Chapitre 9

Estimateurs au maximum de vraisemblance

Avec ce chapitre nous commençons l'étude de quelques outils centraux de la statistique.

9.1 Estimateur

Définition : Soit $n > 0$ un entier. Nous appellerons n-échantillon d'une loi $\mathcal{L}$ toute suite $X_1, \ldots, X_n$ de v.a. indépendantes de loi $\mathcal{L}$.

La statistique-pratique est un ensemble de techniques de traitement de données qui, face à la donnée de n nombres (ou plus généralement vecteurs) $x_1, \ldots, x_n$ produits par "échantillonage" - c'est-à-dire selon un protocole expérimental propre au domaine considéré (sociologie, contrôle de qualité, etc.) - choisit un n-échantillon au sens de la définition ci-dessus pour modèle mathématique suggérant un traitement de ces données.

Prenons l'exemple d'un référendum (ou d'un plébicite) où les électeurs ne peuvent que répondre par "oui" ou "non" (les abstentions étant sans influence sur le résultat, ce qui exclut les cas où il y a un quorum à atteindre). Choisissons $n = 1000$, et posons $x_i = 1$ si la i-ème personne interrogée déclare savoir ce qu'elle ira voter et vouloir voter "oui" (si elle déclare ne pas savoir ou ne pas envisager de voter, on écarte cette réponse de la présente analyse) et $x_i = 0$ si elle déclare vouloir voter "non".

Cette situation simple est généralement modélisée par un 1000-échantillon $X_1, \ldots, X_{1000}$ d'une loi de Bernoulli $\mathcal{B}(1, p)$, et on considère que l'opinion est en faveur du "oui" si et seulement si $p \geq 0.5$.

On est alors confronté au problème "d'estimer" la valeur de p. Dans le modèle condisidéré ici (Bernouilli) la loi des grands nombres vient à notre secours : elle assure que $\lim_{n\to+\infty}(X_1 + \ldots + X_n)/n = \mathbb{E}(X_1) = p$; on dit dans ce cas que $\hat{p} := (X_1 + \ldots + X_n)/n$ est un *estimateur* du paramètre p ; en pratique, on choisit alors $p = p^* := (x_1 + \ldots + x_{1000})/1000$.

Nous nous intéresserons ici à la *statistique paramétrique*, où la loi $\mathcal{L} = \mathcal{L}(\theta)$ retenue peut être caractérisé par un paramètre θ, qui est un nombre ou un vecteur. Ainsi, par exemple, si $X_i \rightsquigarrow \mathcal{B}(1, p)$, alors $\theta = p$ est un nombre, mais si $X_i \rightsquigarrow \mathcal{N}(\mu, \sigma)$, alors $\theta = (\mu, \sigma)$ est un vecteur, tout comme dans le cas d'un dé pipé où l'on peut choisir $\theta = (p_1, \ldots, p_5)$ (et $p_6 = 1 - (p_1 + \ldots + p_5)$) et $p_k := \mathbb{P}_\theta(\{X_i = k\})$.

Définition : On dit que $\hat{\theta} : (x_1, \ldots, x_n) \mapsto \hat{\theta}_n := \hat{\theta}(x_1, \ldots, x_n)$ est un *estimateur* convergeant vers θ si et seulement si , en loi, on a $\theta = \lim_{n\to+\infty} \hat{\theta}(X_1, \ldots, X_n)$ pour toute suite de v.a. X_i indépendantes, de loi $\mathcal{L}(\theta)$.

9.2 Vraisemblance

9.2.1 Heuristique et définition

Nous avons vu que la loi des grands nombres fournit "spontanément" un estimateur de l'espérance d'une loi, mais si l'on recherche une méthode un peu générale pour deviner un estimateur, la *méthode du maximum de vraissemblance* est une stratégie souvent efficace. En voici le principe :

Si un échantillonage a produit la suite finie $x_1^*, \ldots, x_n^*$ de nombres et qu'on a choisit de modéliser cette situation par un n-échantillon $X_1, \ldots, X_n$ de v.a. indépendantes de loi $\mathcal{L}(\theta)$, et si le choix de la

valeur du paramètre θ est le problème auquel on est confronté, on peut considérer l'évènement $E^* = \{X_1 = x_1^*, \ldots, X_n = x_n^*\}$, et plus généralement

$$E(x_1, \ldots, x_n) = \{X_1 = x_1, \ldots, X_n = x_n\} = \{X_1 = x_1\} \cap \ldots \cap \{X_n = x_n\}$$

