Estimations

Introduction

1. **Contexte :** On considère un phénomène aléatoire et on s'intéresse à une variable aléatoire X réelle qui lui est liée.
   La loi de probabilité associée à X appartient à une famille de lois dépendant d'un paramètre θ décrivant un ensemble $\Theta \subset \mathbb{R}$ (ou $\mathbb{R}^2$). *Exemple :* X suit une loi exponentielle de paramètre $\lambda \in [0.2\ ;\ 1]$, associée à l'étude de machines dont la durée de vie espérée est une valeur fixée entre 1 et 5 ans. On cherche estimer cette espérance de vie.

2. **Objectif :** Déterminer la valeur du paramètre θ ou en donner une information partielle.
   Le problème de l'estimation consiste donc à estimer θ ou une fonction $g(\theta)$. Cette dernière pourra représenter, en général, une valeur caractéristique de la loi inconnue -comme son espérance, sa variance, son étendue ...
   Exemple : Dans le problème précédant, on cherche à estimer cette espérance de vie, valant $g(\lambda) = \frac{1}{\lambda}$

   Remarque : Le programme officiel au concours D2 n'évoque que l'estimation d'une moyenne, d'une proportion (que l'on peut voir comme une fréquence) ainsi que d'une variance.
3. **Moyens :** On se donnera un échantillon de données $x_1, \ldots, x_n$ obtenues en observant n fois le phénomène. On supposera que cet échantillon est la réalisation d'une famille de variables aléatoires $X_1, \ldots, X_n$ définies sur un même espace probabilisable $(\Omega\ ;\ \mathcal{A})$ que l'on pourra munir de n'importe quelle probabilité $\mathbb{P}_\theta$ où $\theta \in \Theta$.
   Ces variables aléatoires seront de même loi et mutuellement indépendantes au sein de $\mathbb{P}_\theta$, et ce, pour n'importe quel $\theta \in \Theta$, ce que l'on pourra résumer en disant :

   Les variables aléatoires $X_1,\ \ldots,\ X_n$ sont indépendantes et identiquement distribuées.

   Exemple : Dans le problème précédant, on a observé sur trois machines du même type des durées de fonctionnement respectives de $x_1 = 2$, $x_2 = 2,5$ et $x_3 = 2,25$ années (format décimal).
   On vient de donner une réalisation de X_1, X_2 et X_3 indépendantes et identiquement distribuées de loi exponentielle de paramètre $\lambda \in [0.2\ ;\ 1]$

Estimateur ponctuel

Vocabulaire : Si X est une variable aléatoire donnée, dont la loi $\mathcal{L}$ dépend d'un -voire deux- paramètre(s), on nomme n-échantillon de X, où $n \geq 2$ est un entier naturel, une famille $(X_1\ ;\ \ldots\ ;\ X_n)$ de variables aléatoires indépendantes et identiquement distribuées selon la loi suivie par X.

On pourra éventuellement dire que $(X_1\ ;\ \ldots\ ;\ X_n)$ est un n-échantillon de la loi $\mathcal{L}$, sans se référer à une variable X.

Définition : On appellera *estimateur* de $g(\theta)$ toute variable aléatoire réelle T_n de la forme $T_n = \varphi(X_1; \ldots; X_n)$ où φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ (éventuellement dépendante de n) et indépendante de θ dont la réalisation après expérience est envisagée comme estimation de $g(\theta)$.

Remarques :

1. Si T_n est un estimateur, il peut admettre une espérance, une variance , des moments d'ordre r en tant que variable aléatoire. En revanche, ces valeurs, si elles existent, dépendent de la probabilité $\mathbb{P}_\theta$.
2. Une *estimation* de $g(\theta)$ est donc une réalisation $(x_1 \ldots x_n)$ de $(X_1 \ldots X_n)$.

Notation : Si T_n est un estimateur, on écrira, sous couvert d'existence, $\mathbb{E}_\theta[T_n]$ son espérance et $\mathbb{V}_\theta[T_n]$ sa variance.

Exemple **1** *Moyenne Empirique :*

Si $(X_1\ ;\ \ldots\ ;\ X_n)$ est un n-échantillon de X, on appelle *moyenne empirique* notée $\overline{X}_n$ l'estimateur défini par :

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

Exemple **2** *Variance Empirique :*

Si $(X_1\ ;\ \ldots\ ;\ X_n)$ est un n-échantillon de X, on appelle *variance empirique* notée $\overline{X}_n$ l'estimateur défini par :

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

Exercice 1 (**RàR**) *Paramètres de la moyenne empirique*

On considère X une variable aléatoire suivant une loi $\mathcal{L}$ admettant une espérance μ et une variance σ^2. Pour $n \geq 2$, on se donne un n-échantillon $(X_1\ ;\ \ldots\ ;\ X_n)$ de X.

