L3 : Exercices de Statistique Mathématique

Valère Bitseky-Penda, Marc Hoffmann, Adrian Iuga

Janvier 2014

Table des matières

1 Echantillonnage, modèle statistique.

1.1 Région de confiance et Kolmogorov-Smirnov

On rappelle qui si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes et de même fonction de répartition F continue, alors

$$\sqrt{n} \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right| \xrightarrow{d} \mathbb{B}$$

lorsque $n \to \infty$, où $\mathbb{B}$ est un variable aléatoire qui admet une densité strictement positive sur $(0, \infty)$ et qui ne dépend pas de F et $\widehat{F}_n(x)$ est la fonction de répartition empirique de F construite à partir de $(X_1, \ldots, X_n)$.

1. Pour $\alpha \in (0,1)$, on note $q_{1-\alpha}$ le nombre satisfaisant

   $$\mathbb{P}\big(\mathbb{B} \geq q_{1-\alpha}\big) = \alpha.$$

   Montrer que $q_{1-\alpha}$ est bien défini.

2. Construire deux fonctions $U_{n,\alpha}(x)$ et $L_{n,\alpha}(x)$ à l'aide de $q_{1-\alpha}$, de n et de $\widehat{F}_n(x)$ de sorte que

   $$\mathbb{P}\big(\forall x \in \mathbb{R}, \;\; L_{n,\alpha}(x) \leq F(x) \leq U_{n,\alpha}(x)\big) \longrightarrow 1 - \alpha$$

   lorsque $n \to \infty$.

1.2 Echantillonnage

Soient $X_1, \ldots, X_n$ un n-échantillon de loi inconnue F. On considère la fonctionnelle

$$\vartheta = T(F) = F(b) - F(a),$$

où a, b sont deux réels donnés (avec $a < b$).

1. Calculer l'estimateur $\widehat{\vartheta}_n$ de $\vartheta = T(F)$ par *plug-in*.
2. Montrer que $\widehat{\vartheta}_n$ est asymptotiquement normal, c'est-à-dire

   $$\sqrt{n}(\widehat{\vartheta}_n - \vartheta) \xrightarrow{d} \xi \;\text{ lorsque }\; n \to \infty,$$

   où ξ est une variable aléatoire gaussienne centrée de variance $\sigma(F)^2$ que l'on précisera.

3. Pour $\alpha \in (0,1)$, construire un intervalle de confiance asymptotique de niveau de confiance $1 - \alpha$ pour $\vartheta = T(F)$.

1.3 Modélisation statistique : sondage

Une élection entre deux candidats A et B a lieu : on effectue un sondage à la sortie des urnes. On interroge n votants, n étant considéré comme petit devant le nombre total de votants, et on récolte les nombres n_A et n_B de voix pour A et B respectivement ($n_A + n_B = n$, en ne tenant pas compte des votes blancs ou nuls pour simplifier).

1. Décrire l'observation associée à cette expérience et le modèle statistique engendré par cette observation.
2. Montrer que le modèle statistique engendré par cette observation est identifiable, dominé, et exhiber sa vraisemblance.
3. La situation précédente correspond au cas où la population totale (le nombre N de votants) est $\ll$ très grande $\gg$ devant le nombre n de sondés. Supposons désormais que le nombre total de votants N ne soit pas négligeable devant n. Reprendre la modélisation précédente.

1.4 Modélisation statistique : contrôle de qualité, données censurées

On cherche – en laboratoire – à tester la fiabilité d'un appareil industriel. On fait fonctionner en parallèle n appareils jusqu'à ce qu'ils tombent tous en panne. On note

$$X_1, \ldots, X_n$$

les instants de panne observés. On dispose donc de n observations. On suppose que les temps de panne suivent une loi exponentielle de paramètre $\lambda > 0$.

1. Décrire l'observation associée à cette expérience et le modèle statistique engendré par cette observation.
2. Montrer que le modèle statistique engendré par cette observation est identifiable, dominé et exhiber sa vraisemblance.
3. (Plus difficile et facultatif.) Si les appareils sont fiables, ce qui est réaliste en pratique, la quantité $\max_{i=1,\ldots,n} X_i$ sera souvent hors d'atteinte pour le statisticien. On stoppe l'expérience après un temps terminal T et on observe plutôt

$$X_i^\star = \min\{X_i, T\}, \quad i = 1, \ldots n.$$

Reprendre les deux questions précédentes dans ce contexte.

1.5 Modèle probit et contre-exemple à l'identifiabilité

Nous disposons d'une information relative au comportement de remboursement ou de non-remboursement d'emprunteurs :

$$Y = \begin{cases} 1 & \text{si l'emprunteur rembourse} \\ 0 & \text{si l'emprunteur est défaillant} \end{cases}$$

Afin de modéliser ce phénomène, on suppose l'existence d'une variable aléatoire $Y^\star$ gaussienne, d'espérance m et de variance σ^2, que l'on appellera « capacité de remboursement de l'individu » de sorte que :

$$Y = \begin{cases} 1 & \text{si } Y^\star > 0 \\ 0 & \text{si } Y^\star \leq 0 \end{cases}$$

On note Φ la fonction de répartition de la normale centrée réduite $\mathcal{N}(0,1)$.

