Chapitre14

ESTIMATION

14.1 Introduction

Le point de départ de ce chapitre est l'observation d'un phénomène aléatoire modélisé par une variable aléatoire X qui lui est liée. On supposera la loi de X connue sauf un paramètre, noté θ appartenant à un certain ensemble. Notre objectif sera alors d'**estimer** la valeur de θ à partir d'un échantillon $(X_1, \ldots, X_n)$ de n observations du phénomène toutes supposées indépendantes et de même loi que X.

14.2 Estimation ponctuelle

DÉFINITION 14.1

Soit X une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathscr{A}, P)$.
• On appelle n-**échantillon** de la variable X tout n-uplet

$(X_1, \cdots, X_n)$ de variables aléatoires indépendantes et de même loi que X (on dit aussi i.i.d).

• Pour tout $\omega \in \Omega$ on appelle **réalisation du n- échantillon** $(X_1, ..., X_n)$ le n-uplet $(X_1(\omega), ..., X_n(\omega))$.

REMARQUE **14.1.**
Il faut bien différencier le n-échantillon $(X_1, ..., X_n)$ qui est (vecteur) aléatoire et la réalisation $(X_1(\omega), ..., X_n(\omega))$ qui est un élément de $\mathbb{R}^n$.

DÉFINITION 14.2

Soit $(X_1, ..., X_n)$ un n-échantillon d'une variable aléatoire X dont la loi dépend d'un paramètre θ.
On appelle **estimateur de** θ (ou de $g(\theta)$ où g est une fonction une fonction réelle du paramètre θ) toute variable aléatoire T_n qui est une fonction de l'échantillon

$$T_n = \varphi(X_1, \cdots, X_n)$$

on suppose que θ n'intervient pas dans l'expression de T_n.
• Toute réalisation d'un estimateur est appelée **estimation ponctuelle** de θ.

REMARQUE **14.2.**
La définition d'un estimateur est très générale. Rien ne dit que les réalisations de $T_n = \varphi(X_1, ..., X_n)$ donne une bonne estimation du paramètre θ (ou $g(\theta)$) que l'on souhaite estimer. L'idée est de développer des outils qui permettent de comparer *la qualité* d'un estimateur et de choisir celui qui paraît être le "meilleur" pour un problème donnée.

EXEMPLE. Deux exemples d'estimateurs de $\theta \in [0;1]$: On dispose d'une urne contenant des boules rouges et des boules blanches mais on ne connaît pas la composition de l'urne. Comment estimer la proportion de boules

rouges dans l'urne ? Si on note θ la proportion de boules rouges dans cette urne par exemple. Une bonne idée pour estimer θ est de faire des **tirages successifs avec remises** dans l'urne et de noter X_k la variable valant 1 si la boule tirée au k-ième tirage est rouge et 0 sinon.
Les variables $(X_1,...,X_n)$ forment donc un n-échantillon d'une variable X suivant la loi de Bernoulli $\mathscr{B}(\theta)$ et nous cherchons à estimer ce paramètre $\theta \in [0;1]$. On note alors :

$$T_n = X_1 + X_n \text{ et } \overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Les variables aléatoires T_n et $\overline{X_n}$ sont donc deux estimateurs.
• $T_n(\Omega) = \{0;1;2\}$ ainsi une estimation de θ par T_n ne va pas "bien estimer" le paramètre $\theta \in [0,1]$.
• $\overline{X_n}(\Omega) = \left\{ \frac{k}{n} \quad | \quad k \in [\![0;n]\!] \right\} \subset [0;1]$ donc toute estimation par $\overline{X_n}$ sera cohérente avec θ.

REMARQUE **14.3.**
On y réfléchissant bien, pour estimer correctement θ, il vaut mieux faire un grand nombre de tirages et déterminer la fréquence du nombre de boule rouges (estimateur $\overline{X_n}$) plutôt que de ne tenir compte que du premier et du dernier tirage (estimateur T_n) et de ne même pas faire une moyenne.
On remarquera aussi que lorsque $n \longrightarrow +\infty$, les supports $\overline{X_n}(\Omega)$ contiennent un nombre croissant d'éléments qui se répartissent de façon uniforme dans l'intervalle $[0;1]$.

14.2.1 Biais de l'estimateur

DÉFINITION 14.4 (estimateurs sans biais)

Soit T_n un estimateur de θ admettant une espérance.
• On appelle **biais de l'estimateur** le réel $b_\theta(T_n) = E_\theta(T_n) - \theta$.
•On dit que T_n est **un estimateur sans biais** lorsque son biais $b(T_n)$ est nul.

