Chapitre 5

Théorèmes limites et Estimation Ponctuelle

5.1 Théorèmes limites

On considère n variables aléatoires (discrètes ou continues) $X_1, X_2, \ldots, X_n$ définies sur un même support $X(\Omega)$ (fini ou non).

Exemple 5.1. X_k est le salaire d'un individu tiré au hasard dans la population à qui on attribue l'indice k. X_k est supposé ici être une variable dans $\mathbb{R}_+$.

À partir de celles-ci, on construit une nouvelle variable, notée Y_n, telle que

$$\forall n \geq 1, \quad Y_n = f\left(X_1, X_2, \ldots, X_n\right).$$

Par exemple, on peut considérer simplement $Y_n = X_n$, ou encore la somme des variables $Y_n = X_1 + X_2 + \cdots + X_n$, ou encore la moyenne empirique $Y_n = \left(X_1 + X_2 + \cdots + X_n\right)/n$.

Le but de chapitre est d'étudier le comportement de la suite de variables aléatoires Y_n lorsque la dimension n tend vers l'infini. Est-ce que Y_n est toujours définie comme une variable aléatoire lorsque n tend vers l'infini ? Ou, au contraire, se comporte-t-elle comme une variable *dégénérée* (quantité certaine) ? Quelle est la *loi asymptotique* de cette variable ?

Pour répondre à ces questions, nous allons introduire différents concepts de convergence. La notion de convergence constitue la base de la statistique mathématique et de la théorie des tests. Dans ce cadre, nous présenterons deux résultats fondamentaux : la loi des grands nombres (LGN) et le théorème central limite (TCL). Ces deux théorèmes s'intéressent au comportement asymptotique d'une fonction particulière des variables $X_1, X_2, \ldots, X_n$, à savoir la moyenne empirique

$$Y_n = f(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Pourquoi s'intéresser tout particulièrement au comportement asymptotique de la moyenne empirique ? Dans la pratique, la variable étudiée correspond généralement à un estimateur, souvent fonction de la moyenne empirique, et la dimension n à la taille de l'échantillon. On souhaite alors étudier les propriétés asymptotiques de cet estimateur en faisant tendre la taille de l'échantillon vers l'infini (voir le cours de statistique du second semestre).

Pour illustrer la loi des grands nombres et le théorème central limite, nous terminerons ce chapitre par la méthode de Monte-Carlo : il s'agit d'une méthode numérique qui permet d'approcher $\mathbb{E}[g(X)]$ à partir d'une suite $(X_n)_{n\geq 0}$ de variables aléatoires indépendantes et de même loi que X en utilisant la moyenne empirique $\frac{1}{n}\sum_{k=1}^{n} g(X_k)$. L'efficacité asymptotique de cette méthode est démontrée à l'aide de la LGN et du TCL.

5.1.1 Modes de convergence de suite de variables aléatoires

L'objectif de cette section est d'analyser le comportement d'une *suite de variables aléatoires*, indicées par $n \in \mathbb{N}^*$, c'est-à-dire une famille de variables aléatoires indexée par un entier strictement positif. Une suite de variables aléatoires est généralement notée sous la forme $(Y_n)_{n\in\mathbb{N}^*}$, ou $(Y_n)_{n\geq 1}$, ou simplement (Y_n). Souvent en statistique, cette suite est définie comme une fonction $Y_n = f(X_1, \ldots, X_n)$ d'autres variables aléatoires $X_1, \ldots, X_n$ définies sur $X(\Omega)$. La suite $(Y_n)_{n\in\mathbb{N}^*}$ est définie sur un support $Y(\Omega)$, directement obtenue à partir de la suite $(X_n)_{n\in\mathbb{N}^*}$.

Comme lors de l'étude de suites de fonctions ou de séries classiques, différentes notions de « limites » ou convergence existent et nous nous intéresserons ici aux principales, à savoir

- la convergence presque sûre ;
- la convergence en probabilité ;
- la convergence en loi.

5.1.1.a Convergence presque-sûre

Définition 5.1. *On dit qu'une suite de variables aléatoires* $(Y_n)_{n\in\mathbb{N}^*}$ converge presque-sûrement *(ou au sens fort) vers une constante* $c \in \mathbb{R}$ *si*

$$\mathbb{P}\left(\lim_{n\to+\infty} Y_n = c\right) = 1.$$

On note alors $Y_n \xrightarrow{p.s.} c$.

