École Normale Supérieure
FIMFA — Statistique 2014–2015

Examen

Exercice 1
Dans cet exercice, $X_1, \ldots, X_n$ sont i.i.d. selon une uniforme sur $[a, b]$ ($a < b$ inconnus). On suppose $n > 3$.

1. Préciser le modèle statistique sous-jacent. Est-il identifiable ? dominé ? exponentiel ?
2. Proposer un estimateur au maximum de vraisemblance $(\widehat{a}, \widehat{b})$ de (a, b) si vous avez répondu que le modèle est dominé.
3. Biais et risque quadratique de $\widehat{a}$.
4. Précisez la loi limite de $\widehat{a} - a$ après renormalisation (on attend une loi non-dégénérée).
5. Intervalle de confiance asymptotique pour $\widehat{a}$.
6. Calculer l'information de Fisher dans ce modèle.
7. Proposer un estimateur sans biais de $b - a$. Quel est son risque quadratique ?
8. La borne de Cramer-Rao est-elle pertinente ? Expliquer.
9. Proposer un estimateur $\overline{a}$ de a fondé sur la méthode des moments. Quel est son risque quadratique ? Quelle est sa loi limite après recentrage et renormalisation ?
10. Analyse bayésienne : on suppose désormais qu'on connaît $a = 0$, et que seul b est à estimer. Montrer que la famille des lois de Pareto $Pareto(\alpha, x_m)$ de densité

$$f_{\alpha, x_m}(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \mathbb{I}_{x \geq x_m}$$

est conjugée pour ce modèle. Donner l'espérance a posteriori de b.

Exercice 2
Dans cet exercice $X_1, \ldots, X_n$ sont i.i.d. selon une loi P sur $\mathbb{R}$ (de fonction de répartition F). On veut développer un test pour distinguer

— H_0 : P est une loi uniforme sur un intervalle $[a, b]$ ($a < b$ inconnus).
— H_1 : P n'est pas une loi uniforme sur un intervalle de longueur positive.

On note $X_{(1)} \leq \ldots \leq X_{(n)}$ les statistiques d'ordre de l'échantillon.

1. Si P est la loi uniforme sur $[a, b]$, quelle est la loi de $(X_i - a)/(b - a)$?
2. Proposez une statistique de test S pour H_0 contre H_1. Vous expliquerez comment garantir un niveau $\alpha \in]0, 1[$ et comment calculer la p-valeur (ou dégré de signifcation atteint) à partir de S.

3. Montrer que si, pour tout $x < y$ tel que $P\{]x, y[\} > 0$, la loi conditionnelle de X_1 sachant $X_1 \in]x, y[$ est uniforme sur un sous-intervalle de $]x, y[$, alors P est une loi uniforme sur un intervalle.
4. Le test développé à partir de la statistique de test S est-il consistant ?
5. On considère Q_h uniforme sur $[a, m-h] \cup [m+h, b]$ avec $m = (a+b)/2$ et $0 < h < (b-a)/2$. Quelle est la distance en variation de Q_h aux lois uniformes sur les intervalles ? Quelle est l'information de Kullback de Q_h aux lois uniformes sur les intervalles ?
6. Vous avez développé un test de niveau α pour tester H_0 contre H_1 sur des échantillons de taille n. Minorer la probabilité d'erreur de seconde espèce de ce test, lorsque les données sont distribuées selon Q_h.

Exercice 3

On observe n variables aléatoires binaires identiquement distribuées $(X_1, \ldots, X_n)$ à valeurs dans $\{0, 1\}$; on souhaite tester l'échangeabilité [1] de ces variables. On note N_0 le nombre d'occurrences de la valeur 0 et N_1 le nombre d'occurrences de la valeur 1 :

$$N_0 = \sum_{i=1}^{n} \mathbb{I}_{X_i=0}; \qquad N_1 = \sum_{i=1}^{n} \mathbb{I}_{X_i=1}; \qquad N_0 + N_1 = n.$$

Au vu des observations $x_1, \ldots, x_n$, on appelle *suite* un ensemble d'indices successifs $\{j, j+1, j+2 \ldots, k\}$ tel que

- $x_j = x_{j+1} = \ldots = x_k$
- si $j > 1$ alors $x_j \neq x_{j-1}$
- si $k < n$ alors $x_k \neq x_{k+1}$

et on note S le nombre de suites. Une suite correspond donc à un bloc d'observations successives qui prennent toutes la même valeur. Par exemple, pour les observations

1111001111000001

le nombre de suites est 5.