REMARQUE **14.4.**
La notation $E_\theta(T_n)$ permet de mettre en évidence le réel θ que l'on cherche à estimer mais c'est en réalité tout simplement de $E(T_n)$ qu'il s'agit.

PROPRIÉTÉ 14.5 (Hors prorgamme)

Soit $(X_1,...,X_n)$ un n-échantillon d'une variable X admettant une espérance μ.

L'estimateur moyenne empirique $\overline{X_n}$ est un estimateur sans biais de μ.

REMARQUE **14.5.**
Il faudra donc toujours re-démontrer cette propriété.

EXERCICE 14.1. Soit $(X_1,...,X_n)$ un n-échantillon d'une variable $X \hookrightarrow \mathscr{U}[0,\theta]$.

1. Etude de l'estimateur $\overline{X_n}$.
   (a) Montrer que $\overline{X_n}$ est un estimateur biaisé de θ et préciser $b_\theta(\overline{X_n})$.
   (b) Proposer un estimateur T_n de θ sans biais obtenu comme transformation affine de $\overline{X_n}$.
2. Etude de l'estimateur $T_n = \max(X_1,...,X_n)$.
   (a) Déterminer la fonction de répartition de T_n puis en déduire une densité.
   (b) Montrer que :

$$E_\theta(T_n) = \frac{n\theta}{n+1}.$$

   (c) En déduire un estimateur sans bais de θ.

DÉFINITION 14.6 (asymptotiquement sans biais)

On dit qu'un estimateur de θ est **asymptotiquement sans biais** si :

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

REMARQUE **14.6.**
L'estimateur M_n de l'exercice précédent est un estimateur asymptotiquement sans biais de θ.

EXERCICE 14.2. Estimateur variance empirique
Soit $(X_1,...,X_n)$ un n-échantillon d'une variable X de variance $\sigma^2 > 0$.
On note :

$$T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

où $\overline{X_n}$ désigne l'estimateur moyenne empirique de l'échantillon.

1. Montrer que, pour tout $i \in [\![1;n]\!]$, $\quad E\big((X_i - \overline{X_n})^2\big) = V(X_i - \overline{X_n})$.
2. Montrer que, pour tout $i \in [\![1;n]\!]$, $\quad V(X_i - \overline{X_n}) = \left(1 - \frac{1}{n}\right)^2 V(X_i) + \frac{1}{n^2}\sum_{k\neq i} V(X_k)$.
3. En déduire que
$$V(X_i - \overline{X_n}) = \frac{n-1}{n}\sigma^2.$$
4. En déduire un estimateur sans biais de σ^2.
5. Que dire alors de la qualité de cet estimateur ? Comparez-le à l'estimateur S^2_{n-1} :
$$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

14.2.2 Risque quadratique

Remarquons à présent que rien n'empêche un estimateur sans biais de prendre des valeurs très éloignées de θ car en moyenne les écarts peuvent se compenser.

EXEMPLE. On dispose d'une pièce de monnaie. Comment savoir si la pièce est équilibrée ou non ?
On note $\theta \in [0;1]$ la probabilité que la pièce donne PILE. La pièce est donc équilibrée si $\theta = \frac{1}{2}$. Pour estimer θ, on va faire des lancers successifs avec cette pièce et on va noter X_k la variable aléatoire qui vaut 1 si le k-ième lancer donne PILE et 0 sinon. Les variables $(X_1,...,X_n)$ forment donc un n-échantillon d'une variable X suivant la loi de Bernoulli $\mathscr{B}(\theta)$. On a encore $E(X) = \theta$.
Si on choisit l'estimateur $\overline{X_n}$ alors on a vu que $\overline{X_n}$ est un estimateur sans biais de θ.
Si on choisit l'estimateur $T_n = X_n$. Alors T_n est aussi un estimateur sans biais de θ. Cependant $T_n(\Omega) = \{0,1\}$, donc une estimation de θ par T_n est donc une assez "éloigné" de θ

On a donc besoin d'une mesure supplémentaire pour différencier deux estimateurs sans biais et dire lequel est le "meilleur".

DÉFINITION 14.7 (Hors Programme)

Soit T_n un estimateur de $g(\theta)$ admettant un moment d'ordre 2 pour tout θ.
On appelle **risque quadratique de l'estimateur** le réel :

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

REMARQUE **14.7.**
Le risque quadratique mesure la moyenne de l'écart de T_n à θ au carré. Comme un carré est toujours positif, les écarts relatifs à θ ne peuvent plus se compenser mais se cumulent.
On a donc bien ici une façon de mesurer si T_n est un "bon" estimateur de θ.

