Université Mohamed Khider - Biskra
Département de Mathématiques
Master 2 - Option : Statistique

Examen module : Statistique Non Paramétrique (2020)

Exercice 1 (10 points) : Supposons que la densité f de l'échantillon $X_1, ..., X_n$ d'une variable aléatoire X est deux fois continûment différentiable et s'annule en dehors de l'intervalle $[0,1]$. Choisissons une partition uniforme $C_1, ..., C_m$ de l'intervalle $[0,1[$:

$$C_j = \left[\frac{j-1}{m}, \frac{j}{m}\right[, \quad j = 1, ..., m$$

Comme f est supposée être continue, pour m suffisamment grand, elle est bien approchée par des fonctions en escalier, constantes par morceaux sur les intervalles $\{C_j\}$. On pose $h = 1/m$ et on approche f par la fonction

$$f_h(x) = \sum_{j=1}^{m} \frac{p_j}{h} 1_{C_j}(x), \qquad \text{avec } p_j = \int_{C_j} f(x)\, dx.$$

On ramène ainsi le problème d'estimation de f au problème d'estimation d'un paramètre $m-$dimensionnel $p = (p_1, ..., p_m)$. Ceci peut se faire en utilisant, par exemple la méthode généralisée des moments.
1) Montrer que $\widehat{p}_j = \frac{1}{n}\sum_{k=1}^{n} 1_{C_j}(X_k)$ est un estimateur de p_j.
Par substitution, nous définissons l'estimateur de f par histogramme à m classes comme suit :

$$\widehat{f}_h(x) = \frac{1}{h}\sum_{j=1}^{m} \widehat{p}_j 1_{C_j}(x).$$

2) Vérifier que l'estimateur par histogramme $\widehat{f}_h$ est une densité de probabilité.
Introduisons le risque quadratique de $\widehat{f}_h$ au point $x \in [0,1]$ comme étant la moyenne de l'erreur quadratique :

$$MSE\left(\widehat{f}_h\right) = E\left(\left(\widehat{f}_h - f\right)^2\right) = biais^2\left(\widehat{f}_h\right) + var\left(\widehat{f}_h\right).$$

3) Soit j l'indice de la classe contenant x ; $x \in C_j$. Montrer que

$$\widehat{f}_h(x) = \frac{\widehat{p}_j}{h} = \frac{1}{nh}\sum_{k=1}^{n} 1_{C_j}(X_k) = \frac{Z_j}{nh}$$

où Z_j est un variable aléatoire de loi à déterminer.
4) Caculer pour tout $x \in C_j$, $E\left(\widehat{f}_h\right)$ et $var\left(\widehat{f}_h\right)$.
5) En déduire la $MISE\left(\widehat{f}_h\right)$.
6) Si la fenêtre h de l'estimateur par histogramme $\widehat{f}_h$ est telle que $h_n \to 0$ lorsque $n \to \infty$. Sachant que le risque quadratique intégré peut s'écrire

$$MISE\left(\widehat{f}_h\right) = \frac{h^2}{12}\int_0^1 f'(x)^2\, dx + \frac{1}{nh} + O\left(h^3\right) + O\left(1/n\right), \quad \text{lorsque } n \to \infty.$$

Montrer que la valeur de h_{opt} associée à la $AMISE\left(\widehat{f}_h\right)$ est $h_{opt} = cn^{-1/3}$ avec c est une constante à déterminer.

Exercice 2 (05 points) : Soient les observations de deux variables aléatoires X et Y :

X :	**1.6**	**3.4**	**2.7**	**1.9**	**0.5**	**4.3**	**1.9**	**2.7**	**0.5**	
Y :	**2.5**	**2.3**	**0.9**	**0.8**	**2.7**	**4.4**	**7.1**	**1.6**	**1.9**	**3.8**	**2.1**

Tester à 95% l'hypothèse d'égalité des lois de X et Y :

$$H_0\left(F_X = F_Y\right) \qquad \textit{contre} \qquad H_1\left(F_X \neq F_Y\right)$$

en utilisant le test des sommes des rangs de Wilcoxon, dont la statistique est

$$W_{n,m} = \sum_{i=1}^{n} R\left(X_i\right), \quad \text{sachant que } E\left(W\right) = \frac{n\left(N+1\right)}{2}, \;\; Var\left(W\right) = \frac{nm\left(N+1\right)}{12}.$$

Exercice 3 (05 points) : On veut étudier la liaison entre les caractères : «être fumeur» (plus de 20 cigarettes par jour, pendant 10 ans) et «avoir un cancer de la gorge», sur une population de 1000 personnes, dont 500 sont atteintes d'un cancer de la gorge. Voici les résultats observés :

Observé	**cancer**	**non cancer**
fumeur	**342**	**258**
non fumeur	**158**	**242**

1) Faire un test d'indépendance de Khi -2 à 95% pour établir la liaison entre ces caractères.

