Exercices de Statistiques
Université de Lorraine

Estimation et théorie des tests

Clément Dell'Aiera

1 Tests du χ^2

On considère une variable qualitative X, à valeur dans un ensemble fini $E = \{1, .., d\}$. Les lois de probabilité de telles v.a. sont entièrement décrites par le vecteur de probabilité $(p_1, ..., p_d)^T$, où $p_j = \mathbb{P}(X = j)$. On confondera donc les lois de probabilités de E avec

$$\mathfrak{M}_d = \{p = (p_1, ..., p_d)^T : 0 \geq p_j \geq 1 \text{ et } \sum p_j = 1\}.$$

1. **Test d'adéquation du χ^2.** On observe un n-échantillon de loi p et l'on souhaite tester $p = q$ contre $p \neq q$, où $q \in \mathfrak{M}_d$ est une loi fixée.
   (a) Décrire le modèle statistique.
   (b) On définit les fréquences empiriques

   $$\hat{p}_{n,l} = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j = l} \quad \text{pour } l = 1, ..., d.$$

   Donner la limite du vecteur $\hat{p}_n = (\hat{p}_{n,l})_{l=1,d}^T$ pour la topologie de la convergence en probabilité sous $\mathbb{P}_p$.
   (c) On définit

   $$U_n(p) = \sqrt{n}(\frac{\hat{p}_{n,l} - p_l}{\sqrt{p_l}})_{l=1,d}^T.$$

   Donner la limite en loi de chaque composante de $U_n(p)$ sous $\mathbb{P}_p$. Que peut-on dire a priori de la limite en loi de $U_n(p)$? Pourquoi ?
   (d) On définit

   $$Y_l^j = \frac{1}{\sqrt{p_l}}(1_{X_j = l} - p_l).$$

   Si Y_j désigne le vecteur $(Y_1^j, ..., Y_d^j)$, montrer que $\frac{1}{\sqrt{n}} \sum Y_j = U_n(p)$.
   (e) Calculer $E[Y_l^j]$, et $E[Y_l^j Y_{l'}^j]$. Que valent les composantes de la matrice $V(p) = I_d - \sqrt{p}\sqrt{p}^T$, où $\sqrt{p} = (\sqrt{p_1}, ..., \sqrt{p_d})^T$?
   (f) En déduire la limite en loi sous $\mathbb{P}_p$ de $U_n(p)$ et de $||U_n(p)||^2$, le carré de sa norme euclidienne.

(g) Soient $p, q \in \mathfrak{M}_d$ tels que les coefficients de q soient tous non nuls. On définit :

$$\chi^2(p,q) = \sum_{l=1}^{d} \frac{(p_l - q_l)^2}{q_l}.$$

Cette quantité est appelée "distance du χ^2" bien que ce ne soit pas une distance ! Toutefois, $\chi^2(p,q) = 0$ ssi $p = q$.
Montrer que $n\chi^2(\hat{p}_n, p) = ||U_n(p)||^2$.

(h) On définit, pour $\alpha \in (0,1)$, la zone de rejet

$$\mathcal{R}_{n,\alpha} = \{n\chi^2(\hat{p}_n, p) \geq q^{\chi^2}_{1-\alpha,d-1}\},$$

où $q^{\chi^2}_{1-\alpha,d-1}$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $d-1$ degrés de liberté.
Montrer que le test associé est asymptotiquement de niveau α et est asymptotiquement consistant.

(i) Application numérique. On décrit ici l'expérience de Mendel. Le croisement des pois fait apparaître 4 phénotypes, distibués selon une loi multinomiale de paramètre

$$q = (\frac{9}{16}, \frac{3}{16}), \frac{3}{16}, \frac{1}{16}).$$

Pour $n = 556$ observations, Mendel rapporte les observations suivantes : les phénotypes se répartissent selon $(315, 101, 108, 32)$. Sachant que le quantile d'ordre 0.95 de la loi du χ^2 à 3 degrés de liberté vaut 0.7815, accepter vous le test $p = q$ contre $p \neq q$.

2 Théorème de Cochran et applications

Voici un énoncé simplifié du théorème de Cochran.

Théorème 1 (Cochran)**.** Soit $E = (\mathbb{R}^n, \langle\ ,\ \rangle)$ l'espace euclidien usuel, et F un sous-espace vectoriel de E de dimension $p \leq n$. Notons P la projection orthogonale sur le sous-espace F. Soit $\underline{x}$ un vecteur gaussien de E centré réduit. Les vecteurs $P\underline{x}$ et $P^\perp\underline{x}$ sont indépendants, gaussiens, centrés et de matrice de variance-covariance respectives P et $P^\perp$.
Les variables aléatoires $||P\underline{x}||^2$ et $||P^\perp\underline{x}||^2$ sont indépendantes et suivent une loi du χ^2 à p et $n-p$ degrés de liberté, respectivement.