Cette définition indique que lorsque n tend vers l'infini, les réalisations de la variable aléatoire Y_n sont systématiquement égales à une constante c : on parle alors de convergence *trajectorielle*, c'est la plus forte. On a alors que Y_n tend vers une variable aléatoire « dégénérée », c'est-à-dire une quantité non aléatoire [1].

Cette forme convergence est très contraignante et avec elle il n'y aurait alors « pas beaucoup » de suites « convergentes » On introduit la notion suivante de *convergence en probabilité*, qui intègre les suites convergentes presque-sûrement mais bien d'autres encore.

5.1.1.b Convergence en probabilité

L'idée de la convergence en probabilité est assez similaire à celle de la convergence presque-sûre. Lorsque la dimension n tend vers l'infini, la suite (Y_n) tend vers une constante déterministe c. La différence est que cette convergence n'est pas stricte : la variable Y_n est *presque* dégénérée, mais elle reste toutefois une variable aléatoire même si sa densité est extrêmement concentrée autour de la valeur c.

1. Si Y_n est la moyenne d'une suite de v.a. $(X_n)_{n\in\mathbb{N}}$, alors Y_n converge presque sûrement vers c si toutes les valeurs $Y_1, Y_2, \ldots, Y_n$ sont égales à c lorsque n est assez grand.

Définition 5.2. *On dit qu'une suite de variables aléatoires* $(Y_n)_{n\in\mathbb{N}^*}$ converge en probabilité *(ou au sens faible) vers une constante* c *si pour toute valeur de* $\varepsilon > 0$ *on a*

$$\lim_{n\to+\infty} \mathbb{P}\left(|Y_n - c| > \varepsilon\right) = 0.$$

On note alors $Y_n \xrightarrow{\mathbb{P}} c$.

Remarque 5.1. Cette définition est équivalente à

$$\forall \varepsilon > 0, \quad \lim_{n\to+\infty} \mathbb{P}\left(|Y_n - c| \leq \varepsilon\right) = 1.$$

Avec cette convergence, la variable aléatoire prend des valeurs très proches de c lorsque n tend vers l'infini, mais sans être systématiquement égale à cette valeur (comme ce qu'impose de la convergence presque-sûre). On remarque alors que la convergence presque-sûre implique la convergence en probabilité (la réciproque n'est pas vraie en général).

Sous quelles conditions, une suite de variables aléatoires (Y_n) converge-t-elle en probabilité ? Une condition nécessaire et suffisante à la convergence en probabilité est la suivante.

Propriété 5.1. *Soit* $(Y_n)_{n\in\mathbb{N}^*}$ *une suite de variables aléatoires vérifiant*

$$\lim_{n\to+\infty} \mathbb{E}[Y_n] = c \in \mathbb{R} \quad \textit{et} \quad \lim_{n\to+\infty} \mathbb{V}(Y_n) = 0.$$

Alors la suite $(Y_n)_{n\in\mathbb{N}^*}$ *converge en probabilité vers* c *losque* n *tend vers l'infini :* $Y_n \xrightarrow{\mathbb{P}} c$.

Exemple 5.2. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi de Bernoulli de paramètre $p \in]0,1[$. On pose

$$\forall n \geq 1, \quad Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

la *moyenne empirique* des X_k. Alors

$$\mathbb{E}[Y_n] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[X_k] = \frac{1}{n}\sum_{k=1}^{n} p = p$$

et, par indépendance des X_k,

$$\mathbb{V}(Y_n) = \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{V}(X_k) = \frac{1}{n^2}\sum_{k=1}^{n} p(1-p) = \frac{p(1-p)}{n} \underset{n\to+\infty}{\longrightarrow} 0.$$

Donc $Y_n \xrightarrow{\mathbb{P}} p$.

Voyons maintenant des propriétés supplémentaires de la convergence en probabilité : si une suite de variables aléatoires $(Y_n)_{n\in\mathbb{N}^*}$ converge en probabilité vers une constante c, est-ce qu'une transformation de Y_n converge en probabilité ? Si deux suites de variables aléatoires convergent en probabilité, quelles combinaisons des deux convergent en probabilité ?

