Ecole Normale Supérieure
Gérard Biau
FIMFA — Statistique

# CONTRÔLE TERMINAL — 10/01/2013

**Durée 3h — Aucun document autorisé**

## Exercice 1

Soit $\theta \in ]-1,1[$. On considère un échantillon composé de $n$ variables aléatoires réelles $X_1, \cdots, X_n$, indépendantes et identiquement distribuées, de densité commune par rapport à la mesure de Lebesgue

$$f_\theta(x) = \frac{1}{2}(1+\theta x)\mathbf{1}_{]-1,1[}(x), \quad x \in \mathbb{R}.$$

1. Calculer $\mathbb{E}_\theta(X_1)$.
2. En déduire un estimateur $\hat{\theta}_n$ de $\theta$ par la méthode des moments.
3. Donner le biais de $\hat{\theta}_n$ ainsi que son risque quadratique.
4. Déterminer la loi limite de $\hat{\theta}_n$ (correctement centré et normalisé).
5. Soit $\alpha \in ]0,1[$. Déduire des questions précédentes un intervalle de confiance (bilatère) asymptotique pour $\theta$ de niveau $1-\alpha$.

## Exercice 2

Dans la suite, $\lambda$ désigne la mesure de Lebesgue sur $\mathbb{R}$, $\delta_1$ la mesure de Dirac au point 1 et $\theta$ un paramètre réel strictement plus grand que 1. On considère $n$ variables aléatoires $X_1, \cdots, X_n$, indépendantes et identiquement distribuées, de densité commune **par rapport à** $\lambda + \delta_1$

$$f_\theta(x) = \frac{1}{\theta+1}\mathbf{1}_{]0,\theta]}(x), \quad x \in \mathbb{R}.$$

1. Montrer que pour chaque $\theta > 1$, $f_\theta$ définit bien une densité de probabilité pour la mesure $\lambda + \delta_1$.
2. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de $\theta$.
3. Déterminer la fonction de répartition de $\hat{\theta}_n$.
4. En déduire que $\hat{\theta}_n$ est (faiblement) consistant.
5. Donner la loi limite de $\hat{\theta}_n$ (correctement centré et normalisé).
6. Soit $\alpha \in ]0,1[$. Déduire des questions précédentes un intervalle de confiance (bilatère) asymptotique pour $\theta$ de niveau $1-\alpha$.

## Exercice 3

La concentration dans le sang d'une certaine drogue évolue avec le temps selon le schéma suivant : à chaque instant $t \geq 0$, elle est une réalisation de la variable aléatoire $\theta t + \varepsilon$, où $\theta$ est un paramètre réel inconnu et $\varepsilon$ une variable aléatoire de loi $\mathcal{N}(0,1)$. Des biologistes, qui ont mesuré la concentration aux instants $0 \leq t_1 < \cdots < t_n$, considèrent que ces $n$ valeurs sont des réalisations de variables aléatoires indépendantes.

1. Décrire le modèle statistique et donner son information de Fischer.
2. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de $\theta$.
3. Montrer que $\hat{\theta}_n$ est sans biais et efficace.
4. Proposer un estimateur naturel permettant de prévoir la concentration moyenne dans le sang de la drogue à l'instant $t_{n+1} > t_n$ (justifier ce choix).
5. Soit $\alpha \in ]0,1[$. Construire un intervalle de confiance (bilatère) de niveau $1-\alpha$ pour la concentration moyenne au temps $t_{n+1}$.

## Exercice 4

**Rappel.** Soit $p \in ]0,1[$. On dit qu'une variable aléatoire suit la loi de Rademacher de paramètre $p$ si elle ne prend que deux valeurs, 1 et -1, avec les probabilités respectives $p$ et $1-p$.

On considère le modèle de régression **à design aléatoire**

$$Y_i = \theta X_i + \varepsilon_i, \quad i = 1, \cdots, n,$$

où $X_1, \cdots, X_n$ sont i.i.d. de loi de Rademacher $p \in ]0,1[$ et $\varepsilon_1, \cdots, \varepsilon_n$ sont i.i.d. de loi $\mathcal{N}(0,1)$. On suppose en outre que $(X_1, \cdots, X_n)$ est indépendant de $(\varepsilon_1, \cdots, \varepsilon_n)$.

1. Déterminer $\hat{\theta}_n$, l'estimateur des moindres carrés de $\theta$.
2. Montrer que $\hat{\theta}_n$ converge (fortement) vers $\theta$.
3. Quelle est la loi de $\hat{\theta}_n$ ?
4. On suppose désormais que $p \neq 1/2$ et on considère l'estimateur alternatif
$$\hat{\theta}'_n = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} \mathbf{1}_{[\sum_{i=1}^n X_i \neq 0]}.$$
Montrer que $\hat{\theta}'_n$ converge (fortement) vers $\theta$.
5. Donner la loi limite de $\hat{\theta}'_n$ (convenablement centré et normalisé).
6. Quel estimateur préférez-vous : $\hat{\theta}_n$ ou $\hat{\theta}'_n$ ?

