Examen du 28 Novembre 2008 (Durée 3h)
Calculatrice, cours et poly autorisés

Statistiques - 2008

Le problème et l'exercice qui suivent sont totalement indépendants et peuvent être traités dans l'ordre que vous voulez. L'exercice est purement théorique, le problème nécessite l'usage par moment de la calculatrice. Dans tous les cas, vous pouvez admettre le résultat de certaines questions pour répondre aux suivantes du moment que vous l'indiquez clairement.

Exercice

On observe les réalisations de $X_1, ..., X_n$, n variables aléatoires réelles iid de loi exponentielle de paramètre inconnu $\theta > 0$. On rappelle que la densité de X_1 est donc $\theta e^{-\theta x}\mathbf{1}_{x>0}$. Soit $\alpha \in]0, 1[$.

1. Soit $\mu > \lambda > 0$. Ecrire le test du rapport de vraisemblance au niveau α de H_0: "$\theta = \lambda$" contre H_1: "$\theta = \mu$" sous la forme la plus simple possible.
2. On rappelle que la moyenne empirique est donnée par $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Redémontrer, dans ce cas précis, que le test, Φ, qui rejette H_0 quand $\bar{X} < c_\alpha$ avec $\mathbb{P}_\lambda(\bar{X} < c_\alpha) = \alpha$ est uniformément le plus puissant parmi tous les tests de niveau α de H_0: "$\theta = \lambda$" contre H_1: "$\theta = \mu$".
3. Montrer qu'un test Δ de niveau α est uniformément le plus puissant parmi les tests de niveau α de H_0: "$\theta = \lambda$" contre H_1: "$\theta = \mu$" si et seulement si il est égal p.s. au test Φ.
4. Montrer que Φ est un test uniformément le plus puissant de H_0: "$\theta = \lambda$" contre H_1: "$\theta > \lambda$" parmi les tests de niveau α de ces mêmes hypothèses.
5. Montrer que Φ est un test de H_0: "$\theta \leq \lambda$" contre H_1: "$\theta > \lambda$" de niveau α et qu'il est uniformément le plus puissant parmi les tests de niveau α de ces mêmes hypothèses.
6. Montrer que l'on ne peut pas construire de test de niveau α uniformément le plus puissant parmi les tests de niveau α de H_0: "$\theta = \lambda$" contre H_1: "$\theta \neq \lambda$".

Problème

A une élection deux candidats se présentent : N.S. et S.R.
500 électeurs sont sondés. On leur demande non seulement leur opinion politique (ie leur choix à la future élection) mais aussi leur appartenance sociale (Rentiers/Actifs/Retraités). On obtient les résultats suivants :

	N.S.	S.R.
Rentiers	40	10
Actifs	140	210
Retraités	77	23

1. L'opinion politique dépend-elle de l'appartenance sociale ? *Seront notés : le choix du test, sa présentation claire d'un point de vue théorique, son calcul pratique et la conclusion reposant sur une valeur approchée de la p-valeur correspondante.*

2. On oublie pour cette question seulement l'appartenance sociale des sondés (ie on observe uniquement que 257 sondés se prononcent pour N.S. et 243 se prononcent pour S.R.). On s'intéresse au paramètre inconnu θ qui représente la proportion de gens votant N.S., le jour du scrutin.

   (a) Donner précisément le modèle statistique de Bernoulli dans lequel vous allez travailler.

   (b) Calculer l'estimateur, $\hat{\theta}$, du maximum de vraisemblance dans ce modèle.

   (c) Montrer que $\hat{\theta}$ est sans biais et consistant.

   (d) Montrer qu'il est efficace.

   (e) A l'aide du Lemme de Slutsky, proposer un intervalle de confiance asymptotique au niveau 5% sur θ.

