UPMC - Master de mathématiques et applications
Statistique et apprentissage
Gérard Biau

CONTRÔLE TERMINAL — 23/01/2018

Durée 3h

Exercice 1

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de variables aléatoires réelles de densité commune

$$f_\theta(x) = \frac{1}{2\sqrt{x\theta}} \mathbb{1}_{]0,\theta]}(x),$$

où θ est un paramètre inconnu strictement positif.

1. Expliciter $\hat{\theta}_1$, l'estimateur de θ par la méthode des moments (on utilisera le premier moment).
2. Montrer que $\hat{\theta}_1$ est sans biais et calculer sa variance.
3. Expliciter $\hat{\theta}_2$, l'estimateur de θ par la méthode du maximum de vraisemblance.
4. Calculer $\mathbb{P}(\hat{\theta}_2 \leq t)$ pour tout $t \in \mathbb{R}$.
5. Montrer que $\hat{\theta}_2$ est convergent en moyenne quadratique.
6. Comparer les risques quadratiques de $\hat{\theta}_1$ et $\hat{\theta}_2$. Quel est le meilleur estimateur de ce point de vue ?
7. Préciser la loi asymptotique de $(n/2)(\hat{\theta}_2 - \theta)$.
8. Soit $\alpha \in]0,1[$. Utiliser la question précédente pour proposer un intervalle de confiance asymptotique de niveau $(1-\alpha)$ pour le paramètre $g(\theta) = \ln(\theta)$.

Exercice 2

Soit F la fonction de répartition, supposée continue et strictement croissante, d'une loi symétrique [1] sur $\mathbb{R}$, et $\theta \in \mathbb{R}$ un paramètre inconnu. On dispose d'un échantillon i.i.d. $X_1, \ldots, X_n$ d'observations suivant la loi $F_\theta(x) = F(x-\theta)$ et on considère la statistique

$$\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i>0]}.$$

1. Prouver que, pour tout x dans $\mathbb{R}$, $F(-x) = 1 - F(x)$. Préciser la loi de $n\bar{S}_n$.
2. On souhaite tester l'hypothèse $H_0 : \theta = 0$ contre l'alternative $H_1 : \theta > 0$. On fixe $\alpha \in]0,1[$.
   2.a Proposer un test (non asymptotique) de niveau α basé sur $\bar{S}_n$.
   2.b Proposer un test asymptotique de niveau α basé sur $\bar{S}_n$.
   2.c Montrer que le test précédent est convergent.
3. On suppose dans cette question que F est connue. Donner un intervalle de confiance asymptotique pour θ de niveau $(1-\alpha)$.

Exercice 3

Soit (X, Y) un couple de variables aléatoires à valeurs dans $\mathbb{R}_+ \times \{0, 1\}$. On note $\eta(x) = \mathbb{P}(Y = 1|X = x)$ et on suppose que $\eta(x) = x/(c + x)$, où c est une constante strictement positive.

1. Montrer que l'erreur de Bayes $L^\star$ associée au couple (X, Y) a pour expression
$$L^\star = \mathbb{E}\Big(\frac{\min(c, X)}{c + X}\Big).$$
2. Que vaut $L^\star$ lorsque X suit une loi uniforme sur $[0, \alpha c]$, où α est un paramètre supérieur ou égal à 1 ?
3. Montrer soigneusement qu'il existe une valeur de α maximisant $L^\star$.

1. Rappel : Une variable aléatoire Z a une loi symétrique si Z a la même loi que $-Z$.

Problème

I. Préliminaires. Voici pour commencer des questions indépendantes. Ces résultats vous seront utiles dans les parties qui suivent.

1. Soit Z une variable aléatoire réelle admettant un moment d'ordre 2. Montrer que pour tout $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \frac{\mathrm{Var}(Z)}{\mathrm{Var}(Z) + t^2}.$$

   (On pourra observer que si Z est centrée, alors $t \leq \mathbb{E}((t-Z)\mathbb{1}_{[Z<t]})$.)

2. Soit Z une variable aléatoire binomiale de paramètres $n \in \mathbb{N}^\star$ et $p \in]0,1[$. Prouver que

$$\mathbb{E}\Big(\frac{1}{Z}\mathbb{1}_{[Z>0]}\Big) \leq \frac{2}{(n+1)p}.$$

   (On pourra commencer par majorer $\mathbb{E}(\frac{1}{1+Z})$.)

3. **Attention, question très difficile. Ne l'abordez qu'à la fin, si vous vous ennuyez**. Soit $(p_1, \ldots, p_k)$ un vecteur de probabilités (i.e., $p_i \geq 0$ et $\sum_{i=1}^k p_i = 1$). Montrer que, pour $n \geq 4$, on a

$$\sum_{i=1}^k p_i(1-p_i)^n \leq \frac{k}{en}.$$

II. Enoncé. Soit k un entier strictement positif et (X, Y) un couple de variables aléatoires à valeurs dans $\{1, \ldots, k\} \times \{0, 1\}$. La loi de la variable aléatoire **discrète** X est donc entièrement décrite par le vecteur de probabilités $(p_1, \ldots, p_k)$, où $p_i = \mathbb{P}(X = i)$. On note $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ et on désigne par $L^\star$ l'erreur de Bayes associée à (X, Y).