1. On cherche à estimer $\theta = \mu$. Etablir que la moyenne empirique $\overline{X}_n$ admet une espérance et vérifier que $\mathbb{E}_\theta[\overline{X}_n] = \mu$.
2. On cherche à estimer $\theta = \sigma^2$. Etablir que la moyenne empirique $\overline{X}_n$ admet un moment quadratique et vérifier que $\mathbb{V}_\theta[\overline{X}_n] = \frac{\sigma^2}{n}$.

Exercice 2 *Paramètres de la variance empirique*

On considère X une variable aléatoire suivant une loi $\mathcal{L}$ admettant une espérance μ et une variance σ^2. Pour $n \geq 2$, on se donne un n-échantillon $(X_1\ ;\ \ldots\ ;\ X_n)$ de X.

1. On cherche à estimer $\theta = \sigma^2$.
   Etablir que la variance empirique $S_n{}^2$ admet une espérance et vérifier que $\mathbb{E}_\theta[S_n^2] = \dfrac{n-1}{n}\sigma^2$.
2. Expliquer pourquoi les logiciels de calcul numérique proposent une "variance corrigée" s'écrivant $\sigma^2{}_{n-1} = \frac{n}{n-1}S_n{}^2$

Définition : Soit T_n un estimateur de X. Si pour tout $\theta \in \Theta$, l'estimateur T_n admet une espérance $\mathbb{E}_\theta[T_n]$ alors on nommera *biais* le réel défini par :

$$b_\theta(T_n) = \mathbb{E}_\theta[T_n] - g(\theta)$$

Vocabulaire : Si le biais d'un estimateur T_n est nul ($b_\theta(T_n) = 0$) pour tout $\theta \in \Theta$, on dira que cet estimateur est *sans biais*

Exemple **3** *Moyenne Empirique :*

La moyenne empirique $\overline{X}_n$ d'une variable X est un estimateur sans biais de l'espérance μ de la loi de X.

Exemple **4** *Variance Empirique :*

La variance empirique $S^2{}_n$ d'une variable X est un estimateur qui n'est pas sans biais de la variance σ^2 de la loi de X.

En revance, l'estimateur *variance corrigé* défini comme $\sigma_{n-1} = \dfrac{n}{n-1}S_n{}^2$ est un estimateur sans biais de cette même variance.

On pourra dire *biaisé* au lieu de "n'est pas sans biais".

Exercice 1 *Biais de la variance empirique*

Calculer le biais de la variance empirique $S_n{}^2$, estimateur d'une variable X de loi $\mathcal{L}$ dont la variance est le paramètre σ^2.

Définition : Une suite d'estimateurs $(T_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires telle que chaque T_n est un estimateur associé à un n-échantillon d'une même variable aléatoire X, de la forme $T_n = \varphi(X_1; \ldots; X_n)$.

Exemple **5** Les estimateurs *moyenne empirique* et *variance empirique* d'une variable aléatoire X induisent des suites d'estimateurs $(\overline{X}_n)_{n\geq 1}$ et $(S_n{}^2)_{n\geq 1}$ respectivement, quitte à pose $\overline{X}_1 = X_1$ et $S_1{}^2 = 0$

Définition : Une suite d'estimateurs $(T_n)_{n\in\mathbb{N}}$ de $g(\theta)$ est dite convergente si, pour tout θ de Θ, la suite de variables aléatoire $(T_n)_{n\in\mathbb{N}}$ converge en probabilités vers $g(\theta)$.
On pourra (par abus de langage) dire tout simplement que *l'estimateur est convergent*.

Ainsi, on pourra retenir que cela signifie formellement :

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

Exercice 1 RàR Démontrer que l'estimateur $\overline{X}_n$ de l'espérance μ d'une variable aléatoire X est convergent.
Indication : Utiliser la loi (faible) des grands nombres.

Vocabulaire : On dira qu'un estimateur est *asymptotiquement sans biais* lorsque son biais est de limite nulle pour tout $\theta \in \Theta$ soit encore écrit :

$$\forall \theta \in \Theta \quad \lim_{n \to +\infty} E_\theta(T_n) = g(\theta)$$

Exercice 1 S'assurer de l'équivalence du vocabulaire adopté avec le formalisme proposé.

Exercice 2 Déterminer si l'estimateur biaisé ${S_n}^2$ de la variance σ^2 d'une variable X est asymptotiquement sans biais.

Définition : Si, pour tout $\theta \in \Theta$, l'estimateur T_n admet un moment d'ordre 2, alors on définit son *risque quadratique* par :

$$r_\theta(T_n) = \mathbb{E}_\theta[(T_n - g(\theta))^2]$$

Exercice 1 RàR *Décomposition biais-variance du risque quadratique*

1. Démontrer que, si T_n est un estimateur de $g(\theta)$ et admet un moment quadratique d'ordre 2, établir que l'on a :

$$\forall \theta \in \Theta \; r_\theta(T_n) = b_\theta(T_n)^2 + \mathbb{V}_\theta(T_n)$$

2. Démonter que, dans le cas d'un estimateur sans biais, le risque quadratique est égal à la variance de cet estimateur.

Exercice 2 RàR *Risque quadratique de moyennes empiriques*
Soit X suivant une certaine loi $\mathcal{B}$ dont $p = \theta$ en est un paramètre. On considère la moyenne empirique $\overline{X}_n$ associée à un n-échantillon $(X_1, \ldots, X_n)$.