1. Exprimer la loi de Y en fonction de Φ.
2. On observe un n-échantillon $(Y_1, \ldots, Y_n)$ de même loi que Y. Ecrire le modèle statistique engendré par l'observation $(Y_1, \ldots, Y_n)$. Est-il identifiable ?

1.6 Exemple de modèle qui n'est pas dominé (Exercice facultatif)

Soit X une variable aléatoire de Poisson de paramètre 1 et $Y = \vartheta X$ où $\vartheta \in \Theta = \mathbb{R}_+ = [0, \infty)$ est le paramètre d'intérêt. On observe un n-échantillon $(Y_1, \ldots, Y_n)$ de même loi que Y. Ecrire le modèle statistique engendré par l'observation $(Y_1, \ldots, Y_n)$. Est-il dominé ?

2 Estimateurs, intervalles de confiance.

2.1 Risque quadratique

Etant donné un modèle statistique $\big(\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\vartheta, \vartheta \in \Theta\}\big)$ engendré par un observation Z, on appelle risque quadratique d'un estimateur $\widehat{\vartheta} = \widehat{\vartheta}(Z)$, la fonction

$$\vartheta \in \Theta \rightsquigarrow \mathcal{R}(\widehat{\vartheta}, \vartheta) = \mathbb{E}_\vartheta\big[(\widehat{\vartheta} - \vartheta)^2\big].$$

On dit que l'estimateur $\widehat{\vartheta}^{(1)}$ est préférable à l'estimateur $\widehat{\vartheta}^{(2)}$

$$\forall \vartheta \in \Theta, \;\; \mathcal{R}(\widehat{\vartheta}^{(1)}, \vartheta) \leq \mathcal{R}(\widehat{\vartheta}^{(2)}, \vartheta).$$

On suppose que $Z = (X_1, \ldots, X_n)$ où les X_i sont indépendantes, et de même loi de Poisson de paramètre $\vartheta \in \Theta = [0, \infty)$.

1. Ecrire le modèle statistique associé, montrer qu'il est dominé et écrire sa vraisemblance.
2. On considère les deux estimateurs suivants : $\widehat{\vartheta}^{(1)} = \overline{X}_n^{-1} \sum_{i=1}^n X_i$ et $\widehat{\vartheta}^{(2)} = 1$. Calculer le risque quadratique de chacun de ces estimateurs.
3. L'un des deux estimateurs est-il préférable à l'autre ?
4. De manière générale, peut-on trouver un estimateur préférable à tout autre estimateur (dans ce modèle statistique) ?

2.2 Structure du risque : biais et variance

Soit $\widehat{\vartheta}$ est un estimateur admettant un moment d'ordre deux pour tout $\vartheta \in \Theta$.

1. On appelle biais de $\widehat{\vartheta}$ au point ϑ et on note $b(\vartheta)$ la quantité

   $$b(\vartheta) = \mathbb{E}_\vartheta\big[\widehat{\vartheta}\,\big] - \vartheta.$$

   Exprimer le risque quadratique de $\widehat{\vartheta}$ en fonction de son biais et de la variance de $\widehat{\vartheta}$.
2. On considère le modèle engendré par $Z = (X_1, \ldots, X_n)$, où les X_i sont indépendantes et de même loi normale, centrée et de variance $\vartheta \in \Theta = (0, \infty)$. On considère les deux estimateurs

   $$\begin{aligned} \widehat{\vartheta}^{(1)} &= S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2, \\ \widehat{\vartheta}^{(2)} &= \frac{n-1}{n} S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2. \end{aligned}$$

(a) Ces estimateurs sont-ils sans biais ?
Indication : Si $Y_1, \ldots, Y_n$, n variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$, alors :

$$\sum_i^n Y_i^2 \sim \chi^2(n) \quad et \quad \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \chi^2(n-1).$$

En particulier, si $Z \sim \chi^2(p)$, alors $\mathbb{E}[Z] = p$ et $\mathrm{Var}[Z] = 2p$.

(b) Calculer les risques quadratiques de ces deux estimateurs.

(c) Conclure.

2.3 Intervalle de confiance

Soient $X_1, \ldots, X_n$ le nombre de minutes qu'un groupe d'utilisateurs-test d'internet passent connectés par semaine. Nous modélisons ces variables aléatoires par des lois exponentielles de paramètre λ. On cherche à construire un intervalle de confiance de niveau $1-\alpha$ pour la fonction

$$q(\lambda) = \mathbb{P}_\lambda[X \geq x] = \exp(-\lambda x),$$

la probabilité que les utilisateurs-tests passent plus de x heures connectés dans la semaine.

1. Montrer que si $X \sim \exp(\lambda)$, $2\lambda X_i$ suit une loi du χ^2 à deux degrés de liberté.
2. En déduire un intervalle de confiance de niveau $1-\alpha$ pour λ, puis pour $q(\lambda)$. La loi de $v(X,\lambda) = 2\lambda \sum_{i=1}^n X_i$ dépend-elle de λ ?
3. Comment faire si le nombre d'utilisateurs-tests est important ?

2.4 Intervalle de confiance pour la loi uniforme

Soit $X_1, \ldots, X_n$ n v.a. indépendantes distribuées suivant une loi uniforme sur l'intervalle $[0, \theta]$ et soit $X_{(n)} = \max(X_1, \ldots, X_n)$.

