci-dessous.

Question turquoise [4 points]
On s'intéresse à la variable aléatoire X définie sur $\mathbb{R}$ de densité f proportionnelle à :

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

On note C la constante de proportionalité, c'est-à-dire le facteur défini 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 souhaite générer un échantillon de densité f par acceptation-rejet en partant d'un échantillon généré sous une loi de 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 $Cauchy(0, 1)$ par `dcauchy(x)`.

1. **[2]** Parmi les propositions suivantes, quels couples (g, M) sont valides ? (On pourra utiliser la fonction `optimise`.)

1. $U_{[-10,10]}$ et $M = 24$
2. $N(2.1, 3^2)$ et $M = 11$
3. $Cauchy(0, 1)$ et $M = 40$
4. $\mathcal{E}xp(3)$ et $M = 345$
5. $N(2.1, 3^2)$ et $M = 7$
6. $Student_3$ et $M = 300$

2. **[1]** Parmi les propositions suivantes, quel couple (g, M) est le plus efficace en termes de nombres de simulations ?

1. $U_{[-10,10]}$ et $M = 24$
2. $Student_3$ et $M = 300$
3. $N(2.1, 3^2)$ et $M = 11$
4. $Cauchy(0, 1)$ et $M = 40$
5. $\mathcal{E}xp(3)$ et $M = 345$
6. $N(2.1, 3^2)$ et $M = 7$

On choisit finalement d'utiliser pour g une loi $Cauchy(0, 1)$ et $M = 50$.

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

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

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 maintenant à une v.a. X de densité

$$f(x) = 3\,x^2 \mathbf{I}_{[0,1]}(x)$$

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

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

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 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 permettant d'afficher le graphe de $\widehat{F_n}$.

Question ocre [7 points]
Nous cherchons à évaluer l'intégrale suivante :

$$I = \int_0^3 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

par des méthodes de Monte Carlo

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 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 méthode de Monte Carlo reposant sur la génération d'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 pouvons montrer que I peut être écrite de la façon suivante :

$$I = \int_0^3 \frac{1}{\sqrt{2\pi}} e^{-\frac{(3-x)^2}{2}} dx$$

Par conséquent :

$$I = \frac{1}{2}\int_0^3 \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(3-x)^2}{2}}\right] dx$$

4. **[1.5]** En 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 pour illustrer sur un même graphique, la convergence des 3 méthodes précédentes en fonction de la taille de l'échantillon simulé. On souhaite faire apparaître sur le graphique la valeur exacte de I.

Question sépia [6 points]
On considère le jeu de données `fdeaths` (déjà résident).

1. **[.5]** Reproduire ci-dessous la description des données `fdeaths`.

On cherche à tester l'adéquation de ces données à une loi exponentielle, $\mathcal{E}(\lambda)$, de densité $\lambda \exp(-\lambda x)$ sur $\mathbb{R}_+$. On travaille avec une erreur de Type I égale à 0.05.

2. **[1]** Interpréter le résultat d'un test de Kolmogorov-Smirnov sur l'adéquation de ces données à une loi exponentielle de paramètre $\lambda = 1/550$: les données sont

1. compatibles avec une loi $\mathcal{E}(1/550)$;
2. incompatibles avec une loi $\mathcal{E}(1/550)$.

et indiquer après avoir tracé l'histogramme des données une raison simple pour ce résultat.

On définit l'échantillon `drift` par `drift=fdeaths-min(fdeaths)`.

3. [**1**] Interpréter le résultat d'un test de Kolmogorov-Smirnov sur l'adéquation de `drift` à une loi exponentielle de paramètre `lambda=1/mean(drift)`.

Le fait de choisir λ en fonction des données invalide la p-value fournie par R. On construit ci-dessous une p-value non-biaisée par bootstrap.

4. [**1.5**] Si D est la valeur de la statistique retournée par le test de Kolmogorov-Smirnov sous R, évaluer par un bootstrap paramétrique la probabilité de dépasser le D observé pour `drift` sous l'hypothèse nulle.