PROPRIÉTÉ 14.8

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

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

En particulier si T_n est un estimateur sans biais alors $r_\theta(T_n) = V_\theta(T_n)$.

EXERCICE 14.3. Soit $(X_1,...,X_n)$ un n-échantillon d'une variable $X \hookrightarrow \mathscr{U}[0,\theta]$.
On considère les deux estimateurs suivant :

$$V_n = \frac{2}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad T_n = \frac{n+1}{n}\max(X_1,...,X_n).$$

On a vu dans un exercice précédent que ces deux estimateurs sont sans biais pour l'estimation de θ.
Nous souhaitons donc les comparer.

1. Montrer que $r_\theta(V_n) = \dfrac{\theta^2}{3n}$.
2. Montrer que $E_\theta(T_n^2) = \dfrac{(n+1)^2}{n^2}\displaystyle\int_0^\theta \frac{n t^{n+1}}{\theta^n}\,dt$ puis en déduire que $r_\theta(T_n) = \dfrac{\theta^2}{n(n+2)}$.
3. Quel estimateur aura-t-on tendance à privilégier dans la pratique ?

14.2.3 Estimateur convergent

Comme pour le biais, plus le risque quadratique d'un estimateur est petit, meilleur est cet estimateur. On pourrait donc, comme pour le biais introduire la notion de "sans risque quadratique" et "asymptotiquement sans risque quadratique". La première n'a pas vraiment d'intérêt puisque un estimateur sans risque quadratique est un estimateur avec une variance nulle...
La seconde notion, celle d'estimateur "asymptotiquement sans risque quadratique", serait bonne mais on préfère parler d'**estimateur convergent** qui donne une condition plus fine que simplement faire tendre le risque quadratique vers 0.

DÉFINITION 14.9

On dit qu'une suite $(T_n)_{n\in\mathbb{N}^*}$ d'estimateurs de $g(\theta)$ est **convergente** si pour tout θ, $\forall \varepsilon > 0$,

$$\lim_{n\to+\infty} P_\theta(|T_n - g(\theta)| > \varepsilon) = 0$$

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

REMARQUE **14.8.**
On pense immédiatement à la loi faible des grands nombres que l'on peut maintenant reformulé de la façon suivante :
La moyenne empirique $\overline{X_n}$ d'un n-échantillon est un estimateur sans biais et convergent de l'espérance.

PROPRIÉTÉ 14.10 (Hors-Programme)

Si $\lim_{n\to+\infty} r_\theta(T_n) = 0$, alors l'estimateur T_n de $g(\theta)$ est convergent.

14.2.4 Méthode du maximum de vraisemblance

Cette méthode assez générale permet de déterminer l'expression d'un estimateur ayant de bonnes propriétés pour l'estimation de θ.
On retrouve souvent la moyenne empirique mais pas uniquement.
Principe de la méthode :
• on dispose d'une réalisation $(x_1,...,x_n)$ d'un n-échantillon.
• on fait varier θ dans un intervalle et cherche à maximiser la probabilité :

$$P_\theta\big((X_1 = x_1) \cap ... \cap (X_n = x_n)\big) = \prod_{i=1}^{n} P_\theta(X_i = x_i).$$

DÉFINITION 14.11

Soit $(X_1,...,X_n)$ un n-échantillon d'une loi discrète dont θ est un paramètre que l'on souhaite estimer et $(x_1,...,x_n) \in \mathbb{R}^n$ une réalisation fixée de l'échantillon.
On définie une fonction L_n par :

$$L_n : \theta \mapsto \prod_{i=1}^{n} P_\theta(X_i = x_i).$$

En notant $\theta^* = \varphi(x_1,...,x_n)$ la valeur maximale de L_n, **l'estimateur du maximum de vraisemblance de** θ est l'estimateur défini par :

$$\hat{\theta} = \varphi(X_1,...,X_n).$$

EXERCICE 14.4 (Estimateur du maximum de vraisemblance pour la loi $\mathscr{B}(p)$)**.** Soit $(X_1,...,X_n)$ un n-échantillon de la loi $\mathscr{B}(p)$ et $(x_1,...,x_n) \in \{0;1\}^n$. Le paramètre à estimer est p.
On note $s_n = \sum_{i=1}^{n} x_i$.

1. Montrer que $L_n(\theta) = \theta^{s_n}(1-\theta)^{n-s_n}$.
2. Pour $\theta \in]0;1[$, on pose :
$$h_n(\theta) = \ln(L_n(\theta)).$$
   (a) Montrer que h_n est dérivable et déterminer h_n'.
   (b) Déterminer la valeur θ^* pour laquelle h_n est maximale.
3. En déduire l'estimateur du maximum de vraisemblance.