1. Montrer que $X_{(n)}/\theta$ est une fonction pivotale pour θ.
2. En utilisant cette fonction pivotale, déterminer l'intervalle de confiance de probabilité de niveau de confiance $1-\alpha$ de longueur minimale.

3 Construction et propriétés d'estimateurs

3.1 Maximum de vraisemblance et loi uniforme

On observe $X_1, \cdots, X_n$ indépendantes et de même loi uniforme sur $[0, b]$ où $b > 0$ est le paramètre d'intérêt. On note μ l'espérance commune des X_i.

1. Ecrire le modèle statistique associé et calculer sa vraisemblance
$$\mathcal{L}(b, X_1, \ldots, X_n).$$
2. Déterminer l'estimateur $\widehat{b}_1$ du maximum de vraisemblance de b (c'est-à-dire la quantité $\widehat{b}_1 = \widehat{b}_1(X_1, \ldots, X_n)$ qui maximise la fonction $b \rightsquigarrow \mathcal{L}(b, X_1, \ldots, X_n)$.
3. Déterminer $\widehat{b}_2$ l'estimateur par méthode des moments de b, en se basant que le premier moment.
4. On opère un changement de paramètre : désormais, le paramètre d'intérêt est μ. Déterminer $\widehat{\mu}_1$, l'estimateur du maximum de vraisemblance pour le paramètre μ. (On écrira au préalable la vraisemblance du n-échantillon pour le paramètre μ).
5. Exprimer $\widehat{\mu}_2$, l'estimateur plug-in de μ, obtenu par méthode de moment.
6. Calculer le risque quadratique de $\widehat{\mu}_2$.
7. Etudier le risque quadratique de $\widehat{\mu}_1$.
8. Comparer les estimateurs $\widehat{\mu}_1$ et $\widehat{\mu}_2$: lequel est préférable ?

3.2 Estimateur du maximum de vraisemblance : cas classiques

Calculer l'estimateur du maximum de vraisemblance (s'il existe et s'il est bien défini) et étudier ses propriétés (vitesse de convergence, intervalle de confiance, loi limite, lien avec un estimateur par méthode de moment) dans les cas suivants :

1. Loi de Poisson de paramètre $\lambda > 0$ pour un n-échantillon.
2. Loi binomiale $\mathcal{B}(n, p)$ de paramètre (n, p).
3. Loi normale $\mathcal{N}(\mu, \sigma)$, où $\mu \in \mathbb{R}$ et $\sigma > 0$ pour un n-échantillon.
4. Loi de Pareto translatée de paramètres μ et x_0 de densité
$$f(y) = \frac{\mu x_0^{\mu}}{x^{1+\mu}} 1_{[x_0, +\infty[}$$
pour un n-échantillon.

4 Révisions sur l'estimation

4.1 Rappel de cours

Soit $(X_1, \dots, X_n)$ un n-échantillon de loi $\mathbb{P}_\vartheta$ pour $\vartheta \in \Theta \subset \mathbb{R}$. On suppose le modèle dominé par rapport à une mesure σ-finie μ et on note

$$f(\vartheta, x) = \frac{d\mathbb{P}_\vartheta}{d\mu}(x), \quad x \in \mathbb{R}, \vartheta \in \mathbb{R}$$

la famille de densités ainsi obtenue. On suppose *toutes les propriétés de régularité et d'intégrabilité voulues* pour la famille $(\vartheta, x) \rightsquigarrow f(\vartheta, x)$. On appelle *Information de Fisher* la quantité

$$\mathbb{I}(\vartheta) = \mathbb{E}_\vartheta\big[\big(\partial_\vartheta \log f(\vartheta, X)\big)^2\big] = -\mathbb{E}_\vartheta\big[\partial^2_\vartheta \log f(\vartheta, X)\big].$$

Si l'estimateur du maximum de vraisemblance est bien défini, si $0 < \mathbb{I}(\vartheta) < \infty$ et si l'application $(\vartheta, x) \rightsquigarrow f(\vartheta, x)$ vérifie des conditions de régularité (voir cours), on a

$$\sqrt{n}(\widehat{\vartheta}^{\mathtt{mv}} - \vartheta) \xrightarrow{d} \mathcal{N}\big(0, \tfrac{1}{\mathbb{I}(\vartheta)}\big).$$

4.2 Marqueur d'une infection

N agents infectieux agressent simultanément un organisme, lequel est muni de Q agents de défense. La réponse immunitaire est modélisée de la faccon suivante : chaque agent de défense choisit au hasard un agent infectieux (et un seul) parmi les N agresseurs, indépendamment des autres défenseurs. Un agent de défense a une probabilité $\vartheta \in (0, 1)$ d'annihiler l'agent infectieux choisi pour cible

Pour que l'organisme soit infecté, il suffit qu'un seul agent infectieux ait échappé au système de défense de l'organisme.

1. Montrer que la probabilité qu'un agent infectieux donné contamine l'organisme est
$$p_{Q,N}(\vartheta) = \Big(1 - \frac{\vartheta}{N}\Big)^Q.$$

On répète en laboratoire n scénarios indépendants d'aggression de l'organisme. Dans chaque expérience, on **marque** un agent infectieux donné. Pour l'expérience i, on note $X_i = 1$ si l'agent infectieux a contaminé l'organisme et 0 sinon.