On rappelle que l'erreur de Type II est la probabilité d'accepter à tort l'adéquation à une loi exponentielle. On construit ci-dessous une évaluation bootstrap de l'erreur de Type II.

5.[**2**] Soit $\overline{D}$ le quantile de D à 0.95 obtenu par le bootstrap paramétrique de la question **3.** ci-dessus. Calculer $\overline{D}$ et évaluer par un bootstrap non-paramétrique fondé sur `drift` la probabilité que D soit en dessous de $\overline{D}$, ce qui donne l'erreur de Type II.

Question pourpre [6 points]
On travaille sur des données extraites de la base `faithful` de la manière suivante

`erup=faithful[1:90,1]`.

On suppose d'abord que la loi de distribution f qui a généré les données appartient à la famille paramétrique de lois $\{f_{2,\lambda}\}_\lambda$ où $f_{2,\lambda}$ est la densité d'une loi Gamma$(2,\lambda)$. 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 une moyenne empirique.

(a) 0.578
(b) 0.342
(c) 1.730
(d) 0.087
(e) 1.326

2. [**1**] En déduire l'estimateur $\widehat{f}$ de f associé dans la famille paramétrique (écrire une fonction `f_hat`). Représenter graphiquement cette fonction. Donner les commandes relatives à ces deux questions.

3. [**1**] Le mode d'une distribution de densité f est le point m où $f(x)$ est maximale. Quelle est l'expression du mode de $f_{2,\lambda}$ (on pourra s'aider d'une représentation graphique)

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

4. **[.5]** Calculer un estimateur $\widehat{m}$ du mode de f

(a) 0.754
(b) 1.730
(c) 11.494
(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 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) $[1.61; 1.85]$
(d) $[9.39; 13.47]$
(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 accès à un vecteur d'antécédents et d'images par $\widehat{f}_{NP}$ avec les commandes `d$x` et `d$y`.

7. **[1]** 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) 4.43
(b) 4.32
(c) 4.29
(d) 4.38
(e) 4.41

Université Paris Dauphine — Année 2008-2009
U.F.R. Mathématiques de la Décision / MD3

Examen final du 5 janvier 2009
Séance de 11 heures à 13h45

Préliminaires

Vous devez enregistrer la page Web contenant vos réponses périodiquement pour les valider. La note finale sera attribuée au vu du document informatique seul et l'absence de document enregistré donnera lieu à une note nulle sans possibilité de contestation.
Les questions à choix multiple comportant une plage d'entrée du code R se verront attribuer une note positive uniquement en présence de codes R valides. Chaque question est indépendante des autres questions. L'ordre des questions et des réponses est aléatoire. Les documents disponibles sur votre compte informatique sont autorisés, ainsi que les documents papier du cours et l'aide en ligne de R. L'utilisation de tout service de messagerie ou de mail est interdite et, en cas d'utilisation avérée, se verra sanctionnée par une note nulle pour les deux parties. La copie papier de l'examen doit être rendue à la sortie de la salle informatique.

Question bleue [4 points]
Le code suivant vise à résoudre une grille de Sudoku simple : on rappelle qu'un Sudoku est une grille 9x9 partiellement remplie par des chiffres pour laquelle il existe une seule solution telle qu'un chiffre entre 1 et 9 n'apparaisse qu'une seule fois dans une ligne, dans une colonne ou dans un bloc 3x3. Pour donner la grille de Sudoku, on définit

```
s=matrix(0,ncol=9,nrow=9)
s[1,c(1,2,4,9)]=c(3,1,2,6)
s[2,c(2,3,5,8,9)]=c(5,2,6,7,8)
s[3,c(5,7,8)]=c(3,1,2)
s[4,c(2,4,6,7,9)]=c(3,8,6,2,9)
s[5,c(1,4:6,8,9)]=c(2,7,9,5,1,3)
s[6,c(3,4,6:9)]=c(8,4,3,5,6,7)
s[7,c(1:4,6:8)]=c(5,6,4,3,2,7,9)
s[8,c(1:5,9)]=c(8,7,1,9,5,2)
s[9,c(1,6,8]=c(9,1,8)
```

où les cases vides sont indiquées par des zéros.

