Devoir à rendre pour le 11 Décembre 2007

Statistiques - 2007-2008

Exercice 1 Analyse de la variance à 2 facteurs (ANOVA 2)

On veut mesurer le degré de pollution en nitrates de l'eau en Bretagne. Pour cela on fait des prélèvements dans plusieurs rivières (notées $r \in \{1, ..., R\}$) et en plusieurs endroits où l'on compte à chaque fois le nombre d'exploitations agricoles à proximité (notées $f \in \{0, ..., F-1\}$). On suppose $R > 1$ et $F > 1$.

Pour chaque couple (r, f) on fait m mesures.

Pour chaque mesure k faite à l'endroit indexé par (r, f), on suppose que la concentration en nitrate mesurée $Y_{r,f,k}$ suit une loi gaussienne de moyenne $\theta_{r,f}$ (ie la moyenne ne dépend que de la rivière et du nombre d'exploitations) et de variance σ^2 (ie la variance ne dépend que de l'appareil de mesure).

On suppose de plus que toutes les mesures se font de manière indépendantes.

En prenant l'ordre lexicographique, on note Y le vecteur $Y = (Y_{r,f,k})_{1\leq r\leq R, 0\leq f\leq F-1, 1\leq k\leq m}$ de $\mathbb{R}^n$ et θ le vecteur $\theta = (\theta_{r,f})_{1\leq r\leq R, 0\leq f\leq F-1}$ de $\mathbb{R}^p$.

On note aussi pour tout r, f

$$\overline{Y_{...}} = \frac{1}{n}\sum_{r,f,k} Y_{r,f,k} \quad , \quad \overline{Y_{r..}} = \frac{R}{n}\sum_{f,k} Y_{r,f,k} \quad , \quad \overline{Y_{.f.}} = \frac{F}{n}\sum_{r,k} Y_{r,f,k} \quad , \quad \overline{Y_{rf.}} = \frac{1}{m}\sum_{k} Y_{r,f,k} \quad ,$$

et

$$\overline{\theta_{..}} = \frac{1}{RF}\sum_{r,f} \theta_{r,f} \quad , \quad \overline{\theta_{r.}} = \frac{1}{F}\sum_{f} \theta_{r,f} \quad , \quad \overline{\theta_{.f}} = \frac{1}{R}\sum_{r} \theta_{r,f} \quad .$$

1. Montrer que le modèle envisagé est un modèle linéaire Gaussien avec $Y = m + \varepsilon$ où $m \in V$ sev strict de $\mathbb{R}^n$, $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$. On donnera n et V. Montrer aussi qu'on peut l'écrire $Y = X\theta + \varepsilon$ où X est une matrice injective de $\mathbb{R}^p$ dans $\mathbb{R}^n$. On donnera p et X.
2. Expliquer pourquoi on peut considérer ce modèle comme une analyse de la variance à un facteur (ANOVA 1). Donner les estimateurs de maximum de vraisemblance recentrés de θ et σ^2 ainsi que leur loi.

On veut maintenant étudier les influences séparées de r et f. Pour cela on note

$$E_1 = \{\theta \in \mathbb{R}^p, \forall r, f, \theta_{r,f} = \overline{\theta_{..}}\},$$

$$E_2 = \{\theta \in \mathbb{R}^p, \forall r, f, \theta_{r,f} = \overline{\theta_{r.}}, \sum_r \overline{\theta_{r.}} = 0\},$$

$$E_3 = \{\theta \in \mathbb{R}^p, \forall r, f, \theta_{r,f} = \overline{\theta_{.f}}, \sum_f \overline{\theta_{.f}} = 0\},$$

$$E_4 = \{\theta \in \mathbb{R}^p, \forall r, f, \sum_r \theta_{r,f} = 0, \sum_f \theta_{r,f} = 0\}.$$

