Université de Cergy-Pontoise 2010– 2011
L3 Mathématiques, Statistiques
I. Ignatiouk, E. Löcherbach

Examen. Avril 2011

Les documents ne sont pas autorisés. Durée de l'épreuve : 3 heures.

Exercice 1
Le nombre de demandes hebdomadaires d'un certain produit est une variable aléatoire X de loi de Poisson $\mathcal{P}(\vartheta)$, où $\vartheta > 0$ est inconnu. On note $X_1, \cdots, X_n$ un $n-$échantillon de loi $\mathcal{P}(\vartheta)$ et on considère $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ et $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$.

1.(QCM) **3 points (un demi point par bonne réponse)**

a) $\overline{X}$ est un estimateur sans biais de ϑ : vrai.
b) s_{n-1}^2 est un estimateur sans biais de ϑ : vrai.
c) $\overline{X}$ est un estimateur convergent de ϑ : vrai, par la loi des grands nombres.
d) s_{n-1}^2 est un estimateur convergent de ϑ : vrai, loi des grands nombres.
e) $\overline{X}$ est un estimateur de maximum de vraisemblance pour ϑ : vrai.
f) s_{n-1}^2 est un estimateur de maximum de vraisemblance pour ϑ : faux.
g) $\overline{X}$ est une statistique exhaustive pour ϑ : vrai.
h) s_{n-1}^2 est une statistique exhaustive pour ϑ : faux.

2.(QCM) **2 points (un par bonne réponse)**

a) $S = X_1 + \cdots + X_n$ suit la loi Binomiale $B(n, \vartheta)$: faux.
b) $S = X_1 + \cdots + X_n$ suit la loi de Poisson de paramètre ϑ : faux.
c) $S = X_1 + \cdots + X_n$ suit la loi de Poisson de paramètre $n\vartheta$: vrai.
d) $S = X_1 + \cdots + X_n$ suit une loi Gaussienne (normale) : faux.
e) Lorsque $n \to \infty$, on peut approximer la loi de $\overline{X}$ par la loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec $\mu = \sigma^2 = \vartheta$: faux.
f) Lorsque $n \to \infty$, on peut approximer la loi de $\overline{X}$ par la loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec $\mu = \vartheta$ et $\sigma^2 = \vartheta/n$: vrai.
g) $(n-1)s_{n-1}^2/\vartheta$ suit la loi χ_{n-1}^2 : faux.
h) $Y = \sqrt{n}(\overline{X} - \vartheta)/s_{n-1}$ suit la loi de Student a $n-1$ degrés de liberté : faux.

On souhaite maintenant estimer la probabilité $p = P_\vartheta(X_i = 0)$. On note K le nombre de fois où l'on a observé $X = 0$ dans l'echantillon:

$$K = \sum_{i=1}^n \mathbb{1}_{\{X_i = 0\}}.$$

3. **2 points** Calculer p en fonction de ϑ :

$$p = e^{-\vartheta}.$$

4. **1 + 2 points** Identifier la loi de $Y_i = \mathbb{1}_{\{X_i=0\}}$ et en déduire la loi de K : Y_i suit une loi de Bernoulli de paramètre p et donc K une loi $\mathcal{B}(n,p)$, une loi binomiale.

5. **1 + 2 + 2 points** Montrer que K/n est un estimateur sans biais : **Comme** $K \sim \mathcal{B}(n,p)$, $E(K) = np)$, **donc** $E(K/n) = p$. et convergent : **loi forte des grands nombres!**, de p, puis que :

$$Var_\vartheta(K/n) = e^{-2\vartheta}\ \frac{(e^\vartheta - 1)}{n} :$$

$$\begin{aligned} Var_\vartheta(K/n) &= \frac{1}{n^2} Var_\vartheta(K) \\ &= \frac{1}{n^2} n Var_\vartheta(Y_1) \\ &= \frac{1}{n} p(1-p) = \frac{1}{n} p^2 (\frac{1}{p} - 1). \end{aligned}$$

Comme $p = e^{-\vartheta}$, **cela implique le résultat.**

6. **2 + 2 points** Calculer l'information de Fisher contenue dans l'echantillon $X_1, \ldots, X_n$ sur ϑ. En déduire que l'estimateur K/n n'est pas optimal pour p : On calcule d'abord l'information de Fisher d'un $1-$échantillon. On commence par calculer la log-vraisemblance. Soit $x \in \mathbb{N}$.

$$h(\vartheta, x) = \log P_\vartheta(\{x\}) = \log(e^{-\vartheta} \vartheta^x / x!) = -\vartheta + x \log \vartheta - \log(x!).$$