1. [**1**] Corriger la faute de frappe présente dans cette définition de la grille, fournir la ligne de code corrigée et imprimer la grille dans la case ci-dessous.

On définit une table des valeurs possibles de chaque case par `pool=array(TRUE,dim=c(9,9,9))`, `pool[i,j,]` représentant les valeurs possibles de `s[i,j]` entre 1 et 9 par des `TRUE` et les impossibles par des `FALSE`.

2.[**.5**] Déterminer si l'élimination des cases `s[i,j]` déjà remplies se fait par la boucle commençant par
`for (i in 1:9) for (j in 1:9){`
et continuant par

1. `pool[(s[i,j]>0),-s[i,j]]=FALSE}`
2. `if (s[i,j]>0) pool[i,j,s[i,j]]=FALSE}`
3. `if (s[i,j]>0) pool[i,j,-s[i,j]]=FALSE}`
4. `if (s[i,j]>0) pool[i,j,1:9-s[i,j]]=FALSE}`

Ayant ainsi éliminé les cases remplies de `s`, on parcourt les cases non remplies de `s` par

```
for (t in 1:100){
for (i in sample(1:81)){
    if (s[i]==0){
       a=((i-1)%%9)+1
       b=trunc((i-1)/9)+1
       boxa=3*trunc((a-1)/3)+1
       boxa=boxa:(boxa+2)
       boxb=3*trunc((b-1)/3)+1
       boxb=boxb:(boxb+2)
       for (u in (1:9)[pool[a,b,]]){
         pool[a,b,u]=(sum(u==s[a,])+sum(u==s[,b])+sum(u==s[boxa,boxb]))==0
         }
       if (sum(pool[a,b,])==1){
         s[i]=(1:9)[pool[a,b,]]
  }}}}
```

3. [**.5**] A quoi est égal `s[i]` dans le code ci-dessus ?

1. `s[i,]`
2. `s[a,b]`
3. `s[,i]`
4. `s[b,a]`
5. `s[i,i]`

4. [**1**] La boucle extérieure `for (t in 1:100)` est employée ci-dessus par défaut pour garantir la complétion de toutes les entrées de la grille de Sudoku. Remplacer cette boucle par une contrainte plus efficace utilisant `while` et portant sur la présence ou non de cases de `s` égales à zero.

5. [**1**] Produire la solution de cette grille de Sudoku dans la case ci-dessous.

Question turquoise [4 points]
On s'intéresse à la variable aléatoire X définie sur $\mathbb{R}$ de densité f proportionnelle à :

$$f(x) \propto 0.4\exp(-x^2/2) + 0.6\exp(-x^2/2 + 4(x-2))$$

On note C la constante de proportionalité, c'est-à-dire le facteur défini par l'identité

$$C\int_{\mathbb{R}} \left\{0.4\exp(-x^2/2) + 0.6\exp(-x^2/2 + 4(x-2))\right\}\,\mathrm{d}x = 1,$$

On souhaite générer un échantillon de densité f par acceptation-rejet en partant d'un échantillon généré sous une loi de 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 $Cauchy(0,1)$ par `dcauchy(x)`.

1. [**2**] Parmi les propositions suivantes, quels couples (g, M) sont valides ? (On pourra utiliser la fonction `optimise`.)

1. $U_{[-10,10]}$ et $M = 24$
2. $Student_6$ et $M = 310$
3. $N(2.4, 6^2)$ et $M = 10$
4. $Cauchy(0, 1)$ et $M = 20$
5. $\mathcal{E}xp(3)$ et $M = 345$
6. $N(2.4, 6^2)$ et $M = 7$

2. [**1**] Parmi les propositions suivantes, quel couple (g, M) est le plus efficace en termes de nombres de simulations ?

1. $U_{[-10,10]}$ et $M = 24$
2. $Student_6$ et $M = 310$
3. $N(2.4, 6^2)$ et $M = 10$
4. $Cauchy(0, 1)$ et $M = 20$
5. $\mathcal{E}xp(3)$ et $M = 345$
6. $N(2.4, 6^2)$ et $M = 7$

On choisit finalement d'utiliser pour g une loi $Cauchy(0, 1)$ et $M = 36$.

