

Estimation ponctuelle

I - Introduction

- Les statisticiens connaissent, en général, le type de loi qui décrit un phénomène, mais souvent ils ne connaissent pas tous les paramètres de la dite loi. Ils doivent donc les estimer : c'est l'objectif de ce qu'on appelle la statistique inférentielle.
- On considère une expérience aléatoire E et une variable aléatoire réelle X qui lui est liée. En général, la loi de X n'est pas complètement spécifiée. Le plus souvent, on suppose que cette loi appartient à une famille de loi dépendant d'un paramètre inconnu θ qui appartient à un ensemble Θ de $\mathbb{R}$. Par exemple X suit une loi de poisson de paramètre $\lambda \in \mathbb{R}^+$ inconnu, X suit une loi uniforme sur $[0; \theta]$ de borne θ inconnue, X suit une loi de Bernoulli de paramètre p inconnu,
- Le problème de l'estimation consiste à estimer la vraie valeur du paramètre inconnu à partir d'un échantillon de n réalisations de X , que l'on notera $(x_1, \dots, x_n)$, issues de n réalisations de l'expérience aléatoire E .

Exemple 1

On souhaite estimer la proportion de la population française qui aime les maths.

- On interroge une personne et on note X la variable qui vaut 1 si elle aime les maths et 0 sinon. Alors, X suit une loi de Bernoulli de paramètre p où p est la probabilité d'aimer les maths et on cherche à estimer la valeur de p .
- On a donc ici $\theta = p$ à chercher dans le sous-ensemble $\Theta =]0; 1[$ de $\mathbb{R}$.
- On réalise un sondage et sur 100 personnes interrogées, 27 répondent qu'ils aiment les maths. Ce sondage constitue une estimation du paramètre $\theta = p$ de la variable X . Nous allons voir deux types d'estimations :
 - ★ On pourra chercher une valeur approchée de θ : on parle d'estimation ponctuelle.
 - ★ On pourra chercher un intervalle dans lequel θ a une forte probabilité de se trouver : on parle d'estimation par intervalle de confiance.

II - Estimation ponctuelle

II.1 - n -échantillon

Définition 1

Soit X une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$.

- On appelle n -échantillon de la variable X tout n -uplet $(X_1, \dots, X_n)$ de variables aléatoires qui sont indépendantes et de même loi que X .
On parle parfois d'échantillon i.i.d pour signifier des v.a indépendantes et identiquement distribuées.
- On appelle réalisation du n -échantillon $(x_1, \dots, x_n)$ tout n -uplet $(x_1, x_2, \dots, x_n)$ de $\mathbb{R}^n$ où x_k est la valeur prise par X_k pour $k \in \llbracket 1; n \rrbracket$.

Remarque 1

Bien distinguer le n -échantillon qui est constitué de variables aléatoires de la réalisation de cet échantillon qui sont des valeurs prises par ces variables. Dans la pratique, on obtient la réalisation d'un échantillon en effectuant n épreuves identiques et indépendantes et on note x_k la valeur prise par la variable X_k correspondant à la k -ième expérience.

II.2 - Estimateur et estimation

Définition 2

Soit $(X_1, \dots, X_n)$ un n -échantillon d'une variable X dont la loi dépend d'un paramètre θ .

- On appelle estimateur de θ toute variable aléatoire T_n qui est une fonction des variables $(X_1, \dots, X_n)$ et indépendante de θ :

$$T_n = f(X_1, \dots, X_n)$$

- Si $(x_1, \dots, x_n)$ est une réalisation de l'échantillon $(X_1, \dots, X_n)$, alors $f(x_1, \dots, x_n)$ est une estimation ponctuelle de θ .

Proposition 1 : La moyenne empirique

Soit $(X_1, \dots, X_n)$ un n -échantillon d'une variable X dont la loi dépend d'un paramètre θ .

Alors, la variable $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$ est un estimateur de θ .

Cet estimateur est appelé moyenne empirique de l'échantillon.

Exercice 1

On reprend l'exemple 1 et on note $(X_1, \dots, X_n)$ un n -échantillon de X .