On dérive deux foix par rapport à ϑ :

$$\frac{\partial^2}{\partial \vartheta^2} h(\vartheta, x) = -\frac{x}{\vartheta^2}.$$

On sait que l'information de Fisher du $1-$échantillon est donnée par

$$I(1, \vartheta) = -E_\vartheta \left(\frac{\partial^2}{\partial \vartheta^2} h(\vartheta, X_1) \right) = \frac{1}{\vartheta^2} E_\vartheta(X_1) = \frac{1}{\vartheta}.$$

Puis on utilise que l'information du $n-$échantillon est donnée par

$$I(n, \vartheta) = n I(1, \vartheta) = \frac{n}{\vartheta}.$$

On calcule maintenant la borne de Cramer-Rao. **Attention :** on est en train d'estimer $p = g(\vartheta) = e^{-\vartheta}$. La borne de Cramer-Rao est donc

$$\frac{(g'(\vartheta))^2}{I(n, \vartheta)} = e^{-2\vartheta} \frac{\vartheta}{n},$$

et il est facile de voir que $Var_\vartheta(K/n)$ est plus grand que cette quantité.

7. On va chercher à améliorer cet estimateur en appliquant le théorème de Rao-Blackwell. On pose $T = E(Y_1 \mid S)$. Rappelons qu' étant donné une variable aléatoire Z,

$$E(Z \mid S) = g \circ S$$

où $g(s) = E(Z \mid S = s)$ est l'espérance de Z par rapport à la loi conditionnelle sachant $S = s$.

   i) **2 points** Formuler le théorème de Rao-Blackwell : Cours!

   ii) **2 + 1 points** Calculer $P(X_i = 0 \mid S = s)$ pour tout $s \in \mathbb{N}$ et en déduire que

$$T \;=\; E(Y_i \mid S) \;=\; \left(1 - \frac{1}{n}\right)^S, \quad \forall i = 1, \ldots, n.$$

   Tout d'abord :

$$P(X_i = 0|S = s) = \frac{P(\{X_i = 0\} \cap \{S = s\})}{P(S = s)}.$$

   Puis,

$$P(\{X_i = 0\} \cap \{S = s\}) = P(X_i = 0, \sum_{j=1, j\neq i}^{n} X_j = s) = P(X_i = 0)P(\sum_{j=1, j\neq i}^{n} X_j = s).$$

   Or, $\sum_{j=1,j\neq i}^{n} X_j$ suit une loi de Poisson de paramètre $(n-1)p$, donc

$$P(\sum_{j=1, j\neq i}^{n} X_j = s) = e^{-(n-1)p} \frac{((n-1)p)^s}{s!},$$

   et $P(X_i = 0) = p = e^{-\vartheta}$.
   De la même manière,

$$P(S = s) = e^{-np} \frac{(np)^s}{s!}.$$

   Donc on obtient finalement

$$P(X_i = 0|S = s) = \left(\frac{n-1}{n}\right)^s = \left(1 - \frac{1}{n}\right)^s.$$

   Puis on utilise que Y_i ne prend que deux valeurs 0 et 1, donc

$$E(Y_i|S) = P(Y_i = 1|S) = P(X_i = 0|S) = \left(1 - \frac{1}{n}\right)^S.$$

   iii) **2 points** Vérifier que $T = \mathbb{E}(K/n \mid S)$:

$$E(K/n|S) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i|S) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i|S) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{1}{n}\right)^S = T.$$

   iv) **4 points** Montrer que

$$Var_\vartheta(T) = e^{-2\vartheta}\left(e^{\vartheta/n} - 1\right) \;<\; Var_\vartheta(K/n).$$

   Tout d'abord,

$$Var_\vartheta(T) = E_\vartheta(T^2) - [E_\vartheta(T)]^2.$$

Soit

$$g(t) = E_\vartheta(t^S),$$

la fonction génératrice de S. Comme S suit une loi de Poisson de paramètre $n\vartheta$, nous avons que

$$g(t) = e^{n\vartheta(t-1)}.$$

(Il faut faire le calcul, mais c'est facile.)
Donc :

$$E_\vartheta(T^2) = E_\vartheta\left((1-\frac{1}{n})^{2S}\right) = g((1-\frac{1}{n})^2) = e^{n\vartheta\left[(1-\frac{1}{n})^2-1\right]} = e^{n\vartheta(\frac{1}{n^2}-\frac{2}{n})} = e^{\frac{\vartheta}{n}-2\vartheta}.$$

D'autre part,

$$E_\vartheta(T) = g(1-\frac{1}{n}) = e^{n\vartheta((1-\frac{1}{n})-1)} = e^{-\vartheta},$$

et donc finalement

$$Var_\vartheta(T) = e^{-2\vartheta}\left[e^{\vartheta/n}-1\right].$$

Il est assez facile de voir que

$$e^{-2\vartheta}\left[e^{\vartheta/n}-1\right] \leq e^{-2\vartheta}\ \frac{(e^\vartheta-1)}{n}.$$

(On développe la fonction exponentielle.)

