Examen final du 14 janvier 2011

Préliminaires

Cet examen est à réaliser sur ordinateur en utilisant le langage R et à rendre simultanément sur papier pour les réponses détaillées et sur fichier informatique pour les fonctions R utilisées. Les fichiers informatiques seront à sauvegarder suivant la procédure ci-dessous et seront pris en compte pour la note finale. Toute duplication de fichiers R fera l'objet d'une poursuite disciplinaire. L'absence de document enregistré donnera lieu à une note nulle sans possibilité de contestation.
Pour cet examen, vous devez remettre vos fichiers en ligne sur Intercours, suivant les étapes:

1. **Enregistrez d'abord vos fichiers sur l'ordinateur, sans utiliser d'accents ni d'espace, ni de caractères spéciaux.**
2. **Connectez-vous à Intercours http://intercours.dauphine.fr (ou http://www.ent.dauphine.fr et onglet "cours en ligne" - un clic sur l'image Intercours) Utilisez les identifiants de l'ENT (ceux de votre mail Dauphine)**
3. **Cliquez sur le cours intitulé "Examen (Christian Robert)" (dans la liste des cours à gauche)**
4. **Cliquez sur "Examen" au centre de la page**
5. **Vous allez maintenant soumettre vos fichiers. Pour cela, cliquez sur "Ajouter des pièces jointes" et sélectionnez votre premier fichier. Votre fichier appairait maintenant comme une pièce jointe en dessous du cadre "soumission". Si vous avez plusieurs fichiers à remettre, cliquez de nouveau sur "Ajouter des pièces jointes" pour sélectionner les suivants.**
6. **Une fois que vous aurez soumis vos fichiers, il ne sera plus possible de recommencer la procédure ou de modifier vos fichiers. Vérifiez que vos fichiers apparaissent bien comme des pièces jointes sous le cadre "soumission". Cliquez sur le bouton SOUMETTRE et OK. Un message de confirmation apparaît vous indiquant l'heure de la soumission.**

Aucun document informatique n'est autorisé, seuls les documents papier du cours et les livres de R le sont. L'utilisation de tout service de messagerie ou de mail est interdite et, en cas d'utilisation avérée, se verra sanctionnée.

Les problèmes sont indépendants, peuvent être traités dans n'importe quel ordre et le barème est donné à titre indicatif. L'examen sera noté sur 20.

1 Estimation de constante [3 points]

Soit une densité sur $\mathbb{R}$

$$f(x) \propto \frac{3+\sin^2(x)}{(1+x^2)^2}\,.$$

1. A l'aide d'une méthode d'acceptation rejet, simuler une variable aléatoire de densité f.
2. Déduire de la simulation une approximation de la constante

$$\int_{\mathbf{R}} \frac{3+\sin^2(x)}{(1+x^2)^2}\mathrm{d}x\,.$$

3. Illustrer par un histogramme la pertinence de la simulation en superposant f à l'histogramme.

2 Programmation R [4 points]

Vous devez résoudre par un programme R le problème suivant. Soit une loterie associée à N tickets numérotés de 1 à N. On suppose tous les tickets vendus. Les tickets gagnants sont ceux comportant un 1 et un 3 à droite du 1, comme par exemple 123 et 8135. Trouver l'unique valeur de $999 < N < 9999$ telle que la proportion de billets gagnants soit exactement 10%. *(La bonne réponse sans programme* R *valide ne sera pas acceptée.)*

3 Estimation de densité [3 points]

1. Générer $(X_1, \ldots, X_{100})$ un échantillon de variables aléatoires iid de loi exponentielle $\mathcal{E}xp(1)$.
2. Calculer analytiquement $F(3)$ où F représente la fonction de répartition de la loi exponentielle $\mathcal{E}xp(1)$. Donner un estimateur $\hat{F_n}(3)$ de $F(3)$ fondé sur les X_i et illustrer la convergence de $\hat{F_n}(3)$ vers $F(3)$.
3. On considère à présent une estimation par la méthode du noyau de la densité f de la loi de l'échantillon au point $x = 3$, $\hat{f}(3)$. Donner la valeur obtenue pour cette estimation ainsi qu'un intervalle de confiance bootstrap au niveau 0.95 obtenu à partir de l'échantillon $(X_1, \ldots, X_{100})$ (on pourra utiliser $10,000$ répliques bootstrap).

4 Aire des cercles [5 points]

Soit X une variable aléatoire uniformément distribuée sur $[0,1]$, $X \sim \mathcal{U}(0,1)$.

