Feuille de travaux pratiques - Python #6

Dans ce TP nous nous intéresserons aux estimateurs et à leurs propriétés.

1 Rappels sur les estimateurs

Etant données $(x_1, \cdots, x_n)$, n réalisations indépendantes d'un phénomène aléatoire X de loi inconnue $\mathbb{P}_X$, l'inférence statistique a pour but de donner des informations sur cette loi. On parle

- d'**estimation paramétrique** lorsque l'on s'intéresse à des caractéristiques fini-dimensionnelles de la loi $\mathbb{P}_X$: espérance, variance, etc. Ou si on fait l'hypothèse que la loi inconnue $\mathbb{P}_X$ appartient à une famille de lois connue, indexée par un paramètre. Estimer la loi inconnue $\mathbb{P}_X$ revient alors à estimer la valeur du/des paramètre(s).
- d'**estimation non paramètrique** lorsque les quantités à estimer sont à valeurs dans des espaces fonctionnels de dimension infinie : fonction de répartition, densité (si elle existe), fonction caractéristique, etc.

Notons θ un paramètre à estimer, le but est de construire un estimateur $\hat{\theta}_n$, fonction mesurable de l'échantillon, qui nous donnera des informations sur ce paramètre.

⇝ Comment construire un estimateur ?
Les méthodes les plus courantes sont

- la **méthode des moments** :
  On estime les moments par leur version empirique
- la **méthode du maximum de vraisemblance** :
  La vraisemblance d'un échantillon est la probabilité d'avoir obtenu cet échantillon sous le modèle choisi. Ainsi, plus la probabilité d'avoir obtenu les observations est grande, plus le modèle est proche de la réalité. On choisit donc le modèle pour lequel la vraisemblance est la plus élevée : $\hat{\theta}_n = \arg\max_\theta \mathcal{L}(\theta, (x_1, \cdots, x_n))$.

Bien entendu, tous les estimateurs ne se valent pas, certains ont de meilleurs propriétés statistiques que d'autres...

⇝ Comment juger de la qualité d'un estimateur ?
On cherche à

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

Des critères de performance sont :

- Le **biais** : l'estimateur est dit sans biais si $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$. Il est dit asymptotiquement sans biais si $\mathbb{E}_\theta[\hat{\theta}_n] \underset{n\to+\infty}{\longrightarrow} \theta$. En moyenne l'estimateur se comporte comme le paramètre cible.

- Le **risque quadratique**

$$R(\hat{\theta}_n, \theta) = \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 = \text{Biais}(\hat{\theta}_n, \theta)^2 + \text{Var}_\theta(\hat{\theta}_n).$$

  Cela décrit la variabilité autour de la moyenne. Un estimateur $\hat{\theta}_n^1$ est préférable à un estimateur $\hat{\theta}_n^2$ lorsque son risque quadratique est inférieur.
- La **(forte) consistance** : si $\hat{\theta}_n$ converge (p.s.) en probabilité vers θ.
- La **vitesse de convergence** $(v_n)_n$: si pour tout θ, il existe une loi $l(\theta)$ telle que $v_n(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}/\mathbb{P}_\theta} l(\theta)$.
- L'**asymptotique normalité** si $\sqrt{n}(\hat{\theta}_n - \theta)$ converge en loi vers une variable gaussienne.

Pour contrôler les fluctuations de l'estimateur autour du paramètre, on utilise des intervalles de confiance. Un **intervalle de confiance**, de niveau de confiance $1-\alpha$, 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. C'est pourquoi on s'intéresse souvent aux intervalles de confiance asymptotiques. Un **intervalle de confiance asymptotique**, de niveau de confiance $1-\alpha$, est un intervalle I_n tel que

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

On obtient de tels intervalles à l'aide de la propriété d'asymptotique normalité, du lemme de Slutsky, de la δ-méthode, etc. Pour plus de détails, voir ici

2 Estimation paramétrique

2.1 Sur les estimateurs

Exercice 1. *Support d'une variable uniforme*
Le vecteur de données $\boldsymbol{x}$ 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 $\widehat{\theta}_n$ de θ.*
2. *Illustrer le fait que ces estimateurs sont consistants. Quelle est leur vitesse de convergence ?*
3. *L'estimateur $\widehat{\theta}_n$ est-il asymptotiquement normal ?*
4. *Pouvez-vous donner un intervalle de confiance pour le paramètre θ ?*

Exercice 2. *Quantiles empiriques.*
On considère $(X_1, \ldots, X_n)$ un $n-$échantillon d'une loi à densité f et de fonction de répartition F. On note $(X_{(1)}, \ldots, X_{(n)})$ la statistique d'ordre associée. 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*

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

2.2 Intervalles de confiance exacts

Exercice 3. *Estimation gaussienne.*
Le vecteur de données $\boldsymbol{x}$ *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* $\sigma^{-2}\sum_{k=1}^n (X_k - m)^2$ *? 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*

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

En déduire un intervalle de confiance pour la moyenne m*.*

Exercice 4. *Estimation exponentielle.*
Le vecteur de données $\boldsymbol{y}$ *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*.*

2.3 Intervalle de confiance asymptotique

Exercice 5. *Distance aléatoire.*
On tire uniformément et indépendamment deux points A *et* B *dans le carré* $[0,1]^2$*. On note* X *a 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 $\boldsymbol{z}$ *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*.*

3 Estimation non paramétrique

3.1 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).

Pour les différentes illustrations des théorèmes évoqués plus haut, voir le TP précédent.

3.2 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 $(X_1, \dots, X_n)$. Malheurement, 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

$$K_\varepsilon > 0, \qquad \int_{\mathbb{R}} K_\varepsilon(x)dx = 1, \qquad K_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} \delta_0.$$

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

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

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 $\widehat{f_n}$ avec l'histogramme empirique.*

Références

[CBCC16] Alexandre Casamayou-Boucau, Pascal Chauvin, and Guillaume Connan. *Programmation en Python pour les mathématiques - 2e éd.* Dunod, Paris, 2e édition edition, January 2016.

[CV12] Benoît Cadre and Céline Vial. *Statistique mathématique : cours et exercices corrigés.* Références sciences. Ellipses, Paris, 2012.

[Vig18] Vincent Vigon. *python proba stat.* Independently published, October 2018.