2. On considère l'observation de $(X_1, \dots, X_n)$, où ϑ est les paramètre inconnu et Q et N sont connus. Montrer que la vraisemblance s'écrit
$$\vartheta \rightsquigarrow p_{Q,N}(\vartheta)^{\sum_{i=1}^n X_i}\big(1 - p_{Q,N}(\vartheta)\big)^{n - \sum_{i=1}^n X_i}.$$

3. Montrer que le modèle est régulier et que son information de Fisher vaut
$$\mathbb{I}(\vartheta) = \frac{\big(\partial_\vartheta p_{Q,N}(\vartheta)\big)^2}{p_{Q,N}(\vartheta)\big(1 - p_{Q,N}(\vartheta)\big)}.$$
4. Montrer que l'estimateur du maximum de vraisemblance de ϑ est bien défini, qu'il est asymptotiquement normal et calculer sa variance limite.
5. En déduire un intervalle de confiance asymptotiquement de niveau $\alpha \in (0,1)$ pour ϑ.

On suppose désormais les paramètres N et Q inconnus, et on se place dans la limite $N \approx +\infty$ en supposant $Q = Q_N \sim \kappa N$ pour un $\kappa > 0$ (donc inconnu).

6. En passant à la limite en N dans le modèle précédent, montrer que l'observation de $(X_1, \ldots, X_n)$ permet d'estimer le paramètre $\widetilde{\vartheta} = \kappa\vartheta$ et calculer l'estimateur du maximum de vraisemblance de $\widetilde{\vartheta}$.

4.3 Modèle de régression et variance inhomogène

On observe la variable aléatoire $Z = (X, Y)$ définie par
$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \sigma \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix},$$
où $(\varepsilon_1, \varepsilon_2)$ est un vecteur gaussien centré de matrice de covariance $K = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Le paramètre inconnu est $\mu = (\mu_1, \mu_2)$ et $\sigma > 0$ est connu.

1. Décrire le modèle statistique associé à l'observation Z. Est-il identifiable, dominé ? Si oui, préciser.
2. Montrer qu'il existe une transformation linéaire A de $\mathbb{R}^2$ dans $\mathbb{R}^2$ telle que AZ soit un vecteur gaussien de moyenne $A \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ et de matrice de covariance l'identité sur $\mathbb{R}^2$. Ecrire le modèle linéaire correspondant à l'observation de AZ.
3. *(Facultatif)* Calculer l'estimateur des moindres carrés pour μ.

5 Estimation numérique et maximum de vraisemblance

5.1 Efficacité à un pas

Dans un modèle régulier, l'estimateur du maximum de vraisemblance est « meilleur » que n'importe quel autre Z-estimateur au sens de l'efficacité asymptotique. Pourtant, il est parfois plus facile de mettre en œuvre un estimateur donné (par méthodes des moments par exemple) plutôt que l'estimateur du maximum de vraisemblance.

1. On peut modifier un estimateur $\widehat{\vartheta}_n$ consistant et asymptotiquement normal de sorte qu'il ait asymptotiquement le même comportement que l'estimateur du maximum de vraisemblance. On note $\ell_n(\vartheta) = \frac{1}{n}\sum_{i=1}^n \log f(\vartheta, X_i)$. Si le modèle est régulier et si $\widehat{\vartheta}_n$ est un estimateur asymptotiquement normal, alors l'estimateur modifié *

$$\widetilde{\vartheta}_n = \widehat{\vartheta}_n - \frac{\ell_n'(\widehat{\vartheta}_n)}{\ell_n''(\widehat{\vartheta}_n)}$$

vérifie

$$\sqrt{n}\big(\widetilde{\vartheta}_n - \vartheta\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{\mathbb{I}(\vartheta)}\Big)$$

en loi sous $\mathbb{P}_\vartheta$ et est donc asymptotiquement efficace.

Le choix initial pourra donc être un estimateur consistant et asymptotiquement normal, sans que l'on ait besoin de se soucier (asymptotiquement) de sa variance asymptotique.

2. Donner une esquisse de la démonstration du résultat annoncé précédemment, en identifiant les diffcultés et en faisant les hypothèses techniques voulues. *Indication : on pourra écrire*

$$\begin{aligned}\sqrt{n}\big(\widetilde{\vartheta}_n - \vartheta\big) &= \sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) - \frac{\sqrt{n}\ell_n'(\widehat{\vartheta}_n)}{\ell_n''(\widehat{\vartheta}_n)} \\ &= \sqrt{n}\big(\widehat{\vartheta}_n - \vartheta\big) - \frac{\sqrt{n}\ell_n'(\vartheta) + \sqrt{n}\big(\ell_n'(\widehat{\vartheta}_n) - \ell_n'(\vartheta)\big)}{\ell_n''(\vartheta) + \big(\ell_n''(\widehat{\vartheta}_n) - \ell_n''(\vartheta)\big)}\end{aligned}$$

et on pourra supposer que $\sqrt{n}\big(\ell_n'(\widehat{\vartheta}_n) - \ell_n'(\vartheta)\big) \xrightarrow{\mathbb{P}_\vartheta} 0$ *et* $\big(\ell_n''(\widehat{\vartheta}_n) - \ell_n''(\vartheta)\big) \xrightarrow{\mathbb{P}_\vartheta} 0$.

