Examen du 22 Janvier 2008 (Durée 3h)
Calculatrice, cours et poly autorisés

Statistiques - 2007-2008

Le Problème qui suit est composé de 4 parties totalement indépendantes, que vous pouvez traiter dans l'ordre que vous souhaitez. Dans chaque partie, vous pouvez admettre les résultats des questions précédentes pour traiter les suivantes.

A titre indicatif, la première partie est plus théorique que pratique sans difficulté majeure, la deuxième partie est plus appliquée et plus facile, la troisième partie est appliquée et la quatrième partie est entièrement théorique.

Problème

Un propriétaire de plage privée veut savoir si sa plage est polluée ou pas après que certains de ses clients se sont plaints de divers maux. Il fait prélever à diverses distances de la plage des échantillons de 100 ml d'eau de mer. Dans chaque échantillon prélevé à la distance de x_i mètres du bord de la plage, on compte le nombre de bactéries d '*E. coli*, noté Y_i. Voici les résultats.

distance en mètre x_i	Nombre de bactéries Y_i
1	111
2	114
3	120
4	108
5	107
6	126
7	132
8	115
9	114
10	117
11	102
12	107
13	102
14	101
15	101
16	104
17	104
18	106
19	106
20	105
21	98
22	99
23	92
24	79
25	76
26	84
27	88
28	94
29	94
30	96

Voici quelques valeurs des statistiques correspondantes qui vous éviteront des calculs fastidieux (avec $n = 30$).

$$\bar{Y} = 103.4 \quad , \quad \sqrt{\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2} = 12.67,$$

$$\bar{x} = 15.5 \quad , \quad \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2} = 8.655 \quad , \quad \overline{x^2} = 315.16,$$

$$\overline{xY} = 1517.4.$$

Une eau est polluée si le nombre d'*E. coli* par 100 ml est en moyenne supérieur à 100.

Partie I : Modélisation Poissonienne

On suppose dans cette partie seulement que les Y_i sont des variables de Poisson iid de paramètre inconnu.

1. Donner précisément le modèle statistique et l'ensemble Θ des paramètres.
2. Donner l'estimateur du maximum de vraisemblance dans ce modèle.
3. Soit $N \sim \mathcal{P}(\lambda)$. Calculer pour tout $s \in \mathbb{R}$, $\mathbb{E}(e^{sN})$.
4. Soient $N_1 \sim \mathcal{P}(\lambda_1)$ indépendant de $N_2 \sim \mathcal{P}(\lambda_2)$. Calculer $\mathbb{E}(e^{s(N_1+N_2)})$ et en déduire la loi de $N_1 + N_2$.
5. Soit $N \sim \mathcal{P}(\lambda)$. Soit t fixé. Montrer que $\lambda \to P_\lambda(N > t)$ est une fonction croissante de λ.
6. Donner la loi de $Y_1 + ... + Y_n$.
7. Montrer que si $N \sim \mathcal{P}(\lambda)$ alors $\frac{N-\lambda}{\sqrt{\lambda}} \to^{\mathcal{L}}_{\lambda\to\infty} \mathcal{N}(0,1)$.

   On admettra dorénavant que pour les ordres des grandeur en jeu, l'approximation gaussienne est valable et que donc pour tous les α envisagés par la suite, on peut trouver $t_{\alpha,\lambda}$ tq $P_\lambda(N > t_{\alpha,\lambda}) = \alpha$.
8. Donner un test intuitif de taille $\alpha > 0$ qui se place du côté du propriétaire pour savoir si sa plage est polluée. Que concluez-vous ?

   Il faut donner précisément les hypothèses, la statistique du test, la zone de rejet, vérifier que le test est de taille α et donner la p-valeur approchée.
9. Soit $\lambda_0 = 100$ et $\lambda_1 > 100$. Montrer que le test ci-dessus est aussi le test du rapport de vraisemblance de $\mathrm{H}_0 : \lambda = \lambda_0$ contre $\mathrm{H}_1 : \lambda = \lambda_1$ dans le modèle Poissonien.
10. Montrer que le test fait en (8) est uniformément le plus puissant parmi les tests de niveau α des hypothèses testées en (8).

Partie II : Influence de la distance

On suppose dans cette partie seulement que les Y_i sont des variables gaussiennes indépendantes de variance égale et inconnue, σ^2 et de moyenne dépendant de la distance à la plage, égale à $a + bx_i$ avec a et b inconnus.

