Estimation ponctuelle: Principes de base

Statistique mathématique
M2 santé publique, université Paris-Sud

16 octobre 2018

Estimaton ponctuelle dans les familles paramétriques

Rappelons le point de départ :

- Collection de v.a (un vecteur aléatoire) $X = (X_1, \dots, X_n)$
- $X \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ une famille paramétrique de paramètre $\theta \in \Theta \subseteq \mathbb{R}^d$

Le problème de l'estimation ponctuelle

- Supposons que F est complètement définie par son paramètre θ qui inconnu
- Soit $(x_1, \dots, x_n)$ des réalisations de $X \sim F_\theta$
- Estimer la valeur de θ qui a *généré* les réalisations $(x_1, \dots, x_n)$

Estimaton ponctuelle dans les familles paramétriques

- Approximation de la loi de $g(X_1, \ldots, X_n)$ quand $n \uparrow \infty$.
- L'information apportée par $g(X_1, \ldots, X_n)$.
- Une famille générale de lois (modèles).

Aujourd'hui : Comment estimer θ ? Des procédures générales ?

Estimateur ponctuel

Définition : estimateur

Soit $\left\{F_\theta, \theta \in \Theta\right\}$ un modèle paramétrique où $\Theta \subseteq \mathbb{R}^d$ et soit $\mathbf{X} = \left(X_1, \ldots, X_n\right) \sim F_\theta$ pour un certain $\theta \in \Theta$. Un estimateur ponctuel $\widehat{\theta}$ de θ_0 est une statistique $T : \mathbb{R}^n \mapsto \Theta$, dont le but est d'estimer θ.

Estimateur ponctuel

Définition : estimateur

Soit $\left\{F_\theta, \theta \in \Theta\right\}$ un modèle paramétrique où $\Theta \subseteq \mathbb{R}^d$ et soit $\mathbf{X} = \left(X_1, \ldots, X_n\right) \sim F_\theta$ pour un certain $\theta \in \Theta$. Un estimateur ponctuel $\widehat{\theta}$ de θ_0 est une statistique $T : \mathbb{R}^n \mapsto \Theta$, dont le but est d'estimer θ.

Toute statistique $T : \mathbb{R}^n \mapsto \Theta$ est un estimateur potentiel !
$\hookrightarrow$ Comment mesurer la qualité d'un estimateur ?

- Tout estimateur est une variable aléatoire.
- Un principe général, bon estimateur veut dire :

$$\text{Loi}\left(\widehat{\theta}\right) \text{ se concentre autour de } \theta.$$

$\hookrightarrow$ Une description ∞-dimensionnelle de la qualité !

- A-t-on une mesure plus simple de la qualité ?

Biais et erreur quadratique moyenne

Définition (biais)

Le *biais* d'un estimateur $\widehat{\theta}$ de $\theta \in \Theta$ est défini par

$$\text{biais}\left(\widehat{\theta}\right) = \mathbb{E}_\theta\left[\widehat{\theta}\right] - \theta$$

Cette quantité indique l'écart à la cible quand on utilise $\widehat{\theta}$

Biais et erreur quadratique moyenne

Définition (biais)

Le *biais* d'un estimateur $\widehat{\theta}$ de $\theta \in \Theta$ est défini par

$$\text{biais}\left(\widehat{\theta}\right) = \mathbb{E}_{\theta}\left[\widehat{\theta}\right] - \theta$$

Cette quantité indique l'écart à la cible quand on utilise $\widehat{\theta}$

Définition : estimateur sans biais

Un estimateur $\widehat{\theta}$ de $\theta \in \Theta$ est sans biais si $\mathbb{E}_{\theta}\left[\widehat{\theta}\right] = \theta$, *i.e.* $\text{biais}\left(\widehat{\theta}\right) = 0$.

Nous verrons qu'il ne suffit pas d'avoir un estimateur sans biais ! ! !