## Problème

Dans tout ce qui suit, $\mathcal{P}$ désigne l'ensemble des probabilités **sans atome** sur $\mathbb{R}$. Si $x = (x_1, \cdots, x_\ell)$ sont des nombres réels distincts, on note $R_x(i)$ le **rang** de $x_i$ dans la liste $x$, i.e.

$$R_x(i) = \text{Card}\{j = 1, \cdots, \ell : x_j \leq x_i\}.$$

On désigne par $\mathcal{S}_\ell$ l'ensemble des permutations de $\{1, \cdots, \ell\}$.

1. (**Echauffement 1**) Montrer que lorsque la liste $x = (x_1, \cdots, x_\ell)$ est constituée de nombres réels distincts, on a $R_x \in \mathcal{S}_\ell$.
2. (**Echauffement 2**) Prouver aussi que pour tout $\sigma \in \mathcal{S}_\ell$, $R_x \circ \sigma = R_{x^\sigma}$ si $x^\sigma = (x_{\sigma(1)}, \cdots, x_{\sigma(\ell)})$.

On se donne $(X_1, Y_1), \cdots, (X_n, Y_n)$ des variables aléatoires indépendantes et identiquement distribuées, à valeurs dans $\mathbb{R} \times \mathbb{R}$, chacune de loi inconnue $P \otimes Q$, avec $P, Q \in \mathcal{P}$. Le premier objectif du problème consiste à tester l'hypothèse $H_0 : P = Q$ contre $H_1 : P \otimes Q(T) \neq 1/2$, avec

$$T = \left\{(x, y) \in \mathbb{R}^2 : x \leq y\right\}.$$

En d'autres termes, il s'agit de construire un test permettant de préciser si deux jeux indépendants d'observations indépendantes sont issus (ou pas) de la même loi sans atome. On introduit pour cela la statistique dite de *Mann-Whitney* :

$$U_n = \sum_{i,j=1}^{n} \mathbf{1}_{[X_i < Y_j]}.$$

1. Etablir tout d'abord que le problème de test est bien cohérent, à savoir que $H_0$ et $H_1$ sont disjointes.
2. Soit $Z = (Z_1, \cdots, Z_\ell)$ une suite de variables aléatoires réelles indépendantes et de même loi sans atome. Montrer que $R_Z$ suit la loi uniforme sur $\mathcal{S}_\ell$.
3. Déduire de ce qui précède que $U_n$ est une statistique libre sous $H_0$ (c'est-à-dire que sa loi ne dépend ni de $P$ ni de $Q$).

   **Indication.** On pourra poser $Z = (Z_1, \cdots, Z_{2n})$ avec $Z_i = X_i$ et $Z_{i+n} = Y_i$ pour $i = 1, \cdots, n$, et évaluer $\sum_{j=n+1}^{2n} R_Z(j)$.
4. Quel est le principe de la statistique de Mann-Whitney ?
5. Soit $\alpha \in ]0,1[$. Proposer alors un test de niveau $\alpha$ de $H_0$ contre $H_1$.
6. Donner un exemple de lois $P, Q \in \mathcal{P}$ telles que $P \neq Q$ et $P \otimes Q(T) = 1/2$.
7. Pour quelle raison n'a-t-on pas choisi l'hypothèse alternative $H_1' : P \neq Q$ ?

On désigne maintenant par $\mathcal{P}'$ l'ensemble des probabilités **sans atome** sur $\mathbb{R} \times \mathbb{R}$ et par $\mathcal{E}$ l'ensemble des lois échangeables de $\mathcal{P}'$, i.e. si $(U, V)$ est un couple de variables aléatoires dont la loi est dans $\mathcal{E}$, alors $(U, V)$ a même loi que $(V, U)$. On se donne un échantillon i.i.d. $(X_1, Y_1), \cdots, (X_n, Y_n)$ de loi commune inconnue $P \in \mathcal{P}'$. L'objectif de cette seconde partie est de construire un test non paramétrique de l'hypothèse $H_0 : P \in \mathcal{E}$ contre $H_1 : P \notin \mathcal{E}$.

Pour $i = 1, \cdots, n$, on note $Z_i = X_i - Y_i$ et $|Z| = (|Z_1|, \cdots, |Z_n|)$. On désigne par $R_{|Z|}(k)$ le rang de $|Z_k|$ dans la liste $|Z|$, et on introduit la statistique de *Wilkoxon*

$$W = \min\left(T^+, n(n+1)/2 - T^+\right),$$

avec

$$T^+ = \sum_{k=1}^{n} k I_k, \quad \text{où } I_k = \mathbf{1}_{[Z_{R_{|Z|}(k)} > 0]}.$$

1. Montrer que, sous $H_0$, les variables aléatoires $I_1, \cdots, I_n$ sont indépendantes et de même loi de Bernoulli de paramètre $1/2$.
2. Quel est le principe de la statistique de Wilkoxon ?
3. Soit $\alpha \in ]0,1[$. Construire alors un test de niveau $\alpha$ de $H_0$ contre $H_1$.
4. Imaginer une situation concrète dans laquelle ce test peut être utilisé.