Université de Cergy-Pontoise 2011– 2012
L3 Mathématiques, Statistiques
E. Löcherbach

Correction de l'examen d'avril 2013

Les documents ne sont pas autorisés. Durée de l'épreuve : 3 heures.
Seulement les réponses qui ne sont pas des questions de cours sont développées
Questions de cours

1. Donner la définition d'un modèle statistique régulier.

2. Donnner l'énoncé et la preuve du théorème de Cramer-Rao.

3. Soit $X_1, \ldots, X_n$ un $n-$échantillon de la loi de Bernoulli $B(\vartheta)$.

   (a) Calculer l'information de Fisher et montrer que $\bar{X}_n$ est efficace.

   (b) Rappeler l'inégalité de Hoeffding pour des variables aléatoires qui suivent une loi de Bernoulli et en donner la preuve.

   (c) Utiliser l'inégalité de Hoeffding pour construire un intervalle de confiance au niveau de confiance α pour ϑ.

Exercice 1
Soit $X_1, \ldots, X_n$ un $n-$échantillon de la loi de $\vartheta + Y$ où Y suit une loi exponentielle de paramètre 1, $\vartheta \in \Theta =]0, \infty[$.

1. Montrer que la variable X_1 admet une densité de probabilité qui est donnée par

$$f_\vartheta(x) = e^{-(x-\vartheta)} 1_{x > \vartheta}.$$

   Solution : Soit g une fonction test (borelienne, positive), alors

$$E_\vartheta(g(X_1)) = E(g(\vartheta + Y)) = \int_0^\infty e^{-t} g(\vartheta + t) dt = \int_\vartheta^\infty e^{-(x-\vartheta)} g(x) dx = \int_0^\infty g(x) f_\vartheta(x) dx,$$

   ou nous avons fait un changement de variables $x = t + \vartheta$.

2. Soit $\bar{X}_n$ la moyenne empirique. Déterminer la constante c telle que

$$\bar{X}_n - c \text{ est un estimateur sans biais de } \vartheta.$$

   Nous avons que $E_\vartheta(\bar{X}_n) = E_\vartheta(X_1) = E(\vartheta + Y) = \vartheta + E(Y) = \vartheta + 1$, **donc il faut prendre** $c = 1$.

3. Un autre estimateur possible est $\hat{\vartheta}_n = \min\{X_1, \ldots, X_n\}$. (Pourquoi ?) Montrer, **sans faire le calcul**, que $\hat{\vartheta}_n$ est un estimateur qui est biaisé.
   Solution : $X_i \stackrel{\mathcal{L}}{=} \vartheta + Y > \vartheta$ **p.s. (car Y suit une loi exponentielle). Cela implique que** $\hat{\vartheta}_n > \vartheta$ **p.s., autrement dit :**

$$\hat{\vartheta}_n - \vartheta > 0.$$

Supposons maintenant que l'estimateur soit sans biais, donc

$$E_\vartheta(\hat{\vartheta}_n - \vartheta) = 0,$$

alors cela implique que

$$\hat{\vartheta}_n - \vartheta = 0 \quad p.s.$$

(une variable positive qui a espérance nulle doit elle même être nulle p.s.), ce qui est une contradiction.

4. Montrer que $\hat{\vartheta}_n$ converge en probabilité vers ϑ lorsque $n \to \infty$.
   Solution : Comme $\hat{\vartheta}_n \geq \vartheta$,

   $$P_\vartheta(|\hat{\vartheta}_n - \vartheta| \geq \epsilon) = P_\vartheta(\hat{\vartheta}_n \geq \vartheta + \epsilon) = P_\vartheta(\forall i = 1, \ldots, n : X_i \geq \vartheta + \epsilon) = (P_\vartheta(X_1 \geq \vartheta + \epsilon))^n .$$

   Or,

   $$P_\vartheta(X_1 \geq \vartheta + \epsilon) = P(Y \geq \epsilon) = e^{-\epsilon}.$$

   Donc

   $$P_\vartheta(|\hat{\vartheta}_n - \vartheta| \geq \epsilon) = e^{-\epsilon n} \to 0$$

   lorsque $n \to \infty$.

5. On veut maintenant construire un intervalle de confiance en utilisant l'estimateur $\hat{\vartheta}_n$. Lequel des intervalles suivants est préférable (donnez une justification de votre réponse)?

   $$I_1 = [\hat{\vartheta}_n - \delta, \hat{\vartheta}_n + \delta]$$

   ou

   $$I_2 = [\hat{\vartheta}_n - \delta, \hat{\vartheta}_n]$$

   ou

   $$I_3 = [\hat{\vartheta}_n, \hat{\vartheta}_n + \delta],$$

   pour un $\delta > 0$ à choisir.
   Solution : Le deuxième : car $\hat{\vartheta}_n$ est toujours plus grand que le vrai paramètre, donc ce n'est pas la peine de considérer des valeurs plus grandes que $\hat{\vartheta}_n$.