Propriété 5.2. *1. Si* $Y_n \xrightarrow{\mathbb{P}} c$ *et si* g *est continue sur* $\mathbb{R}$ *alors* $g(Y_n) \xrightarrow{\mathbb{P}} g(c)$.
En particulier,

$$Y_n \xrightarrow{\mathbb{P}} c \implies aY_n + b \xrightarrow{\mathbb{P}} ac + b, \ \forall a, b \in \mathbb{R}.$$

2. On suppose que $Y_n \xrightarrow{\mathbb{P}} c$ *et que* $Z_n \xrightarrow{\mathbb{P}} d$.

Linéarité : $\forall a, b \in \mathbb{R}$, $aY_n + bZ_n \xrightarrow{\mathbb{P}} ac + bd$.

Produit : $Y_n Z_n \xrightarrow{\mathbb{P}} cd$.

Quotient : *Si* $\mathbb{P}(d = 0) = 0$, *alors* $\dfrac{Y_n}{Z_n} \xrightarrow{\mathbb{P}} \dfrac{c}{d}$.

3. La convergence presque-sûre est plus forte que la convergence en probabilité : $Y_n \xrightarrow{p.s.} c \implies Y_n \xrightarrow{\mathbb{P}} c$.

Exemple 5.3. Soient $(X_n^1)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi de Bernoulli de paramètre $p_1 \in]0,1[$ et $(X_n^2)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes et de même loi de Bernoulli de paramètre $p_2 \in]0,1[$. On pose $Y_n = \frac{1}{n}\sum_{k=1}^n X_k^1$ et $Z_n = \frac{1}{n}\sum_{k=1}^n X_k^2$. D'après l'exemple précédent, on a $Y_n \xrightarrow{\mathbb{P}} p_1$ et $Z_n \xrightarrow{\mathbb{P}} p_2$. Alors, pour tous $a, b \in \mathbb{R}$, $aY_n + b \xrightarrow{\mathbb{P}} ap_1 + b$, $aZ_n + b \xrightarrow{\mathbb{P}} ap_2 + b$, $aY_n + bZ_n \xrightarrow{\mathbb{P}} ap_1 + bp_2$, $Y_n Z_n \xrightarrow{\mathbb{P}} p_1 p_2$ et $\frac{Y_n}{Z_n} \xrightarrow{\mathbb{P}} \frac{p_1}{p_2}$ car $p_2 \neq 0$.

5.1.1.c Convergence en loi (ou en distribution)

La notion de convergence en loi (ou en distribution) est fondamentalement différente des deux notions de convergence étudiées précédemment (presque-sûre et en probabilité). Pour ces deux notions, nous avons vu qu'une suite de variables aléatoires, indicée par n, converge vers une constante (quantité déterministe), lorsque la dimension n tend vers l'infini. Au contraire, dans le cadre de la convergence en loi, une suite de variables aléatoires converge vers une autre variable aléatoire, ne dépendant pas de la dimension n. Il s'agit alors de considérer la convergence simple des fonctions de répartitions de Y_n vers la fonction de répartition d'une autre variable aléatoire Z, définie sur le même support $Y(\Omega)$.

Définition 5.3. *On dit que la suite* $(Y_n)_{n\in\mathbb{N}^*}$ converge en loi *vers* Z, *lorsque*

$$\lim_{n\to+\infty} \mathbb{E}[g(Y_n)] = \mathbb{E}[f(Z)] \ \textit{pour toute fonction } g \textit{ continue bornée.}$$

On écrit alors $Y_n \xrightarrow{\mathcal{L}} Z$.

Au lieu d'écrire $Y_n \xrightarrow{\mathcal{L}} Z$, on peut directement faire le lien avec la loi de Z : par exemple $Y_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ lorsque $Z \sim \mathcal{N}(0,1)$.

