Université Paris-Dauphine Année 2011-2012
Département de Mathématique

Examen NOISE, sujet A

Résoudre trois et uniquement trois exercices au choix.

Exercice 1

On considère la variable aléatoire X de densité

$$f(x) = C\,x\,e^{-2x^2/3},\ x \in \mathbb{R}_+ . \tag{1}$$

1. Écrire une fonction `fAR1()` qui permette de simuler des réalisations de X par acceptation-rejet et qui retourne un n-échantillon $\{X_i\}_{i=1,\dots,n}$ et le taux d'acceptation. (On pourra par exemple majorer $x\exp\{-x^2/3\}$ sur $\mathbb{R}_+$ et utiliser une loi normale appropriée et restreinte à $\mathbb{R}_+$. Ou bien utiliser une loi exponentielle.)
2. À l'aide de `fAR1()` generer $n = 10^4$ réalisations de $X \sim f$. Calculer la constante de normalisation C et donner un intervalle de confiance sur C au niveau 95%.
3. Calculer la fonction de répartition F de X et écrire une deuxieme fonction `fAR2()` qui permette de simuler des réalisations de $X \sim F$ par inversion générique.
4. En utilisant la commande `par(mfrow=c(1,2))`, tracer les histogrammes des réalisations obtenues par `fAR1()` et `fAR2()` et en y superposant à chaque fois la densité théorique f.
5. Quelle comparaison entre `fAR1()` et `fAR2()` fait-elle sens ? Choisir l'algorithme que vous pensez être “le meilleur”.
6. Utilisant cet algorithme, donner une estimation de Monte-Carlo de $\mathbb{E}[X]$, $\mathbb{E}[X^2]$ et $P(X > 2)$, ainsi que les intervalles de confiance correspondant au niveau 95%.

Exercice 2

On considère le jeu de données AirPassengers qui fournit le nombre de passagers sur des lignes internationales (en milliers) par mois entre 1949 et 1960.

1. Afficher les données
   ```
   > data(AirPassengers)
   > AirPassengers
   ```
   Représenter graphiquement les données. Donnez une explication rationnelle aux pics de voyageurs chaque année.
2. On s'intéresse au nombre de voyageurs par an :
   ```
   > M = matrix(AirPassengers,ncol=12,nrow=12,byrow=T)
   > Y = apply(M,1,sum)
   > t = seq(1949,1960,by=1)
   ```
3. On cherche à modéliser le nombre de passagers Y comme une fonction linéaire du temps :
   $$Y_i = a + bt_i + E_i$$
   où les E_i sont des variables aléatoires indépendantes, centrées de variance σ^2.

(a) Dans une nouvelle figure, tracer le nuage de points $(t_i, Y_i)_{i=1\ldots n}$

(b) Posons

$$\overline{Y_n} = \frac{\sum_{i=1}^n Y_i}{n} \quad , \quad \overline{t_n} = \frac{\sum_{i=1}^n t_i}{n} \quad , \quad S_{Yt} = \frac{1}{n}\sum_{i=1}^n Y_i t_i - \overline{Y_n}\,\overline{t_n} \quad , \quad S_t^2 = \frac{1}{n}\sum_{i=1}^n t_i^2 - \overline{t_n}^2$$

On sait que les estimateurs des moindres carrés de a et b sont :

$$\widehat{b} = \frac{S_{Yt}}{S_t^2} \quad , \quad \widehat{a} = \overline{Y_n} - \widehat{b}\,\overline{t_n}$$

Ajouter sur le graphe la droite d'équation $y = \widehat{a} + \widehat{b}t$.

(c) Estimer par procédure bootstrap le biais de l'estimateur $\widehat{a}$.

(d) Donner un intervalle de confiance bootstrap à 95% pour $\widehat{b}$.

(e) Donner une estimation du nombre de passagers en 1962.

Exercice 3

On considère la distribution de Weibull généralisée de paramètres $(k,\ \lambda,\ \theta)$ de densité

$$g(x; k, \lambda, \theta) = \frac{k}{\lambda}\left(\frac{x-\theta}{\lambda}\right)^{k-1} e^{-\left(\frac{x-\theta}{\lambda}\right)^k}\mathbb{I}_{x>\theta}\,,$$

On admet le résultat que le mode de g est $m = \theta + \lambda\left(\frac{k-1}{k}\right)^{1/k}$. Dans la suite on considèrera $\theta = 1$, $k = 3$ et $\lambda = 7$.

1. Proposer un algorithme d'acceptation-rejet pour la simulation de $X \sim g$ reposant sur un échantillon de loi $\mathcal{N}(m, s^2)$ où m est le mode de g et s^2 est un paramètre choisi de sorte à maximiser le taux d'acceptation (on pourra le déterminer numériquement). Écrire la fonction `fAR(...)` ayant comme paramètres de sortie le vecteur des n réalisations ainsi que le taux d'acceptation.
2. Proposer une deuxième méthode de simulation reposant sur le principe d'inversion générique à partir de réalisations d'une loi $\mathcal{U}(0, 1)$. Écrire la fonction `fIG(...)` ayant comme paramètres de sortie le vecteur des n réalisations.
3. Simuler deux n-échantillons `xAR` et `xIG` à l'aide des deux differentes méthodes, et, en utilisant la commande `par(mfrow=c(1,2))`, tracer les histogrammes des réalisations obtenues par `fAR()` et `fIG()` et y superposer la densité théorique f. [*Utiliser* `par(mfrow=c(1,1))` *pour revenir à la situation d'un graphique par fenêtre.*]
4. Donner le code `R` qui permet de tracer sur un même graphique les fonctions de répartition empirique et théorique.
5. Choisir l'algorithme de simulation paraissant le plus avantageux et donner une estimation Monte-Carlo de $\mathbb{E}[X]$, $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

