Examen final du 11 janvier 2010
Séance de 8 heures à 10h45

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 sera pénalisée par un zéro. 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.

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.

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

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

et fonction de répartition

$$G(x;k,\lambda,\theta) = \begin{cases} 1-\exp\left\{-\left(\frac{x-\theta}{\lambda}\right)^{k}\right\} & x>\theta \\ 0 & x\leq\theta \end{cases}.$$

Remarques. Si $X \sim W(k,\lambda,\theta)$ alors

$$E[X] = \lambda\Gamma(1+\frac{1}{k})+\theta \quad , \quad Var[X] = \lambda^2\Gamma(1+\frac{2}{k}) - \left(\lambda\Gamma(1+\frac{1}{k})\right)^2 \quad , \quad q_2 = \lambda(\ln 2)^{1/k}+\theta\,,$$

où Γ est la fonction Gamma (`gamma` dans `R`) et q_2 est la médiane. Attention, la fonction logarithme népérien (ln) est donnée par `log` dans `R`.

1 Simulation de réalisations de la loi de Weibull

1. **Méthode 1: à partir de réalisations d'une loi $\mathcal{U}(0,1)$.**
   (i) Proposer une méthode de simulation de cette loi reposant sur le *principe d'inversion générique.*
   (ii) Ecrire une fonction `rweibull1` ayant pour argument d'entrée n le nombre de réalisations et les paramètres (k, λ, θ) et fournissant en sortie le vecteur des n réalisations.
   (iii) Simuler un 1000-échantillon $X^{(1)}$ avec cette fonction. Montrer la pertinence de la fonction `rweibull1` en traçant l'histogramme de $X^{(1)}$ ainsi que la vraie densité.

2. **Méthode 2: à partir de réalisations d'une loi $\mathcal{N}(0,1)$.** On remarque que si $X \sim W(k, \lambda, \theta)$ alors $Y = 2\left(\frac{X-\theta}{\lambda}\right)^k \sim \chi^2_2$.
   (i) En déduire une deuxième méthode de simulation de réalisations d'une loi de Weibull de paramètres (k, λ, θ) reposant sur des simulations de lois normales centrées réduites[1].
   (ii) Ecrire une fonction `rweibull2` ayant pour argument d'entrée n le nombre de réalisations et les paramètres (k, λ, θ) et fournissant en sortie le vecteur des n réalisations.
   (iii) Simuler un 1000-échantillon $X^{(2)}$ avec cette fonction. Montrer la pertinence de la fonction `rweibull2` en traçant l'histogramme de $X^{(2)}$ ainsi que la vraie densité.

3. **Comparaison des méthodes:** Démontrer analytiquement que les deux méthodes précédentes sont parfaitement équivalentes.

Le package `stat` comporte une fonction permettant de simuler des variables aléatoires suivant une loi de Weibull avec $\theta = 0$

Dans la suite on utilisera cette fonction.

```
> library(stats)
> X = theta + rweibull( ... à compléter par vos soins)
```

2 Utilisation des réalisations de la loi de Weibull

Soit f la densité de probabilité définie par

$$f(x) = C(x+2)(\cos(x+2))^2 \exp\{-(x+2)^2\}\mathbb{I}_{x>-2}\,,$$

où C est la constante de normalisation.

1. Proposer un ensemble de paramètres (k, λ, θ) et une constante M tels que $\forall x > -2$,
$$(x+2)(\cos(x+2))^2 e^{-(x+2)^2} \leq Mg(x; k, \lambda, \theta)\,.$$

2. En déduire une méthode de simulation reposant sur un n-échantillon de loi de Weibull. On notera `fAR` la fonction correspondante.

3. Illustrer graphiquement la pertinence de votre méthode de simulation.

3 Calcul d'une intégrale

On cherche à calculer la constante de normalisation C.

1. Proposer une méthode de Monte Carlo reposant sur l'utilisation d'un échantillon de taille $n = 10000$ de loi de Weibull, permettant de calculer C.

2. Illustrer la convergence de votre méthode.

3. Founir un intervalle de confiance à 95% pour C.

[1] On rappelle que, si $Z_1 \dots Z_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0,1)$, alors $X = \sum_{i=1}^n Z_i^2$ est distribuée selon une loi du χ^2 à n degrés de liberté.

4 Fonction de répartition empirique

On suppose que $\theta = 0$, $\lambda = 2$, $k = 3$.

1. Tracer sur un même graphe les fonctions de répartition issues d'échantillons de tailles respectives $n = 30$, $n = 100$, $n = 1000$, $n = 10000$.
2. Ajouter sur le graphe la fonction de répartition théorique.
3. Commentez votre graphique.
4. A partir de l'échantillon de taille $n = 10000$ estimer la probabilité :

$$p = P(X \leq q_2) = P(X \leq (\lambda \ln 2)^{1/k} + \theta)\,.$$

   Fournir un intervalle de confiance pour p et comparer à la vraie valeur.

5. A partir de l'échantillon de taille $n = 10000$, fournir une estimation du quantile à 63% i.e. q tel que $P(X < q) = 0.63$. Quel lien peut-on faire entre ce quantile et λ ?

5 Bootstrap

Télécharger le jeu de données `lynx` :

```
> data(lynx)
> x = lynx
```

On suppose que les observations sont les réalisations d'un échantillon $\mathbf{X_n} = (X_1 \ldots X_n)$ de taille $n = 114$ d'une variable aléatoire X. On suppose que les $X_i \sim W(k, \lambda, \theta)$ avec (k, λ, θ) inconnus.

Notons $Q_{0.5}(\mathbf{X_n})$ la médiane empirique de l'échantillon $\mathbf{X_n}$ et $Q_{0.63}(\mathbf{X_n})$ le quantile empirique de l'échantillon $\mathbf{X_n}$ à 63%. On s'intéresse aux trois statistiques d'échantillonnage suivantes:

$$T = \min_{i=1\ldots n}(\mathbf{X_n}) \quad , \quad L = Q_{0.63}(\mathbf{X_n} - T) \quad \text{et} \quad K = \frac{\ln(\ln 2)}{\ln((Q_{0.5}(\mathbf{X_n}) - T)/L)}\,.$$

T, L et K sont des estimateurs respectifs de θ, λ et k.

1. Donner un intervalle de confiance à 95% par bootstrap pour (k, λ, θ) (On réalisera $B = 1000$ tirages bootstrap).
2. Afficher l'histogramme des données et tracer sur l'histogramme la densité de Weibull avec les paramètres estimés. Penser à régler le nombre de classes de l'histogramme de façon à optimiser la représentation.
3. Afficher également une estimation non-paramétrique de la densité, en décrivant votre choix de noyau.
4. Que pensez-vous de l'hypothèse selon laquelle les données observées suivent une loi de Weibull ?

6 Compréhension d'un code R

Dans le code ci-dessous, on cherche à simuler (x, y) tel que $x \sim \mathcal{N}(0, 1)$ et $y|x < -2 \sim \mathcal{N}(-2, 1)$, $y|x > -2 \sim \mathcal{N}(2, 1)$. Identifier les erreurs de programmation et corriger ce code pour obtenir le résultat voulu.

```
simulator=function(n==1){

   x=y=NUL
   for (t in 1:n){

     z=rnorm(1)
     if (z<-2) w=rnorm(-2)
     else w=rnorm(2)
     x=c(c,z)
     y=c(y,w)
     }
   list(x,y)
   }
```

En déduire une approximation de la moyenne et de la variance de y.