Final SY02 Automne 2009

Nom :	**Signature :**
Prénom :	

Répondre sur ce document, en ne reportant que les grandes lignes du raisonnement et les résultats (faire d'abord les calculs au brouillon). La qualité de la présentation sera prise en compte dans la notation. Aucune copie supplémentaire ne sera acceptée. Aucun document n'est autorisé, à l'exception du recueil de tables. Les calculettes sont autorisées à condition qu'elles ne contiennent aucune information relative au cours de sy02.

Exercice 1 (9 points)

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de variable parente X, de densité

$$f(x) = \frac{x}{\theta^2} \exp\left(-\frac{x}{\theta}\right) 1_{[0,+\infty[}(x),$$

θ étant un paramètre positif.

1. Montrer qu'il existe un estimateur efficace de θ. On le notera $\widehat{\theta}$.

Le support de la loi de X est $\mathbb{R}^+$ et ne dépend donc pas du paramètre θ ; les conditions de Cramer-Rao sont donc vérifiées.

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{x_i}{\theta^2} \exp\left(-\frac{x_i}{\theta}\right) = \frac{\prod_{i=1}^{n} x_i}{\theta^{2n}} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right) = \frac{\prod_{i=1}^{n} x_i}{\theta^{2n}} \exp\left(-\frac{n\overline{x}}{\theta}\right)$$

$$\log L(\theta; x_1, \ldots, x_n) = -2n \log\theta - \frac{n\overline{x}}{\theta} + cste$$

$$\frac{\partial \log L}{\partial \theta}(\theta; x_1, \ldots, x_n) = -\frac{2n}{\theta} + \frac{n\bar{x}}{\theta^2} = \frac{2n}{\theta^2}\left(\frac{\overline{x}}{2} - \theta\right)$$

On obtient donc la factorisation :

$$\frac{\partial \log L}{\partial \theta}(\theta; X_1, \ldots, X_n) = \frac{2n}{\theta^2}\left(\frac{\overline{X}}{2} - \theta\right)$$

$\widehat{\theta} = \frac{\overline{X}}{2}$ est donc l'estimateur efficace du paramètre θ.

2. En déduire les espérances et les variances de $\widehat{\theta}$ et de X.

$$\mathbb{E}(\widehat{\theta}) = \theta \qquad \text{(Un estimateur efficace est sans biais)}$$

$$\mathrm{Var}(\widehat{\theta}) = \frac{1}{\frac{2n}{\theta^2}} = \frac{\theta^2}{2n} \qquad \text{(corollaire de la factorisation)}$$

$$\mathbb{E}(X) = \mathbb{E}(\overline{X}) = 2\mathbb{E}(\widehat{\theta}) = 2\theta$$

$$\mathrm{Var}(X) = n\mathrm{Var}(\overline{X}) = 4n\mathrm{Var}(\widehat{\theta}) = 4n\frac{\theta^2}{2n} = 2\theta^2$$

3. Déterminer une fonction asymptotiquement pivotale pour θ que l'on exprimera en fonction de $\widehat{\theta}$.

Un estimateur efficace est asymptotiquement gaussien. En utilisant les résultats de la question précédente, on obtient donc

$$\widehat{\theta} \xrightarrow{L} \mathcal{N}\left(\theta, \frac{\theta^2}{2n}\right)$$

et la fonction asymptotiquement pivotale

$$\frac{\widehat{\theta} - \theta}{\theta/\sqrt{2n}} \xrightarrow{L} \mathcal{N}(0,1).$$

4. On considère le problème de test $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$ avec $\theta_1 > \theta_0$. Montrer que la région critique W du test le plus puissant pour ce problème au niveau α^* s'exprime en fonction de $\widehat{\theta}$, puis donner une approximation de W en supposant n grand.

Expression de W en fonction de $\widehat{\theta}$
Il s'agit d'un problème de comparaison de deux hypothèses simples. Le test le plus puissant est donné par le théorème de Neyman-Pearson.

$$\frac{L(\theta_1; x_1, \ldots, x_n)}{L(\theta_0; x_1, \ldots, x_n)} = \left(\frac{\theta_0}{\theta_1}\right)^{2n} \exp\left(n\overline{x}\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right)\right) = \left(\frac{\theta_0}{\theta_1}\right)^{2n} \exp\left(2n\widehat{\theta}\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right)\right)$$

Comme $\frac{1}{\theta_0} - \frac{1}{\theta_1} > 0$, ce rapport est une fonction croissante de $\widehat{\theta}$. La région critique du test de Neyman-Pearson, de la forme $W = \left\{\frac{L(\theta_1; x_1, \ldots, x_n)}{L(\theta_0; x_1, \ldots, x_n)} > k\right\}$, peut donc s'écrire

$$W = \{\widehat{\theta} > k'\}$$

pour une certaine constante k'.