2) Quelle est la mesure d'association qu'on peut utilisée ici ? Que vaut sa valeur ? Qu'en déduisez-vous?

$$z_{0.95} = 1.64, \qquad z_{0.975} = 1.96,$$

$$\chi^2_{(0.95)}\left(1ddl\right) = 0,0039, \qquad \chi^2_{(0.95)}\left(2ddl\right) = 0,1026, \qquad \chi^2_{(0.95)}\left(3ddl\right) = 0,3518$$

Bonne chance

Master 2 : Statistique
Corrigé-Type Examen module : Statistique Non Paramétrique

Exercice 1 (10 points) : 1) On a

$$p_j = \int_{C_j} f(x)\,dx = \int_R 1_{C_j} f(x)\,dx = E\left(1_{C_j(X)}\right), \quad \text{donc } \hat{p}_j = \frac{1}{n}\sum_{k=1}^{n} 1_{C_j}(X_k)$$

2) Il est clair que $\hat{f}_h \succ 0$. De plus,

$$\begin{aligned}\int \hat{f}_h(x)\,dx &= \int \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j 1_{C_j}(x)\,dx = \sum_{j=1}^{m}\frac{\hat{p}_j}{h}\int_{C_j} dx = \sum_{j=1}^{m}\frac{\hat{p}_j}{h}\frac{1}{m} = \sum_{j=1}^{m}\hat{p}_j \\ &= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{n} 1_{C_j}(X_k) = \frac{n}{n} = 1. \text{ Donc, } \hat{f}_h \text{ est une densité.}\end{aligned}$$

3) Remarquons que, $\forall x \in C_j$:

$$\hat{f}_h(x) = \frac{\hat{p}_j}{h} = \frac{1}{nh}\sum_{k=1}^{n} 1_{C_j}(X_k) = \frac{Z_j}{nh}$$

où Z_j est une variable aléatoire de loi Binomiale $\beta(n, p_j)$, car Z_j est la somme de n variables indépendantes de loi de Bernoulli de paramètre

$$P\left(1_{C_j}(X) = 1\right) = P(X \in C_j) = \int_{C_j} f(x)\,dx = p_j.$$

4) Alors, pour tout $x \in C_j$:

$$E\left(\hat{f}_h\right) = \frac{p_j}{h}, \qquad var\left(\hat{f}_h\right) = \frac{p_j(1-p_j)}{nh^2}.$$

5) On a $MISE\left(\hat{f}_h\right) = \int MSE\left(\hat{f}_h\right) = \int biais^2\left(\hat{f}_h\right)dx + \int var\left(\hat{f}_h\right)dx$, avec

$$\begin{aligned}\int biais^2\left(\hat{f}_h\right)dx &= \int\left(E\left[\hat{f}_h\right] - f\right)^2 dx = \sum_{j=1}^{m}\int_{C_j}\left(\frac{p_j}{h} - f(x)\right)^2 dx \\ &= \sum_{j=1}^{m}\frac{p_j^2}{h} - 2\frac{p_j}{h}\int_{C_j} f(x)\,dx + \int f^2(x)\,dx = \int f^2(x)\,dx - \sum_{j=1}^{m}\frac{p_j^2}{h}.\end{aligned}$$

et

$$\int var\left(\hat{f}_h\right)dx = \sum_{j=1}^{m}\int_{C_j} var\left(\hat{f}_h\right)dx = \sum_{j=1}^{m}\frac{p_j(1-p_j)}{nh} = \frac{1}{nh} - \frac{1}{nh}\sum_{j=1}^{m} p_j^2.$$

6) Sachant que le risque quadratique intégré peut s'écrire

$$MISE\left(\hat{f}_h\right) = \frac{h^2}{12}\int_0^1 f'(x)^2\,dx + \frac{1}{nh} + O\left(h^3\right) + O(1/n), \quad \text{lorsque } n \to \infty.$$

Alors, par minimisation,

$$h_{opt} = cn^{-1/3} = \left(\frac{1}{6}\int_0^1 f'(x)^2\,dx\right)^{-1/3} n^{-1/3}.$$

Exercice 2 (05 points) : Le vecteur (X, Y) ordonné est :

$$0.5, 0.5, 0.8, 0.9, 1.6, 1.6, 1.9, 1.9, 1.9, 2.1, 2.3, 2.5, 2.7, 2.7, 2.7, 3.4, 3.8, 4.3, 4.4, 7.1$$