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

1. 0.88
2. 0.026
3. 0.069
4. 0.82
5. 0.34

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 maintenant à une v.a. X de densité

$$f(x) = \frac{\exp(x)}{\exp(1) - 1} \mathbf{I}_{[0,1]}(x)$$

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

1. $X = \exp(x) \Big/ (\exp(1) - 1)$
2. $X = \log[(\exp(1) - 1)U + 1]$
3. $X = (\exp(x) - 1) \Big/ (\exp(1) - 1)$
4. $X = \log[(\exp(1) - 1)U]$

5. $X = \log[(\exp(1)-1)U] \Big/ (\exp(1)-1)$

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 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 permettant d'afficher le graphe de $\widehat{F_n}$.

Question ocre [7 points]
Nous cherchons à évaluer l'intégrale suivante :

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

par des méthodes de Monte Carlo

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 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 méthode de Monte Carlo reposant sur la génération d'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 pouvons montrer que I peut être écrite de la façon suivante :

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

Par conséquent :

$$I = \frac{1}{2}\int_0^1 \left[e^{\frac{-x^2}{2}} + e^{\frac{-(1-x)^2}{2}}\right] dx$$

4. **[1.5]** En déduire une nouvelle méthode d'estimation de I reposant sur des variables aléatoires uniformes. Fournir un intervalle de confiance à 95% pour I pour $n = 500$.
5. **[1]** Fournir les codes pour illustrer sur un même graphique, la convergence des 3 méthodes précédentes en fonction de la taille de l'échantillon simulé. On souhaite faire apparaître sur le graphique la valeur exacte de I.

Question sépia [6 points]
On considère le jeu de données `fdeaths` (déjà résident).

1. **[.5]** Reproduire ci-dessous la description des données `fdeaths`.

On cherche à tester l'adéquation de ces données à une loi exponentielle, $\mathcal{E}(\lambda)$, de densité $\lambda \exp(-\lambda x)$ sur $\mathbb{R}_+$. On travaille avec une erreur de Type I égale à 0.05.

2. [**1**] Interpréter le résultat d'un test de Kolmogorov-Smirnov sur l'adéquation de ces données à une loi exponentielle de paramètre $\lambda = 1/540$: les données sont

1. compatibles avec une loi $\mathcal{E}(1/540)$;
2. incompatibles avec une loi $\mathcal{E}(1/540)$.

et indiquer après avoir tracé l'histogramme des données une raison simple pour ce résultat.

On définit l'échantillon `drift` par `drift=fdeaths-min(fdeaths)`.

3. [**1**] Interpréter le résultat d'un test de Kolmogorov-Smirnov sur l'adéquation de `drift` à une loi exponentielle de paramètre `lambda=1/sqrt(var(drift))`.

Le fait de choisir λ en fonction des données invalide la p-value fournie par R. On construit ci-dessous une p-value non-biaisée par bootstrap.

4. [**1.5**] Si D est la valeur de la statistique retournée par le test de Kolmogorov-Smirnov sous R, évaluer par un bootstrap paramétrique la probabilité de dépasser le D observé pour `drift` sous l'hypothèse nulle.

On rappelle que l'erreur de Type II est la probabilité d'accepter à tort l'adéquation à une loi exponentielle. On construit ci-dessous une évaluation bootstrap de l'erreur de Type II.

5.[**2**] Soit $\overline{D}$ le quantile de D à 0.95 obtenu par le bootstrap paramétrique de la question **3.** ci-dessus. Calculer $\overline{D}$ et évaluer par un bootstrap non-paramétrique fondé sur `drift` la probabilité que D soit en dessous de $\overline{D}$, ce qui donne l'erreur de Type II.

Question pourpre [6 points]
On travaille sur des données extraites de la base `faithful` de la manière suivante

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

On suppose d'abord que la loi de distribution 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 une moyenne empirique.

