

FEUILLE DE TRAVAUX PRATIQUES #6

L'objet de la statistique (inférentielle) est le suivant : on observe n fois un phénomène aléatoire X de loi inconnue \mathbb{P}_X et on recueille ainsi des données (x_1, \dots, x_n) . On fait alors l'hypothèse que les données x_i sont les réalisations de variables aléatoires indépendantes X_i de même loi que la loi inconnue, c'est-à-dire $x_i = X_i(\omega)$ où $\mathbb{P}_{X_i} = \mathbb{P}_X$. On souhaite alors déterminer quelle est la loi \mathbb{P}_X , ou plus modestement d'estimer certaines de ses caractéristiques (moyenne, variance etc.).

Ce document a pour but d'illustrer, à l'aide de Scilab, quelques résultats de base concernant l'estimation inférentielle. Les exercices à traiter en priorité sont indiqués en rouge.

1 Estimation paramétrique

On parle d'estimation paramétrique lorsque l'on ne s'intéresse précisément qu'à certaines caractéristiques fini-dimensionnelles de la loi P_X , ou encore lorsque l'on fait l'hypothèse supplémentaire que la loi inconnue \mathbb{P}_X appartient à une famille de lois connue, famille indexée par un paramètre à valeurs dans un espace de dimension finie. Par exemple, la loi inconnue peut être une loi de Bernoulli $\mathcal{B}(p)$, pour un certain réel $p \in [0, 1]$, elle peut être une exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda > 0$, ou encore une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$. Estimer la loi inconnue \mathbb{P}_X revient alors à estimer la valeur du/des paramètre(s).

1.1 Sur les estimateurs

On rappelle qu'un estimateur θ_n d'un paramètre inconnu θ de la loi \mathbb{P}_X est simplement une fonction mesurable des données (x_1, \dots, x_n) , ou par extension une fonction mesurable de l'échantillon (X_1, \dots, X_n) . Bien entendu, tous les estimateurs ne se valent pas et on cherche en général :

- à minimiser la distance (à préciser) entre estimateur et paramètre à estimer,
- à maximiser la vitesse de convergence de l'estimateur vers sa limite,
- à contrôler les fluctuations autour de la limite pour obtenir des intervalles de confiance.

On introduit en particulier les notions d'estimateur

- sans biais si $\mathbb{E}[\theta_n] = \theta$,
- (fortement) consistant si θ_n converge (p.s.) en probabilité vers θ ,
- asymptotiquement normal si $\sqrt{n}(\mathbb{E}[\theta_n] - \theta)$ converge vers une variable gaussienne.

Exercice 1 Support d'une variable uniforme

Le vecteur de données \mathbf{x} téléchargeable [ici](#) correspond à $n = 1000$ réalisations de variables indépendantes, de loi commune uniforme dans un intervalle $[0, \theta]$, où $\theta > 0$ est inconnu.

1. Expliciter l'estimateur empirique $\bar{\theta}_n$ et l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ .
2. Illustrer le fait que ces estimateurs sont consistants. Quelle est leur vitesse de convergence ?
3. L'estimateur $\hat{\theta}_n$ est-il asymptotiquement normal ?
4. Pouvez-vous donner un intervalle de confiance pour le paramètre θ ?

Exercice 2 Quantiles empiriques.

On considère (X_1, \dots, X_n) un n -échantillon d'une loi à densité f et de fonction de répartition F . On note $(X_{(1)}, \dots, X_{(n)})$ la statistique d'ordre associée, que l'on pourra obtenir grâce à la fonction de tri `gsort`. Pour tout $p \in]0, 1[$, le quantile d'ordre p de la loi sous-jacente, noté $k(p)$ est alors défini par $k(p) := F^{-1}(p)$. Le quantile empirique d'ordre p associé à l'échantillon est lui défini par $\widehat{k}_n(p) := X_{(\lfloor np \rfloor + 1)}$.

1. Dans le cas où la loi sous-jacente est la loi normale centrée réduite, illustrer le fait que le quantile empirique est un estimateur fortement consistant du quantile (théorique), i.e. que pour tout $p \in]0, 1[$, lorsque n tend vers l'infini

$$\widehat{k}_n(p) \xrightarrow{ps} k(p).$$

2. Illustrer le fait que le quantile empirique est asymptotiquement normal, autrement dit pour tout $p \in]0, 1[$, lorsque n tend vers l'infini

$$\sqrt{n} \left(\widehat{k}_n(p) - k(p) \right) \xrightarrow{d} \mathcal{N}(0, \sigma_p^2), \quad \text{où } \sigma_p^2 = \frac{p(1-p)}{f(k(p))^2}.$$

1.2 Intervalles de confiance exacts

On rappelle qu'un intervalle de confiance exact, de niveau de confiance $1 - \alpha$, pour une caractéristique θ de la loi inconnue \mathbb{P}_X , est un intervalle I_n tel que

$$\mathbb{P}_X(\theta \in I_n) \geq 1 - \alpha.$$

Un tel intervalle est rarement explicitable, sauf si, par exemple, l'estimateur de θ a une loi connue.

Exercice 3 Estimation gaussienne.

Le vecteur de données `x` téléchargeable [ici](#) correspond à $n = 1000$ réalisations indépendantes de variables $\mathcal{N}(m, \sigma^2)$.