1. Soit A l'aire du disque de rayon X centré en 0. Déterminer la fonction de répartition de A. Calculer sa densité. Donner la ligne de code R permettant de simuler une réalisation de la variable A.
2. Calculer l'espérance et la variance de A soit par le calcul, soit par une expérience de Monte-Carlo (auquel cas il faudra fournir la précision de l'évalaution).
3. Soient $A_1, \ldots, A_{100}$ des variables i.i.d. de même loi que A. On pose

$$\bar{A} = \frac{1}{100}\sum_{i=1}^{100} A_i$$

En utilisant le théorème de la limite centrale, donner une approximation de la probabilité

$$\mathbb{P}\left(\bar{A} \leq \frac{\pi}{3} + 1.64\frac{\pi\sqrt{5}}{75}\right) \tag{4.1}$$

4. Approcher la probabilité (4.1) par la méthode du bootstrap avec une précision de 2 décimales.

5 Loi de Gumbel [5 points]

La fonction de répartition de la loi de Gumbel $Gu(\mu, \beta)$, est donnée sur $\mathbb{R}$ par

$$F(t; \mu, \beta) = \exp\{-\exp((\mu - t)/\beta)\} .$$

1. Écrire une fonction R permettant de générer par la méthode d'inversion un échantillon de taille n de la loi $Gu(\mu, \beta)$, pour des valeurs données des paramètres μ et β.
2. On fixe $\mu = 3$ et $\beta = 4$. Donner le code R permettant de tracer simultanément
   - La fonction de répartition F de $Gu(\mu, \beta)$;
   - les fonctions de répartition empiriques $\widehat{F}_n$ d'échantillons $(X_1, \ldots, X_n)$ simulés suivant la loi $Gu(\mu, \beta)$, pour $n = 10, 100, 1\,000$.
3. Pour $t \in \mathbb{R}$ fixé, donner l'expression de l'intervalle de confiance au niveau α sur $F(t; \mu, \beta)$, basé sur le théorème de la limite centrale. On notera dans la suite cet intervalle: $IC_\alpha(t) = [q_\alpha^{(1)}(t), q_\alpha^{(2)}(t)]$.
4. Montrer que, pour tout t, la longueur p de cet intervalle est majorée par:
   $$p \leq q_{1-\alpha/2}/\sqrt{n},$$
   où q_β est le quantile d'ordre β de la loi normale $\mathcal{N}(0, 1)$.
5. On fixe $\alpha = 0.05$. Déterminer une taille $n^\star$ d'échantillon pour laquelle la longueur de l'intervalle de confiance au niveau α de $F(t)$ est inférieure à 10^{-1} quel que soit t.

6 Troncation [7 points]

On dénote par Φ la fonction de répartition de la loi normale centrée réduite, soit `pnorm` en R.

1. En utilisant la commande R `data`, créez un vecteur $\mathbf{X_M}$ qui contient $M = 100$ mesures de débit du Nil disponible sous R comme `Nile`.
2. Donner un histogramme de $\mathbf{X_M}$ et sur le même graphe superposer la densité de la loi normale estimée à partir de ce vecteur, en précisant les choix de vos estimateurs de la moyenne et de la variance. Cette loi normale vous semble-t-elle pertinente pour représenter les variations de débit ?
3. Soit le vecteur $\mathbf{Y_n}$ composé des $n = 15$ premières valeurs de $\mathbf{X_M}$,
   $$\mathbf{Y_n} = (y_1, \ldots, y_{15}) .$$
   On considère la loi sur θ définie par la densité
   $$\pi_\mu(\theta|\mathbf{Y_n}) \quad \propto \quad f_\mu(\theta|\mathbf{Y}_n) \;=\; \exp\left\{-\frac{n}{300}\left[\theta - \frac{1}{n+1}\left(\mu + \sum_{k=1}^{n} y_k\right)\right]^2\right\}\{1 - \Phi(\{1000 - \theta\}/150)\}$$
4. Proposer une méthode d'acceptation-rejet de $\pi_\mu(\theta|\mathbf{Y_n})$ fondée sur une densité instrumentale $\rho_1(\theta)$ de loi normale dont vous préciserez les paramètres acceptables.
5. Écrire la fonction R rposterior correspondante qui
   - simule un échantillon iid de taille T selon $\pi_\mu(\theta|\mathbf{Y_n})$;
   - retourne un estimateur de Monte Carlo du taux d'acceptation et un intervalle de confiance à 90% sur ce taux;
   - retourne une estimation de la constante d'intégration de cette densité.
6. En utilisant rposterior pour $\mu = 800$, produire 1000 tirages suivant π et représenter l'histogramme de ces tirages avec la véritable densité $\pi_\mu(\theta|\mathbf{Y_n})$ en superposition.
7. Estimer le taux d'acceptation (en indiquant la précision) lorsque $\mu = \{700, 800, 900, 1000\}$ et en déduire les constantes de normalisation des densités correspondantes.

7 Mélanome [6 points]

On définit le coefficient de correlation entre deux variables aléatoires X et Y par

$$\rho_{X,Y} \;=\; \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \;=\; \frac{\mathbb{E}[\,(X-\mathbb{E}[X])\,(Y-\mathbb{E}[Y])\,]}{\sigma_X \sigma_Y}\,,$$

où σ_X et σ_Y sont les écart-types des deux variables aléatoires.