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

2. [**1**] En déduire l'estimateur $\widehat{f}$ de f associé dans la famille paramétrique (écrire une fonction `f_hat`). Représenter graphiquement cette fonction. Donner les commandes relatives à ces deux questions.

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

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

4. **[.5]** Calculer un estimateur $\widehat{m}$ du mode de f

(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 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 accès à un vecteur d'antécédents et d'images par $\widehat{f}_{NP}$ avec les commandes `d$x` et `d$y`.

7. **[1]** 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

Université Paris Dauphine — Année 2008-2009
U.F.R. Mathématiques de la Décision / MD3

Examen final du 5 janvier 2009
Séance de 14 heures à 16h45

Préliminaires

Vous devez enregistrer la page Web contenant vos réponses périodiquement pour les valider. La note finale sera attribuée au vu du document informatique seul et l'absence de document enregistré donnera lieu à une note nulle sans possibilité de contestation.
Les questions à choix multiple comportant une plage d'entrée du code R se verront attribuer une note positive uniquement en présence de codes R valides. Chaque question est indépendante des autres questions. L'ordre des questions et des réponses est aléatoire.
Les documents disponibles sur votre compte informatique sont autorisés, ainsi que les documents papier du cours et l'aide en ligne de R. L'utilisation de tout service de messagerie ou de mail est interdite et, en cas d'utilisation avérée, se verra sanctionnée par une note nulle pour les deux parties. La copie papier de l'examen doit être rendue à la sortie de la salle informatique.

Question bleue [4 points]
Le code suivant vise à résoudre une grille de Sudoku simple : on rappelle qu'un Sudoku est une grille 9x9 partiellement remplie par des chiffres pour laquelle il existe une seule solution telle qu'un chiffre entre 1 et 9 n'apparaisse qu'une seule fois dans une ligne, dans une colonne ou dans un bloc 3x3. Pour donner la grille de Sudoku, on définit

```
s=matrix(0,ncol=9,nrow=9)
s[1,c(1,2,4,9)]=c(1,0,2,6)
s[2,c(1,2,5,9)]=c(8,3,1,9)
s[3,c(2,6,7)]=c(2,9,3)
s[4,c(3:5,7)=c(3,4,7,6)
s[5,c(4:6,8,9)]=c(5,6,8,3,1)
s[6,c(1,3,6,9)]=c(6,1,2,4)
s[7,c(1,8,9)]=c(7,8,3)
s[8,c(1,2,4,6,7)]=c(5,4,1,7,9)
s[9,c(2:7,9)]=c(1,9,8,2,6,7,5)
```

où les cases vides sont indiquées par des zéros.

1. [**1**] Corriger la faute de frappe présente dans cette définition de la grille, fournir la ligne de code corrigée et imprimer la grille dans la case ci-dessous.

On définit une table des valeurs possibles de chaque case par `pool=array(TRUE,dim=c(9,9,9))`, `pool[i,j,]` représentant les valeurs possibles de `s[i,j]` entre 1 et 9 par des `TRUE` et les impossibles par des `FALSE`.

2.[**.5**] Déterminer si l'élimination des cases `s[i,j]` déjà remplies se fait par la boucle commençant par
`for (i in 1:9) for (j in 1:9){`
et continuant par

1. `pool[(s[i,j]>0),-s[i,j]]=FALSE}`
2. `if (s[i,j]>0) pool[i,j,s[i,j]]=FALSE}`
3. `if (s[i,j]>0) pool[i,j,-s[i,j]]=FALSE}`
4. `if (s[i,j]>0) pool[i,j,1:9-s[i,j]]=FALSE}`

Ayant ainsi éliminé les cases remplies de `s`, on parcourt les cases non remplies de `s` par

```
for (t in 1:100){
for (i in sample(1:81)){
    if (s[i]==0){
       a=((i-1)%%9)+1
       b=trunc((i-1)/9)+1
       boxa=3*trunc((a-1)/3)+1
       boxa=boxa:(boxa+2)
       boxb=3*trunc((b-1)/3)+1
       boxb=boxb:(boxb+2)
       for (u in (1:9)[pool[a,b,]]){
         pool[a,b,u]=(sum(u==s[a,])+sum(u==s[,b])+sum(u==s[boxa,boxb]))==0
         }
       if (sum(pool[a,b,])==1){
         s[i]=(1:9)[pool[a,b,]]
  }}}}
```

3. [**.5**] A quoi est égal `s[i]` dans le code ci-dessus ?