1. On suppose dans cette question que $m = 1$ et que σ est inconnue. Quelle est- alors la loi de $\sigma^{-2} \sum_{k=1}^n (X_k - m)^2$? À l'aide de la fonction `cdfchi`, en déduire un intervalle de confiance pour la variance σ^2 .
2. On suppose maintenant que m et σ sont inconnues. On désigne par \bar{X}_n la moyenne empirique de l'échantillon. Quelle est la loi de $\sigma^{-2} \sum_{k=1}^n (X_k - \bar{X}_n)^2$? En déduire un intervalle de confiance pour la variance σ^2 .
3. On suppose encore que m et σ sont inconnues. Quelle est la loi de la variable ci-dessous

$$\sqrt{n-1} \frac{\sum_{k=1}^n (X_k - m)}{\sum_{k=1}^n (X_k - \bar{X}_n)} ?$$

En déduire un intervalle de confiance pour la moyenne m .

Exercice 4 Estimation exponentielle.

Le vecteur de données `y` téléchargeable [ici](#) correspond à des réalisations indépendantes de variables exponentielles $\mathcal{E}(\lambda)$ pour un $\lambda > 0$ inconnu.

1. Quel est l'estimateur du maximum de vraisemblance de la moyenne $1/\lambda$?
2. Quel est sa loi?
3. En déduire un intervalle de confiance exact de niveau 95% pour le paramètre λ .

1.3 Intervalle de confiance asymptotique

On rappelle qu'un intervalle de confiance asymptotique, de niveau de confiance $1 - \alpha$, pour une caractéristique θ de la loi inconnue \mathbb{P}_X , est un intervalle I_n tel que

$$\lim_{n \rightarrow +\infty} \mathbb{P}_X(\theta \in I_n) \geq 1 - \alpha.$$

En vertu du théorème limite central, ou de ses variantes, un tel intervalle est plus facilement explicitable qu'un intervalle de confiance exact.

Exercice 5 Distance aléatoire.

On tire uniformément et indépendamment deux points A et B dans le carré $[0, 1]^2$. On note X la distance euclidienne entre A et B .

1. Exprimer la distance moyenne $\mathbb{E}[X]$ comme une intégrale multiple.
2. Estimer $\mathbb{E}[X]$ par la méthode de Monte-Carlo.
3. Montrer que $\mathbb{E}[X^2] = 1/3$ et en déduire un intervalle de confiance asymptotique de niveau 95% pour la distance moyenne $\mathbb{E}[X]$.

Exercice 6 Référendum.

Le vecteur de données \mathbf{z} téléchargeable [ici](#) correspond à des réalisations indépendantes de variables de Bernoulli de paramètre p inconnu.

1. Quel est l'estimateur du maximum de vraisemblance de p .
2. Donner un intervalle de confiance asymptotique de niveau 95% pour p .

2 Estimation non paramétrique

Par rapport à l'estimation paramétrique où l'on ne s'intéresse qu'à certaines caractéristiques fini-dimensionnelles de la loi P_X , ou l'on fait une hypothèse a priori sur la nature de la loi, l'objet de l'estimation non paramétrique est d'estimer la loi P_X elle-même, via par exemple sa fonction de répartition, sa densité si elle existe, sa fonction caractéristique etc., autant de quantités à valeurs dans des espaces fonctionnels de dimension infinie.

2.1 Fonction de répartition empirique

Si l'on souhaite estimer la fonction de répartition d'un n -échantillon (X_1, \dots, X_n) de loi inconnue, le choix le plus naturel consiste à considérer la fonction de répartition empirique qui est définie pour tout $x \in \mathbb{R}$ par

$$F_n(x) := \frac{\#\{1 \leq k \leq n, X_k \leq x\}}{n}.$$

Le théorème de Glivenko–Cantelli garantit alors que, uniformément en x , $F_n(x)$ est un estimateur consistant de $F(x)$. Par ailleurs, sous des hypothèses de régularité, le théorème de Kolmogorov–Smirnov permet d'obtenir facilement une région de confiance pour F (pour la norme infinie).

On renvoie au TP sur les tests non paramétriques pour les différentes illustrations des théorèmes évoqués plus haut.

2.2 Estimateur à noyau d'une densité

On dispose de données (x_1, \dots, x_n) dont on suppose qu'elles sont les réalisations d'un échantillon (X_1, \dots, X_n) où la loi de X_1 est inconnue, tout au plus sait-on qu'elle admet une densité f . Comment estimer cette densité? Une idée naturelle consiste à utiliser la fonction de répartition empirique F_n associée à l'échantillon (X_1, \dots, X_n) . Malheureusement, la fonction F_n n'est pas dérivable et ne peut donc pas considérer sa dérivée f_n qui serait un candidat naturel pour estimer la densité inconnue f . Cependant, on peut régulariser F_n par une suite de noyau. Soit en effet une famille de noyaux $(K_\varepsilon)_{\varepsilon>0}$ tels que

$$K_\varepsilon > 0, \quad \int_{\mathbb{R}} K_\varepsilon(x) dx = 1, \quad K_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} \delta_0.$$

Par exemple, on peut considérer des familles de noyaux du type $K_\varepsilon(x) \propto K(x/\varepsilon)$ où

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{ou encore} \quad K(x) = \frac{3}{4} (1 - x^2) \mathbb{1}_{[-1,1]}(x).$$

On introduit alors l'estimateur à noyau

$$\hat{f}_n = \frac{1}{n\varepsilon_n} \sum_{k=1}^n K\left(\frac{x - X_k}{\varepsilon_n}\right),$$

où la suite ε_n est à calibrer. On admettra que le choix $\varepsilon_n = n^{-1/5}$ est opérant.

Exercice 7 Estimation d'une densité.

Les données téléchargeables [ici](#) correspondent à des réalisations d'un n -échantillon (X_1, \dots, X_n) de variables à densité, densité inconnue que l'on souhaite estimer.

1. Tracer l'histogramme empirique associé aux données pour obtenir l'allure de la densité.
2. Implémenter la méthode à noyau décrite ci-dessus pour estimer f . Superposer le graphe de l'estimateur \hat{f}_n avec l'histogramme empirique.