Remarque 5.2. Une définition alternative, plus pratique, issue de la précédente, est la suivante. Notons F_{Y_n} la fonction de répartition de Y_n pour tout $n \in \mathbb{N}^*$ et F_Z celle de Z. Alors on dit que la suite $(Y_n)_{n\in\mathbb{N}^*}$ converge en loi vers Z lorsque

$$\forall y \in Y(\Omega), \quad \lim_{n\to+\infty} F_{Y_n}(y) = F_Z(y).$$

Ainsi, la convergence en loi signifie que la variable aléatoire Y_n lorsque n est assez grand est distribué identiquement à Z (qui suit par exemple une loi normale centrée réduite, ou autre).

Remarque 5.3. Une suite de variables aléatoires discrètes peut converger en loi vers une variable aléatoire discrète ou continue, tandis qu'une suite de variables aléatoires continues peut converger en loi vers une variable aléatoire à densité mais aussi vers une variable aléatoire discrète.

Exemple 5.4. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires uniformes sur $\{1,\dots,n\}$. On pose $Y_n = a + \dfrac{b-a}{n}X_n$, $n \geq 1$. Alors, pour toute fonction continue g sur $[a,b]$, on a, en utilisant les sommes de Riemann,

$$\begin{aligned}
\mathbb{E}[g(Y_n)] &= \mathbb{E}\left[g\left(a + \frac{b-a}{n}X_n\right)\right] \\
&= \frac{1}{n}\sum_{k=1}^{n} g\left(a + \frac{b-a}{n}k\right) \\
&= \frac{1}{b-a}\left(\frac{b-a}{n}\sum_{k=1}^{n} g\left(a + \frac{b-a}{n}k\right)\right) \\
&\underset{n\to+\infty}{\longrightarrow} \frac{1}{b-a}\int_a^b g(x)dx = \mathbb{E}[g(Z)], \quad Z \sim \mathcal{U}_{[a,b]}.
\end{aligned}$$

Regardons maintenant le lien entre la convergence en loi et celle en probabilité : la convergence en probabilité entraîne la convergence en loi, *mais la réciproque est fausse sauf si la limite est une variable aléatoire dégénérée (non aléatoire).*

Propriété 5.3. *1.* $Y_n \xrightarrow{\mathbb{P}} Z \implies Y_n \xrightarrow{\mathcal{L}} Z$.

2. $Y_n \xrightarrow{\mathcal{L}} c \in \mathbb{R} \implies Y_n \xrightarrow{\mathbb{P}} c$.

On remarque donc que la convergence en loi est plus faible (moins contraignante) que la convergence en probabilité. De ce fait, il n'y a pas unicité de la limite lors de la convergence en loi (puisqu'il n'y a pas unicité de la limite lors de la convergence en probabilité) et il n'y a pas non plus compatibilité de l'espérance avec la convergence en loi (puisque c'est aussi le cas lors de la convergence en probabilité). La convergence en loi n'est pas compatible, sauf cas particuliers, avec les opérations sur les variables aléatoires (au contraire de la convergence en probabilité).

Cas particulier de la convergence en loi dans le cas de v.a. discrètes.
Si, pour tout $n \in \mathbb{N}^*$, Y_n et Z sont toutes des variables aléatoires discrètes, on a la caractérisation suivante

$$Y_n \xrightarrow{\mathcal{L}} Z \iff \forall k \in Y(\Omega),\ \lim_{n\to\infty} \mathbb{P}(Y_n = k) = \mathbb{P}(Z = k).$$

C'est ainsi que l'on approche des lois binomiales par des lois de Poisson pour p petit.

Exemple 5.5. Soit $(Y_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans $\{0,1\}$ telles que

$$\forall n \geq 1, \quad \mathbb{P}(Y_n = 1) = \frac{1}{n} \quad \text{et} \quad \mathbb{P}(Y_n = 0) = 1 - \frac{1}{n}.$$

On a alors

$$\lim_{n\to+\infty} \mathbb{P}(Y_n = 1) = \lim_{n\to+\infty} \frac{1}{n} = 0 \quad \text{et} \quad \lim_{n\to+\infty} \mathbb{P}(Y_n = 0) = \lim_{n\to+\infty} 1 - \frac{1}{n} = 1,$$

donc $Y_n \xrightarrow{\mathcal{L}} 0$ et par conséquent $Y_n \xrightarrow{\mathbb{P}} 0$.

