# Examen final du 11 janvier 2010
# Séance de 14 heures à 16h45

### 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 Pareto de paramètres $(\alpha,\, \theta) \in \mathbb{R}^{*+} \times \mathbb{R}^{*+}$, densité

$$g(x;\alpha,\theta) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}\mathbb{I}_{x>\theta}\,,$$

et fonction de répartition

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

**Remarques.** Si $X \sim Par(\alpha,\theta)$ alors

$$E[X] = \frac{\alpha\theta}{\alpha-1} \quad , \quad Var[X] = \frac{\theta^2\alpha}{(\alpha-1)^2(\alpha-2)} \quad , \quad q_2 = \theta 2^{1/\alpha}$$

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

# 1 Simulation de réalisations de la loi de Pareto

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 `rpareto1` ayant pour argument d'entrée $n$ le nombre de réalisations et les paramètres $(\alpha, \theta)$ et fournissant en sortie le vecteur des $n$ réalisations.
   (iii) Simuler un 1000-échantillon avec cette fonction pour $\theta = 1$ et $\alpha = 5$. Tracer l'histogramme des réalisations ainsi que la vraie densité.

2. **Méthode 2: à partir de réalisations d'une loi exponentielle.** On remarque que si $Y \sim Par(\alpha, \theta)$ alors $X = \ln\left(\frac{Y}{\theta}\right) \sim \mathcal{E}(\alpha)$.
   (i) En déduire une deuxième méthode de simulation de réalisations d'une loi de Pareto de paramètres $(\alpha, \theta)$ reposant sur des simulations de lois exponentielles d'espérance $\frac{1}{\alpha}$ (donc de paramètres $\alpha$).
   (ii) Ecrire une fonction `rpareto2` ayant pour argument d'entrée $n$ le nombre de réalisations et les paramètres $(\alpha, \theta)$ et fournissant en sortie le vecteur des $n$ réalisations.
   (iii) Simuler un 1000-échantillon avec cette fonction pour $\theta = 1$ et $\alpha = 5$. Tracer l'histogramme des réalisations ainsi que la vraie densité.

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

Dans la suite on utilisera le code suivant pour simuler des réalisations i.i.d de loi $Par(\alpha, \theta)$.

```
> X = theta * exp(rexp(n, alpha))
```

# 2 Utilisation des réalisations de la loi de Pareto

Soit $f$ la densité de probabilité

$$f(x) = C \times (\cos(x+1))^2 \frac{1}{x^4} \mathbb{I}_{x>2}\,,$$

où $C$ est la constante de normalisation.

1. Proposer un ensemble de paramètres $(\alpha, \theta)$ et une constante $M$ tels que $\forall x > 2$,

$$(\cos(x+1))^2 \frac{1}{x^4} \leq M g(x; \alpha, \theta)\,.$$

2. En déduire une méthode de simulation reposant sur un $n$-échantillon de loi de Pareto. 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 $n = 20000$ de loi de Pareto, permettant de calculer $C$.

2. Illustrer la convergence de votre méthode.

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

# 4 Fonction de répartition empirique

On suppose que $\theta = 1$, $\alpha = 5$

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

$$p = P(X \leq q_2) = P(X \leq \theta 2^{\frac{1}{\alpha}}) .$$

   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 95% i.e. $q$ tel que $P(X \leq q) = 0.95$.

# 5 Bootstrap

Télécharger le jeu de données `lynx` :

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

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

On s'intéresse aux deux statistiques d'échantillonnage suivantes:

$$T = \min_{i=1\dots n} (X_1, \dots, X_n) \quad , \quad A = \frac{\overline{X_n}}{\overline{X_n} - T}$$

où $\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$. $T$ et $A$ sont des estimateurs respectifs de $\theta$ et $\alpha$.

1. Construire $B = 1000$ échantillons bootstrap et en déduire un intervalle de confiance à 95% pour $(\alpha, \theta)$.
2. Afficher l'histogramme des données et tracer sur l'histogramme la densité de Pareto 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 Pareto?

# 6 Compréhension du code R

Le code ci-dessous cherche à obtenir un intervalle de confiance bootstrap sur la médiane d'un échantillon de `5*n` points, lorsqu'on observe `n=55` valeurs d'un échantillon dénoté `obs`. Identifier les erreurs de programmation et corriger ce code pour obtenir une évaluation correcte des bornes de l'intervalle de confiance.

```
bootranj=function(V==10^4){

 inta=rep(0,lenf=V)
 for (T in 1:V)
   boot=sample(ord,55*n,rep=T)
   inta[T]=mediane(boot)
```

```
  }

int=c(quantile(inta,.025),quantile(inta,.975)
list(int[1],int[2]
}
```

En simulant `obs` par `obs=rnorm(n)`, en déduire l'intervalle de confiance. Répéter l'expérience avec `obs=rcauchy(n)`.