Examen médian : OS02 - Théorie de la décision - 2 h

20 novembre 2008

Documents autorisés : Polycopiés distribués et un formulaire.

Exercice 1 (9 points) *Soient $\xi_1, \ldots, \xi_n$ des variables aléatoires i.i.d. dont la loi admet la densité de probabilité $f_\theta(x)$:*

$$f_\theta(x) = \begin{cases} \dfrac{2(\theta - x)}{\theta^2} & si \quad x \in [0, \theta], \\ 0 & si \quad x \notin [0, \theta] \end{cases}$$

On veut tester l'hypothèse de base $\mathcal{H}_1 = \{f = f_1\}$ et l'hypothèse concurrente $\mathcal{H}_2 = \{f = f_{\theta > 1}\}$.

Introduisons le test («ad hoc») suivant :

$$\delta(\xi_1, \ldots, \xi_n) = \begin{cases} \mathcal{H}_2 & si \quad \xi_{(n)} > c \\ \mathcal{H}_1 & si \quad \xi_{(n)} \leq c \end{cases}$$

où $\xi_{(n)} = \max\{\xi_1, \ldots, \xi_n\}$.

1. *Calculer les risques α_1, α_2 du test δ en fonction du seuil $c \in \mathbb{R}$. Dessiner (approximativement !) les courbes $\alpha_i = \alpha_i(c)$.*
2. *Déterminer le risque α_2 en fonction de α_1 : $\alpha_1 \mapsto \alpha_2(\alpha_1)$ et dessiner (approximativement !) sa courbe representative en coordonnées $\alpha_1 O \alpha_2$.*

Réponses :

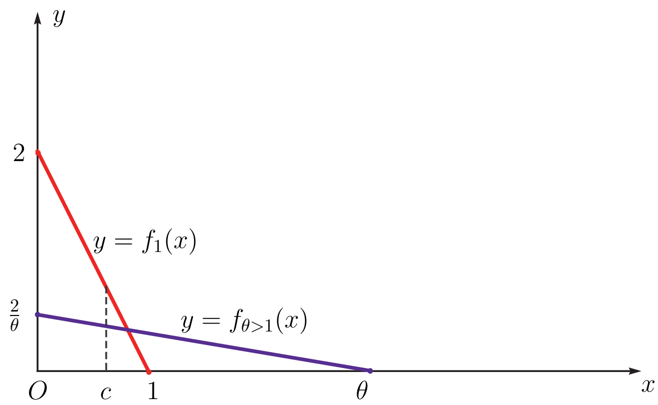

FIG. 1 – Les densités de probabilité sous $\mathcal{H}_1 = \{f = f_1\}$ et $\mathcal{H}_2 = \{f = f_{\theta>1}\}$.

1) Les densités de probabilité sous $\mathcal{H}_1$ et $\mathcal{H}_2$ et le seuil c sont graphiquement représentés sur la figure 1. On calcule le risque α_1. Soit $0 \leq c \leq 1$.

$$\begin{aligned} \alpha_1 &= \mathbb{P}_1(\max\{\xi_1, \ldots, \xi_n\} > c) = 1 - \mathbb{P}_1(\xi_1 \leq c \text{ et } \xi_2 \leq c \ldots \text{et } \xi_n \leq c) \\ &= 1 - \prod_{i=1}^{n} \mathbb{P}_1(\xi_i \leq c) = 1 - [\mathbb{P}_1(\xi \leq c)]^n = 1 - \left[\int_0^c 2(1-x)dx\right]^n = 1 - c^n(2-c)^n. \end{aligned}$$

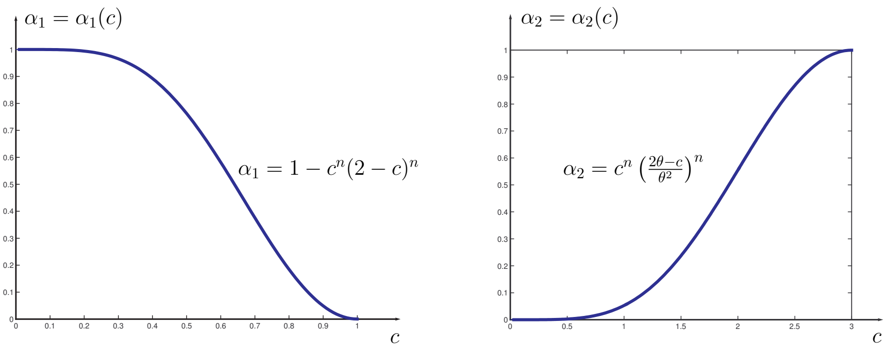

FIG. 2 – Les courbes $\alpha_i = \alpha_i(c)$.