Biais et erreur quadratique moyenne

Définition (biais)

Le *biais* d'un estimateur $\widehat{\theta}$ de $\theta \in \Theta$ est défini par

$$\text{biais}\left(\widehat{\theta}\right) = \mathbb{E}_\theta\left[\widehat{\theta}\right] - \theta$$

Cette quantité indique l'écart à la cible quand on utilise $\widehat{\theta}$

Définition : estimateur sans biais

Un estimateur $\widehat{\theta}$ de $\theta \in \Theta$ est sans biais si $\mathbb{E}_\theta\left[\widehat{\theta}\right] = \theta$, *i.e.* $\text{biais}\left(\widehat{\theta}\right) = 0$.

Nous verrons qu'il ne suffit pas d'avoir un estimateur sans biais ! ! !

Définition : erreur quadratique moyenne (MSE)

l'*erreur quadratique moyenne* d'un estimateur $\widehat{\theta}$ de $\theta \in \Theta \subseteq \mathbb{R}$ est définie par

$$MSE\left(\widehat{\theta}\right) = \mathbb{E}_\theta\left[\left(\widehat{\theta} - \theta\right)^2\right]$$

Biais et MSE

On utilise les notions de biais et de variance au lieu de concentration autour du paramètre :

- Le biais donne une indication sur la position de la loi de $\widehat{\theta}$ par rapport à θ (si on considère la moyenne comme un bon estimateur de la position).
- Le MSE est une mesure de dispersion/position de $\widehat{\theta}$ autour de θ.
- Si $\widehat{\theta}$ est sans biais pour $\theta \in \mathbb{R}$ alors $MSE\left(\widehat{\theta}\right) = \mathrm{Var}\left(\widehat{\theta}\right)$.
- Pour $\Theta \subseteq \mathbb{R}^d$, nous avons $MSE\left(\widehat{\theta}\right) = \mathbb{E}_\theta \left\| \widehat{\theta} - \theta \right\|^2$

Exemple

Exemple

Soit $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ et soit $\widehat{\mu} = \bar{X}$. Alors

$$\mathbb{E}(\widehat{\mu}) = \mu \text{ et } MSE(\widehat{\mu}) = \mathrm{Var}(\widehat{\mu}) = \frac{\sigma^2}{n}.$$

Sur cet exemple, le biais et le MSE nous donnent une description complète de la concentration de la loi de $\widehat{\mu}$ autour de μ. $\widehat{\mu}$ est gaussien donc sa loi est complètement définie par sa moyenne et sa variance.

Décomposition biais-variance du MSE

Décomposition biais-variance du MSE

$$MSE\left(\widehat{\theta}\right) = \mathrm{Var}\left(\widehat{\theta}\right) + \mathrm{biais}^2\left(\widehat{\theta}\right)$$

- Une relation fondamentale et simple.
- Un petit MSE n'implique pas forcément une absence de biais.
- Parfois, des estimateurs biaisés sont meilleurs que des estimateurs sans biais.
- Compromis biais/variance (exemple : régression non paramétrique).

Illustration

Convergence

On peut aussi définir la qualité d'un estimateur quand n augmente ! ! !

Convergence (on dit aussi consistance *Consistency*)

Une suite d'estimateur $\left\{\widehat{\theta}_n\right\}_{n\geq 1}$ de $\theta \in \Theta$ est dite convergente si

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$$

- Un estimateur convergent (consistant) se concentre asymptotiquement autour de la vraie valeur de θ quand la taille de l'échantillon n augmente.
- Une étude détaillée de la *qualité asymptotique* de $\widehat{\theta}_n$ nécessite l'étude de la loi de $\widehat{\theta}_n$ quand $n \uparrow \infty$.

Estimateur plug-in (substitution) ou Méthode des moments

On cherche une procédure générale pour construire des estimateurs.

$\hookrightarrow$ idée : $\theta \mapsto F_\theta$ est une bijection sous l'hypothèse d'identifiabilité ? du modèle

- Sous l'hypothèse d'identifiabilité $\nu(F_\theta) = q(\theta)$ pour une certaine fonction q (penser à $q(x) = x$ ou $q(x) = x^2$ ...)