Exercice 2
Une étude portant sur l'IQ d'enfants de cinq ans donne les valeurs suivantes :

103 112 97 98 111 85 113 97 102

On suppose que ces données sont la réalisation d'un $n-$échantillon $X_1, \ldots, X_n$ d'une loi $N(m, \sigma^2)$. On pose $\overline{X} = (X_1 + \cdots X_n)/n$ et $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Aide au calculs : pour l'échantillon observé, $\overline{x} = (x_1 + \cdots x_n)/n = 102$ et $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2 = 1045/8 = 130,625$.

1.(QCM) **2 points**

a) La variable aléatoire $\overline{X}$ suit la loi $\mathcal{N}(m, \sigma^2)$: faux
b) La variable aléatoire $Y = \frac{1}{\sigma/\sqrt{n}}(\overline{X} - m)$ suit la loi de Student à n degrés de liberté : faux.
c) La variable aléatoire $Z = \frac{\overline{X}-m}{s_{n-1}\sqrt{n}}$ suit la loi de Student à $n-1$ degrés de liberté : faux.
d) La variable aléatoire $W = \frac{\overline{X}-m}{s_{n-1}/\sqrt{n}}$ suit la loi de Student à n degrés de liberté : faux, c'est à $n-1$ degrés de liberté!
e) La variable aléatoire $(n-1)s_{n-1}^2/\sigma^2$ suit la loi de χ^2 à $n-1$ degrés de liberté : vrai !

On souhaite estimer le paramètre m par un intervalle de confiance au risque de 5%.

2.(QCM) **1 + 2 points**

a) Il s'agit d'un intervalle $[a, b]$ avec $a, b \in \mathbb{R}$ tels que $\mathbb{P}(m \in [a, b]) = 0,95$: **ce n'est pas tout à fait vrai, car tel que la réponse est formulée, a et b sont des constantes, donc déterministes. La réponse serait vrai si on remplaçait $a, b \in \mathbb{R}$ par a, b aléatoires. 1 point pour ceux qui ont répondu vrai.**

b) On utilisera la table de la loi $\mathcal{N}(0,1)$: faux.

c) On utilisera la table de la loi de Student à 9 degrés de liberté : faux, il faut utiliser celle de 8 ddl.

d) On utilisera la table de la loi de Student à 8 degrés de liberté : vrai !

e) On utilisera la table de la loi χ^2 à 8 degrés de liberté : faux pour cette question. La table de la loi du χ^2 sera utilisée pour estimer la variance!

3.(QCM) **1 + 2 points** On considère les variables aléatoires Y, Z et W définies dans la question 1. En utilisant les tables, on cherchera

a) $u > 0$ tel que $\mathbb{P}(Y \geq u) = \mathbb{P}(Y \leq -u) = 0,025$: vrai (1 points), si on suppose que σ est connu.

b) $u > 0$ tel que $\mathbb{P}(Y \geq u) = \mathbb{P}(Y \leq -u) = 0,05$: faux.

c) $u > 0$ tel que $\mathbb{P}(|Z| \leq u) = 0,95$: faux, Z n'est pas la bonne variable.

d) $u > 0$ et $v > 0$ tels que $\mathbb{P}((n-1)s_{n-1}^2/\sigma^2 \leq u) = \mathbb{P}((n-1)s_{n-1}^2/\sigma^2 \leq v) = 0,025$: faux : on ne veut pas estimer la variance.

e) $u > 0$ tel que $\mathbb{P}(W \leq u) = 0,975$: vrai (2 points)

4.(QCU) **3 points** On déduit un intervalle de confiance $[a,b]$ pour m au risque de 5%:

a) $[a,b] = [102 - 0,6534\sigma;\ 102 + 0,6534\sigma]$: faux.

b) $[a,b] = [102 - 0,5483\sigma;\ 102 + 0,5483\sigma]$: faux.

c) $[a,b] = [102 - 6,918s; 102 + 6,918s]$:faux.

d) $[a,b] = [102 - 0,754s;\ 102 + 0,754s]$: faux.

e) $[a,b] = [102 - 0,7687s;\ 102 + 0,7687s]$: vrai (3 points).