Pour $c \in \mathbb{R}$ on obtient :

$$\alpha_1 = \begin{cases} 1 & \text{si} \quad c < 0 \\ 1 - c^n(2-c)^n & \text{si} \quad 0 \le c < 1 \\ 0 & \text{si} \quad c \ge 1 \end{cases}$$

Cette function représentée sur la figure 2. On calcule le seuil c comme fonction de α_1 :

$$c^2 - 2c + \sqrt[n]{1-\alpha_1} = 0.$$

D'où on a les solutions de l'équation quadratique

$$c_{1,2} = 1 \pm \sqrt{1 - \sqrt[n]{1-\alpha_1}}.$$

La solution c appartenant à l'intervalle $[0,1]$ est $c_1 = 1 - \sqrt{1 - \sqrt[n]{1-\alpha_1}}$. On calcule le risque α_2. Soit $0 \le c \le \theta$, où $\theta > 1$.

$$\begin{aligned} \alpha_2 &= \mathbb{P}_2(\max\{\xi_1, \ldots, \xi_n\} \le c) = \mathbb{P}_2(\xi_1 \le c \,\text{et}\, \xi_2 \le c \ldots \text{et}\, \xi_n \le c) \\ &= \prod_{i=1}^{n} \mathbb{P}_2(\xi_i \le c) = [\mathbb{P}_2(\xi \le c)]^n = \left[\int_0^c \frac{2(\theta - x)}{\theta} dx\right]^n = c^n \left(\frac{2\theta - c}{\theta^2}\right)^n. \end{aligned}$$

Pour $c \in \mathbb{R}$ on obtient :

$$\alpha_2 = \begin{cases} 0 & \text{si} \quad c < 0 \\ c^n \left(\dfrac{2\theta - c}{\theta^2}\right)^n & \text{si} \quad 0 \le c < \theta \\ 1 & \text{si} \quad c \ge \theta \end{cases}$$

2) Finalement, en utilisant le paramétrage $\alpha_1 = \alpha_2(c)$ et $\alpha_2 = \alpha_2(c)$, le risque α_2 comme fonction de α_1 est définie de la façon suivante :

$$\alpha_2 = \left(1 - \sqrt{1 - \sqrt[n]{1-\alpha_1}}\right)^n \left(\frac{2\theta - 1 + \sqrt{1 - \sqrt[n]{1-\alpha_1}}}{\theta^2}\right)^n \quad \text{pour} \quad \alpha_1 \in [0, 1[$$

et le reste est représenté graphiquement sur la figure 3.

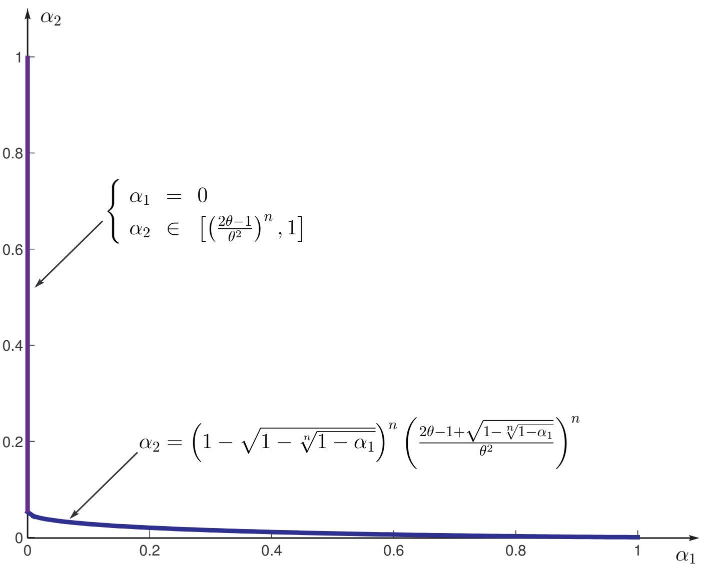

FIG. 3 – La courbe $\alpha_2 = \alpha_2(\alpha_1)$.

Exercice 2 (6 points) *Soit ξ une observation (l'échantillon de taille $n = 1$) de distribution $\mathcal{F}$, dont la densité de probabilité est f. On considère l'hypothèse de base $\mathcal{H}_1 = \{f = f_1\}$ et l'hypothèse concurrente $\mathcal{H}_2 = \{f = f_2\}$. Les densités f_1 et f_2 sont définies de la façon suivante :*

$$f_1(x) = \begin{cases} e^{1-x} & si \quad x \geq 1, \\ 0 & si \quad x < 1 \end{cases} \qquad f_2(x) = \begin{cases} 2e^{2(1-x)} & si \quad x \geq 1, \\ 0 & si \quad x < 1 \end{cases}$$

