Estimation

ECE 2 Lycée international de Valbonne

Table des matières

I Introduction et vocabulaire

I.1 Mise en place du problème

Exemple : On possède une pièce truquée, dont la probabilité d'obtenir pile, notée p, est inconnue. On sait que la loi associée à cette pièce est par définition une loi de Bernoulli, mais on en ignore le paramètre. On cherche à **estimer** ce paramètre. Pour ce faire on peut lancer un très grand nombre de fois la pièce, on compte 1 à chaque fois que l'on obtient pile et on fait la moyenne des résultats. On "sait" que l'on va obtenir une valeur approchée de p.

De façon plus générale On considère un phénomène aléatoire et on s'intéresse à une variable aléatoire réelle X qui lui est liée, dont on suppose que la loi de probabilité n'est pas complètement spécifiée et appartient à une famille de lois $\mathcal{P}_\theta$ dépendant d'un paramètre θ décrivant un sous-ensemble Θ de $\mathbb{R}$ (éventuellement de $\mathbb{R}^2$). Le paramètre θ est une quantité inconnue, fixée dans toute l'étude, que l'on cherche à déterminer ou pour laquelle on cherche une information partielle. Le problème de l'estimation consiste alors à estimer la vraie valeur du paramètre θ ou de $g(\theta)$ (fonction à valeurs réelles du paramètre θ), à partir d'un échantillon de données $x_1, ..., x_n$ obtenues en observant n fois le phénomène. Cette fonction du paramètre représentera en général une valeur caractéristique de la loi inconnue comme son espérance, sa variance, son étendue…

I.2 Vocabulaire

Dans toute la suite n est un entier naturel strictement plus grand que 1 et toutes les variables aléatoires sont définies sur un même espaces probabilisable $(\Omega, \mathcal{A})$ et munit d'une famille de loi de probabilité $(\mathbb{P}_\theta)_{\theta\in\Theta}$.

Dans notre exemple

$$\Theta =]0;\ 1[$$

car p est la probabilité d'obtenir "Pile", on cherche a estimer p est la famille de probabilité est la famille de loi de Bernoulli de paramètre p.

Définition 1 (Échantillon et réalisation)**.**
Soit X une variable aléatoire de loi $\mathcal{P}_\theta$.

*Un n **échantillon** de la loi X est un n-uplet $(X_1, X_2, \ldots, X_n)$ de variables mutuellement indépendantes [1] et suivant la même loi que X.*

*Une **observation** (ou échantillon observé) est un n-uplet $(x_1, x_2, \ldots, x_n)$ de réels tel que x_1, x_2,…, x_n sont les valeurs respectivement prises par X_1, X_2, …X_n.*

Exemple : Revenons à notre pièce truquée. Ici la loi que l'on cherche à déterminer est donnée par le paramètre $p \in]0;\ 1[$ et $X \hookrightarrow \mathcal{B}(p)$. Un échantillon est par exemple (X_1, X_2, X_3) où X_1, X_2 X_3 sont des variables mutuellement indépendantes et suivant la même loi que X. (on décide de répéter trois fois le lancer) et une observation est $(1, 1, 0)$ (qui

1. Plus précisement indépendantes pour toutes les lois $\mathbb{P}_\theta$

correspond à Pile, Pile , Face).
Exemple : On **suppose** que la note des étudiants à une épreuve suit une loi normale de paramètres $10, 2$.

Un échantillon de taille n est $(X_1, X_2, \dots, X_n)$ où pour $i \in [\![1,\ n]\!]$, $X_i \hookrightarrow \mathcal{N}(10, 2)$ et où les variables aléatoires sont mutuellement indépendantes. Une observation est par exemple $(10.1, 12.3, 14.5, \dots)$

II Estimation ponctuelle

II.1 Définitions et exemples

Définition 2 (Estimateur et estimation)**.**
*Soit $(X_1, X_2, \dots, X_n)$ un échantillon de la loi de X un **estimateur** de $g(\theta)$ est une variable aléatoire de la forme $T_n = \varphi(X_1, X_2, \dots, X_n)$.*
où φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ indépendante de θ mais qui peut dépendre de n.
*Une **estimation** de $g(\theta)$ est une valeur $\varphi(x_1, x_2, \dots, x_n)$ où $x_1, x_2, \dots, x_n$ est une réalisation.*

Notation [2] Dans le cas où ils existent on note $E_\theta(T_n)$ et $V_\theta(T_n)$ l'espérance et la variance de T_n pour la loi $\mathbb{P}_\theta$

II.1.a Moyenne empirique

Définition 3 (Moyenne empirique)**.**
Soit $(X_1, X_2, \dots, X_n)$ un échantillon de la loi X. On note

$$\overline{X_n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \sum_{k=1}^{n}$$

Exercice 1.
Quelle est la fonction φ utilisée içi ?

