Examen du 9 février 2010

Durée : 2 heures. Tous les documents ainsi que les calculatrices sont autorisés.

Exercice 1 On a mesuré à 20 endroits différents la hauteur de la neige ainsi que les valeurs de certaines variables environnementales (altitude, rugosité, pente, orientation). On souhaite déterminer un modèle linéaire qui explique la hauteur de la neige en fonction de ces 4 variables :

$$\boxed{\texttt{H\_neige} = \alpha + \beta_1 \times \texttt{altitude} + \beta_2 \times \texttt{rugosité} + \beta_3 \times \texttt{pente} + \beta_4 \times \texttt{orientation} + \epsilon.}$$

La sortie de l'instruction

```
> lm(H_neige ~ altitude + rugosité + pente + orientation)
```

est donnée ci-dessus :

```
Call:       lm(formula = H_NEIGE ~ altitude + rugosite + pente + orient.)
Coefficients:           Estimate Std. Error t value Pr(>|t|)
            (Intercept) 80.87490   74.33521   1.088    0.294
            altitude    -0.01954    0.01673  -1.168    0.261
            rugosite     1.16557    1.95858   0.595    0.561
            pente       -9.93698   22.08235  -0.450    0.659
            orient.      0.06935    0.06504   1.066    0.303
Residual standard error: 28.67 on 15 degrees of freedom
Multiple R-squared: 0.6199,     Adjusted R-squared: 0.5186
```

1. Quelles sont les valeurs estimées des coefficients $(\alpha, \beta_1, \beta_2, \beta_3, \beta_4)$ et de la variance des résidus σ^2 ?
2. Donner la valeur du coefficient de détermination R^2.
3. Selon le modèle estimé, quelle serait la prédiction pour la hauteur de la neige lorsque `altitude`$=$ 3000, `rugosité`$=$ 200, `pente`$=$ 20 et `orientation`$=$ 100 ?
4. Si l'on effectue un test d'utilité sur chacune des 4 variables explicatives, laquelle des 4 est la variable la moins utile ? la deuxième variable la moins utile ? Est-il correct d'affirmer que ces deux variables sont inutiles ? Expliquez votre réponse.
5. On supprime la variable `pente` et re-estime les coefficients. On obtient

```
Call:  lm(formula = H_NEIGE ~ altitude + rugosite + orient.)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 87.38353   71.07410   1.229  0.23667
altitude    -0.02098    0.01600  -1.311  0.20838
rugosite     0.28493    0.07670   3.715  0.00188 **
orient.      0.07791    0.06063   1.285  0.21705
Residual standard error: 27.95 on 16 degrees of freedom
Multiple R-squared: 0.6148,     Adjusted R-squared: 0.5426
```

Ce nouveau modèle est-il moins bon que le modèle précédent ? Laquelle des 3 variables explicatives influence le plus la variable `H_NEIGE` ? Répondez à la question 3 en utilisant ce nouveau modèle.

Exercice 2 L'écrivain bien connu Ephrem Michon prétend être l'auteur d'un ouvrage attribué à un auteur inconnu. Les écrits déjà publiés d'Ephrem Michon avaient été analysés selon la longueur des phrases. Afin de déterminer si Ephrem Michon peut être l'auteur de l'ouvrage en question, on a analysé un échantillon de 2000 phrases tirées de cet ouvrage. On a alors le tableau suivant :

	Proportion de phrases	
Nombre de mots	Ephrem Michon	Auteur inconnu
≤ 3	0.020	0.015
4 – 5	0.020	0.025
6 – 8	0.041	0.042
9 – 12	0.102	0.084
13 – 16	0.263	0.250
17 – 20	0.279	0.203
21 – 24	0.118	0.148
25 – 27	0.105	0.136
28 – 29	0.032	0.077
≥ 30	0.020	0.020

On veut tester l'hypothèse que Ephrem Michon est l'auteur du nouvel ouvrage.

1. Laquelle des trois valeurs suivantes est nécessaire pour effectuer le test au niveau α : $q_\alpha(\chi^2_{1999})$, $q_{1-\alpha}(\chi^2_{1999})$, $q_{1-\alpha}(t_9)$ ou $q_{1-\alpha}(\chi^2_9)$?
2. Sachant que pour $\alpha = 1\%$, on a $q_\alpha(\chi^2_{1999}) = 21490$, $q_{1-\alpha}(\chi^2_{1999}) = 18549$ et $q_{1-\alpha}(\chi^2_9) = 21,67$, peut-on affirmer au seuil de 1% que Ephrem Michon n'est pas l'auteur du nouvel ouvrage ?

Exercice 3

Les coûts annuels de rénovation pour les appartements d'une société de gestion d'un parc immobilier sont distribués selon la loi :

$$p(\theta; x) = \frac{2x}{\theta(1+x^2)^{1+1/\theta}} \mathbb{1}_{\mathbf{R}_+}(x),$$

où θ est un paramètre réel strictement positif. Les données historiques sur les 20 dernières années du 20$^{\text{ème}}$ siècle ont permis à la société d'établir ce modèle avec un paramètre $\theta = 3$.
En 2010, la société décide de faire une étude sur les dix dernières années pour vérifier la pertinence du modèle. Sans remettre en cause la forme de la densité $p(\theta; x)$, la société souhaite vérifier si la valeur du paramètre θ a évolué. Pour cela, elle prélève les coûts annuels de rénovation de n appartements au cours des 10 dernières années et obtient ainsi une réalisation $x_1, \dots, x_n$ d'un échantillon $X_1, \dots, X_n \overset{\text{iid}}{\sim} p(\theta; x)$.

1. Vérifier que $p(\theta; x)$ est bien une densité de probabilité.
2. Montrer qu'une variable aléatoire X_1 qui a pour densité $p(\theta; x)$ admet une espérance finie si et seulement si $\theta < 2$. Quelles sont les valeurs de θ pour lesquelles X_1 est de carré intégrable ?
3. Ecrire la log-vraisemblance du modèle et déterminer l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ. Cet estimateur est-il sans biais ? Calculer son risque quadratique.
4. Démontrer que $\hat{\theta}_n$ est un estimateur fortement convergent de θ et qu'il est asymptotiquement normal de variance limite θ^2.
5. Construire un intervalle de confiance de niveau asymptotique 95% pour le paramètre θ. (Le quantile d'ordre 97,5% de la loi normale centrée réduite est égal à 1,96.)
6. Proposer un test de niveau asymptotique 5% de l'hypothèse nulle $H_0 : \theta = 3$ contre l'alternative $H_1 : \theta \neq 3$.
   Application numérique : sur un échantillon de taille $n = 100$, on a observé $x_1, \dots, x_n$ tel que

   $$\frac{1}{n} \sum_{i=1}^{n} \log(1 + x_i^2) = 3,5.$$

   Ces observations sont-elles suffisantes pour rejeter H_0 au seuil de 5%.

7. Vérifier que si X a pour densité $p(\theta; x)$, alors la variable aléatoire $Y = \theta^{-1} \log(1 + X^2)$ suit la loi exponentielle dont on déterminera le paramètre. En déduire que la loi de la variable aléatoire $\hat{\theta}_n / \theta$ est indépendante de θ.

8. On admet que la variable aléatoire $\zeta_n = 2n\hat{\theta}_n / \theta$ suit la loi χ^2_{2n}. A l'aide des quantiles $q_{\alpha/2}(\chi^2_{2n})$ et $q_{1-\alpha/2}(\chi^2_{2n})$ proposer un test de niveau (non asymptotique) α pour l'hypothèse $H_0 : \theta = 3$ contre $H_1 : \theta \neq 3$.