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 $F(x) = \mathbb{P}(X \leq x)$.
   (a) On pose $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{X_i \leq x}$. 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 $t > 0$,

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

   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

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

   (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 Neyman-Pearson : loi exponentielle

On observe un n-échantillon $\underline{x} = (X_1, ..., X_n)$ de variables iid de loi exponentielle de paramètre $\lambda > 0$, de densité

$$x \mapsto \lambda \exp(-\lambda x) 1_{x \leq 0}.$$

1. Rappeler l'espérance et la variance (les calculer si besoin) d'une loi exponentielle de paramètre λ. On rappelle que $2\lambda \sum_{j=1}^{n} X_j$ suit alors une loi du χ^2 à $2n$ degrés de liberté.
2. Ecrire le modèle statistique engendré par l'observation $\underline{x}$.
3. Calculer l'estimateur du maximum de vraisemblance $\hat{\lambda}^{MV}$ de λ.
4. Montrer que $\hat{\lambda}^{MV}$ est asymptotiquement normal, et calculer sa variance limite.
5. Soient $0 < \lambda_0 < \lambda_1$. Construire un test d'hypothèse de

   $$H_0 : \lambda = \lambda_0 \text{ contre } H_1 : \lambda = \lambda_1$$

   de niveau α et uniformément plus puissant. Expliciter le choix du seuil définissant la région critique. Montrer que le test est consistant, i.e. que l'erreur de seconde espèce du test tend vers 0 lorsque $n \to \infty$.

9 Intervalles de confiance

1. Soient $\mu \in \mathbb{R}$, $\sigma > 0$ et $\alpha \in (0, 1)$. On observe un n-échantillon $\underline{x} = (X_1, ..., X_n)$ de variables iid de loi normale $\mathcal{N}(\mu, \sigma^2)$.
   (a) Donner un intervalle de confiance de niveau α pour μ, si σ^2 est connu, puis si σ^2 est inconnu.
   (b) Donner un intervalle de confiance de niveau α pour σ^2, si μ est connu, puis si μ est inconnu.
2. On observe un n-échantillon $\underline{Y} = (Y_1, ..., Y_n)$ de variables iid suivant une loi de Bernoulli de paramètre $0 < p < 1$ inconnu. Donner un intervalle de confiance pour p au niveau α.

10 Estimateur du maximum de vraisemblance

On observe un n-échantillon $\underline{x} = (X_1, ..., X_n)$ de variables iid suivant une loi $\mathbb{P}_\theta$ de densité :

$$f_\theta(x) = \frac{2}{\sqrt{\pi}\theta^{3/2}} x^2 e^{-\frac{x^2}{\theta}}, \text{ pour } \theta > 0.$$

1. Décrire le modèle statistique engendré par $\underline{x}$.
2. Calculer un estimateur du maximum de vraisemblance, que l'on notera $\hat{\theta}$.
3. Examiner les qualités suivantes de $\hat{\theta}$: efficacité, biais et convergence.

On observe un n-échantillon $\underline{x} = (X_1, ..., X_n)$ de variables iid suivant une loi exponentielle de paramètre $\lambda > 0$, de densité

$$g_\lambda(x) = \lambda e^{-\lambda x} 1_{x \geq 0}.$$

1. Décrire le modèle statistique engendré par $\underline{x}$.
2. Calculer la vraismeblance du modèle.
3. Donner un estimateur du maximum de vraisemblance pour λ.

11 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 $p \neq q$.

12 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 $\sqrt{n}\frac{\overline{X}_n - \mu}{s_n^2}$.

13 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.

14 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.

15 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