5.1.2 Loi des grands nombres et TCL

Dans cette section, nous allons nous intéresser au comportement asymptotique de la moyenne empirique de la suite $(X_n)_{n\geq 1}$: nous allons donc poser

$$\forall n \geq 1, \quad Y_n = \frac{X_1 + \cdots + X_n}{n} := \overline{X}_n.$$

La notation de la moyenne empirique $\overline{X}_n$ est la notation classique en statistique.

Étudions alors les modes de convergence de cette suite $\left(\overline{X}_n\right)_{n\geq 1}$.

5.1.2.a Lois des Grands Nombres (LGN)

Nous commençons par étudier le mode de convergence de la suite $\left(\overline{X}_n\right)_{n\geq 1}$ vers une constante. Intuitivement, si les variables aléatoires sont i.i.d. (indépendantes et identiquement distribuées), la moyenne empirique $\overline{X}_n$ devrait converger vers l'espérance des variables aléatoires $X_1, \ldots, X_n$ supposée identique $m = \mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n]$.

Nous verrons que selon les hypothèses sur la suite $(X_n)_{n\geq 1}$, la convergence aura lieu en probabilité (au sens *faible*) ou presque-sûrement (au sens *fort*). C'est pourquoi l'on distingue la loi *faible* des grands nombres de la loi *forte*. Le terme de *loi* est ici à prendre au sens de *théorème* et non au sens probabiliste.

Loi faible des Grands Nombres (LfGN). Une des principales applications de la notion de convergence en probabilité est la loi faible des grands nombres. Elle a été énoncée par Khintchine (1878-1959) quand les X_k sont des variables aléatoires i.i.d., mais le résultat est valable pour des hypothèses plus faibles précisées dans l'énoncé suivant.

Théorème 5.1 (LfGN)**.** *Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires* indépendantes *d'espérance commune m finie, alors la moyenne empirique $\overline{X}_n$ converge en probabilité vers m :*

$$\overline{X}_n \xrightarrow{\mathbb{P}} m \iff \forall \varepsilon > 0, \ \lim_{n\to\infty} \mathbb{P}\big(|\overline{X}_n - m| \leq \varepsilon\big) = 1.$$

Exemple 5.6. On lance une centaine de fois une pièce de monnaie équilibrée ($m = p = \frac{1}{2}$), on s'attend à avoir « en moyenne » une cinquantaine de Piles ($m = np = 100 \times 1/2 = 50$).

Dans le cas où la variance σ^2 est finie, son obtention est une conséquence directe de l'inégalité de Bienaymé-Tchebyschev (ou encore de la propriété 4.1). Par linéarité de l'espérance, on a en effet

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mathbb{E}[X_1] = m,$$

tandis que par indépendance

$$\mathbb{V}\big(\overline{X}_n\big) = \frac{1}{n^2}\big(\mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n)\big) = \frac{n\mathbb{V}(X_1)}{n^2} = \frac{\sigma^2}{n} < +\infty \quad \text{si } \sigma^2 < +\infty.$$

Soit $\varepsilon > 0$. D'après l'inégalité de Bienaymé-Tchebyschev, on a donc

$$\mathbb{P}\big(|\overline{X}_n - m| > \varepsilon\big) \leq \frac{\sigma^2}{\varepsilon^2 n},$$

qui tend bien vers 0 quand n tend vers $+\infty$, quelle que soit la différence ε fixée.

Forte des Grands Nombres (LFGN). Lorsque la convergence précédente est presque-sûre au lieu d'être en probabilité, on dit que la *Loi Forte des Grands Nombres (LFGN)* est vérifiée (ou le théorème de Kolmogorov). Sa preuve est plus difficile et nous nous contenterons de l'énoncer. Dans le cadre d'hypothèses qui est le nôtre, elle s'applique à la moyenne empirique de la même manière que la loi faible.

Théorème 5.2 (LFGN)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires* indépendantes*, de* même loi *et* intégrables *($\mathbb{E}[|X_k|] < +\infty$, $1 \leq k \leq n$), d'espérance m, alors la moyenne empirique $\overline{X}_n$ converge presque-sûrement vers m :*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \quad \xrightarrow[n\to+\infty]{p.s.} \quad m.$$

5.1.2.b Théorème Central Limite (TCL)

Le théorème central limite (TCL) permet d'étudier la convergence en loi d'une transformée de la moyenne empirique de variables aléatoires indépendantes. C'est sans conteste le théorème fondamental de la statistique mathématique.