1. Un estimateur de $\rho_{X,Y}$ fondé sur un échantillon $(X_1,Y_1),,...,(X_n,Y_n)$ est le coefficient empirique

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^n (X_i-\bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i-\bar{Y})^2}}$$

   En répétant 500 fois la simulation de deux échantillons normaux $\mathcal{N}(0,1)$ indépendants de taille 100, $X_1,...,X_{100}$ et $Y_1,,...,Y_{100}$, tracer l'histogramme des $\hat{\rho}$ ainsi simulés et superposer à l'histogramme la courbe de la densité normale la mieux adaptée.

2. On considère la série `melanoma` du taux d'incidence du mélanome dans le Connecticut (US) sur la période $1936-1972$. Accédez à la série sous R par la commande

```
data(melanoma, package="lattice")
```

   On applique un modéle de régression linéaire à ces données, reliant le taux d'incidence Y (`y=melanoma[,2]`) au temps X (`x=melanoma[,1]`) par

$$Y_i \;=\; \alpha \;+\; \beta\, X_i \;+\; \epsilon_i, \qquad i=1,\ldots,n\,, \tag{7.1}$$

   et son estimation par la *méthode des moindres carrées*.

   À partir des donnees `melanoma`, donner les estimations $\hat{\alpha}$ et $\hat{\beta}$ de l'intercept α et de la pente β de la droite (7.1). Tracer dans un même graphique les données comme un nuage de points par `plot` et la droite

$$\hat{Y} \;=\; \hat{\alpha} \;+\; \hat{\beta} X\,,$$

   résultant de cette estimation.

3. Par rééchantillonage (bootstrap) des résidus estimés $\hat{\epsilon}_i$, donner des estimations (fondées sur des échantillons bootstrap) des densités de $\hat{\alpha}_{BS2}$, de $\hat{\beta}_{BS2}$ et de $\hat{\rho}_{BS2}$, estimateurs des paramètres α, β et $\rho_{X,Y}$, ainsi que des intervalles de confiances de niveau 95% pour les trois estimateurs.

8 Loi normale bi-dimensionnelle [9 points]

Soit la matrice définie positive symétrique 2×2

$$\Sigma = \begin{pmatrix} 1 & 2 \\ 2 & 6 \end{pmatrix}\,,$$

que l'on notera `Sig` dans le code R.

1. Laquelle des deux commandes `chol(Sig)%*%t(chol(Sig))` ou `t(chol(Sig))%*%(chol(Sig))` redonne `Sig`?

2. Considérons la loi normale sur $\mathbb{R}^2$ de moyenne 0 et de matrice de variance covariance Σ, $\mathcal{N}_2(0,\Sigma)$. Utilisez `rnorm(2)` et `chol(Sig)` pour donner la ligne de code R permettant de simuler un vecteur de loi $\mathcal{N}_2(0,\Sigma)$. Vous vérifierez par une expérience de Monte-Carlo que votre code est correct en trouvant des approximations convergentes des variances (1 et 6) et de la covariance (2) de votre vecteur normal ainsi simulé.

3. On considère à présent la loi sur $\mathbb{R}^2$ de densité

$$\pi(x) \propto \exp\left(-x^{\mathrm{T}}\Sigma^{-1}x/2\right) \Big/ \sqrt{x^{\mathrm{T}}\Sigma^{-1}x}\,.$$

(a) Montrez que la distribution est bien définie quand Σ est la matrice identité, $\Sigma = I_2$, c'est-à-dire que $\int_{\mathbb{R}^2} \pi(x)\mathrm{d}x < \infty$ en utilisant un changement de variables en polaires.

(b) Expliquer pourquoi l'utilisation de l'échantillonnage d'importance fondé sur la distribution instrumentale normale $\mathcal{N}_2(0, \Sigma)$ pour approcher

$$\mathbb{P}\left[X^{\mathrm{T}}\Sigma^{-1}X < 1\right] \tag{8.1}$$

n'est satisfaisante ni d'un point de vue théorique ni en pratique. Illustrer les difficultés pratiques de cette approximation de (8.1) par échantillonnage d'importance par une expérience de Monte-Carlo traçant l'évolution de l'estimateur d'importance.

(c) Dans l'expérience de Monte-Carlo ci-dessus, donnez un intervalle de confiance sur votre estimation finale de (8.1) fondé sur (a) le théorème de la limite centrale et (b) une évaluation bootstrap.

(d) On considère à présent l'échantillonnage d'importance alternatif fondé sur la représentation polaire dans $\mathbb{R}^2$, par une distribution gamma $\mathcal{G}a(1, 1/10)$ sur le carré du rayon $\eta = X^{\mathrm{T}}X$ et par une distribution uniforme sur l'angle. (*Note:* 1 est la valeur du paramètre `shape` et 1/10 est la valeur du paramètre `rate` dans les fonctions `dgamma` et `rgamma`.) De même que dans les questions précédentes, vous mettrez en œuvre une expérience de Monte-Carlo tracant la convergence de l'estimateur d'importance de (8.1) et donnant un intervalle de confiance sur votre estimation finale de (8.1) fondé sur une évaluation bootstrap.

(e) En utilisant la même expérience de Monte Carlo, donner une évaluation de la constante de proportionalité de la densité $\pi(x)$ avec sa précision.