Réponse

On reprend l'exemple de la pièce truquée. On peut choisir comme estimateur de p la moyenne empirique.
Exemple : On veut déterminer le paramètre p d'une loi géométrique $\mathcal{G}(p)$ avec $p \in\ ]0;\ 1[$. Cette fois ci la moyenne empirique est un estimateur de $1/p$. La fonction g apparaissant dans la question précédente est g :

II.2 Biais

Dans la définition qui précède on ne distingue pas les "bons" estimateurs des "mauvais" estimateurs.
Exemple : On suppose que les notes à une épreuve suivent une loi normale $\mathcal{N}(m, \sigma^2)$ on peut estimer le paramètre m avec l'estimateur $T_n = \max(X_1, X_2, \dots, X_n)$ c'est à dire en posant

2. On pratique on n'écrit pas l'indice θ.

$$\varphi\left(x_1, x_2, \ldots, x_n\right) \mapsto \qquad\qquad et \quad g :$$

C'est un estimateur mais on se doute que c'est n'est pas un bon estimateur pour le paramètre recherché puisque cela revient à considérer que la moyenne est égale à la meilleure des notes.

Définition 4 (Biais)**.**
Si pour tout paramètre θ de Θ, T_n admet une espérance, on appelle ***biais*** *de T_n le réel*

$$b_\theta(T_n) = E_\theta(T_n) - g(\theta)$$

Définition 5 (Estimateur sans biais)**.**
L'estimateur T_n de $g(\theta)$ est ***sans biais*** *si et seulement si $E_\theta(T_n) =$ pour tout paramètre θ de Θ.*

Exemple important : la moyenne empirique Soit un échantillon $(X_1, X_2, \ldots X_n)$, on cherche à évaluer l'espérance μ de la loi en posant $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ que l'on nomme moyenne empirique

Remarque : La fonction g n'est pas explicitée elle dépend de la loi . Par exemple si les variables aléatoires suivent des loi de Bernouilli , alors $g(p) = p$, si elle suivent des loi géométriques alors $g(p) = \frac{1}{p}$.

On sait que

$$E(\overline{X_n}) = \frac{1}{n}\sum_{k=1}^{n} E(X_i) = \frac{n}{n}\mu = \mu$$

donc cette estimateur de l'espérance est sans biais, en moyenne on obtient ce que l'on chercher à estimer

II.3 Risque quadratique

Définition 6 (risque quadratique)**.**
Si pour tout paramètre θ de Θ, T_n admet un moment d'ordre 2, on appelle ***risque quadratique*** *de T_n le réel*

$$r_\theta(T_n) = E_\theta\left[(T_n - g(\theta))^2\right]$$

Proposition 1 (Décomposition biais-variance du risque quadratique d'un estimateur.)**.**

Si pour tout paramètre θ de Θ, T_n admet un moment d'ordre 2,

$$r_\theta(T_n) = b_\theta(T_n)^2 + V_\theta(T_n)$$

Démonstration :

Exemple : Cas de la moyenne empirique

II.4 Estimateurs asymptotiquement sans biais

Il arrive que l'estimateur que l'on étudie à un biais mais que celui devient très petit quand on prend un échantillon suffisamment grand

Définition 7 (Suite d'estimateurs asymptotiquement sans biais)**.**
Une suite $(T_n)_{n>1}$ d'estimateurs de $g(\theta)$ est ***asymptotiquement sans biais*** *si et seulement si pour tout θ de Θ,*

$$\lim_{n\to+\infty} E_\theta(T_n) =$$

Par abus de langage on dit aussi que ***l'estimateur est asymptotiquement sans biais****.*

Premier estimateur de la variance empirique On note m l'espérance de X et σ^2 sa variance.

$$S_n = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$$

où $\overline{X}_n$ est la moyenne empirique

On va étudier le biais de cet estimateur de la variance.