Etant donné un échantillon $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ de variables aléatoires indépendantes et de même loi que (X, Y), on considère la règle de classification naturelle $g_n^\star$ définie pour $x \in \{1, \ldots, k\}$ par

$$g_n^\star(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^n \mathbb{1}_{[X_i = x]}\mathbb{1}_{[Y_i=1]} > \sum_{i=1}^n \mathbb{1}_{[X_i=x]}\mathbb{1}_{[Y_i=0]} \\ 0 & \text{sinon.} \end{cases}$$

(Par convention, une somme vide est nulle.) On suppose bien entendu que $\mathscr{D}_n$ et (X, Y) sont indépendants et on pose

$$L(g_n^\star) = \mathbb{P}(g_n^\star(X) \neq Y | \mathscr{D}_n).$$

L'objectif principal du problème est de montrer que, pour $n \geq 4$,

$$\mathbb{E}L(g_n^\star) \leq L^\star + \sqrt{\frac{k}{2(n+1)}} + \frac{k}{en}. \tag{1}$$

A. Echauffement.

1. Prouver en une ligne (cours) que $\mathbb{E}L(g_n^\star) \to L^\star$ lorsque n tend vers l'infini.
2. Montrer que
$$L(g_n^\star) \geq \sum_{x:\sum_{i=1}^n \mathbb{1}_{[X_i=x]}=0} \eta(x)p_x.$$
3. En déduire que
$$\mathbb{E}L(g_n^\star) \geq \sum_{x=1}^k \eta(x)p_x(1-p_x)^n.$$
4. On suppose dans cette question (et seulement dans cette question) que $\eta(x) = 1$ pour tout x.
   4.a Que vaut $L^\star$?
   4.b Trouver un vecteur $(p_1, \ldots, p_k)$ tel que $\mathbb{E}L(g_n^\star) \geq 1/2$ pour $k \geq 2n$.
   4.c Qu'en conclure ?

B. Preuve de l'inégalité (1).

1. Dans toute la suite, on note $N(x) = \sum_{i=1}^n \mathbb{1}_{[X_i=x]}$. Ecrire $g_n^\star(x)$ en faisant intervenir $N(x)$ (avec la convention $0/0 = 0$).
2. Quelle est, conditionnellement à $\mathbb{1}_{[X_1=x]}, \ldots, \mathbb{1}_{[X_n=x]}$, la loi de la variable aléatoire $Z(x) = \sum_{i=1}^n \mathbb{1}_{[X_i=x]}Y_i$?
3. Prouver alors que
$$\mathbb{E}L(g_n^\star) = \sum_{x=1}^k p_x\big(\eta(x) + (1-2\eta(x))\mathbb{P}(\mathrm{Bin}(N(x),\eta(x)) > N(x)/2)\big),$$
où la notation $\mathrm{Bin}(N(x),\eta(x))$ désigne une variable aléatoire binomiale de paramètres $N(x)$ et $\eta(x)$ (nulle par convention si $N(x) = 0$).
4. En déduire que
$$\mathbb{E}L(g_n^\star) \leq \sum_{x=1}^k p_x\big(\xi(x) + \big(1-2\xi(x)\big)\mathbb{P}(\mathrm{Bin}(N(x),\xi(x)) \geq N(x)/2)\big),$$
où l'on a posé $\xi(x) = \min(\eta(x), 1-\eta(x))$. (On pourra utiliser le fait que $\mathbb{P}(\mathrm{Bin}(m,p) \leq m/2) = \mathbb{P}(\mathrm{Bin}(m,1-p) \geq m/2)$.)

5. Montrer alors que

$$\mathbb{E}L(g_n^\star) \leq L^\star + \sum_{x=1}^{k} p_x(1-2\xi(x))\mathbb{E}\Big(\frac{1}{1+(1-2\xi(x))^2 N(x)}\Big).$$

6. Poursuivre en établissant que

$$\mathbb{E}L(g_n^\star) \leq L^\star + \sum_{x=1}^{k} p_x \mathbb{E}\Big(\frac{1}{2\sqrt{N(x)}}\mathbb{1}_{[N(x)>0]} + (1-2\xi(x))\mathbb{1}_{[N(x)=0]}\Big).$$

7. Prouver finalement que

$$\mathbb{E}L(g_n^\star) \leq L^\star + \sum_{x=1}^{k} p_x(1-p_x)^n + \frac{1}{2}\sum_{x=1}^{k} p_x \sqrt{\mathbb{E}\Big(\frac{1}{N(x)}\mathbb{1}_{[N(x)>0]}\Big)}.$$

8. Etablir alors l'inégalité (1).

C. Le cas multivarié à composantes indépendantes. On suppose dans cette dernière partie que X est multivariée à valeurs dans $\{0,1\}^d$. On note $X = (X^{(1)}, \ldots, X^{(d)})$ (chaque $X^{(j)}$ prend donc ses valeurs dans $\{0,1\}$) et on suppose que les composantes $X^{(1)}, \ldots, X^{(d)}$ sont **indépendantes conditionnellement** à $Y = 1$, et aussi **indépendantes conditionnellement** à $Y = 0$. On note

$$p(j) = \mathbb{P}(X^{(j)} = 1 | Y = 1), \quad q(j) = \mathbb{P}(X^{(j)} = 1 | Y = 0)$$

et $p = \mathbb{P}(Y = 1)$, en supposant toutes ces quantités strictement comprises entre 0 et 1.

1. Pour $x = (x^{(1)}, \ldots, x^{(d)})$, que valent $\mathbb{P}(X = x | Y = 1)$ et $\mathbb{P}(X = x | Y = 0)$?
2. Donner alors l'expression de la règle de Bayes $g^\star$ associée au couple (X, Y).
3. En posant

$$\alpha_0 = \ln\Big(\frac{p}{1-p}\Big) + \sum_{j=1}^{d} \ln\Big(\frac{1-p(j)}{1-q(j)}\Big)$$

et

$$\alpha_j = \ln\Big(\frac{p(j)}{q(j)} \cdot \frac{1-q(j)}{1-p(j)}\Big), \quad 1 \leq j \leq d,$$

écrire $g^\star$ en fonction de α_0 et des α_j.

4. Pourquoi ce résultat est-il intéressant ?