et sa probabilité

$$L(x_1, \ldots, x_n\,;\theta) := \mathbb{P}_\theta(E(x_1, \ldots, x_n)) = \mathbb{P}_\theta(\{X_1 = x_1\} \cap \ldots \cap \{X_n = x_n\}) = \mathbb{P}_\theta(\{X_1 = x_1\}) \cdot \ldots \cdot \mathbb{P}_\theta(\{X_n = x_n\}),$$

cette dernière égalité résultant de l'hypothèse d'indépendance des v.a. X_i. L'idée très heuristique est alors que le choix θ^* qu'il convient d'effecteur pour θ, est celui pour lequel cette probabilité est maximale pour les valeurs x_1^*, …, x_n^* obtenues, et donc de poser

$$\theta^* = \text{Argmax}_{\,\theta}\{L(x_1^*, \ldots, x_n^*\,;\theta)\},$$

c'est-à-dire la valeur (si elle existe et est unique) de θ pour laquelle la fonction $\theta \mapsto L(x_1^*, \ldots, x_n^*\,;\theta)$ est maximale. Souvent, ceci peut se ramener à résoudre en θ l'équation $\frac{\partial L}{\partial \theta}(x_1^*, \ldots, x_n^*\,;\theta) = 0$.

Définition : La fonction $L_n : (x_1, \ldots, x_n\,;\theta) \mapsto \boxed{L_n(x_1, \ldots, x_n\,;\theta) = \prod_{i=1}^n \mathbb{P}_\theta(\{X_i = x_i\})}$ pour des $X_i \rightsquigarrow \mathcal{L}(\theta)$ s'appelle la vraisemblance de la loi $\mathcal{L}$.

La v.a. obtenue en appliquant la fonction $(x_1, \ldots, x_n) \mapsto \text{Argmax}_{\,\theta}\{L(x_1, \ldots, x_n\,;\theta)\}$ appliquée au n-échantillon $(X_1, \ldots, X_n)$ s'appelle l'*estimateur au maximum de vraisemblance* du paramètre θ de la loi discrète $\mathcal{L}(\theta)$.

9.2.2 Exemples

Referendum

Reprenons l'exemple où les X_i suivent une loi de Bernoulli $\mathcal{B}(1, p)$, et donc $\theta = p$. Introduisons la notation $s := x_1 + \ldots + x_{1000}$ pour la somme des valeurs observées sur l'échantillon x_1, …, x_{1000}, c'est-à-dire le nombre de personnes interrogées qui ont déclaré qu'elles voterons "oui". Nous avons donc $L_n(x_1, \ldots, x_n\,;\theta) = \prod_{i=1}^n \mathbb{P}_\theta(\{X_i = x_i\}) = \theta^s(1-\theta)^{n-s}$, pour $\theta = p$, $n = 1000$, et $s = x_1 + \ldots + x_n$, puisque $\theta = p = \mathbb{P}_\theta(\{X_i = 1\})$ et $1 - \theta = 1 - p = \mathbb{P}_\theta(\{X_i = 0\})$.

Les extrémités de l'intervalle $[0, 1]$ auquel appartient θ ne peuvent être des extrema (sauf si $s = 0$ ou $s = n$) et le maximum θ^* de la fonction concave $\theta^s(1-\theta)^{n-s}$ est donc un zéro de la dérivée $\frac{\partial}{\partial \theta} L_n(x_1, \ldots, x_n\,;\theta) = \theta^{s-1}(1-\theta)^{n-s-1}(s - n\theta)$, d'où $\theta^* = \frac{s}{n} = \frac{x_1 + \ldots + x_n}{n}$. En d'autres termes, l'estimateur au maximum de vraisemblance $\hat{p}$ de p est donc $\boxed{\hat{\theta} := \frac{X_1 + \ldots + X_n}{n}}$, c'est à dire le même estimateur que l'estimateur de l'espérance $\mathbb{E}(X_1)$ trouvé en appliquant la loi des grands nombres, ce qui convient, puisque $p = \mathbb{E}(X_i)$.