Démontrer que le risque quadratique de l'estimateur $\overline{X}_n$ existe et vaut $\dfrac{p(1-p)}{n}$

Propriété : Si $(T_n)_{n\in\mathbb{N}}$ est une suite d'estimateurs de $g(\theta)$ convergente et $f : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue, alors $(f(T_n))_{n\in\mathbb{N}}$ est une suite d'estimateurs convergente vers $f(g(\theta))$.

Démonstration : Utiliser $\lim_{n\to+\infty} f(X_n) = f(\lim_{n\to+\infty} X_n)$ fourni par la continuité de f.

Propriété : [*condition suffisante de convergence*]
Si, pour tout $\theta \in \Theta$ on a $\lim_{n\to+\infty} r_\theta(T_n) = 0$ alors la suite d'estimateurs $(T_n)_{n\in\mathbb{N}}$ de $g(\theta)$ est convergente.

Remarque : Ceci suppose que T_n admet bien un moment d'ordre 2 (pour tous θ et n à considérer).

Démonstration :

Exercice 1 1. Justifier que, si $\lim_{n\to+\infty} r_\theta(T_n) = 0$ alors $\mathbb{V}_\theta(T_n)$ et $b_\theta(T_n)$ convergent vers 0 (en tant que suites indexées par l'entier n).
2. En utilisant l'inégalité de Markov, établir que :

$$\mathbb{P}[|T_n - g(\theta)| \geq \varepsilon] \leq \frac{1}{\varepsilon^2}\mathbb{E}_\theta[(T_n - g(\theta))^2]$$

3. Conclure

***Exemple* 6** La moyenne empirique $\overline{X}_n$ d'un n-échantillon de loi Bernoulli de paramètre p que l'on souhaite estimer dans $\Theta =]0;1[$ est sans biais, convergent comme son risque quadratique est $r_p(\overline{X}_n) = \dfrac{p(1-p)}{n}$ (de limite nulle lorsque n tend vers $+\infty$).

Compléments : résultats utilisés

Dans cette section, nous proposons de détailler les propriétés souvent fournies et démontrées qui permettent de mieux considérer les démarches de ce chapitre :

Inégalité de Markov

Soit X une variable aléatoire réelle positive admettant une espérance (bien que...). On a, pour tout $a > 0$:

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

Démonstration : Faite en exercice dans le TD 18, exercice 14 qui plus est corrigé

Inégalité de Bienaymé-Tchebychev

Soit X une variable aléatoire réelle admettant des moments d'ordres 1 et 2. On a, pour tout $\varepsilon > 0$:

$$\mathbb{P}[|X - \mathbb{E}[X]| > \varepsilon] \leq \frac{\mathbb{V}[X]}{\varepsilon^2}$$

Démonstration : Conséquence de l'inégalité de Markov en prenant astucieusement la lettre Z au lieu de X et en posant $Z = |X - \mathbb{E}[X]|^2$ et $a = \varepsilon^2$. Voir le TD 18, exercice 14 (qui plus est corrigé encore une fois)

Convergence en probabilités

On dit que la suite de VAR (X_n) converge en probabilités lorsqu'il existe une VAR notée X définie sur le même espace que les X_n vérifiant :

$$\forall \varepsilon > 0 \quad \lim_{n \to +\infty} \mathbb{P}[|X_n - X| \geq \varepsilon] = 0$$

On écrira alors $X_n \longrightarrow^{\mathbb{P}} X$
Remarque : L'inégalité de Bienaymé-Tchebytchev est souvent liée à l'obtention d'un tel résultat, en particulier lorsqu'il est apparent que $\lim_{n \to +\infty} \mathbb{V}[X_n] = 0$ (faire le lien avec un risque quadratique asymptotiquement nul pour des estimateurs asymptotiquement sans biais)

Convergence en loi

On dit que la suite de VAR (X_n) converge en loi vers X définie sur le même espace que les X_n lorsque :

$$\forall a \in \mathbb{R} \quad \lim_{n \to +\infty} F_{X_n}(a) = F_X(a)$$

avec F_Z notation désignant la fonction de répartition de la variable aléatoire Z (en toute généralité)
Remarque : La fonction de répartition d'une VAR caractérise sa loi. On demande donc la convergence *point par point* de la suite de fonctions de répartitions (F_{X_n}) vers la fonction F afin de caractériser une (nouvelle) loi : celle de X (la limite)