Exercices de Statistiques
Université de Lorraine

Estimation et théorie des tests

Clément Dell'Aiera

1 Généralités sur l'estimateur du maximum de vraisemblance

1. Rappeler les propriétés de l'EMV.
2. Soient X_j des variables exponentielles indépendantes de paramètre $\theta > 0$, non-observées, et T un instant de censure. Soit $\mathcal{E}^n$ l'expérience engendrée par l'observation du n-échantillon $X_j^* = \min\{T, X_j\}$. Donner une mesure qui domine le modèle et calculer sa vraisemblance.
3. Montrer que l'estimateur du maximum de vraisemblance ne dépend pas du choix de la mesure dominante.

2 Exemples de calculs de maximum de vraisemblance

Pour chaque loi, on considère un n-échantillon tiré de façon i.i.d selon cette loi. Proposer un espace des paramètres donnant un modèle identifiable. Donner une mesure dominante si possible. Calculer la vraisemblance du modèle, ainsi que la log-vraisemblance, donner les équations de vraisemblance et déterminer , s'il existe, un estimateur du maximum de vraisemblance.

1. Modèle gaussien standard, de densité par rapport à la mesure de Lebesgue
$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(x-\mu)^2 \quad , \theta = (\mu, \sigma).$$
2. Modèle de Bernoulli
$$\mathbb{P}_\theta(X = 1) = 1 - \mathbb{P}(X = 0) = \theta.$$
3. Modèle de Laplace, où $\sigma > 0$ est connu, de densité par rapport à la mesure de Lebesgue
$$f_\theta(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\theta|}{\sigma}\right).$$

4. Modèle uniforme, de densité par rapport à la mesure de Lebesgue

$$f_\theta(x) = \frac{1}{\theta} 1_{[0,\theta]}(x).$$

5. Modèle de Cauchy, de densité par rapport à la mesure de Lebesgue

$$f_\theta(x) = \frac{1}{\pi(1-(x-\theta)^2)}.$$

6. Modèle de translation. On considère la densité

$$h(x) = \frac{e^{-\frac{|x|}{2}}}{2\sqrt{2\pi|x|}}.$$

Le modèle de translation par rapport à la densité h est le modèle dominé par la mesure de Lebesgue sur $\mathbb{R}$ de densités

$$f_\theta(x) = h(x-\theta) \quad , x \in R, \theta \in \mathbb{R}.$$

3 Méthode des moments

1. Calculer des estimateurs des moments d'ordre 1 et 2 pour l'expérience engendrée par l'observation d'un n-échantillon de variables exponentielles de paramètre $\theta > 0$. Donner l'asymptotique des ces deux estimateurs.
2. On considère le modèle de translation associé à la famille des lois de Cauchy :

$$f_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)} \quad , x \in \mathbb{R}.$$

On note g la fonction signe, qui vaut 1 si $x > 0$, -1. Trouver un estimateur pour $\theta \mapsto \mathbb{E}[g(X_1)]$ et donner ses propriétés.

4 Estimation de la fonction de répartition

On se donne un n-échantillon X_1,..., X_n i.i.d suivant une loi donnée par la même fonction de répartition (f.d.r) F sur $\mathbb{R}$. $\mathcal{F}$ dénote l'ensemble des fonctions de répartition sur $\mathbb{R}$.

1. Décrire l'expérience statistique.
2. Le modèle est-il dominé ?
3. On veut estimer $F(x) = \mathbb{P}(X \le x)$.
   (a) On pose $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{X_i \le x}$. Calculer $\mathbb{E}[\hat{F}_n(x)]$ et $V[\hat{F}_n(x)]$.
   (b) Montrer que $\hat{F}_n(x)$ converge presque-sûrement vers $F(x)$.
   (c) Montrer que, si $l(x,y) = (x-y)^2$ est la perte quadratique,

$$\sup_{F \in \mathcal{F}} \mathbb{E}[l(\hat{F}_n(x), F(x))] = \frac{1}{4n}.$$

   (d) En déduire que $\hat{F}_n(x)$ converge uniformément en norme $\mathcal{L}^2$ vers $F(x)$, et donc en probabilité.

4. (a) Montrer que

$$\mathbb{P}(|\hat{F}_n(x) - F(x)| > t) \leq \frac{1}{t^2} Var[\hat{F}_n(x)] \leq \frac{1}{4nt^2}$$

(b) Soit $\alpha \in]0;1[$. Déterminer

$$t_{\alpha,n} = \inf\{t > 0 \ : \ \frac{1}{4nt^2} \leq \alpha\}$$

et en déduire un intervalle de confiance pour $F(x)$ de niveau $1-\alpha$.

(c) Comment interpréter $I_{n,\alpha}$? Quelle est sa précision ?

5. On pose $\xi_n = \sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}}$.

(a) Déterminer la limite en loi de ξ_n.

(b) On note $J_{n,\alpha}$ l'intervalle $[-\phi^{-1}(1-\frac{\alpha}{2}); \phi^{-1}(1-\frac{\alpha}{2})]$.Calculer la limite de $\mathbb{P}(\xi_n \in J_{n,\alpha})$ lorque n tend vers ∞.

