Université de Biskra
Facultés SESNV
Département de Mathématique

07/03/2021
Master 2 (Stat)
08h00–09h00

****

Examen – Module Statistique Non Paramétrique

****

Exercice 1 (10 points) : Soit X une variable aléatoire de densité inconnue f. Supposons que nous avons n observations $X_1, X_2, ..., X_n$ provenant de X. Soit $K : R \to R$ une fonction (un noyau), $h > 0$ (une fenêtre), l'estimateur à noyau de la densité f est

$$f_n(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right).$$

Considérons le problème d'estimation de la première dérivée de la densité :

$$f'(x) = f^{(1)}(x) = \frac{d}{dx}f(x).$$

Un estimateur naturel est donné en fonction de la dérivée de l'estimateur à noyau de f

$$f_n^{(1)}(x) = \frac{d}{dx}f_n(x) = \frac{d}{dx}\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right) = \frac{1}{nh^2}\sum_{j=1}^{n} K^{(1)}\left(\frac{X_j - x}{h}\right),$$

avec $K^{(1)}(t) = \frac{d}{dx}K(t) = K'(t)$.

De plus, la fonction noyau K est supposé (densité, bornée, symétrique, de moment d'ordre 4 fini) et que $K^{(1)}(t)$ existe et différent de 0, avec

$$\int t^2 K(t)dt < \infty, \quad \int t^2 |K(t)|\, dt < \infty \quad \text{et} \quad \int K^{(1)}(t)^2\, dt < \infty.$$

Nous supposons de plus, que f est 4 fois continûment différentiable. Alors,

$$\begin{aligned} i)\ \ Ef_n^{(1)}(x) &= f^{(1)}(x) + \frac{h^2}{2}f^{(3)}(x)\,\mu_2(K) + o\left(h^2\right). \\ ii)\ \ Var(f_n^{(1)}(x)) &= \frac{1}{nh^3}f(x)R(K^{(1)}) + O(\frac{1}{n}), \end{aligned}$$

avec $\mu_2(K) := \int t^2 K(t)dt$ et $R(g) := \int g^2(t)\, dt$.

1) Quelles sont les conditions que doit vérifier la fenêtre h, pour que $f_n^{(1)}$ converge vers $f^{(1)}$.

2) Donner l'expression de l'erreur quadratique moyenne asymptotique ($AMSE$) de cet estimateur $f_n^{(1)}$. En déduire l'$AMISE$.

3) Donner l'expression de la fenêtre optimale locale, notée h_{opt}.

4) Donner l'expression de la fenêtre optimale globale, notée h^*_{opt}.

5) Quelle est la vitesse de convergence de $f_n^{(1)}$. Comparer la avec celle de f_n qui vaut $O\left(n^{-4/5}\right)$.

Exercice 2 (10 points) :
On souhaite tester l'homogéniété et l'association entre deux variables aléatoires X et Y, dont on dispose de l'échantillon:

X : 45 33 38 30 32 47 54 60 82 79

Y : 34 39 29 44 37 62 55 74 101 87 65

Nous utilisons tout d'abord l'approche consistant à situer les valeurs avec la médiane empirique, calculée sur la globalité de l'échantillon.

1) Montrer qu'il faut utiliser un test non paramétrique.

2) Rappeler la statistique $M_{n,m}$ de la médiane et l'expression de $E(M_{n,m})$ et $Var(M_{n,m})$.

3) Appliquer le test de la médiane et conclure ($z_{0.975} = 1.96$).

Nous souhaitons mantenant, étudié l'association entre X et Y. Pour cela nous regroupons les données en 2 sous populations (A: données $\leq$ *médiane*) et (B : données $>$ *médiane*). Nous formons alors le tableau de contingence suivant:

	X	Y
A	a	b
B	c	d

4) Donner les valeurs de a, b, c et d, puis calculer le coefficient ϕ de Pearson et conclure.

Exercice 1 (10 points) : On a,

$$\begin{aligned} i)\ \ Ef_n^{(1)}(x) &= f^{(1)}(x) + \frac{h^2}{2} f^{(3)}(x)\,\mu_2(K) + o\left(h^2\right). \\ ii)\ \ Var(f_n^{(1)}(x)) &= \frac{1}{nh^3} f(x) R(K^{(1)}) + O(\frac{1}{n}), \end{aligned}$$

avec $\mu_2(K) := \int t^2 K(t)dt$ et $R(g) := \int g^2(t)\,dt$.

1) Pour que $f_n^{(1)}$ converge vers $f^{(1)}$, il faut que

$$Biais\left(f_n^{(1)}(x)\right) \to 0 \ \text{ et } Var(f_n(x)) \to 0 \quad \text{quand } n \to \infty \qquad \textbf{(1pt)}$$

Donc, la fenêtre h doit vérifier la condition

$$h \to 0, \quad nh^3 \to \infty \ \text{ quand } n \to \infty. \qquad \textbf{(1pt)}$$

2) Expression de l'$AMSE$:

$$\begin{aligned} AMSE\left(f_n^{(1)}(x)\right) &= ABiais^2\left(f_n^{(1)}(x)\right) + AVar(f_n^{(1)}(x)) \\ &= \frac{h^4}{4} f^{(3)}(x)^2 \mu_2^2(K) + \frac{1}{nh^3} f(x) R(K^{(1)}). \qquad \textbf{(1pt)} \end{aligned}$$

En déduire l'$AMISE$:

$$\begin{aligned} AMISE\left(f_n^{(1)}(x)\right) &= \int AMSE\left(f_n^{(1)}(x)\right) dx \\ &= \frac{h^4}{4} \mu_2^2(K) R(f^{(3)}) + \frac{1}{nh^3} R(K^{(1)}). \qquad \textbf{(1pt)} \end{aligned}$$