∗. Il faut bien sûr que le dénominateur du terme de correction soit non nul. On admettra que l'événement sur lequel il est bien défini a une $\mathbb{P}_\vartheta$-probabilité qui tend vers 1 si le modèle est régulier.

5.2 Modèle de Cauchy

On observe un n-échantillon de $(X_1, \ldots, X_n)$ de variables aléatoires de Cauchy de densité

$$f(\vartheta, x) = \frac{1}{\pi\big(1 + (x - \vartheta)^2\big)}, \quad x \in \mathbb{R}$$

par rapport à la mesure de Lebesgue.

1. Montrer que la densité $f(\vartheta, \cdot)$ n'a pas de moment d'ordre k pour $k \geq 1$. (Le choix $g(x) = x^k$ avec k entier ne s'applique pas ici pour la méthode des moments).
2. Prenons $g(x) = \text{signe}(x)$, avec

$$\text{signe}(x) = \begin{cases} -1 & \text{si} \quad x \leq 0 \\ 1 & \text{si} \quad x > 0. \end{cases}$$

   Montrer que

$$\mathbb{E}_\vartheta\big[g(X_1)\big] = \int_{\mathbb{R}} \text{signe}(x) f(\vartheta, x) dx = 1 - 2F(-\vartheta),$$

   où

$$F(t) = \frac{1}{\pi} \int_{-\infty}^{t} \frac{dt}{1 + t^2} = \frac{1}{\pi} \text{Arctg}(t) + \frac{1}{2}.$$

   En déduire l'estimateur d'où l'estimateur

$$\widehat{\vartheta}_n = \text{tg}\left(\frac{\pi}{2n} \sum_{i=1}^{n} \text{signe}(X_i) \right).$$

3. On admettra que est consistant et asymptotiquement normal. En évaluant $\frac{\ell_n'(\widehat{\vartheta}_n)}{\ell_n''(\widehat{\vartheta}_n)}$, mettre en oeuvre l'efficacité à un pas. Pour cela, étant donné un estimateur $\widehat{\vartheta}_n$ et un entier n, on simule un n-échantillon et on calcule $(\widehat{\vartheta}_n - \vartheta)^2$, pour un choix de ϑ (par exemple $\vartheta = 1$). On répète M fois ce procédé (par exemple $M = 1000$), obtenant ainsi $\widehat{\vartheta}_n^{(j)}, j = 1, \ldots, M$, puis on calcule

$$e(\widehat{\vartheta}_n) = M^{-1} \sum_{j=1}^{M} \big(\widehat{\vartheta}_n^{(j)} - \vartheta\big)^2.$$

   On représente sur le même graphe l'évolution de cette quantité, pour l'estimateur par moment et l'estimateur corrigé lorsque n augmente, que l'on compare à l'information de Fisher du modèle que l'on aura préalablement calculé.
4. Justifier cette comparaison

5.3 (Facultatif.) Emission de particules

Une source émet des particules de type A avec probabilité ϑ et de type B avec probabilité $1-\vartheta$, où $\vartheta \in \Theta = (0,1)$. On mesure l'énergie des particules, qui est distribuée selon une densité f_0 connue pour les particules de type A et f_1 pour les particules de type B. Si l'on détecte n particules avec des énergies $X_1, \ldots, X_n$, quelle est la valeur de ϑ ? En postulant que l'observation est un n-échantillon, la fonction de vraisemblance de l'expérience statistique engendrée par l'observation s'écrit

$$\mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \prod_{i=1}^{n} \big(\vartheta f_0(X_i) + (1-\vartheta) f_1(X_i)\big),$$

de sorte que

$$\partial_\vartheta \log \mathcal{L}_n(\vartheta, X_1, \ldots, X_n) = \sum_{i=1}^{n} \frac{f_0(X_i) - f_1(X_i)}{\vartheta f_0(X_i) + (1-\vartheta) f_1(X_i)}.$$

La résolution de l'équation de vraisemblance associée est d'autant plus difficile que n est grand. Supposons que $\int_{\mathbb{R}} \big(F_0(x) - F_1(x)\big)^2 dx < +\infty$, où $F_i(x) = \int_{-\infty}^{x} f_i(t)dt$, $i = 1, 2$. Soit $\widehat{\vartheta}_n$ l'estimateur qui minimise

$$a \rightsquigarrow \int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_a(x)\big)^2 dx,$$

avec

$$F_a(x) = aF_0(x) + (1-a)F_1(x),$$

et $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \le x}$ désigne la fonction de répartition empirique de F. En dérivant par rapport à la variable a, on obtient

$$\int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_a(x)\big)\big(F_0(x) - F_1(x)\big)dx = 0,$$

d'où

$$\widehat{\vartheta}_n = \frac{\int_{\mathbb{R}} \big(\widehat{F}_n(x) - F_0(x)\big)\big(F_0(x) - F_1(x)\big)dx}{\int_{\mathbb{R}} \big(F_0(x) - F_1(x)\big)^2 dx}.$$