1. `s[i,]`
2. `s[a,b]`
3. `s[,i]`
4. `s[b,a]`
5. `s[i,i]`

4. [**1**] La boucle extérieure `for (t in 1:100)` est employée ci-dessus par défaut pour garantir la complétion de toutes les entrées de la grille de Sudoku. Remplacer cette boucle par une contrainte plus efficace utilisant `while` et portant sur la présence ou non de cases de `s` égales à zero.

5. [**1**] Produire la solution de cette grille de Sudoku dans la case ci-dessous.

Question turquoise [4 points]
On s'intéresse à la variable aléatoire X définie sur $\mathbb{R}$ de densité f proportionnelle à :

$$f(x) \propto 0.2\exp(-x^2/2) + 0.8\exp(-x^2/2 + 2(x-1))$$

On note C la constante de proportionalité, c'est-à-dire le facteur défini par l'identité

$$C\int_{\mathbb{R}} \left\{0.2\exp(-x^2/2) + 0.8\exp(-x^2/2 + 2(x-1))\right\}\,\mathrm{d}x = 1,$$

On souhaite générer un échantillon de densité f par acceptation-rejet en partant d'un échantillon généré sous une loi de 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 $Cauchy(0,1)$ par `dcauchy(x)`.

1. [**2**] Parmi les propositions suivantes, quels couples (g, M) sont valides ? (On pourra utiliser la fonction `optimise`.)

1. $U_{[-10,10]}$ et $M = 24$
2. $N(2.4, 6^2)$ et $M = 18$
3. $Cauchy(0, 1)$ et $M = 11$
4. $\mathcal{E}xp(3)$ et $M = 345$
5. $N(2.4, 6^2)$ et $M = 12$
6. $Student_8$ et $M = 45$

2. [**1**] Parmi les propositions suivantes, quel couple (g, M) est le plus efficace en termes de nombres de simulations ?

1. $U_{[-10,10]}$ et $M = 24$
2. $N(2.4, 6^2)$ et $M = 18$
3. $Cauchy(0, 1)$ et $M = 11$
4. $\mathcal{E}xp(3)$ et $M = 345$
5. $N(2.4, 6^2)$ et $M = 12$
6. $Student_8$ et $M = 45$

On choisit finalement d'utiliser pour g une loi $Cauchy(0, 1)$ et $M = 17$.

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

1. 0.85
2. 0.058
3. 0.147
4. 0.74
5. 0.34

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 maintenant à une v.a. X de densité

$$f(x) = \sin(x)\mathbf{I}_{[0,\pi/2]}(x)$$

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

1. $X = \sin(U)$
2. $X = \arcsin(U)$
3. $X = \cos(U)$
4. $X = \arccos(1 - U)$
5. $X = \cos(U) + \pi/2$

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 calculant $\widehat{F_n}(t)$ et donner une estimation de $F(t)$ au point $t = 1$ en fonction de cet échantillon.

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

Question ocre [7 points]
Nous cherchons à évaluer l'intégrale suivante :

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

par des méthodes de Monte Carlo

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 reposant sur la génération d'un n-échantillon de variables aléatoires de loi de Cauchy. Fournir un intervalle de confiance à 95% pour I pour $n = 1000$.
3. **[1.5]** Proposer une méthode de Monte Carlo reposant sur la génération d'un n-échantillon de variables aléatoires de loi uniforme. Fournir un intervalle de confiance à 95% pour I pour $n = 1000$. Parmi ces 2 méthodes, laquelle est la meilleure ?