Principe de substitution

Soit $\nu = q(\theta) = \nu(F_\theta)$ le paramètre d'intérêt d'un modèle paramétrique $\{F_\theta\}_{\theta\in\Theta}$. Si on peut construire un estimateur $\widehat{F_\theta}$ de F_θ basé sur l'échantillon $\mathbf{X} = (X_1, \ldots, X_n)$, alors on peut estimer $\nu(F_\theta)$ par $\nu(\widehat{F_\theta})$. Un tel estimateur est appelé estimateur par substitution *plug-in*.

Estimateur plug-in

- En quelque sorte, on a inversé notre point de vue : on considère le paramètre θ comme une fonction de F_θ au lieu de voir F_θ comme fonction de θ ! ! !
- Notons que $\theta = \theta\big(F_\theta\big)$ si q est la fonction identité.
- En pratique, un tel principe est utilisable quand on peut explicitement décrire **la fonctionnelle** $F_\theta \mapsto \nu\big(F_\theta\big)$.

Quelques paramètres comme fonctionnelles de F

Exemples de paramètres fonctionnelles :

- La moyenne :
$$\mu(F) = \int_{-\infty}^{+\infty} x dF(x)$$
- La variance :
$$\sigma^2(F) = \int_{-\infty}^{+\infty} \left[x - \mu(F)\right]^2 dF(x)$$
- La médiane : $\mathrm{med}(F) = \inf\{x : F(x) \geq 1/2\}$
- On peut définir un paramètre $\theta(F)$ indirectement
$$\int_{-\infty}^{+\infty} \psi\Big(x - \theta(F)\Big) dF(x) = 0.$$
- La densité (lorsque elle existe) en x_0 : $\theta(F) = \frac{d}{dx} F(x)|_{x=x_0}$

La fonction de répartition empirique (la loi empirique)

Principe de plug-in

Passer d'un problème d'estimation de θ à un problème d'estimation de F.
Mais comment ?

On considère le cas où $\mathbf{X} = \big(X_1, \ldots, X_n\big)$ iid. On peut définir la version empirique de la fonction de répartition $F_X(\cdot)$ par

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\Big\{X_i \leq x\Big\}.$$

La fonction de répartition empirique (la loi empirique)

Principe de plug-in

Passer d'un problème d'estimation de θ à un problème d'estimation de F.
Mais comment ?

On considère le cas où $\mathbf{X} = \big(X_1, \ldots, X_n\big)$ iid. On peut définir la version empirique de la fonction de répartition $F_X(\cdot)$ par

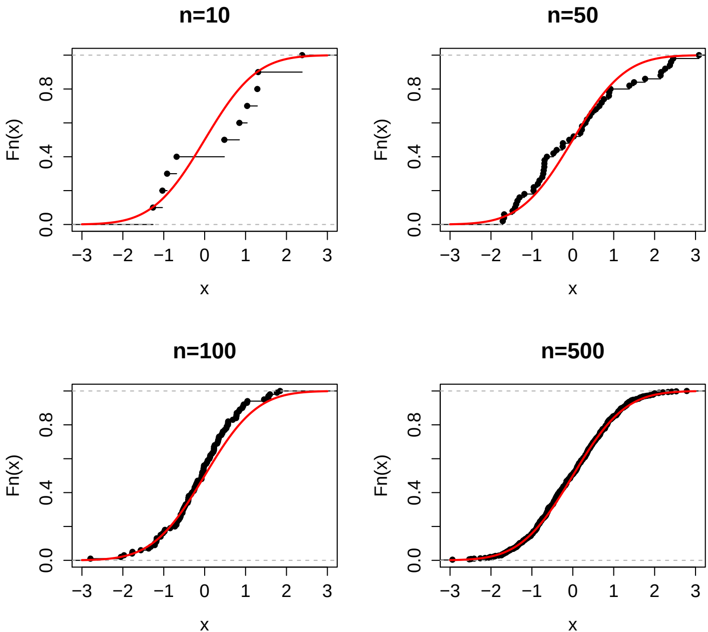
$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\Big\{X_i \leq x\Big\}.$$

- Affecte une masse de $1/n$ pour chaque observation
- Loi forte des grands nombres $\Rightarrow \widehat{F}_n(x) \xrightarrow{p.s} F(x)$ pour tout $x \in \mathbb{R}$

$\hookrightarrow$ car les $\mathbf{1}\Big\{X_i \leq x\Big\}$ sont des variables aléatoires iid Bernoulli$\big(F(x)\big)$

Cela suggère d'utiliser $\nu\big(\widehat{F}_n\big)$ comme estimateur de $\nu\big(F\big)$ ...