On admettra que $\widehat{\vartheta}_n$ est asymptotiquement normal. Alors l'estimateur modifié

$$\widetilde{\vartheta}_n = \widehat{\vartheta}_n - \frac{\partial_\vartheta \log \mathcal{L}_n(\widehat{\vartheta}_n, X_1, \ldots, X_n)}{\partial^2_\vartheta \log \mathcal{L}_n(\widehat{\vartheta}_n, X_1, \ldots, X_n)}$$

où

$$\partial^2_\vartheta \log \mathcal{L}_n(\widehat{\vartheta}_n, X_1, \ldots, X_n) = -\sum_{i=1}^{n} \frac{\big(f_0(X_i) - f_1(X_i)\big)^2}{\big(\vartheta f_0(X_i) + (1-\vartheta) f_1(X_i)\big)^2}$$

est asymptotiquement efficace, et sa variance asymptotique est l'information de Fisher du modèle

$$\mathbb{I}(\vartheta) = \int_{\mathbb{R}} \frac{\big(f_0(x) - f_1(x)\big)^2}{\vartheta f_0(x) + (1-\vartheta) f_1(x)} dx.$$

Verifier numériquement ce résultat avec la même méthodologie que pour l'exercice précédent.

6 Introduction aux tests

6.1 Neyman-Pearson : loi exponentielle

Soit $n \geq 1$ un entier. On observe

$$X_1, \ldots, X_n$$

où les variables aléatoires X_i sont indépendantes, de même loi exponentielle de paramètre $\lambda > 0$, c'est-à-dire de densité

$$x \rightsquigarrow \lambda \exp\big(-\lambda x\big) 1_{\{x \geq 0\}}.$$

1. Ecrire le modèle statistique engendré par l'observation de $(X_1, \ldots, X_n)$.
2. Calculer l'estimateur du maximum de vraisemblance $\widehat{\lambda}_n^{\mathtt{mv}}$ de λ.
3. Montrer que $\widehat{\lambda}_n^{\mathtt{mv}}$ est asymptotiquement normal et calculer sa variance limite.
4. Soient $0 < \lambda_0 < \lambda_1$. Construire un test d'hypothèse de

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

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

6.2 Contrôle de qualité

On s'intéresse au nombre d'objets défectueux fabriqués dans une usine. On effectue un contrôle de qualité en prélevant au hasard sur des chaînes de fabrication similaires, n objets que l'on classifie en objet défectueux ou non. On code par 1 l'objet défectueux et par 0 l'objet non défectueux. On note X_i le résultat du contrôle sur l'objet i et $S_n = \sum_{i=1}^n X_i$.

Le fabricant de ces objets désire tester si la norme de ϑ_0, au plus, de proportion d'objets défectueux, stipulée dans son contrat a été respectée.

1. Ecrire le modéle statistique correspondant.
2. Ecrire les hypothèses testées par le fabricant.
3. Proposer l'allure de la région de rejet $\mathcal{R}$ en fonction de S_n et d'un seuil déterministe c que l'on fixera dans la suite.
4. Exprimer l'erreur de première espèce du test ainsi construit.
5. Pour toute loi de l'alternative H_1, exprimer la probabilité d'accepter H_0 à tort, c'est-à-dire l'erreur de seconde espèce du test.
6. Montrer † que pour c fixé, la fonction : $\vartheta \rightsquigarrow \mathbb{P}_\theta(S_n \geq c)$ est croissante.

†. En notant $f_\theta(k)$ la probabilité qu'une variable de loi Binomiale de paramètre (ϑ, n) soit égale à k, on pourra étudier le rapport $\frac{f_\vartheta(k)}{f_{\vartheta'}(k)}$ lorsque $\vartheta' < \vartheta$.

7. En déduire une expression de l'erreur de première espèce et de l'erreur de seconde espèce.
8. On cherche c minimisant la somme de ces deux erreurs. Qu'obtient-on ? Conclure.

6.3 Neyman-Pearson et loi discrète

On considre l'expérience statistique engendrée par l'observation d'une seule variable aléatoire X de loi de Poisson de paramètre $\vartheta > 0$. On teste $H_0 : \vartheta = \vartheta_0$ contre $H_1 : \vartheta = \vartheta_1$, avec $\vartheta_0 \neq \vartheta_1$.

1. Ecrire l'expérience statistique associée.
2. Montrer que la forme de la zone de rejet du test de Neyman-Pearson au niveau α, s'il existe, doit s'écrire

$$\mathcal{R}_{n,\alpha} = \Big\{ \exp\big(-(\vartheta_1 - \vartheta_0)\big)(\vartheta_1\vartheta_0^{-1})^X \geq c(\alpha)\Big\},$$

   où le choix de $c(\alpha)$ garantit que le test est exactement de niveau α.
3. Montrer que l'équation déterminant $c(\alpha)$ n'a pas de solution en général.
4. Proposer un choix de $c(\alpha)$ de sorte que l'erreur de première espèce du test soit inférieure ou égale à α.
5. Dans quel sens le test ainsi construit est-il optimal ?
6. On suppose que $\vartheta_1 > \vartheta_0$. Donner la règle de décision explicite au niveau $\alpha = 5\%$ pour $\vartheta_0 = 0.5$. A.N.

$$\mathbb{P}_{\vartheta_0}\big[X > 9\big] = 0,032, \quad \text{et} \quad \mathbb{P}_{\vartheta_0}\big[X > 8\big] = 0,068.$$

7 Tests : calcul numérique de la puissance

7.1 Mesure par Monte-Carlo de la puissance d'un test

On considère un échantillon $X_1, \dots, X_n$ de loi $\mathcal{N}(\mu, \sigma^2)$, où le paramètre inconnu est $\vartheta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$.