1. Montrer que si les (X_i) sont échangeables alors la loi conditionnelle de S sachant N_0 et N_1 est libre de loi de X. (On ne demande pas d'expliciter cette loi ; elle est tabulée ci-dessous pour certaines valeurs.)
2. En déduire un test de H_0 : "les variables $(X_1, \ldots, X_n)$ sont échangeables" contre l'hypothèse H_1 de votre choix. Ce test s'appelle test des suites de Wald-Wolfowitz.
3. *Application 1 :* Sur les 30 jours du mois d'avril 2013, on a noté 1 si on a observé de la pluie à la station Saint-Germain-des-Prés à Paris, et 0 si on n'a pas observé de pluie. Les données collectées sont :

   000110001111100001110000010001

   soit 12 jours de pluie au total. À quel niveau rejetez-vous l'hypothèse d'échangeabilité ?
   (Source des données : Météo Paris)

1. Une collection de variables aléatoires $X_1, \ldots, X_n$ est échangeable si pour toute permutation π agissant sur $\{1, \ldots, n\}$, $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ a même loi que $X_1, \ldots, X_n$.

4. On se donne deux échantillons de variables aléatoires continues $Y_1, \ldots, Y_m$ et $Z_1, \ldots, Z_p$. On souhaite adapter le test de Wald-Wolfowitz pour tester l'hypothèse H_0 : "Y et Z proviennent de la même distribution". Pour cela, on note $W_1, \ldots, W_{m+p}$ l'agrégation des deux échantillons (Y_i) et (Z_j), et on ordonne les W_k. On note alors

$$X_k = \begin{cases} 0 & \text{si } \exists i : W_{(k)} = Y_i \\ 1 & \text{si } \exists j : W_{(k)} = Z_j \end{cases}$$

On construit la v.a. S comme précédemment. Justifier l'utilisation de S pour tester l'hypothèse H_0.

5. *Application 2 :* On souhaite comparer l'effet de deux somnifères, notés A et B. Pour ce faire, on fait appel à 17 cobayes. Chaque cobaye se voit administrer l'un des deux somnifères, et on mesure l'augmentation (ou la diminution) du temps de sommeil en heures par rapport à une nuit sans somnifère. Les observations sont :

Traitement A	-1.6	-1.2	-0.2	-0.1	0.0	0.7	2.0	3.4	3.7
Traitement B	0.1	0.8	1.1	1.6	1.9	4.4	4.6	5.5	

(Source des données : Cushny & Peebles 1905).

On souhaite tester l'hypothèse H_0 : "les deux traitements ont le même effet" à l'aide du test de Wald-Wolfowitz. Au niveau $\alpha = 0.05$, rejetez-vous H_0 ?

Tables statistiques

On donne ci-dessous la fonction de répartition pour :
- $S_{18,12}$, la statistique de Wald-Wolfowitz pour $N_0 = 18$ et $N_1 = 12$
- $S_{9,8}$, la statistique de Wald-Wolfowitz pour $N_0 = 9$ et $N_1 = 8$

x	$P(S_{18,12} \leq x)$	$P(S_{9,8} \leq x)$
4	0.000	0.005
5	0.000	0.021
6	0.000	0.069
7	0.001	0.158
8	0.004	0.318
9	0.011	0.499
10	0.030	0.701
11	0.066	0.842
12	0.132	0.939
13	0.233	0.980
14	0.364	0.996
15	0.514	0.999
16	0.661	1
17	0.791	1
18	0.994	-
19	0.947	-
20	0.978	-
21	0.993	-
22	0.998	-
23	0.999	-