1. *Déterminer la statistique $\Lambda(\xi)$ du test δ du rapport de vraisemblance pour choisir entre $\mathcal{H}_1$ et $\mathcal{H}_2$. Écrire le test du rapport de vraisemblance (RV).*
2. *Calculer les risques α_1, α_2 en fonction du seuil h du test RV. La fonction $c \mapsto R(c) = \mathbb{P}_1(\Lambda(\xi) \geq c)$, est-elle continue sur $c > 0$?*
3. *Déterminer le risque α_2 en fonction de α_1 : $\alpha_1 \mapsto \alpha_2(\alpha_1)$ et dessiner (approximativement !) sa courbe representative en coordonnées $\alpha_1 O \alpha_2$.*
4. *Ce test $\delta(\xi)$, est-il optimal dans la classe $\mathcal{K}_\alpha$ en un sens quelconque ? Si la réponse est «oui», préciser dans quel sens ce test est optimal.*
5. *On suppose que les hypothèses simples $\mathcal{H}_1$ et $\mathcal{H}_2$ sont des événements aléatoires : $\mathbb{P}(\mathcal{H}_1) = p$, $\mathbb{P}(\mathcal{H}_2) = q$ et $p + q = 1$. Déterminer un test $\delta_Q(\xi)$ ($Q = (p, q)$) qui minimise le coût de Bayes $J_Q = p\alpha_1 + q\alpha_2$, où $\alpha_j = \mathbb{P}_j(\delta_Q \neq \mathcal{H}_j)$.*
6. *Déterminer le coût de Bayes $J_Q(p)$ en fonction de p.*

Réponses :

1) Le test du RV est

$$\delta(\xi) = \begin{cases} \mathcal{H}_2 & \text{si} \quad \Lambda(\xi) = 2\exp(1 - \xi) & \geq & h \\ \mathcal{H}_1 & \text{si} \quad \Lambda(\xi) & < & h \end{cases}$$

2) On commence avec le risque α_1

$$\begin{aligned}\alpha_1 &= \mathbb{P}_1\left(\Lambda(\xi) \geq h\right) = \mathbb{P}_1\left(2\exp\left(1-\xi\right) > h\right) = \mathbb{P}_1\left(2\exp\left(1-\xi\right) \geq h\right) = \mathbb{P}_1\left(1-\xi \geq \log\frac{h}{2}\right)\\ &= \mathbb{P}_1\left(\xi \leq 1-\log\frac{h}{2}\right) = \int_1^{1-\log\frac{h}{2}} \exp(1-x)dx = 1-\exp\left(\log\frac{h}{2}\right) = 1-\frac{h}{2}\end{aligned}$$

ensuite

$$\begin{aligned}\alpha_2 &= \mathbb{P}_2\left(\Lambda(\xi) < h\right) = \mathbb{P}_2\left(2\exp\left(1-\xi\right) < h\right) = \mathbb{P}_2\left(2\exp\left(1-\xi\right) < h\right) = \mathbb{P}_2\left(1-\xi < \log\frac{h}{2}\right)\\ &= \mathbb{P}_2\left(\xi > 1-\log\frac{h}{2}\right) = \int_{1-\log\frac{h}{2}}^{\infty} 2\exp[2(1-x)]dx = \exp\left[\log\left(\frac{h}{2}\right)^2\right] = \left(\frac{h}{2}\right)^2.\end{aligned}$$

La fonction $c \mapsto R(c) = \mathbb{P}_1(\Lambda(\xi) \geq c) = 1-\frac{c}{2}$ pour $c \in [0,2]$ est continue sur l'intervalle $[0,2]$.

3) Le risque α_2 en fonction de α_1 est défini par l'équation $\alpha_2(\alpha_1) = (1-\alpha_1)^2$. La courbe representative de cette fonction est une parabole.

4) C'est une «question - cours» : $\delta(\xi)$ est un test le plus puissant.

5) Le test de Bayes minimise le coût de Bayes

$$\delta(\xi) = \begin{cases} \mathcal{H}_2 & \text{si} \quad \Lambda(\xi) = 2\exp\left(1-\xi\right) \quad \geq \quad \dfrac{p}{q} \\ \mathcal{H}_1 & \text{si} \quad \Lambda(\xi) \quad < \quad \dfrac{p}{q} \end{cases}$$

6) Le coût de Bayes est

$$J_Q(p) = \begin{cases} p\left(1-\dfrac{p}{2(1-p)}\right) + (1-p)\left(\dfrac{p}{2(1-p)}\right)^2 = p - \dfrac{p^2}{4(1-p)} & \text{si} \quad p \in \left[0, \frac{2}{3}\right[\\ 1-p & \text{si} \quad p \in \left[\frac{2}{3}, 1\right] \end{cases}$$