Preuve de $e)$ **:** Tout d'abord il faut chercher u tel que

$$P(W > u) = 0,025.$$

Comme W suit une loi de Student à 8 ddl, on trouve dans la table :

$$u = 2,31.$$

Rappelons que

$$W = \frac{\overline{X} - m}{s_{n-1}/\sqrt{n}} = 3\frac{\bar{X} - m}{s}.$$

Donc,

$$W > u \Leftrightarrow \bar{X} - m > \frac{s \cdot u}{3} = 0,77 \cdot s.$$

Donc $e)$ est vrai, car $\bar{X} = 102$.

Maintenant, on veut tester l'hypothèse $H_0 : m = 100$ contre $H_1 : m = 108$ au risque de premier espèce $\alpha = 0,1$.

1.(QCM) **3 points**

a) Le risque de première espèce α est égale à la probabilité de se tromper en rejettant l'hypothèse H_0 : vrai.

b) Le risque de première espèce α est égale à la probabilité de se tromper lorsque H_0 est vraie : vrai.

c) Le risque de première espèce α est égale à la probabilité de se tromper en acceptant l'hypothèse H_0 : faux.

d) Le risque de deuxième espèce β est égale à la probabilité de se tromper en rejettant l'hypothèse H_0 : faux.

e) Le risque de deuxième espèce β est égale à la probabilité de se tromper en acceptant l'hypothèse H_0 : vrai.

2. **2 points** Rappeler la définition de la région critique W d'un test : la décision d'un test se fait usuellement à l'aide d'une statistique de test T et de la région critique W de telle manière à ce que lorsque $\{T \in W\}$ alors le test se décide pour H_1, sinon pour H_0.

3.(QCU) **2 points** Pour résoudre le problème du test on utilisera une région critique W du type $W = \{\overline{X} - m \in I\}$

a) avec $I =]-\infty, A]$, $A \in \mathbb{R}$, et $m = 102$:faux.

b) avec $I =]-\infty, A]$, $A \in \mathbb{R}$, et $m = 100$: faux.

c) avec $I =]-\infty, A]$, $A \in \mathbb{R}$, et $m = 108$: faux.

d) avec $I = [A, +\infty[$, $A \in \mathbb{R}$, et $m = 100$: vrai.

e) avec $I = [A, +\infty[$, $A \in \mathbb{R}$, et $m = 102$: faux.

f) avec $I = [A, +\infty[$, $A \in \mathbb{R}$, et $m = 108$: faux.

4.(QCM) **2 + 2 points**

a) D'après le test, on doit accepter l'hypothèse H_0 au risque $\alpha = 0,1$: vrai.

b) On calcule le risque de deuxième espèce β en utilisant la table de la loi $\mathcal{N}(0,1)$: faux.

c) $0,2 \leq \beta \leq 0,3$: vrai.

d) $0,4 \leq \beta \leq 0,6$: faux.

e) $\beta < 0,2$: faux.

f) Si on change la valeur de α en prenant $\alpha = 0,05$ alors la valeur de β va diminuer : faux.

Preuve : On cherche donc une région critique du type $\{\bar{X} - 100 > A\}$. Pour contrôler le risque de premier espèce il faut calculer

$$P_{100}(\bar{X} - 100 > A) = P_{100}\left(\frac{\bar{X} - 100}{s/\sqrt{n}} > \frac{\sqrt{n}}{s}A\right) = P(W > \frac{3}{s}A).$$

Il faut choisir A tel que

$$P(W > \frac{3}{s}A) \leq 0,1.$$

D'après la table,

$$\frac{3}{s}A = 1,4.$$

Donc

$$A = \frac{s}{3} \cdot 1,4 = 5,334.$$

Dans notre cas, $\bar{x} - m = 2$ et 2 n'est pas plus grand que A, donc le test ne rejette pas H_0. Donc a) est vrai.

On va maintenant calculer β :

$$\begin{aligned}
\beta &= P_{108}(\bar{X} - 100 < 5,334) \\
&= P_{108}(\bar{X} - 108 < 5,334 - 8) \\
&= P_{108}(\bar{X} - 108 < -3,334) \\
&= P(W < -\frac{3}{s} \cdot 3,334) \\
&= P(W > 0,875).
\end{aligned}$$

Ici, W suit une loi de Student de 8 ddl. Nous avons que d'une part

$$P(W > 0,875) \leq P(W > 0,71) = 0,25$$

et d'autre part

$$P(W > 0,875) \geq P(W > 0,89) = 0,2,$$

donc

$$\beta \in [0,2, 0,25).$$

Finalement, pour voir que f) est faux, prenons $\alpha = 0,05$. Donc d'après la table,

$$\frac{3}{s} A = 1,86, \text{ donc } A = 7,086.$$

Du coup

$$\beta = P(W < \frac{3}{s}(A - 8)) = P(W > \frac{3}{s} \cdot 0,91)$$

est plus grand que dans le cas $\alpha = 0,1$.