Exercice 4
Dans cet exercice, on considère un modèle exponentiel univarié où

$$p_\theta(x) = \exp\left(\theta T(x) - \eta(\theta)\right)$$

est la densité de P_θ par rapport une mesure de probabilité ν sur $\mathbb{R}$ (T est une fonction monotone, donc mesurable de $\mathbb{R}$ dans $\mathbb{R}$). On définit Θ comme l'intervalle d'intérieur non vide sur lequel $\int_{\mathbb{R}} \exp(\theta T(x))\nu(\mathrm{d}x) < \infty$. Dans la suite θ_0 appartient à l'intérieur de Θ.
Dans la suite $X_1, \ldots, X_n$, sont i.i.d. selon P_θ, $\theta \in \Theta$ inconnu.
On veut développer un test pour distinguer $H_0 : \theta \leq \theta_0$ et $H_1 : \theta > \theta_0$.

1. Proposer une statistique de test S pour distinguer P_{θ_0} de $P_{\theta'}$ ($\theta' > \theta_0$). Préciser la forme de la région critique choisie pour obtenir un niveau donné.
2. Si maintenant vous souhaitez utiliser la statistique S pour distinguer H_0 de H_1, comment choisir une région critique pour obtenir un niveau donné ?
3. Si $\theta_1 > \theta'$, pouvez-vous concevoir un test pour distinguer P_{θ_0} de P_{θ_1} de niveau α, mais plus puissant en θ_1 que le test générique développé pour tester H_1 contre H_0 ? Justifier.

Exercice 5
Dans cet exercice $\theta \in \Theta =]0,1[$, et P_θ est la loi de Bernoulli de paramètre θ. Θ est muni d'une loi a priori de densité π. Si Q est une loi sur $\{0,1\}^n$, le risque entropique cumulé de Q par rapport à P_θ, est

$$D(P_\theta^{\otimes n}, Q) = \mathbb{E}_{P_\theta^{\otimes n}}\left[\log \frac{P_\theta^{\otimes n}(X_1, \ldots, X_n)}{Q(X_1, \ldots, X_n)}\right],$$

c'est à dire l'entropie relative de $P_\theta^{\otimes n}$ par rapport à Q. Le risque moyen de Q sous la loi a priori π est

$$\int_{[0,1]} D(P_\theta^{\otimes n}, Q)\pi(\theta)\mathrm{d}\theta\,.$$

1. Montrer que pour l'a priori π, le risque moyen est minimisé par Q_n^* définie par $Q_n^*(x_1, \ldots, x_n) = \int_{[0,1]} P_\theta^n(x_1, \ldots, x_n)\pi(\theta)\mathrm{d}\theta$ pour $x_1 \ldots, x_n \in \{0,1\}^n$.
2. La distribution $\widehat{Q}_n$ est définie par

$$\widehat{Q}_n(x_1, \ldots, x_n) = \frac{\sup_{\theta\in[0,1]} P_\theta^{\otimes n}\{x_1, \ldots, x_n\}}{Z_n}$$

où $Z_n = \sum_{x_1,\ldots,x_n} \sup_{\theta\in[0,1]} P_\theta^{\otimes n}\{x_1, \ldots, x_n\}$.
Pour $\theta \in \Theta$, sous $P_\theta^{\otimes n}$, quelle est la limite de $D(P_\theta^{\otimes n}, \widehat{Q}_n) - \log Z_n$ lorsque n tends vers l'infini ?
3. Montrer que, toujours sous $P_\theta^{\otimes n}$, $\log Z_n \sim \frac{1}{2}\log n$ lorsque $n \leftarrow \infty$ [2].

2. On peut utiliser l'encadrement de Stirling

$$(n/\mathrm{e})^n\sqrt{2\pi n} \leq n! \leq (n/\mathrm{e})^n\sqrt{2\pi n}\exp(1/(12(n+1)))$$