Fonction de répartition empirique



Fonction de répartition empirique

On peut faire mieux que la convergence ponctuelle ....

Glivenko-Cantelli

Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes, de loi F. Alors, $\widehat{F}_n(y) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\left\{X_i \leq y\right\}$ convergen uniformément vers F avec probabilité 1, *i.e.*

$$\sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right| \xrightarrow{p.s} 0$$

Exemples

Exemple (moyenne d'une fonction)

On considère $\theta(F) = \int_{-\infty}^{+\infty} h(x) dF(x)$. Un estimateur plug-in basé sur la fonction de répartition empirique est donné par

$$\widehat{\theta} = \theta(\widehat{F}_n) = \int_{-\infty}^{+\infty} h(x) d\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

Exemples

Exemple (moyenne d'une fonction)

On considère $\theta(F) = \int_{-\infty}^{+\infty} h(x)dF(x)$. Un estimateur plug-in basé sur la fonction de répartition empirique est donné par

$$\widehat{\theta} = \theta(\widehat{F}_n) = \int_{-\infty}^{+\infty} h(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

Exemple (variance)

On considère maintenant $\sigma^2(F) = \int_{-\infty}^{+\infty} \left[x - \mu(F)\right]^2 dF(x)$. Nous avons

$$\sigma^2(\widehat{F}_n) = \int_{-\infty}^{+\infty} x^2 d\widehat{F}_n(x) - \left(\int_{-\infty}^{+\infty} x d\widehat{F}_n(x)\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

Méthode des moments

Une méthode ancienne (Karl Pearson, fin 1800's)

Méthode des moments

Soit $X_1, \ldots, X_n$ un échantillon iid issu de $F_\theta, \theta \in \mathbb{R}^p$. L'estimateur $\widehat{\theta}$ par la méthode des moments de θ est la solution en θ des p équations

$$\int_{-\infty}^{+\infty} x^{k_j} \widehat{F}_n(x) = \int_{-\infty}^{+\infty} x^{k_j} F_\theta(x), \quad \{k_j\}_{j=1}^{p} \subset \mathbb{N}$$

- C'est en quelque sorte un estimateur plug-in car on estime les moments théoriques par les moments empiriques et ensuite on estime θ.
- Utile lorsqu'on en a pas de formule explicite de la fonctionnelle $\theta(F)$.

Le problème des moments

Théorème

Supposons que F est une loi déterminer par ses moments. Soit $\{F_n\}_n$ une suite de loi telles que $\int x^k dF_n(x) < \infty$ pour tout n et k. Alors,

$$\lim_{n\to\infty} \int x^k dF_n(x) = \int x^k F(x), \quad \forall k \geq 1 \Rightarrow F_n \xrightarrow{w} F.$$

Le problème des moments

Théorème

Supposons que F est une loi déterminer par ses moments. Soit $\{F_n\}_n$ une suite de loi telles que $\int x^k dF_n(x) < \infty$ pour tout n et k. Alors,

$$\lim_{n\to\infty} \int x^k dF_n(x) = \int x^k F(x), \quad \forall k \geq 1 \Rightarrow F_n \xrightarrow{w} F.$$

Mais toutes les lois ne sont pas déterminées par les moments ! ! ! !

Lemme

La loi de X est déterminée par ses moments s'il existe un voisinage ouvert A contenant 0 tel que

$$M_X(u) = \mathbb{E}\left[e^{-\langle u,X\rangle}\right] < \infty, \quad \forall u \in A.$$

Exemple (loi exponentielle)

Exemple

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$. Alors $\mathbb{E}\left[X_i^r\right] = \lambda^{-r}\Gamma(r+1)$.