3. Les sondeurs se servent de strates (ici l'appartenance sociale) pour améliorer leur intervalle de confiance : on parle de sondage stratifié. D'un point de vue théorique, on dispose de k strates numérotées de 1 à k et l'on observe n couples iid (Y_i, X_i) où Y_i est à valeurs dans $\{0, 1\}$ et X_i est à valeurs dans $\{1, ..., k\}$. Pour le sondage effectué ici, Y_i vaut 1 si le $i$ème sondé vote N.S. et $X_i = j$ si le $i$ème sondé appartient à la strate j (par exemple $j = 2$ correspond à la strate "Actifs"). On pose $w_j = \mathbb{P}(X_i = j)$, la proportion supposée **connue** de gens dans la strate j et $\rho_j = \mathbb{P}(Y_i = 1|X_i = j)$, l'intention de vote de la strate j supposée **inconnue**. On note ρ le vecteur colonne formé des ρ_j.

   (a) Montrer que la log-vraisemblance s'écrit dans ce modèle :

$$\ell(\rho, (Y_1, X_1), ...(Y_n, X_n)) = \sum_{j=1}^{k} \left[C_{n,j} \log(\rho_j) + (N_{n,j} - C_{n,j}) \log(1 - \rho_j) + N_{n,j} \log(w_j)\right],$$

   où $C_{n,j} = \sum_{i=1}^{n} \mathbf{1}_{Y_i=1, X_i=j}$ et $N_{n,j} = \sum_{i=1}^{n} \mathbf{1}_{X_i=j}$.

   (b) Calculer $\hat{\rho}$ l'estimateur du maximum de vraisemblance de ρ.

   (c) Montrer que

$$\sqrt{n}(\hat{\rho} - \rho) \to^{\mathcal{L}}_{n\to\infty} \mathcal{N}_k\left(0, \Sigma\right)$$

   où Σ est la matrice diagonale de termes diagonaux $\Sigma_{j,j} = \frac{\rho_j(1-\rho_j)}{w_j}$.
   On pourra remarquer sans le redémontrer que le Lemme de Slutsky reste vrai pour des vecteurs. On pourra aussi introduire des variables gaussiennes indépendantes $\mathcal{N}_j$ pour $j = 1, ..., k$ centrées et de variance $\rho_j(1 - \rho_j)w_j$.

   (d) On cherche maintenant un estimateur de θ, l'intention de vote globale pour N.S. Donner le lien entre θ, ρ et les w_j.

   (e) Proposer un estimateur $\tilde{\theta}$, estimateur consistant de θ, construit à partir de $\hat{\rho}$.

(f) Donner la loi limite de $\sqrt{n}(\tilde{\theta} - \theta)$.

(g) Montrer que sa variance limite est plus faible que celle de $\hat{\theta}$, l'estimateur construit sans connaître les strates.

4. La proportion de rentiers dans la population totale est de 15%, celles des actifs de 65 % et celle des retraités de 25%. En utilisant de nouveau le lemme de Slutsky, donner un intervalle de confiance asymptotique au niveau 5% sur θ construit avec $\tilde{\theta}$ et $\hat{\rho}$ et comparer le résultat à l'intervalle de confiance obtenu à la question 2(e).

Table de quantiles

Pour vous aider voici un tableau de valeurs qui donne t tel que $P(U > t) = a$

	a=0.10	a=0.05	a=0.025	a=0.01	a=0.005	0.001
U$\sim \mathcal{N}(0,1)$	1.282	1.645	1.960	2.326	2.576	3.090
U$\sim \chi^2(1)$	2.706	3.841	5.024	6.635	7.879	10.827
U$\sim \chi^2(2)$	4.605	5.991	7.378	9.210	10.597	13.815
U$\sim \chi^2(3)$	6.251	7.815	9.348	11.345	12.838	16.268
U$\sim \chi^2(4)$	7.779	9.488	11.143	13.277	14.860	18.465
U$\sim \chi^2(5)$	9.236	11.070	12.833	15.086	16.750	20.517
U$\sim \chi^2(6)$	10.645	12.592	14.449	16.812	18.548	22.457