Ce vecteur contient des exo-equax, alors le vecteur des rangs est

$$1, 1, 3, 4, 5.5, 5.5, 8, 8, 8, 10, 11, 12, 14, 14, 14, 16, 17, 18, 19, 20$$

Le vecteur des rangs de X est :

$$R(X_i) : 1, 1, 5.5, 8, 8, 14, 14, 16, 18$$

en utilisant le test des sommes des rangs de Wilcoxon,

$$W_{n,m} = \sum_{i=1}^{n} R(X_i) = 85.5.$$

avec

$$E(W) = \frac{n(N+1)}{2} = \frac{9 \times 21}{2} = 94.5, \qquad Var(W) = \frac{nm(N+1)}{12} = \frac{9 \times 11 \times 21}{12} = 173.25$$

$$\frac{W_{n,m} - E(W)}{\sigma(W)} = \frac{85.5 - 94.5}{\sqrt{173.25}} = 0.68 \prec z_{0.975} = 1.96$$

Alors, à 95% l'hypothèse d'égalité des lois de X et Y est acceptée $H_0(F_X = F_Y)$.

Exercice 3 (05 points) : 1) Test d'indépendance de Khi -2 à 95% :

$$\begin{aligned} Q_{ind} &= \sum\sum \frac{\left(\frac{N_{i*}N_{*j}}{n} - N_{ij}\right)^2}{\frac{N_{i*}N_{*j}}{n}} \\ &= \frac{\left(\frac{600\times500}{1000} - 342\right)^2}{\frac{600\times500}{1000}} + \frac{\left(\frac{600\times500}{1000} - 258\right)^2}{\frac{600\times500}{1000}} + \frac{\left(\frac{400\times500}{1000} - 158\right)^2}{\frac{400\times500}{1000}} + \frac{\left(\frac{400\times500}{1000} - 242\right)^2}{\frac{400\times500}{1000}} \\ &= 29.4 \succ \chi^2_{(0.95)}(1ddl) = 0,0039. \end{aligned}$$

On décide donc de rejeter H_0. Ainsi, il y'a dépendance à 95% entre «être fumeur» et «avoir un cancer de la gorge».

2) La mesure d'association qu'on peut utilisée ici est le coefficient ϕ, car il s'agit d'un tableau 2×2.

$$\phi = \sqrt{\frac{Q_{ind}}{n}} = \sqrt{\frac{29.4}{1000}} = 0.17.$$

Il s'agit donc, d'une faible dépendance positif.

Université Mohamed Khider - Biskra
Département de Mathématiques
Master 2 - Option : Statistique

Examen de rattrapage S1 module : Statistique Non Paramétrique (2020)

Exercice 1 (06 points) : Considérons l'échantillon $X_1, ..., X_n$ d'une variable aléatoire $X \in \mathbb{R}$, de densité f.
1) Montrer que l'estimateur par noyau $\widehat{f}_n$ de la densité de probabilité f est une densité.

2) Soit la variable $Y = G(X)$, où G est une distribution (inversible) de densité g. Donner l'estimateur par noyau de la densité de Y en fonction de celle de X.

Exercice 2 (07 points) : Supposons que la densité f de l'échantillon $X_1, ..., X_n$ d'une variable aléatoire X est deux fois continûment différentiable et s'annule en dehors de l'intervalle $[0, 1]$. Choisissons une partition uniforme $C_1, ..., C_m$ de l'intervalle $[0, 1[$:

$$C_j = \left[\frac{j-1}{m}, \frac{j}{m}\right[, \quad j = 1, ..., m$$

Nous définissons l'estimateur de f par histogramme à m classes comme suit :

$$\widehat{f}_h(x) = \frac{1}{h}\sum_{j=1}^{m} \widehat{p}_j 1_{C_j}(x), \qquad h = 1/m \to 0 \text{ lorsque } n \to \infty.$$

avec $\widehat{p}_j = \frac{1}{n}\sum_{k=1}^{n} 1_{C_j}(X_k)$ est un estimateur de $p_j = \int_{C_j} f(x)\,dx$.
Montrer que le risque quadratique intégré de $\widehat{f}_h$ est

$$MISE\left(\widehat{f}_h\right) = \frac{h^2}{12}\int_0^1 f'(x)^2\,dx + \frac{1}{nh} + O\left(h^3\right) + O(1/n), \quad \text{lorsque } n \to \infty.$$

En déduire la valeur de h_{opt} associée à $AMISE\left(\widehat{f}_h\right)$

Exercice 3 (07 points) : On veut comparer le salaire des femmes et des hommes dans une grande entreprise. On prend un échantillon de 120 hommes et un échantillon de 150 femmes puis on note le salaire selon "faible", "moyen" et "élevé". On observe:

	faible	moyen	élevé
Homme	**10**	**70**	**40**
Femme	**30**	**60**	**60**

Peut-on dire avec un niveau de 5% que les hommes et les femmes ont un niveau de salaire différent.