3) Expression de la fenêtre optimale locale, notée h_{opt} :

$$h_{opt} = Arg \min_h AMSE\left(f_n^{(1)}(x)\right), \qquad \textbf{(0.5 pt)}$$

c'est la solution de l'equation $\frac{d}{dh} AMSE\left(f_n^{(1)}(x)\right) = 0$,

$$\begin{aligned} h^3 f^{(3)}(x)^2 \mu_2^2(K) - \frac{3}{nh^4} f(x) R(K^{(1)}) &= 0 \Rightarrow h^7 f^{(3)}(x)^2 \mu_2^2(K) = \frac{3}{n} f(x) R(K^{(1)}) \\ &\Rightarrow h^7 = \frac{3 f(x) R(K^{(1)})}{f^{(3)}(x)^2 \mu_2^2(K)} n^{-1} \\ &\Rightarrow h_{opt} = \left(\frac{3 f(x) R(K^{(1)})}{f^{(3)}(x)^2 \mu_2^2(K)}\right)^{1/7} n^{-1/7}. \qquad \textbf{(1.5 pt)} \end{aligned}$$

4) Expression de la fenêtre optimale globale, notée h_{opt}^*. De même,

$$h_{opt}^* = Arg \min_h AMISE\left(f_n^{(1)}(x)\right) = \left(\frac{3R(K^{(1)})}{R(f^{(3)}) \mu_2^2(K)}\right)^{1/7} n^{-1/7}. \qquad \textbf{(1.5 pt)}$$

5) Vitesse de convergence de $f_n^{(1)}$: Injectons h_{opt}^* dans la $AMISE\left(f_n^{(1)}(x)\right)$,

$$\begin{aligned}
AMISE\left(f_n^{(1)}(x)\right) &\leq AMISE\left(f_n^{(1)}(x), h_{opt}^*\right) = \frac{h_{opt}^{*4}}{4}\mu_2^2(K)R(f^{(3)}) + \frac{1}{nh_{opt}^{*3}}R(K^{(1)}) \\
&= \left(\frac{3R(K^{(1)})}{R(f^{(3)})\mu_2^2(K)}\right)^{4/7} n^{-4/7}\mu_2^2(K)R(f^{(3)}) + n^{-1}R(K^{(1)})\left(\frac{3R(K^{(1)})}{R(f^{(3)})\mu_2^2(K)}\right)^{-3/7} n^{3/7} \\
&= \left\{3R(K^{(1)})^4 R(f^{(3)})^3\mu_2^2(K)^3\right\}^{1/7} n^{-4/7} \\
&= O\left(n^{-4/7}\right).
\end{aligned}$$

(1.5 pt)

Comme celle de f_n vaut $O\left(n^{-4/5}\right) < O\left(n^{-4/7}\right)$. Alors, f_n converge plus rapidement que $f_n^{(1)}$. **(1 pt)**

Exercice 2 (10 points) :

I) Test de la médiane:

1) Il faut utiliser un test non paramétrique car la distribution des données est inconnue. **(1 pt)**

2) La statistique $M_{n,m}$ de la médiane est:

$$M_{n,m} = \frac{1}{n}\sum_{j=1}^{n} 1_{\left(R(X_j) > \frac{N+1}{2}\right)}.$$ **(1 pt)**

telle que $R(X_j)$ est le rang de la $j^{ème}$ observation de X. De plus, sous H_0 et pour $N = 21 = 2k+1$ (*impair*):

$$E(M_{n,m}) = \frac{N-1}{2N} = \frac{20}{42} = 0.476, \qquad Var(M_{n,m}) = \frac{n(N+1)}{2mN^2} = \frac{10(22)}{2(11)(21)^2} = 0.023.$$ **(2 pts)**

3) Application du test de la médiane: calcul des rangs :

$$\begin{aligned}
(X,Y)_{(i)} &= 29, \mathbf{30}, \mathbf{32}, \mathbf{33}, 34, 37, \mathbf{38}, 39, 44, \mathbf{45}, \quad \mathbf{47}, \quad \mathbf{54}, 55, \mathbf{60}, 62, 65, 74, \mathbf{79}, \mathbf{82}, 87, 101 \\
R(X) &= 2,\ 3,\ 4,\ 7,\ 10,\ 11,\ 12,\ 14,\ 18,\ 19
\end{aligned}$$ **(1 pt)**

Alors,

$$M_{n,m} = \frac{1}{10}\sum_{j=1}^{10} 1_{(R(X_j)>11)} = \frac{4}{10} = 0.4$$ **(1 pt)**

et

$$\frac{M_{n,m} - E(M_{n,m})}{\sqrt{Var(M_{n,m})}} = \frac{0.4 - 0.476}{\sqrt{0.023}} = 0.5 < z_{0.975} = 1.96$$

Donc H_0 est acceptée (les variables sont homogènes). **(1 pt)**

II) Association entre X et Y :

1) Tableau de contingence (A: données $\leq Med = 47$) et (B : données $> Med = 47$):

	X	Y
A	$a = 6$	$b = 5$
B	$c = 4$	$d = 6$

(2 pts)

Le coefficient de ϕ de Pearson est

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(b+d)(c+d)(a+c)}} = \frac{6(6) - 5(4)}{\sqrt{11(11)(10)(10)}} = 0.145.$$ **(0.5 pt)**

Ce qui implique l'absence de liaison (association) entre les deux groupes A et B. **(0.5 pt)**