Approximation de W pour n grand
On obtient une expression approchée pour k' en imposant que le niveau du test soit approximativement égal à une valeur α^*.

$$\mathbb{P}_{H_0}(\widehat{\theta} > k') = \alpha^* \Leftrightarrow \mathbb{P}_{H_0}\left(\frac{\widehat{\theta} - \theta_0}{\theta_0/\sqrt{2n}} > \frac{k' - \theta_0}{\theta_0/\sqrt{2n}}\right) = \alpha^*$$

D'après la question 2, $\frac{\widehat{\theta} - \theta_0}{\theta_0/\sqrt{2n}}$ suit approximativement pour H_0, pour n grand, une loi normale centrée-réduite. On en déduit

$$\frac{k' - \theta_0}{\theta_0/\sqrt{2n}} \approx u_{1-\alpha^*}$$

$$\Leftrightarrow k' \approx \theta_0\left(1 + \frac{u_{1-\alpha^*}}{\sqrt{2n}}\right).$$

5. On considère maintenant le problème de test suivant $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Existe-t-il un test UPP pour ce problème ?

Soit le problème de test $H_0 : \theta = \theta_0$ contre $h_1 : \theta = \theta_1$ avec $\theta_1 \neq \theta_0$.
On a vu que, si $\theta_1 > \theta_0$, la RC du test de NP est de la forme $W = \{\widehat{\theta} < k'\}$.
En revanche, si $\theta_1 < \theta_0$, le rapport des vraisemblances est fonction décroissante de $\widehat{\theta}$, et la RC du test de NP est donc de la forme $W = \{\widehat{\theta} > k''\}$.
La région critique du test de NP dépend donc de l'hypothèse simple h_1 considérée : par conséquent, il n'existe pas de test UPP.

6. Calculer la statistique du rapport de vraisemblance λ, exprimée en fonction de $\widehat{\theta}$, pour le problème de test de la question 5.

$$\begin{aligned}
\lambda &= \frac{sup_{\theta \in H_0} L(\theta; x_1, \ldots, x_n)}{sup_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n)} = \frac{L(\theta_0; x_1, \ldots, x_n)}{L(\widehat{\theta}; x_1, \ldots, x_n)} \\
&= \frac{\frac{1}{\theta_0^{2n}} \exp(-\frac{n\overline{x}}{\theta_0})}{\frac{1}{\widehat{\theta}^{2n}} \exp(-\frac{n\overline{x}}{\widehat{\theta}})} = \frac{\frac{1}{\theta_0^{2n}} \exp(-\frac{2n\widehat{\theta}}{\theta_0})}{\frac{1}{\widehat{\theta}^{2n}} \exp(-2n)} \\
&= \left(\frac{\widehat{\theta}}{\theta_0}\right)^{2n} \exp\left(2n\left(1 - \frac{\widehat{\theta}}{\theta_0}\right)\right)
\end{aligned}$$

7. En utilisant la statistique $\ln \lambda$ et en supposant que n est grand, proposer une région critique pour le test de la question 5. Quelle décision prendra-t-on si $\theta_0 = 2$, $n = 50$, $\sum_i x_i = 115$ et $\alpha^* = 0.05$.

Région critique
La fonction $-2 \ln \lambda$ suit asymptotiquement, sous H_0, une loi du χ^2_1. La région critique du test du RV est donc

$$W = \{-2 \ln \lambda \geq c\} \qquad \text{avec} \qquad c \simeq \chi^2_{1;1-\alpha^*}.$$

Application numérique
Région critique : $W = \{-2 \ln \lambda \geq \chi^2_{1;0.95} = 3.84\}$
Calcul de $-2 \ln \lambda$:

$$\begin{aligned}
-2 \ln \lambda &= -2\left(2n(\ln \widehat{\theta} - \ln \theta_0) + 2n\left(1 - \frac{\widehat{\theta}}{\theta_0}\right)\right) \\
&= -4n\left(\ln \widehat{\theta} - \ln \theta_0 + 1 - \frac{\widehat{\theta}}{\theta_0}\right) \\
&= 25.67
\end{aligned}$$

On rejette donc l'hypothèse H_0.

Exercice 2 (3 points)

On a examiné 220 giroflées, au point de vue de la morphologie de leurs fleurs et de leurs feuilles. Les 220 giroflées se répartissent suivant le tableau suivant.

	Fleurs simples	Fleurs doubles
Feuilles dentées	74	81
Feuilles non dentées	41	24

Peut-on considérer, au niveau de signification de 5 %, que les deux critères de classification sont indépendants dans la population totale de référence ?

Pour tester l'indépendance des deux critères, on peut effectuer le test du χ^2 de contingence :
La région critique est $W : \{d^2 > \chi^2_{(r-1)(s-1);1-\alpha^*}\}$, où r et s sont les nombres de modalités des deux facteurs et

$$D^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

	Fleurs simples	Fleurs doubles	Total
Feuilles dentées	74 (81.03)	81 (73.98)	155
Feuilles non dentées	41 (33.98)	24 (31.02)	65
Total	115	105	220

D'où $d^2 = 4.32$ et $\chi^2_{1;0.95} = 3.84$.
Par conséquent, on rejette l'hypothèse H_0 d'indépendance entre les deux facteurs.