1. On pose $\widehat{\sigma}_n = \Big((n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2\Big)^{1/2}$. Montrer que le test de zone de rejet
$$\mathcal{R}_{n,\alpha} = \Big\{ \frac{\big|\bar{X}_n - \mu_0\big|}{\widehat{\sigma}_n} \geq t_{n,\alpha} \Big\}$$
est de niveau $\alpha \in (0,1)$ pour tester $\mu = \mu_0$ contre $\mu \neq \mu_0$ pour un seuil $t_{n,\alpha}$ que l'on déterminera.
2. On suppose $\sigma = 1$. Donner une expression de la puissance du test en tout point $\vartheta = (\mu, 1)$ avec $\mu \neq \mu_0$.
3. Tracer graphiquement la fonction de puissance $(\mu, 1) \rightsquigarrow \mathbb{P}_{\mu,1}(\mathcal{R}_{n,\alpha})$ pour $\mu \neq \mu_0$ pour plusieurs valeurs de n ($n = 10, 100$ et 10^4) et un choix de μ_0. Quelle est sa limite lorsque $\mu \to \mu_0$? Est-ce surprenant ?
4. En répétant l'expérience M fois, calculer M fois la p-valeur du test sous l'hypothèse (ce qui fournit des données $p_0^{(1)}, \dots, p_0^{(M)}$) et sur un point de l'alternative où la puissance est "assez grande" (ce qui fournit des données $p_1^{(0)}, \dots, p_M^{(M)}$). Tracer les deux histogrammes des distributions des $p_0^{(i)}$ d'une part, et des $p_1^{(i)}$ d'autre part. Commenter.
5. On suppose toujours $\sigma = 1$. Proposer un autre test de niveau α et démontrer numériquement que sa puissance est meilleure que celle du test précédent.
6. On suppose désormais que
$$X_i = \mu + \sigma(\xi_i^2 - 1)$$
où les ξ_i sont des variables aléatoires gaussiennes centrées réduites. On suppose pour simplifier $\sigma = 1$. Evaluer numériquement (par méthode de Monte-Carlo ‡) l'erreur de première espèce et la fonction de puissance du test précédent dans ce contexte, avec les mêmes valeurs que précédemment pour n. Que conclure ?

7.2 Test du signe (Facultatif)

On considère le modèle statistique engendré par l'observation du vecteur aléatoire $Z^n = \big(X^n, Y^n\big)$ de $\mathbb{R}^{2n}$ défini par
$$X^n = \big(X_1, \dots, X_n\big),$$

‡. c'est-à-dire en répétant M fois l'expérience en en comptant le nombre de succés relatifs pour évaluer une probabilité.

où les variables aléatoires X_i sont indépendantes, de même loi ayant une fonction de répartition F continue, et

$$Y^n = \big(Y_1, \dots, Y_n\big),$$

où les variables aléatoires Y_i sont indépendantes, de même loi ayant une fonction de répartition G continue. On suppose que les vecteurs X^n et Y^n sont indépendants. On considère le test d'hypothèse

$$H_0 \ : \ F = G \quad \text{contre} \quad H_1 \ : \ F \neq G.$$

1. Montrer que
$$\mathbb{P}\big[X_i = Y_i\big] = 0$$
et en déduire que si $F = G$,
$$\mathbb{P}\big[X_i > Y_i\big] = \frac{1}{2}.$$
2. On pose
$$N\big(Z^n\big) = \sum_{i=1}^{n} 1_{\big\{X_i > Y_i\big\}}.$$
Quelle est la loi de N sous H_0 ?
3. En déduire que le test simple défini par la zone de rejet
$$\mathcal{R}(c) = \Big\{\Big|N(Z^n) - \frac{n}{2}\Big| \geq c\Big\}$$
permet de construire un test de niveau $\alpha \in (0,1)$ de H_0 contre H_1 pour un choix $c = c(\alpha) > 0$ que l'on précisera. Parmi tous les choix possibles de $c(\alpha)$, lequel préférer ?
4. Donner un équivalent de $c(\alpha) = c_n(\alpha)$ lorsque $n \to \infty$.
5. Montrer que néanmoins le test n'est pas consistant.
6. Mettre en oeuvre numériquement le test du signe pour des distributions cibles que l'on choisira.