Exercice 3 (5 points) *Soient $\Xi = (\xi_1, \ldots, \xi_n)$ des variables aléatoires i.i.d. issue d'une loi normale $\mathcal{N}(a,1)$ et $\Phi = (\phi_1, \ldots, \phi_m)$ des variables aléatoires i.i.d. issue d'une loi normale $\mathcal{N}(b,1)$. On veut tester l'hypothèse de base*

$$\mathcal{H}_1 = \{a = b\} \ \ \textit{contre l'hypothèse concurrente (composée !)} \ \ \mathcal{H}_2 = \{|a-b| > 1\}.$$

Introduisons le test («ad hoc») suivant :

$$\delta(\Xi, \Phi) = \begin{cases} \mathcal{H}_2 & si \quad |\overline{\xi} - \overline{\phi}| \quad > \quad 1 \\ \mathcal{H}_1 & si \quad |\overline{\xi} - \overline{\phi}| \quad \leq \quad 1 \end{cases}$$

où $\overline{\xi} = \frac{1}{n}\sum_{i=1}^n \xi_i$ et $\overline{\phi} = \frac{1}{m}\sum_{i=1}^m \phi_i$.

1. *Calculer les risques α_1 et α_2.*
2. *Étudier la limite de α_2 lorsque $n, m \to \infty$.*

Réponses :

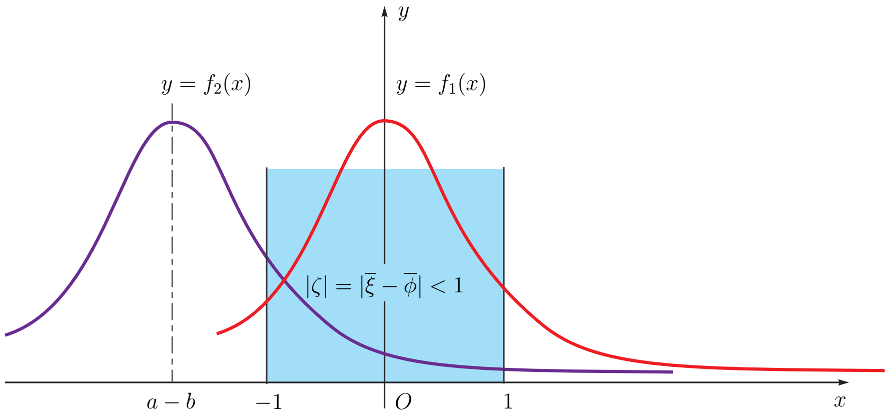

FIG. 4 – Les densités de probabilité sous $\mathcal{H}_1$ et $\mathcal{H}_2$ et le domaine d'acceptation de $\mathcal{H}_1$.

1) La statistique $\zeta = \overline{\xi} - \overline{\phi}$ suit une loi normale $\mathcal{N}(a-b, \frac{1}{n} + \frac{1}{m})$. Les densités de probabilité sous $\mathcal{H}_1$ et $\mathcal{H}_2$ et le domaine d'acceptation de $\mathcal{H}_1$ sont graphiquement représentés sur la figure 4. Le risque α_1 est calculé de la façon suivante :

$$\alpha_1 = \mathbb{P}_1(|\zeta| > 1) = \mathbb{P}_1\left(\left|\frac{\zeta}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\right| > \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\right) = 2\left[1 - \Phi\left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\right)\right]$$

Le risque α_2 est

$$\begin{aligned}\alpha_2 &= \mathbb{P}_{|a-b|>1}(|\zeta| < 1) = \mathbb{P}_{|a-b|>1}\left(\frac{-a+b-1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \leq \frac{\zeta - a + b}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \leq \frac{-a+b+1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\right) \\ &= \Phi\left(\frac{-a+b+1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\right) - \Phi\left(\frac{-a+b-1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\right).\end{aligned}$$

2) Soit $\varepsilon = |a-b| - 1 > 0$, alors $0 \leq \alpha_2 \leq 1 - \Phi\left(\frac{\varepsilon}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\right)$. Soit $k = \min\{n, m\}$. On a

$$0 \leq \alpha_2 \leq 1 - \Phi\left(\frac{\varepsilon}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\right) \leq 1 - \Phi\left(\frac{\varepsilon}{\sqrt{\frac{2}{k}}}\right).$$

On calcule la limite de $\Phi\left(\frac{\varepsilon}{\sqrt{\frac{2}{k}}}\right)$ lorsque $k \to \infty$. On obtient $\lim_{k\to\infty} \Phi\left(\frac{\varepsilon}{\sqrt{\frac{2}{k}}}\right) = 1$ et donc $\lim_{k\to\infty} \alpha_2(k) = 0$.