14.3 Estimation par intervalle de confiance

S'il existe des critères pour juger des qualités d'un estimateur ponctuel T_n de θ (biais, risque, convergence), aucune certitude ne peut jamais être apportée quant au fait que l'estimation donne la vraie valeur à estimer. La démarche de l'estimation par intervalle de confiance consiste à trouver un intervalle aléatoire qui contient θ avec une probabilité minimale donnée.

DÉFINITION 14.12 (Intervalle de confiance)

Soient $(X_1,\cdots,X_n)$ un n-échantillon,$\alpha \in]0;1[$ et U_n, V_n deux estimateurs de $g(\theta)$ tels que $U_n \leqslant V_n$
On dit que $[U_n, V_n]$ est un **intervalle de confiance de** θ **au niveau de confiance** $1-\alpha$
si :

$$P_\theta(\theta \in [U_n, V_n]) = P_\theta(U_n \leqslant g(\theta) \leqslant V_n) \geqslant 1-\alpha$$

14.3.1 Avec Bienaymé-Tchebychev

Déterminer un intervalle de confiance avec Bienaymé-Tchebychev

1. On montre que l'estimateur est sans biais et admet un moment d'ordre 2.
2. On applique Bienaymé-Tchebychev et (après transformations) on obtient pour tout $\varepsilon > 0$,
$$P\big(|T_n - \theta| \geqslant \varepsilon\big) \leqslant \frac{V(T_n)}{\varepsilon^2}$$
3. On choisi $\varepsilon > 0$ pour que $\dfrac{V(T_n)}{\varepsilon^2} = \alpha$ puis par passage au complémentaire, on obtient,
$$P\big(|T_n - g(\theta)| \leqslant \varepsilon\big) \geqslant 1 - \alpha$$
4. On transforme l'intérieur de la probabilité pour présenter le résultat sous la forme
$$P(U_n \leqslant \theta \leqslant V_n) \geqslant 1 - \alpha$$

EXERCICE 14.5. Soit X une variable aléatoire qui suit une loi de Bernoulli dont on veut estimer le paramètre p avec la moyenne empirique $\overline{X}_n$ d'un échantillon.
En utilisant l'inégalité de Bienaymé-Tchebychev et la majoration $p(1-p) \leqslant \dfrac{1}{4}$, déterminer un intervalle de confiance de p au niveau de confiance 0,95.

14.3.2 Avec le théorème central limite

On a vu que l'utilisation de l'inégalité de Bienaymé-Tchebychev permettait d'obtenir des intervalles de confiance. Néanmoins, afin d'élargir notre palette d'outils, nous allons voir une méthode utilisant le théorème de la limite centrale pour quasiment obtenir un intervalle de confiance. On obtient en réalité ce qu'on appelle un intervalle de confiance asymptotique.

DÉFINITION 14.13 (Intervalle de confiance asymptotique)

Soient $(X_1, \cdots, X_n)$ un n-échantillon, $\alpha \in]0;1[$ et U_n, V_n deux estimateurs de $g(\theta)$ tels que $U_n \leqslant V_n$
L'intervalle $[U_n, V_n]$ est un **intervalle de confiance asymptotique de** $g(\theta)$ **au niveau de confiance** $1-\alpha$ si pour tout $n \in \mathbb{N}$,
$$\lim_{n \longrightarrow +\infty} P_\theta(U_n \leqslant \theta \leqslant V_n) \geqslant 1 - \alpha \quad \text{avec } \alpha \in]0;1[.$$

Déterminer un intervalle de confiance avec le TLC

1. On vérifie que le paramètre à estimer intervient dans l'expression
$$\overline{X}_n^* = \sqrt{n}\frac{\overline{X}_n - E(X_n)}{\sqrt{V(X_n)}}$$
2. On applique le TLC à notre n-échantillon, on obtient, que pour tout $x > 0$,
$$\lim_{n \longrightarrow +\infty} P(-x < \overline{X}_n^* \leqslant x) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1$$
3. On choisit x pour que $2\Phi(x) - 1 \geqslant 1 - \alpha$
4. On transforme l'intérieur de la probabilité pour présenter le résultat sous la forme
$$\lim_{n \to +\infty} P(U_n \leqslant \theta \leqslant V_n) \geqslant 1 - \alpha$$

EXERCICE 14.6. Reprenons l'exercice précédent.

2. En utilisant cette fois le TLC et la majoration $p(1-p) \leqslant \frac{1}{4}$, déterminer un intervalle de confiance asymptotique de p au niveau de confiance 0,95.
3. Comparer les deux résultats obtenus