Exemple (loi exponentielle)

Exemple

Soit $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$. Alors $\mathbb{E}\left[X_i^r\right] = \lambda^{-r}\Gamma(r+1)$.
On définit une classe d'estimateurs de λ qui dépendent de r donnés par

$$\widehat{\lambda} = \left[\frac{1}{n\Gamma(r+1)} \sum_{i=1}^{n} X_i^r\right]^{-\frac{1}{r}}.$$

Nous verrons qu'on peut choisir r pour avoir le "meilleur estimateur" ....

Exemple (la loi gamma)

La loi gamma

Soit $X_1, \ldots, X_n$ un échantillon issu d'une loi gamma de paramètres α et β,

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases} \quad \text{où } \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

On sait que $\alpha > 0$, $\beta > 0$, $\mathbb{E}(X) = \alpha\beta$ et $\mathrm{Var}(X) = \alpha\beta^2$.

Exemple (la loi gamma)

La loi gamma

Soit $X_1, \ldots, X_n$ un échantillon issu d'une loi gamma de paramètres α et β,

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases} \text{ où } \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

On sait que $\alpha > 0$, $\beta > 0$, $\mathbb{E}(X) = \alpha\beta$ et $\mathrm{Var}(X) = \alpha\beta^2$.

$$\widehat{\alpha} = \frac{\bar{X}^2}{\widehat{\sigma}^2} \text{ et } \widehat{\lambda} = \frac{\bar{X}}{\widehat{\sigma}^2}.$$

Exemple : loi uniforme discrète

Exemple

Soit $X_1, \ldots, X_n$ iid de loi $\mathcal{U}\{1, \ldots, \theta\}$ où $\theta \in \mathbb{N}$. Calculer $\mathbb{E}(X)$ et proposer un estimateur $\widehat{\theta}$.

Exemple : loi uniforme discrète

Exemple

Soit $X_1, \ldots, X_n$ iid de loi $\mathcal{U}\{1, \ldots, \theta\}$ où $\theta \in \mathbb{N}$. Calculer $\mathbb{E}(X)$ et proposer un estimateur $\widehat{\theta}$.

Avantage de la méthode des moments : elle peut être généralisée à des données non-iid.

La fonction de vraisemblance

Définition

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ des variables aléatoires de densité jointe $f(\mathbf{x}; \theta)$ (ou fonction de masse) où $\theta \in \Theta \subseteq \mathbb{R}^p$. La fonction de vraisemblance $L(\theta)$ est la fonction aléatoire

$$L(\theta) = f(\mathbf{X}; \theta).$$

Notons qu'on considère L comme une fonction de θ et pas de $\mathbf{X}$.
Interprétation dans le cas discret ?

La fonction de vraisemblance

Définition

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ des variables aléatoires de densité jointe $f(\mathbf{x}; \theta)$ (ou fonction de masse) où $\theta \in \Theta \subseteq \mathbb{R}^p$. La fonction de vraisemblance $L(\theta)$ est la fonction aléatoire

$$L(\theta) = f(\mathbf{X}; \theta).$$

Notons qu'on considère L comme une fonction de θ et pas de $\mathbf{X}$.
Interprétation dans le cas discret ?
Lorsque $\mathbf{X}$ est iid de densité $f(\cdot; \theta)$, alors la vraisemblance est

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

Estimateur du maximum de vraisemblance

Définition (estimateur du maximum de vraisemblance)

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon aléatoire issu de F_θ, et soit $\widehat{\theta}$ tel que

$$L(\widehat{\theta}) \geq L(\theta), \quad \forall \theta \in \Theta.$$

Alors $\widehat{\theta}$ est appelé *un estimateur du maximum de vraisemblance* de θ.

Estimateur du maximum de vraisemblance

Définition (estimateur du maximum de vraisemblance)

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ un échantillon aléatoire issu de F_θ, et soit $\widehat{\theta}$ tel que

$$L(\widehat{\theta}) \geq L(\theta), \quad \forall \theta \in \Theta.$$

Alors $\widehat{\theta}$ est appelé *un estimateur du maximum de vraisemblance* de θ.

On appelle $\widehat{\theta}$ l' estimateur du maximum de vraisemblance, s'il est l'unique maximum de $L(\theta)$,

$$\widehat{\theta} = \underset{\theta \in \Theta}{\text{argmax}} L(\theta).$$