1. Montrons que $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \overline{X}_n^2$

2. Montrons que $E(S_n) = \frac{n-1}{n}\sigma^2$.

3. En déduire le biais de cette estimateur de σ^2. Est il sans biais ? asymptotiquement sans biais ?

II.5 Estimateurs convergents

Définition 8 (Convergence d'une suite d'estimateur)**.**
*Estimateur convergent. Une suite d'estimateurs $(T_n)_{n>1}$ de $g(\theta)$ est **convergente** si pour tout paramètre θ de Θ,*

$$\forall \varepsilon > 0 \qquad \lim_{n \to +\infty} \mathbb{P}_\theta([|T_n - g(\theta)| > \varepsilon]) = 0$$

*Par abus de langage on dit aussi que **l'estimateur** T_n **est convergent**.*

Remarque : On rapprochera cette définition de la loi faible des grands nombres vue au chapitre précédent. Cette définition est rarement demandée, les exercices se contentant d'utiliser la proposition suivante.

Théorème 1 (Condition suffisante de convergence d'un estimateur.)**.**
Si pour tout paramètre θ de Θ, la limite $\lim_{n \to +\infty} r_\theta(T_n) = 0$, alors la suite d'estimateurs $(T_n)_{n>1}$ de $g(\theta)$ est convergente.

Démonstration :

Corollaire 1 (Cas d'un estimateur sans biais ou asymptotiquement sans biais)**.**
Si T_n est un estimateur sans biais (ou asymptotiquement sans biais) et si pour tout paramètre θ de Θ

$$\lim_{n \to +\infty} V_\theta(T_n) = 0$$

alors T_n est convergent.

Démonstration :

Exemple : Estimation du paramètre d'une loi uniforme par le maximum. On suppose que $X \hookrightarrow \mathcal{U}([0;\, \theta])$ où θ est un réel positif que l'on cherche à déterminer avec l'estimateur :

$$T_n = \max(X_1, X_2, \ldots, X_n)$$

1. Si $n = 3$ et que l'échantillon est $(1.2, 3, 0.4)$ quelle est l'estimation de θ ?

2. Calculer une densité de T_n.

3. Calculer l'espérance de T_n. Est ce un estimateur sans biais de θ ?

4. Calculer la variance de T_n et son risque quadratique.

5. Que peut on dire de cet estimateur de θ ?

III Estimation par intervalle de confiance

III.1 Définitions

Dans tout ce paragraphe $(U_n)_{n>1}$ et $(V_n)_{n>1}$ désigneront des suites d'estimateurs de $g(\theta)$ tels que pour tout $\theta \in \Theta$ et pour tout $n >1$, $\mathbb{P}_\theta([U_n \leqslant V_n]) = 1$.

Définition 9 (Intervalle de confiance, niveau de confiance.)**.**
*On dit que $[U_n, V_n]$ est un **intervalle de confiance** de $g(\theta)$ au **niveau de confiance** $1-\alpha$ (où $\alpha \in [0;\, 1]$) si pour tout $\theta \in \Theta$,*

$$\mathbb{P}_\theta(U_n \leqslant g(\theta) \leqslant V_n) > 1-\alpha$$

Exemple : On veut estimer la masse m d'un certain objet. Pour cela, on effectue des pesées successives et l'on note m la moyenne obtenue. On admet que la variable aléatoire renvoyant le résultat d'une pesée de l'objet étudié suit une loi de variance σ^2 avec $\sigma = 0.1$. On effectue une suite de pesées et on note X_i le résultat le la i-ième pesée.

1. . On note $\overline{X_n} = \dfrac{1}{n}\sum_{k=1}^{n} X_k$. Quelle est la variance et l'espérance de $\overline{X_n}$?

2. Montrer que pour tout e > 0 on a $\mathbb{P}(|\overline{X_n} - m| < \varepsilon) \leqslant 1 - \dfrac{\sigma^2}{n\varepsilon^2}$

3. En déduire un intervalle de confiance au niveau de confiance 0.9.

Définition 10 (Intervalle de confiance asymptotique.)**.**
*On appelle **intervalle de confiance asymptotique** de $g(\theta)$**au niveau de confiance** $1-\alpha$ une suite $([U_n, V_n])_{n>1}$ vérifiant pour tout θ de Θ : il existe une suite de réels (α_n) à valeurs dans $[0;\, 1]$, de limite α, telle que pour tout $n > 1$,*

$$\mathbb{P}_\theta([U_n \leqslant g(\theta) \leqslant V_n]) > 1-\alpha_n$$

Par abus de langage on dit aussi que $[U_n, V_n]$ est un intervalle de confiance asymptotique.