6. Choisir un des intervalles et constuire un intervalle de confiance (théorique) au niveau α, en se basant sur l'estimation du point 4 (c'est-à-dire déterminer δ en fonction de n et de α).
   Solution : On choisit I_2. Donc nous voulons controler

   $$1 - \alpha = P_\vartheta(\vartheta \notin I_2) = P_\vartheta(\vartheta < \hat{\vartheta}_n - \delta) = P_\vartheta(\hat{\vartheta}_n > \vartheta + \delta) = e^{-\delta n},$$

   ou nous avons utilisé point 4. Donc il faut choisir

   $$\delta = -\frac{1}{n}\log(1 - \alpha).$$

Exercice 2
Un appareil de guidage radar récolte des informations dans une direction de l'espace. Celles-ci se présentent sous la forme d'un vecteur d'observation $(Z_1, \ldots, Z_n)$ où les Z_i sont $i.i.d$ et de loi $\mathcal{N}(0, \sigma^2)$ s'il n'y a pas d'obstacle et de loi $\mathcal{N}(m, \sigma^2)$ s'il y en a un, où σ^2 est connu. Si le radar détecte un obstacle, on recalcule la trajectoire, sinon on ne change rien. On souhaite tester

H_0 : "$m \neq 0$" contre H_1 : "$m = 0$". Comment interpréter dans ce cas les risques de première et de deuxième espèce ?
Solution : Erreur de première espèce: en réalité il y a un obstacle, mais on ne le découvre pas.
Erreur de deuxième espèce : en réalité il n'y a pas d' obstacle, mais le test en découvre un, et donc on recalcule la trajectoire inutilement.

Exercice 3
Une société d'assurances a comptabilisé parmi ses 500 assurés ceux qui ont déclaré un ou plusieurs sinistres au cours de l'année écoulée. Cela a donné le tableau suivant :

- 0 sinistre : 171 assurés,
- 1 sinistre : 202 assurés,
- 2 sinistres : 80 assurés,
- 3 sinistres : 36 assurés,
- ≥ 4 sinistres : 11 assurés.

Nous souhaitons tester l'hypothèse H_0 : "la répartition des assurés suit une loi de Poisson de paramètre 1" en utilisant un test du χ^2 d'adéquation.

1. Rappeler le cadre général du test du χ^2 d'adéquation et expliquer intuitivement pourquoi il doit "marcher".
   Solution : On observe un $n-$échantillon d'une loi inconnue à valeurs dans un espace E. On partitionne
   $$E = E_1 \cup \ldots \cup E_q$$
   et on compte le nombre d'observations qui tombent dans chaque E_i : N_n^i. On voudrait tester si la loi du $n-$échantillon est une loi précise μ_0 ou pas. Pour ce faire on introduit
   $$p_i = \mu_0(E_i), 1 \leq i \leq q$$
   et
   $$D_n = \sum_{i=1}^{q} \frac{(N_n^i - np_i)^2}{np_i}.$$
   Sous la loi μ_0, par le TCL, D_n va converger en loi vers une variable aléatoire limite qui est finie (et en fait : qui suit une loi du chi-deux à $q-1$ ddl). Par contre, si ce n'est pas la loi μ_0 qui a engendré les données, en général, D_n va exploser (converger vers ∞) - car les p_i ne sont pas les bonnes quantités pour centraliser, c'est-à-dire ce ne sont pas les bonnes espérances. Donc le test va se décider pour $H_1 =$ "la loi n'est pas μ_0" si et seulement si $D_n > K$. On choisit K tel que
   $$P(U > K) = \alpha,$$
   avec $U \sim \chi^2(q-1)$.

2. Donner q et la partition $E_1, \ldots, E_q$ dans le cadre du présent exercice, ainsi que les valeurs $p_1, \ldots, p_q$.
Solution : $E = \{0, 1, \ldots\}, E_1 = \{0\}, E_2 = \{1\}, \ldots, E_4 = \{3\}, E_5 = \{n : n \geq 4\}$. **Donc** $q = 5$. **Les valeurs** $p_1, \ldots, p_5$ **sont données par la loi de Poisson avec paramètre** 1 :

$$p_1 = e^{-1}, p_2 = e^{-1}, p_3 = e^{-1} \times \frac{1}{2}, p_4 = e^{-1} \times \frac{1}{6}, p_5 = 1 - (p_1 + \ldots + p_4) = 1 - e^{-1} \times 2,\bar{6}.$$

3. Calculer la statistique du test et conclure.
La statistique du test est donnée par

$$\begin{aligned} D_n &= \frac{(171 - 500 \cdot e^{-1})^2}{500e^{-1}} + \frac{(202 - 500 \cdot e^{-1})^2}{500e^{-1}} + \frac{(80 - 250 \cdot e^{-1})^2}{250e^{-1}} + \\ &\quad \frac{(36 - 250 \cdot \frac{1}{3} \cdot e^{-1})^2}{250 \cdot \frac{1}{3}e^{-1}} + \frac{(11 - 500 \cdot (1 - 2,\bar{6}e^{-1}))^2}{500(1 - 2,\bar{6}e^{-1})}. \end{aligned}$$

Un calcul pénible donne

$$D_n \approx 5,41.$$

Nous avons $q - 1 = 4$ **et donc** $K = 9,4877$. **Donc :** $D_n \leq K$, **le test ne se décide donc pas pour** H_1 : **la répartition des assurés est donnnée par une loi de Poisson de paramètre** 1.