Exercice 3 (8 points)

Quinze veaux ont été répartis au hasard en trois lots, alimentés chacun de façon différente. Les gains de poids observés au cours d'une même période et exprimés en kg étant les suivants :

lot 1 :	41.2	41.0	40.0	40.1	40.6
lot 2 :	39.8	39.9	42.5	41.1	39.8
lot 3 :	46.0	44.9	44.7	45.7	47.0

le but de l'étude est de mettre en évidence une relation entre l'alimentation et la croissance des veaux. Les moyennes et les variance de ces 3 distributions sont $\bar{x}_1 = 40.58$, $\bar{x}_2 = 40.62$, $\bar{x}_3 = 45.66$, $s_1^{*\,2} = 0.282$, $s_2^{*\,2} = 1.407$ et $s_3^{*\,2} = 0.853$.

1. Tester la normalité des données correspondant au premier lot au niveau $\alpha^* = 0.05$.

Le tableau du test de Stephens donne :

x_i	$\widehat{F}(x_i)$	$z_i = \frac{x_i - \overline{x_1}}{s_1^*}$	$F_0(x_i) = \Phi(z_i)$	$\lvert\widehat{F}(x_i^-) - F_0(x_i)\rvert$	$\lvert\widehat{F}(x_i) - F_0(x_i)\rvert$
40.0	0.2	-1.09	0.14	0.14	0.06
40.1	0.4	-0.90	0.18	0.02	0.22
40.6	0.6	0.04	0.52	0.12	0.08
41.0	0.8	0.79	0.79	0.19	0.01
41.2	1	1.17	0.88	0.08	0.12

On a donc $D_n^* = 0.22$; or, la région critique du test de Stephens, au niveau $\alpha^* = 0.95$, est :

$$D_n^*\left(\sqrt{n} + \frac{0.85}{\sqrt{n}} - 0.01\right) > 0.895.$$

Ici, $D_n^*\left(\sqrt{n} + \frac{0.85}{\sqrt{n}} - 0.01\right) = 0.565$: on ne rejette donc pas l'hypothèse de normalité des données, au niveau $\alpha^* = 0.95$.

On supposera pour la suite que l'hypothèse de normalité peut être acceptée pour les deux autres lots.

2. Peut-on considérer que les variances des trois échantillons sont égales, au niveau de signification $\alpha^* = 0.05$?

On utilise le test de Bartlett dont la région critique est $W = \left\{B > \chi^2_{K-1,1-\alpha^*}\right\}$ avec

$$B = (N-K)\ln(MSW) - \sum_k \left((n_k - 1)\ln {s_k^*}^2\right)$$

où $N = 15$ et $K = 3$.
On obtient

$$SSW = \sum_k (n_k - 1){s_k^*}^2 = 4.(0.282 + 1.407 + 0.853) = 10.17$$

$$MSW = \frac{SSW}{N-K} = 0.85$$

et

$$B = 2.35.$$

Sachant que $\chi^2_{K-1,1-\alpha^*} = \chi^2_{2,0.95} = 5.99$, on ne rejette pas l'hypothèse d'égalité des variances au niveau $1 - \alpha^* = 0.95$.

3. Montrer que le type d'alimentation a un effet significatif sur la croissance des veaux. On prendra $\alpha^* = 0.05$.

On procède au test d'analyse de la variance.
La statistique de test utilisée est la statistique $F = \frac{MSB}{MSW}$ qui, sous H_0, suit une loi de Fisher à $K-1$ et $N-K$ ddl. La région critique est donc

$$W = \{F > f_{K-1,N-K;1-\alpha^*}\}.$$

Si on note $\overline{x}$ la moyenne empirique de l'échantillon total, on a

$$SSB = \sum_k (n_k \overline{x}_k^2) - N\overline{x}^2 = 85.35.$$

On en déduit :

$$MSB = \frac{SSB}{K-1} = 42.67$$

et

$$F = 50.36.$$

Sachant que $f_{K-1,N-K;1-\alpha^*} = f_{2,12,0.95} = 3.885$, on rejette l'hypothèse d'égalité des espérances au niveau $\alpha^* = 0.95$.

4. Préciser pour quels types d'alimentation il existe des différences significatives.

On applique la procédure de comparaison multiple LSD de Fisher. Pour chaque couple (k, ℓ) de niveaux, on calcule la statistique

$$t_{k,\ell} = \frac{|\overline{x}_k - \overline{x}_\ell|}{\sqrt{MSW}\sqrt{1/n_k + 1/n_\ell}}.$$

Si cette quantité est supérieure à $t_{N-K;1-\alpha/2}$, on déclare qu'il y a une différence significative entre μ_k et μ_ℓ. (remarque : cela ressemble au test de Student, sauf que l'on utilise l'estimateur MSW de la variance calculé à l'aide des K échantillons). Les résultats peuvent être présentés dans un tableau

	1	2	3
1			
2	0.07		
3	**8.73**	**8.66**	

Ces valeurs sont à comparer à $t_{N-K;0.975} = t_{12;0.975} = 2.18$. Les valeurs significatives au niveau 5% sont indiquées en gras.