On considère $T = X_1$ et $\bar{X}_n$ la moyenne empirique de $(X_1, \dots, X_n)$.

- Vérifier que T et $\bar{X}_n$ sont deux estimateurs de p .
- On interroge 100 personnes. La première répond "non" et sur le total, 27 répondent "oui". Donner les estimations correspondants aux estimateurs T et $\bar{X}_{100}$.

II.3 - Qualité de l'estimateur (Hors Programme)

II.3.a - Biais et risque quadratique

Définition 3 : Hors Programme

Soit T_n un estimateur de θ admettant une espérance.

- On appelle biais de l'estimateur T_n le réel $b(T_n) = E(T_n) - \theta$.
- On dit que T_n est un estimateur sans biais de θ lorsque

$$b(T_n) = 0, \text{ soit } E(T_n) = \theta$$

- On dit que l'estimateur T_n est asymptotiquement sans biais lorsque

$$\lim_{n \rightarrow +\infty} b(T_n) = 0, \text{ soit } \lim_{n \rightarrow +\infty} E(T_n) = \theta$$

Remarque 2

- Le biais mesure l'écart moyen entre les valeurs prises par l'estimateur et le paramètre θ à estimer.
- Si l'estimateur est sans biais, les valeurs de l'estimateur sont en moyenne très proches de θ . Cependant, comme le montre l'estimateur T de l'exercice 1 (qui est clairement sans biais), une estimation peut donner une valeur aberrante du paramètre inconnu. On a donc besoin d'une mesure supplémentaire pour différencier deux estimateurs sans biais et dire lequel est le « meilleur ». Il s'agit du risque quadratique.

II.3.b - Risque quadratique

Définition 4 : Hors Programme

Soit T_n un estimateur de θ admettant un moment d'ordre 2.

On appelle risque quadratique de l'estimateur le réel : $r_\theta(T_n) = E((T_n - \theta)^2)$.

Remarque 3

Entre deux estimateurs sans biais de θ , on choisira celui dont le risque quadratique est le plus faible.

Théorème 1 : *Hors Programme*

Soit T_n un estimateur de θ admettant un moment d'ordre 2. Alors on a

$$r_\theta(T_n) = V(T_n) + (b(T_n))^2$$

En particulier, si T_n est un estimateur sans biais ($b(T_n) = 0$), alors : $r_\theta(T_n) = V(T_n)$.

Exercice 2

On reprend l'exercice 1 et on note $(X_1, \dots, X_n)$ un n -échantillon de X .

1. Calculer le biais et le risque quadratique de l'estimateur $T_n = X_1$ puis de l'estimateur $\overline{X_n}$.
2. Quel estimateur de p choisir?

II.4 - Estimateur convergent**Définition 5**

On dit qu'une suite $(T_n)_{n \in \mathbb{N}^*}$ d'estimateurs de θ est convergente si :

$$\forall \epsilon > 0, \lim_{n \rightarrow +\infty} P(|T_n - \theta| > \epsilon) = 0.$$

On dit aussi, par abus de langage, que l'estimateur T_n est convergent.

Exercice 3

On considère une variable aléatoire X d'espérance m et $(X_1, \dots, X_n)$ un n -échantillon de X .

Montrer que $\overline{X_n} = \frac{X_1 + \dots + X_n}{n}$ est un estimateur sans biais et convergent de m .

Théorème 2 : *Hors Programme*

Soit T_n un estimateur de θ admettant un moment d'ordre 2.

Si $\lim_{n \rightarrow +\infty} r_\theta(T_n) = 0$, alors T_n est un estimateur convergent de θ .

Exercice 4

Soit T_n un estimateur de θ admettant un moment d'ordre 2.

1. Montrer, en utilisant l'inégalité de Markov, que : $\forall \epsilon > 0, P(|T_n - \theta| \geq \epsilon) \leq \frac{E((T_n - \theta)^2)}{\epsilon^2} = \frac{r_\theta(T_n)}{\epsilon^2}$
2. Démontrer alors le théorème 2.