8 Tests asymptotiques

8.1 Test du signe

On considère le modèle statistique engendré par l'observation du vecteur aléatoire $Z^n = \big(X^n, Y^n\big)$ de $\mathbb{R}^{2n}$ défini par

$$X^n = \big(X_1, \ldots, X_n\big),$$

où les variables aléatoires X_i sont indépendantes, de même loi ayant une fonction de répartition F continue, et

$$Y^n = \big(Y_1, \ldots, Y_n\big),$$

où les variables aléatoires Y_i sont indépendantes, de même loi ayant une fonction de répartition G continue. On suppose que les vecteurs X^n et Y^n sont indépendants. On considère le test d'hypothèse

$$H_0 \ : \ F = G \quad \text{contre} \quad H_1 \ : \ F \neq G.$$

1. Montrer que
$$\mathbb{P}\big[X_i = Y_i\big] = 0$$
et en déduire que si $F = G$,
$$\mathbb{P}\big[X_i > Y_i\big] = \frac{1}{2}.$$

2. On pose
$$N\big(Z^n\big) = \sum_{i=1}^{n} 1_{\{X_i > Y_i\}}.$$
Quelle est la loi de N sous H_0 ?

3. En déduire que le test simple défini par la zone de rejet
$$\mathcal{R}(c) = \left\{ \left| N(Z^n) - \frac{n}{2} \right| \geq c \right\}$$
permet de construire un test de niveau $\alpha \in (0,1)$ de H_0 contre H_1 pour un choix $c = c(\alpha) > 0$ que l'on précisera. Parmi tous les choix possibles de $c(\alpha)$, lequel préférer ?

4. Donner un équivalent de $c(\alpha) = c_n(\alpha)$ lorsque $n \to \infty$.

5. Montrer que néanmoins le test n'est pas consistant.

8.2 Observations inhomogènes

On observe n variables aléatoires $(X_1, \ldots, X_n)$ indépendantes et identiquement distribuées de loi $\mathbb{P}$. Après une certaine action, on observe $m = m_n$ variables aléatoires $(Y_1, \ldots, Y_{m_n})$ indépendantes des X_i, qui mesurent l'effet de cette action. Les Y_i sont indépendantes, identiquement distribuées, de loi $\mathbb{Q}$.

On suppose

$$\lim_{n\to\infty} \frac{n}{m_n} = \gamma \in (0, \infty),$$

et on veut tester

$$H_0 : \mathbb{P} = \mathbb{Q} \text{ (absence d'effet de l'action) contre } H_1 : \mathbb{P} \neq \mathbb{Q}.$$

8.2.1 Comparaison des moyennes

On suppose que les quantités

$$v(\mathbb{P}) = \mathrm{Var}[X_1] \text{ et } v(\mathbb{Q}) = \mathrm{Var}[Y_1]$$

sont bien définies et connues. On pose

$$\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i \text{ et } \overline{Y}_{m_n} = {m_n}^{-1} \sum_{i=1}^{m_n} Y_i.$$

1. On se place sous l'hypothèse nulle H_0. Quelle est la loi limite lorsque $n \to \infty$ de
$$\sqrt{n}\big(\overline{X}_n - \overline{Y}_{m_n}\big) \ ?$$
2. En déduire un test asymptotiquement de niveau $\alpha \in (0, 1)$ de H_0 contre H_1 que l'on explicitera.
3. Le test est-il consistant ?
4. On suppose que $\mathbb{P}$ est la loi exponentielle de paramètre λ et que $\mathbb{Q}$ est la loi exponentielle de paramètre λ^2, pour $\lambda > 0$. Reprendre les questions 1 et 2 dans ce contexte. Le test devient-il consistant ?

8.2.2 Approche par maximum de vraisemblance

On considère l'expérience $\mathcal{E}^n$ engendrée par $Z^n = (X_1, \ldots, X_n, Y_1, \ldots, Y_{m_n})$, où les X_i suivent la loi exponentielle de paramètre λ et les Y_i la loi exponentielle de paramètre λ^2. On note $\mathbb{P}_\lambda^n$ la loi de Z^n sur $\mathbb{R}_+^{n+m_n}$.

5. Montrer que $\mathcal{E}^n$ est dominée par la mesure de Lebesgue sur $\mathbb{R}_+^{n+m_n}$ et expliciter sa fonction de vraisemblance.
6. Montrer que l'estimateur du maximum de vraisemblance $\widehat{\lambda}_n^{\mathtt{mv}}$ est bien défini et le calculer.

7. Calculer l'information de Fisher

$$\mathbb{I}_n(\lambda) = \mathbb{E}_\lambda^n\big[\big(\partial_\lambda \log f(\lambda; X_1, \dots, X_n, Y_1, \dots, Y_{m_n})\big)^2\big],$$

où $\mathbb{E}_\lambda^n$ désigne l'espérance pour la loi de $Z^n = (X_1, \dots, X_n, Y_1, \dots, Y_{m_n})$ lorsque le paramètre est λ.

8. (*Facultatif.*) Montrer que

$$\sqrt{n}\big(\widehat{\lambda}_n^{\mathtt{mv}} - \lambda\big) \xrightarrow{d} \mathcal{N}\big(0, \tfrac{\lambda^2}{1+4\gamma^{-1}}\big).$$

9. En déduire un test asymptotiquement de niveau $\alpha \in (0,1)$ de H_0 : $\lambda = 1$ contre $H_1 : \lambda \neq 1$ et consistant (on prendra soin de démontrer que le test ainsi construit est bien consistant).
10. (*Facultatif.*) Le test basé sur $\widehat{\lambda}_n^{\mathtt{mv}}$ est-il asymptotiquement plus puissant que le test construit dans la Question 4 ?