Variables poissoniennes

Supposons que le tirage d'un n-échantillon $X_1, \ldots, X_n$ de v.a. suivant une loi de Poisson $\mathcal{P}(\lambda)$, $\lambda > 0$ inconnu, ait produit l'échantillon $x_1, \ldots, x_n$. Ici $\theta = \lambda$, et $\mathbb{P}_\theta(\{X_i = x_i\}) = e^{-\theta}\frac{\theta^{x_i}}{x_i\,!}$; la vraisemblance de l'échantillon $x_1, \ldots, x_n$ est donc ici $L_n(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n e^{-\theta}\frac{\theta^{x_i}}{x_i\,!} = e^{-n\theta}\frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i\,!}$, et donc $L_n(x_1, \ldots, x_n; \theta) = e^{-n\theta}\frac{\theta^s}{\prod_{i=1}^n x_i\,!}$, où l'on a une nouvelle fois posé $s := x_1 + \ldots + x_n$. Il est un peu plus commode de calculer avec le logarithme de cette expression est comme ln est une fonction croissante, il nous suffit de rechercher le maximum θ^* de

$$l_n(x_1, \ldots, x_n; \theta) = \ln(L_n(x_1, \ldots, x_n; \theta)) = -n\theta + s\ln(\theta) - \sum_{i=1}^n \ln(x_i\,!).$$

Cette fonction est concave et son extremum θ^* est donc le zéro de la dérivée $\frac{\partial}{\partial \theta} l_n(x_1, \ldots, x_n; \theta) = -n + \frac{s}{\theta}$, c'est à dire $\theta^* = \frac{s}{n}$.

Nous trouvons donc une nouvelle fois $\boxed{\hat{\theta} := \frac{X_1 + \ldots + X_n}{n}}$ comme estimateur de λ, ce qui convient, puisque $\lambda = \mathbb{E}(X_i)$ pour toute v.a. $X_i \rightsquigarrow \mathcal{P}(\lambda)$.

9.3 Cas d'une loi continue

9.3.1 Heuristique et définition

Si la loi $\mathcal{L}(\theta)$ suivie par les X_i est une loi continue, comme $\mathcal{U}_{[a,b]}$ ou $\mathcal{N}(\mu,\sigma)$, on a $\mathbb{P}_\theta(\{X_i = x_i\}) = 0$, et la vraisemblance que nous avons considérée jusqu'ici est tout bonnement (ou plutôt "mauvaisement") nulle, et tous les θ sont des extrema, ce qui ne nous avance guère. L'idée est alors de remplacer $\mathbb{P}_\theta(\{X_i = x_i\})$ par $\mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\})$ pour une $\varepsilon > 0$ suffisamment petit, puis de rechercher θ_ε maximisant $\prod_{i=1}^n \mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\})$. On peut se débarasser du ε qui est arbitraire par la remarque suivante dans le cas où la densité $x \mapsto f_\theta(x)$ caractérisant la loi $\mathcal{L}(\theta)$ est une fonction continue au point x_i : dans ce cas le théorème de la moyenne assure l'existence de fonctions $\varepsilon \mapsto \alpha_{i,\theta}(\varepsilon)$ telles que $\mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\}) = 2\varepsilon(f_\theta(x_i) + \alpha_{i,\theta}(\varepsilon))$, avec $\lim_{\varepsilon\to 0} \alpha_{i,\theta}(\varepsilon) = 0$; le (ou les) θ_ε rendant maximal $\prod_{i=1}^n \mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\})$ sont les même que ceux maximisant

$$\prod_{i=1}^n \frac{1}{2\varepsilon}\mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\}) = \prod_{i=1}^n (f_\theta(x_i) + \alpha_{i,\theta}(\varepsilon)) \ ;$$

en faisant tendre ε vers 0, cette expression devient

$$L_n(x_1,\ldots,x_n;\theta) := \prod_{i=1}^n f_\theta(x_i) \tag{9.1}$$

que nous adoptons comme vraisemblance dans ce cas :

Définition : Si la loi $\mathcal{L}(\theta)$ des X_i est une loi continue de densité f_θ, on appelle vraisemblance de l'échantillon $(x_1,\ldots,x_n)$ pour la loi continue $\mathcal{L}(\theta)$ la fonction $\boxed{L_n(x_1,\ldots,x_n;\theta) := \prod_{i=1}^n f_\theta(x_i)}$.

9.3.2 Exemples

Distribution uniforme

On suppose que l'échantillon x_1, ..., x_n est tiré de manière uniforme entre 0 et a, mais a et b sont inconnus. On modélise donc le problème par une loi uniforme $\mathcal{U}[a,b]$ dont la densité est $f_{(a,b)} := \frac{1}{b-a}\mathbb{I}_{[a,b]}$ et on va chercher un estimateur de $\theta = (a,b)$ par la méthode du maximum de vraisemblance. La vraisemblance de l'échantillon $x_1,\ldots,x_n$ est donc $L_n(x_1,\ldots,x_n;\theta) := \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{1}{b-a}\mathbb{I}_{[a,b]}(x_i) = \frac{1}{(b-a)^n}$ si tous les $x_i \in [a,b]$ et vaut 0 si un des $x_i \notin [a,b]$. On voit donc que $L_n(x_1,\ldots,x_n;\theta)$ est maximal si $\theta = \theta^* = (a^*,b^*) = (\mathrm{Min}\,\{x_1,\ldots,x_n\}, \mathrm{Max}\,\{x_1,\ldots,x_n\})$, puisque ceci nous donne la plus petite valeur de $b-a$ sans annuler la vraisemblance. Ceci nous conduit à considérer l'estimateur