1. Démontrer le théorème.
2. Soient X_j un n-échantillon gaussien i.i.d d'espérance μ et de variance σ^2. On note

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \text{ et } s_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2.$$

Démontrer que $\overline{X}_n$ suit un loi normale $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ et que $(n-1)\frac{s_n^2}{\sigma^2}$ suit une loi du χ^2 à $n-1$ degrés de liberté. En déduire la loi de $\sqrt{n}\frac{\overline{X}_n - \mu}{s_n^2}$.

3 Un modèle non-linéaire

Soit $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ une fonction de classe $\mathcal{C}^2$ que l'on suppose connue. Soit le modèle

$$y = f(X, \alpha) + \epsilon \text{ et } \epsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

On cherche à estimer

$$\theta = (\alpha, \sigma^2) \in \Theta \subset \mathbb{R}^{k+1}.$$

On note $L(\alpha) = ||y - f(X, \alpha)||^2$.

1. Définir le modèle, et calculer la vraisemblance.
2. Montrer que maximiser la vraisemblance est équivalent à minimiser $L(\alpha)$.
3. Calculer l'information de Fisher du modèle.

4 Maximum de vraisemblance et séries temporelles

Soient $\lambda \in \mathbb{R}$ tel que $|\lambda| < 1$, $c \in \mathbb{R}$ et $\sigma^2 > 0$. On observe un échantillon $\{Y_t\}_{t \leq T}$ que l'on pense suivre le modèle $AR(1)$

$$Y_t = c + \lambda Y_{t-1} + \epsilon_t \text{ où les } \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

sont des variables i.i.d.
On cherche à estimer

$$\theta = (c, \lambda, \sigma^2)^T \in \Theta \subset \mathbb{R}^3.$$

1. Calculer $\mathcal{L}(Y_1; \theta)$, $\mathcal{L}(Y_t | Y_{t-1}; \theta)$, et en déduire $\mathcal{L}(Y_2, Y_1; \theta)$.
2. Calculer la vraisemblance du modèle $\mathcal{L}(Y_1, ..., Y_n | \theta)$.
3. Calculer la matrice de variance-covariance du processus $AR(1)$ gaussien. On la note Ω.
4. Réécrire la log-vraisemblance du modèle en utilisant Ω. Quel est le lien avec la question 2 ?
5. Déterminer un estimateur du maximum de vraisemblance.
6. Refaire l'exercice pour le modèle $MA(1)$ gaussien

$$Y_t = c + \epsilon_t - \theta \epsilon_{t-1} \text{ où } \epsilon_t \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

7. En cas de forme de motivation extrême, le faire pour le modèle $ARMA(p, q)$ gaussien

$$Y_t = c + \sum_{j=1}^{p} \lambda_j Y_{t-j} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

avec $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

5 Test du χ^2

1. Soit $(X_k, Y_k)_{k=1,...,n}$ un n-échantillon d'une loi $Q = (q_{ij})_{(i,j)\in\{1,...,I\}^2}$ sur $\{1, ..., I\}^2$, dont les marginales sont égales. Soit, pour tout $k = 1, ..., n$, le vecteur aléatoire

$$Z_k = (1_{X_k=i} - 1_{Y_k=i})_{1\leq i\leq I}.$$

   (a) Quelle est la matrice de covariance Γ de Z_k ?

   (b) On suppose Γ inversible et on note son inverse

$$\Gamma^{-1} = (\Gamma^{ij})_{1\leq i,j\leq I-1}.$$

   Soient, pour tout $i, j = 1, ..., I$,

$$\begin{array}{l} N_{ij} = \sum_{k=1}^{n} 1_{X_k=i,Y_k=j} \\ N_{i.} = \sum_{k=1}^{n} 1_{X_k=i} \\ N_{.j} = \sum_{k=1}^{n} 1_{Y_k=j} \end{array}$$

   Quelle est la loi asymptotique de

$$\frac{1}{n} \sum_{1\leq i,j\leq I} (N_{i.} - N_{.i})(N_{j.} - N_{.j})\Gamma^{ij} \, ?$$

2. On ne suppose plus a priori que les marginales soient égales. On observe (X_k, Y_k) décrit comme ci-dessus. Soit V la matrice $(V_{ij})_{1\leq i,j\leq I-1}$ définie par

$$nV_{ii} = N_{i.} + N_{.i} - 2N_{ii},$$

   et pour tout $i \neq j$,

$$nV_{ij} = -(N_{ij} + N_{ji}).$$

   (a) Montrer que, sous l'hypothèse d'égalité des marginales, V converge vers Γ.

   (b) Soit $(V^{ij})_{1\leq i,j\leq I-1}$ l'inverse de V . Montrer que

$$\Delta = \frac{1}{n} \sum_{i,j} (N_{i.} - N_{.i})(N_{j.} - N_{.j})V_{ij}$$

   converge vers une loi $\chi^2(I - 1)$.

3. Quel test peut-on construire ?

4. Appliquer ce test aux données suivantes. On évalue le degré de vision des deux yeux de 7477 femmes agées de 30 à 40 ans en le classifiant en 4 groupes (1 à 4, du meilleur au pire). On obtient

oeildroit — oeilgauche	1	2	3	4
1	1520	266	124	66
2	234	1512	432	78
3	117	362	1772	205
4	36	82	179	492