(c) Donner un intervalle de confiance asymptotique pour $J_{n,\alpha}$, ainsi que sa précision asymptotique.

6. Soient Y_j des variables aléatoires réelles indépendantes centrées : $\mathbb{E}Y_j = 0$ et bornées : $a_j \leq Y_j \leq b_j$. On veut démontrer ce que l'on appelle l'*inégalité de Hoeffding* : pour tout $t > 0$,

$$\mathbb{P}(\sum Y_j < t) \leq e^{-st} \prod e^{\frac{s^2(b_j - a_j)^2}{8}} \quad \forall s > 0.$$

On pose $\Phi_Y(s) = \log \mathbb{E}[e^{s(Y - \mathbb{E}Y)}]$.

(a) Montrer que

$$\Phi_Y''(s) = e^{-\Phi_Y(s)}\mathbb{E}[Y^2 e^{sY}] - e^{-2\Phi_Y(s)}(\mathbb{E}[Ye^{sY}])^2.$$

(b) On définit une nouvelle mesure de probabilité par $\mathbb{Q}(A) = e^{-\Phi_Y(s)}\mathbb{E}[e^{sY}1_A]$ pour tout borélien A. Comment interpréter $\Phi_Y''(s)$ dans ce cadre ?

(c) Montrer alors que $\Phi_Y(s) \leq s^2\frac{(b-a)^2}{8}$.

(d) En déduire l'inégalité de Hoeffding.

7. (a) Soient X_j des v.a. de Bernoulli de paramètre p et $\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$, montrer que

$$\mathbb{P}(|\overline{X}_n - p| > t) \leq 2e^{-2nt^2}.$$

(b) En déduire un intervalle de confiance de niveau $1-\alpha$ pour $F(x)$.

8. Comparer les différents intervalles de confiance que vous avez obtenu.

5 Principe de Neyman : décision à 2 points

1. Soit f la densité d'une loi de probabilité sur $\mathbb{R}$, et $\mathcal{E}$ l'expérience statistique engendré par un n-échantillon de loi $p_\theta(x) = f(x-\theta)$. On suppose que $\Theta = \{0, \theta_0\}$ avec $\theta_0 \neq 0$. On veut tester $H_0|\theta = 0$ contre $H_1|\theta = \theta_0$.

(a) Décrire l'expérience statistique et donner la vraisemblance du modèle.

(b) Donner la zone de rejet du test de Neyman-Pearson de niveau α associé à H_0 et H_1.

2. L'expérimentateur observe une seule réalisation d'une v.a. X de loi de Poisson de paramètre $\theta > 0$. On veut tester $H_0|\theta = \theta_0$ contre $H_1|\theta = \theta_1$, où $\theta_0 \neq \theta_1$.
   (a) Donner la zone de rejet du test de Neyman-Pearson de niveau α associé.
   (b) Sachant que $\mathbb{P}_{\theta_0}(X > 9) = 0.032$ et $\mathbb{P}_{\theta_1}(X > 8) = 0.068$, donner une zone de rejet explicite pour $\alpha = 0.05 = 5\%$. Le test est-il optimal ?

6 Neyman-Pearson : familles à rapport de vraisemblance monotone

1. Soit $\mathcal{E}$ l'expérience statistique engendrée par un n-échantillon de loi normale $\mathcal{N}(\theta, \sigma^2)$, où σ^2 est connu, et $\theta \in \Theta = \mathbb{R}$. On souhaite tester $H_0|\theta = \theta_0$ contre $H_1|\theta = \theta_1$, où $\theta_0 < \theta_1$.
   (a) Décrire le modèle ainsi que la vraisemblance. On choisira la mesure de Lebesgue comme mesure dominante.
   (b) Calculer le rapport de vraisemblance
   $$\frac{f(\theta_1, Z)}{f(\theta_0, Z)}.$$
   (c) Donner la zone de rejet pour le test de Neyman-Pearson associé.
2. Pour la même expérience statistique, on a un test optimal (uniformément plus puissant) de H_0 contre H_1 donné par la région de rejet
   $$\mathcal{R} = \{\overline{X}_n > c\}$$
   où c est solution de $\mathbb{P}_{\theta_0}(\overline{X}_n > c) = \alpha$.
   (a) Calculer explicitement la valeur de la constante $c = c(\theta_0, \alpha)$.
   (b) Calculer la puissance de ce test.

7 Exercice

L'expérimentateur observe 2 échantillons indépendants $X_1, ..., X_n$ et $Y_1, ..., Y_m$ de tailles distinctes $n \neq m$, de lois respectives $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$. Il souhaite tester

$$H_0 \ : \ \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \ \mu_1 \neq \mu_2.$$

Si $s_{n,1}^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \overline{X}_n)^2$ et $s_{m,2}^2 = \frac{1}{m}\sum_{j=1}^m (Y_j - \overline{Y}_m)^2$, construire un test basé sur la statistique

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{s_{n,1}^2 + s_{m,2}^2}}$$

et étudier sa consistance.

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

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

   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 $p \neq q$.