$$\boxed{\hat\theta = (\hat a,\hat b) = (\mathrm{Min}\,\{X_1,\ldots,X_n\}, \mathrm{Max}\,\{X_1,\ldots,X_n\})}.$$

Il reste à montrer que si X_1, ..., X_n est un n-échantillon de loi $\mathcal{U}[a,b]$, alors $\mathrm{Min}\,\{X_1,\ldots,X_n\}$ converge bien, en probabilité, vers a et que $\mathrm{Max}\,\{X_1,\ldots,X_n\}$ converge en probabilité vers b. Considérons par exemple le cas de $\mathrm{Min}\,\{X_1,\ldots,X_n\}$. On a $\{a+\varepsilon < \mathrm{Min}\,\{X_1,\ldots,X_n\}\} = \{a+\varepsilon < X_1,\ldots,a+\varepsilon < X_n\}$, d'où, comme les X_i sont indépendants, $\mathbb{P}(\{a+\varepsilon < \mathrm{Min}\,\{X_1,\ldots,X_n\}\}) = \mathbb{P}(\{a+\varepsilon < X_1\} \cap \ldots \cap \{a+\varepsilon < X_n\}) = \mathbb{P}(\{a+\varepsilon < X_1\}) \cdot \ldots \cdot \mathbb{P}(\{a+\varepsilon < X_n\}) = \left(\frac{b-a-\varepsilon}{b-a}\right)^n$, qui tend bien vers 0 lorsque n tend vers $+\infty$. On montrerais de même que $\mathrm{Max}\,\{X_1,\ldots,X_n\}$ converge en probabilité vers b.

Variables normales

On suppose à présent que l'échantillon x_1, ..., x_n est tiré de manière normale avec une espérance μ et un écart-type σ, mais μ et σ sont inconnus. On modélise donc le problème par une loi normale $\mathcal{N}(\mu,\sigma)$ dont la densité est $f_{(\mu,\sigma)}(x) := \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ et on va chercher un estimateur de $\theta = (\mu,\sigma)$ par la méthode du maximum de vraisemblance. La vraisemblance de l'échantillon x_1, ..., x_n est donc

$$\begin{aligned} L_n(x_1,\ldots,x_n;\theta) : &= \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}}. \end{aligned}$$

Ici, il est une nouvelle fois plus agréable de considérer la log-vraisemblance

$$l_n(x_1,\dots,x_n;\theta) := \ln(L_n(x_1,\dots,x_n;\theta)) = -n(\ln(\sigma)+\ln(\sqrt{2\pi})) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Pour que $\theta^* = (\mu^*,\sigma^*)$ soit un extremum sur $\mathbb{R}\times\mathbb{R}_*^+$ il faut que les deux dérivées $\frac{\partial}{\partial\mu}l_n(x_1,\dots,x_n;\theta) = \frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\mu) = \frac{1}{\sigma^2}(s-n\mu)$ où $s=\sum_{i=1}^n x_i$, et $\frac{\partial}{\partial\sigma}l_n(x_1,\dots,x_n;\theta) = -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^n(x_i-\mu)^2$ s'annulent pour $\theta=\theta^*$, ce qui implique que $\mu = \frac{s}{n} = \frac{1}{n}\sum_{i=1}^n x_i$, et $\sigma^2 = \frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2$. Ceci nous conduit donc à envisager l'estimateur

$$\boxed{\hat\theta = (\hat\mu,\hat\sigma) = \left(\frac{1}{n}\sum_{i=1}^n X_i, \left(\frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2\right)^{\frac{1}{2}}\right)}.$$

En ce qui concerne la première composante $\hat\mu$, nous retrouvons une nouvelle fois la moyenne comme estimateur de l'espérance $\mu=\mathbb{E}(X_i)$, quant-à la seconde composante, nous trouvons

$$\boxed{\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2)}$$

dont nous verrons qu'il s'agit bien, pour toute loi, d'un estimateur de la variance σ^2.