Exercice 4

En probabilité, le *kurtosis* mesure l'aplatissement de la densité de probabilité d'une variable aléatoire définie sur les nombre réels. Le kurtosis se note β_1 et est calculé par la formule suivante :

$$\beta_1 = E\left[\left(\frac{X - E(X)}{\sigma}\right)^4\right]$$

où $\mu = E[X]$ et σ est l'écart type de X.
Soit $X_1, \ldots, X_n$ un n-échantillon. On propose l'estimateur des moments suivant pour β_1 :

$$B_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X_n}}{S_n}\right)^4$$

où $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ est un estimateur non biaisé de σ^2

1. On considère le cas où $X \sim \mathcal{U}_{[-1,1]}$
   (a) Calculer sur feuille la vraie valeur de β_1
   (b) Proposer une méthode de Monte Carlo permettant d'estimer le biais de l'estimateur B_1 de β_1 pour un échantillon de taille $n = 10$, $n = 100$ puis $n = 1000$
2. Calculer sur votre feuille β_1 pour une variable aléatoire gaussienne.
   On rappelle que si $X \sim \mathcal{N}(0,1)$, $E[X] = 0$, $E[X^2] = 1$, $E[X^3] = 0$, $E[X^4] = 3$
3. On considère les données faithfull :

```
> data(faithful)
> X = faithful$eruptions
```

   (a) Tracer l'histogramme des données
   (b) Estimer β_1 par l'estimateur qui vous semble le meilleur. Comparer à ce que vous auriez obtenu si les données avaient été gaussiennes.
   (c) Par bootstrap, fournir un intervalle confiance à 95% pour β_1. 3 appartient-il à l'intervalle de confiance ? Que pouvez-vous en conclure sur la distribution des données faithful$eruptions ?
   (d) Refaire la même chose (intervalle de confiance et conclusion) en prenant pour données les éruptions de durée supérieure à 3 min.

```
> X = faithful$eruptions
> Y  = X[X>3]
```

Exercice 5
On considère le jeu de données `faithful` disponible sous R et plus particulièrement

```
> x=faithful[,1]
```

a. Donner la valeur de la médiane empirique $\hat{\theta}(x)$ de l'échantillon, estimateur de la médiane θ de la distribution théorique correspondant aux données `x`.

b. On cherche à présent à déterminer si cet estimateur $\hat{\theta}$ est biaisé. Construire un échantillon bootstrap `bootmed` de taille 500 de médianes empiriques $\hat{\theta}(x^\star)$ et en déduire un intervalle de confiance à 95% du biais $\hat{\theta}(x^\star) - \hat{\theta}(x)$, évaluation bootstrap de l'intervalle de confiance sur le biais $\hat{\theta}(x) - \theta$.

c. A partir de l'échantillon bootstrap de biais `bootmed` ci-dessous, construire un second intervalle de confiance à 95% fondé sur l'hypothèse de normalité de la distribution de ce biais. Les deux intervalles conduisent-ils à la même conclusion sur l'existence d'un biais, i.e. sur la présence ou non de zéro dans cet intervalle ?

d. On cherche maintenant à tester la normalité de l'échantillon bootstrap de biais `bootmed`, c'est à dire à savoir si la loi de `bootmed` est une loi normale $\mathcal{N}(\mu, \sigma^2)$. Utiliser la fonction non-paramétrique `ks.test` pour réaliser ce test. (En cas de message d'erreur impliquant

```
Warning message:
  cannot compute correct p-values with ties
```

on pourra remplacer l'échantillon bootstrap `bootmed` par sa version randomisée `jitter(bootmed)`.)

e. Reprendre la question d en testant une normalité avec moyenne μ nulle.

Exercice 6

Lorsque l'on appelle la fonction de test `ks.test` pour évaluer l'adéquation d'un échantillon `x` à une famille de lois, par exemple $\mathcal{N}(\mu, \sigma^2)$, la solution fournie par

```
> ks.test(x,"pnorm",mean=mean(x),sd=sd(x))

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1219, p-value = 0.01283
alternative hypothesis: two-sided
```

ne donne pas la bonne `p-value`. Cet exercice le démontre par simulation.

a. Créer 1000 échantillons x de taille 50 et de loi $\mathcal{N}(0,1)$ et sauvegarder les valeurs correspondantes de `p-value` lors de l'exécution de

```
> ks.test(x,"pnorm")
```

b. Tracer l'histogramme des p-values ainsi obtenues et tester par `ks.test` l'adéquation de cet échantillon de p-values à une loi uniforme $\mathcal{U}(0,1)$ (codée en R par `"punif"`). (L'adéquation devrait être acceptée car la loi théorique de la p-value est bien $\mathcal{U}(0,1)$ dans ce cas.)

c. Reprendre l'expérience de la question a. en exécutant à présent le test d'adéquation de ces échantillons x de taille 50 et de loi $\mathcal{N}(0,1)$ à la famille des lois normales

```
> x=rnorm(50)
> ks.test(x,"pnorm",mean=mean(x),sd=sd(x))
```

et en sauvegardant les valeurs correspondantes de `p-value`.

d. Tester par `ks.test` si, à nouveau, l'adéquation de cet échantillon de p-values à une loi uniforme $\mathcal{U}(0,1)$ est acceptée. Conclure.