1. Montrer que le modèle est un modèle linéaire gaussien dont on donnera précisément les paramètres.
2. Donner l'écriture formelle des estimateurs recentrés du maximum de vraisemblance de a, b, σ^2. Calculer $\hat{a}$ et $\hat{b}$.

   Votre logiciel de calcul préféré renvoie $\sqrt{\hat{\sigma^2}} = 7.62$.
3. Donner un intervalle de confiance de coefficient de sécurité 95% sur le degré de pollution (i.e. le nombre de bactéries) à 30 m de la plage.

Partie III : Est-ce Gaussien ?

On suppose dans cette partie que les Y_i sont iid mais qu'on ne sait rien sur leur loi commune.

1. On veut tester que les variables Y_i sont des $\mathcal{N}(100, 100)$. Faire le test de Kolmogorov-Smirnov. *La valeur de la statistique de Kolmogorov-Smirnov est ici 1.131.*
2. Peu satisfait de la p-valeur du test ci-dessus, un statisticien décide de compter le nombre de valeurs inférieures à 100 et supérieures à 100. A l'aide d'un test du khi-deux, montrer que si les Y_i sont bien des $\mathcal{N}(100, 100)$ alors la probabilité de voir ces données est plus petite que 7%.
3. De manière formelle, quel test peut-on faire pour tester que les Y_i sont Gaussiens de moyenne et variance inconnues ?

Partie IV : Le "meilleur" estimateur ?

On suppose dans cette partie que les Y_i sont des variables normales iid de variance connue qu'on supposera égale à 1 (quitte à renormaliser) et de moyenne inconnue notée m ($m \in \mathbb{R}$).

1. Montrer que $\bar{Y}$ est l'estimateur du maximum de vraisemblance et qu'il est sans biais.
2. Montrer que $\bar{Y}$ atteint la borne de Cramer-Rao.
3. On suppose que m est lui-même aléatoire de loi $\mathcal{N}(\eta, \tau^2)$. Montrer que $(m, Y_1, ..., Y_n)$ est un vecteur Gaussien et donner la densité de ce vecteur par rapport à la mesure de Lebesgue.
4. En utilisant la forme de cette densité, montrer que la loi de m sachant $X_1, ..., X_n$ est $\mathcal{N}\left(\frac{\eta+n\bar{Y}\tau^2}{n\tau^2+1}, \frac{\tau^2}{n\tau^2+1}\right)$.
5. On se place dans le cadre bayésien avec comme loi a priori sur m, $\mathcal{N}(\eta, \tau^2)$ et comme risque bayésien de l'estimateur T, qui ne dépend que de $Y_1, ..., Y_n$, $\mathbb{E}((T-m)^2)$ pour la loi de $(m, Y_1, ..., Y_n)$. Donner l'estimateur bayésien $T^{\eta,\tau}$ ainsi que la valeur de son risque bayésien $r^{\eta,\tau}$.
6. Montrer que le risque bayésien $r^{\eta,\tau}$ tend vers $\sup_{m\in\mathbb{R}} \mathbb{E}_m((\bar{Y}-m)^2)$ quand $\tau^2 \to \infty$ et en déduire que $\bar{Y}$ est minimax.

Tables de quantiles

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

	a=0.15	a=0.10	a=0.07	a=0.05	a=0.025	a=0.01	a=0.005
U$\sim \mathcal{N}(0,1)$	1.036	1.282	1.476	1.645	1.960	2.326	2.576
U$\sim$T(2)	1.386	1.886	2.383	2.920	4.303	6.965	9.925
U$\sim$ T(28)	1.056	1.313	1.519	1.701	2.048	2.467	2.763
U$\sim$ T(30)	1.055	1.310	1.516	1.697	2.042	2.457	2.750
U$\sim$ KS(asymptotique)	1.148	1.223	nc	1.358	1.518	1.629	nc
U$\sim \chi^2(1)$	2.072	2.706	3.283	3.841	5.024	6.634	7.879
U$\sim \chi^2(2)$	3.794	4.605	5.318	5.991	7.377	9.210	10.596
U$\sim \chi^2(3)$	5.317	6.251	7.060	7.815	9.348	11.345	12.838