3. Montrer que E_1 est en somme directe orthogonale avec E_2. Montrer que $E_1 \overset{\perp}{\oplus} E_2$ est l'ensemble des θ ne dépendant que des rivières. De même montrer que E_1 est en somme directe orthogonale avec E_3. Montrer que $E_1 \overset{\perp}{\oplus} E_3$ est l'ensemble des θ ne dépendant que du nombre d'exploitations.
4. Montrer que $E_2 \overset{\perp}{\oplus} E_3$ et montrer que E_4 est le supplémentaire orthogonal de $E_1 \overset{\perp}{\oplus} E_2 \overset{\perp}{\oplus} E_3$ dans $\mathbb{R}^p$.
5. En déduire que pour tout θ de $\mathbb{R}^p$, il existe une unique constante μ, un unique vecteur α de $\mathbb{R}^R$, un unique vecteur β de $\mathbb{R}^F$ et un unique vecteur γ de $\mathbb{R}^p$ tels que
$$\forall r, f, \theta_{r,f} = \mu + \alpha_r + \beta_f + \gamma_{r,f},$$
avec $\sum_r \alpha_r = \sum_f \beta_f = \sum_r \gamma_{rf} = \sum_f \gamma_{rf} = 0$. Donner une expression de $\mu, \alpha, \beta, \gamma$ en fonction de $\overline{\theta_{..}}$, $\overline{\theta_{r.}}$, $\overline{\theta_{.f}}$ et $\theta_{r,f}$.
6. Donner les dimensions de E_1, E_2, E_3, E_4.
7. On note $(.|.)_j$ le produit scalaire classique de $\mathbb{R}^j$. Montrer qu'il existe λ tel que pour tout $\theta_1, \theta_2 \in \mathbb{R}^p$,
$$(X\theta_1 | X\theta_2)_n = \lambda(\theta_1|\theta_2)_p.$$
8. En déduire une décomposition en sev $\overset{\perp}{\oplus}$ de V. Donner alors les estimateurs par maximum de vraisemblance de μ, α, β et γ.
9. On suppose que pour tout r, f $\gamma_{r,f} = 0$, c'est-à-dire qu'il n'y a pas d'influence croisée entre la rivière et le nombre d'exploitations. Monter qu'alors $Y = m + \varepsilon$ avec $m \in W$ où W est sev strict de V dont on donnera la dimension.
10. Dans le modèle où $m \in W$, donner les estimateurs par maximum de vraisemblance de m puis de μ, α, β et de σ^2.
11. Construire un test au niveau 5% de H_0 :"$m \in W$"contre H_1 :"$m \in V \backslash W$". Donner sa puissance.
12. Dans le plus grand modèle ($m \in V$), donner un intervalle de confiance sur μ avec coefficient de sécurité 95%.

Exercice 2 Un agent immobilier se demande s'il lui faut, à certains mois de l'année, solliciter auprès des écoles des environs l'envoi de stagiaires (mal rémunérés) pour l'aider. L'opinion qu'il s'est forgée par l'expérience est en effet que la fréquence des ventes d'appartements (plus exactement, des signatures de promesses de vente) n'est pas uniforme au cours de l'année.

1. Ses tables donnent, mois par mois sur l'année écoulée, le nombre de mandats qui se sont conclus par une signature :

A	M	J	J	A	S	O	N	D	J	F	M
6	6	5	3	1	2	1	2	2	1	3	4

   Ces données confirment-elles son sentiment empirique ?
2. L'agent veut maintenant préciser l'inhomogénéité de la répartition. Son hypothèse est qu'il y a deux fois plus de signatures au printemps (mois d'avril, mai, juin) qu'en été ou en hiver (juillet, août, septembre, d'une part, janvier, février, mars, d'autre part). Il suspecte donc que le taux annuel de signatures par saison est de la forme $(2\theta, \theta, 1-4\theta, \theta)$.

(a) On note N_1, N_2, N_3, N_4 le nombre de signatures en printemps, été, automne et hiver. Supposons que l'hypothèse est vraie : proposez un bon estimateur de θ, par exemple l'estimateur du maximum de vraisemblance.

(b) Effectuez alors un test d'adéquation à la forme postulée pour la loi. A quel point peut-on retenir l'hypothèse qu'il y a deux fois plus de signatures au printemps qu'en été et hiver ?

Données : voici une table de quantiles $\chi^2_{k,\beta}$. On rappelle que $\chi^2_{k,\beta}$ est le réel positif tel que $\mathbb{P}(\chi^2_k \leq \chi^2_{k,\beta}) = \beta$.

β	0.40	0.60	0.80	0.95	0.98	0.99
$k=2$	1.02	1.83	3.22	5.99	7.82	9.21
$k=3$	1.87	2.95	4.64	7.81	9.84	11.34
$k=4$	2.75	4.04	5.99	9.49	11.67	13.28