Nous cherchons à améliorer la méthode précédente. Pour cela, nous pouvons montrer que I peut être écrite de la façon suivante :

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

Par conséquent :

$$I = \frac{1}{2}\int_0^2 \left[\frac{1}{x^2+1} e^{-x} + \frac{1}{(2-x)^2+1} e^{-(2-x)}\right] dx$$

4. **[1.5]** En déduire une nouvelle méthode d'estimation de I reposant sur des variables aléatoires uniformes. Fournir un intervalle de confiance à 95% pour I pour $n = 1000$.
5. **[1]** Fournir les codes pour illustrer sur un même graphique, la convergence des 4 méthodes précédentes en fonction de la taille de l'échantillon simulé.

Question sépia [6 points]
On considère le jeu de données `fdeaths` (déjà résident).

1. **[.5]** Reproduire ci-dessous la description des données `fdeaths`.

On cherche à tester l'adéquation de ces données à une loi exponentielle, $\mathcal{E}(\lambda)$, de densité $\lambda \exp(-\lambda x)$ sur $\mathbb{R}_+$. On travaille avec une erreur de Type I égale à 0.01.

2. **[1]** Interpréter le résultat d'un test de Kolmogorov-Smirnov sur l'adéquation de ces données à une loi exponentielle de paramètre $\lambda = 1/530$: les données sont

1. compatibles avec une loi $\mathcal{E}(1/530)$;
2. incompatibles avec une loi $\mathcal{E}(1/530)$.

et indiquer après avoir tracé l'histogramme des données une raison simple pour ce résultat.

On définit l'échantillon `drift` par `drift=fdeaths-min(fdeaths)`.

3. [**1**] Interpréter le résultat d'un test de Kolmogorov-Smirnov sur l'adéquation de `drift` à une loi exponentielle de paramètre `lambda=1/mean(drift)`.

Le fait de choisir λ en fonction des données invalide la p-value fournie par R. On construit ci-dessous une p-value non-biaisée par bootstrap.

4. [**1.5**] Si D est la valeur de la statistique retournée par le test de Kolmogorov-Smirnov sous R, évaluer par un bootstrap paramétrique la probabilité de dépasser le D observé pour `drift` sous l'hypothèse nulle.

On rappelle que l'erreur de Type II est la probabilité d'accepter à tort l'adéquation à une loi exponentielle. On construit ci-dessous une évaluation bootstrap de l'erreur de Type II.

5.[**2**] Soit $\overline{D}$ le quantile de D à 0.99 obtenu par le bootstrap paramétrique de la question **3.** ci-dessus. Calculer $\overline{D}$ et évaluer par un bootstrap non-paramétrique fondé sur `drift` la probabilité que D soit en dessous de $\overline{D}$, ce qui donne l'erreur de Type II.

Question pourpre [6 points]

On travaille sur des données extraites de la base `faithful` de la manière suivante

`erup=faithful[181:270,1]`.

On suppose d'abord que la loi de distribution f qui a généré les données appartient à la famille paramétrique de lois $\{f_{4,\lambda}\}_\lambda$ où $f_{4,\lambda}$ est la densité d'une loi Gamma$(4,\lambda)$. 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 une moyenne empirique.
(a) 0.578
(b) 0.342
(c) 1.730
(d) 1.138
(e) 1.326

2. [**1**] En déduire l'estimateur $\widehat{f}$ de f associé dans la famille paramétrique (écrire une fonction `f_hat`). Représenter graphiquement cette fonction. Donner les commandes R relatives à ces deux questions.

3. [**1**] Le mode d'une distribution de densité f est le point m où $f(x)$ est maximale. Quelle est l'expression du mode de $f_{4,\lambda}$ (on pourra s'aider d'une représentation graphique).
(a) λ
(b) $\frac{3}{\lambda}$

(c) $\frac{2}{\lambda}$
(d) $\frac{1}{\lambda}$
(e) $\frac{\lambda}{2}$

4. **[.5]** Calculer un estimateur $\widehat{m}$ du mode de f
(a) 0.754
(b) 1.730
(c) 2.636
(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 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) $[2.47; 2.80]$
(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 accès à un vecteur d'antécédents et d'images par $\widehat{f}_{NP}$ avec les commandes `d$x` et `d$y`.

7. **[1]** 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) 4.43
(b) 4.32
(c) 4.30
(d) 4.34
(e) 4.27