Pourquoi une transformation de la moyenne empirique et non elle-même ? Si la moyenne empirique vérifie la loi faible des grands nombres, c'est-à-dire

$$\overline{X}_n \xrightarrow{\mathbb{P}} m,$$

alors lorsque n tend vers l' infini, la moyenne empirique tend à être une variable aléatoire « dégénérée » : elle se réduit « presque » à une quantité déterministe égale à m (constante) puisque sa variance tend vers 0. Sous des conditions supplémentaires sur les variables X_k, la moyenne empirique vérifie la loi forte des grands nombres, soit

$$\overline{X}_n \xrightarrow{p.s.} m,$$

et alors, lorsque n tend vers l' infini, la moyenne empirique n'est plus une variable aléatoire. Ce résultat est problématique : lorsque la dimension n tend vers l'infini, la distribution de la moyenne empirique est dégénérée et il n'est pas possible de construire une théorie de l'estimation (ou inférence) à partir de cette distribution.

La solution consiste à *transformer* la variable $\overline{X}_n$ de sorte à ce que la variable transformée converge en loi vers une *distribution non dégénérée*, c'est-à-dire une distribution dont la variance ne tende ni vers 0, ni vers l'infini (auquel cas la densité serait non définie). Comme nous allons le découvrir dans l'énoncé du théorème central limite, cette transformation est de la forme

$\sqrt{n}\left(\overline{X}_n - m\right)$. Dans cette transformation, l'élément le plus important est le terme $\sqrt{n}$ qui détermine la *vitesse de convergence* de la variable transformée. La *loi asymptotique* que nous allons alors obtenir est la *loi normale*, que l'on avait déjà utilisée dans l'approximation des lois binomiales.

Il existe plusieurs versions du théorème central limite, nous présentons ici celle énoncée par Lindeberg-Lévy (1920).

Théorème 5.3 (TCL). *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de moyenne commune m finie et de variance commune σ^2 également finie. Alors*

$$\sqrt{n}\left(\overline{X}_n - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

D'autres énoncés équivalents de ce théorème sont possibles :

- si l'on réduit $\sqrt{n}\left(\overline{X}_n - m\right)$, on obtient

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, 1);$$

- si l'on multiplie par n le numérateur et le dénominateur de la fraction de l'énoncé précédent et que l'on simplifie le dénominateur avec la vitesse $\sqrt{n}$, on obtient

$$\frac{\sum_{k=1}^{n} X_k - n \cdot m}{\sigma\sqrt{n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Remarque 5.4. Dans la pratique, on considère que n est « suffisamment grand » lorsque n atteint quelques dizaines : par exemple lorsque $n \geq 30$, mais cela dépend de la nature du problème, de la population et du contexte de l'étude.

Cette loi est très générale et ne s'utilise pas seulement pour approcher la loi binomiale. Par exemple (parmi tant d'autres), nous avons

1. Si le paramètre λ est supérieur à 10, on peut approcher la loi de Poisson $\mathcal{P}(\lambda)$ par une loi normale de même espérance et variance $\mathcal{N}(\lambda, \lambda)$ en remplaçant $\mathbb{P}(X = k)$ par $\mathbb{P}(k - 0,5 < Z < k + 0,5)$ (correction de continuité), où $Z \sim \mathcal{N}(\lambda; \lambda)$.
2. Quelles que soient les lois des variables $X_1, \ldots, X_n$ i.i.d., d'espérance m et de variance σ^2 finies, leur moyenne empirique $\overline{X}_n$ suit asymptotiquement une loi normale $\mathcal{N}(m, \sigma^2)$ si n tend vers l'infini.
3. Quand le nombre de degré de liberté k tend vers l'infini, la loi de Student $\mathcal{S}t(k)$ tend vers la loi normale centrée réduite $\mathcal{N}(0, 1)$.

5.2 Estimation Ponctuelle

L'objectif de l'estimation est de révéler de l'information sur une caractéristique de la population à partir d'un échantillon. Dans ce cadre, on cherche à révéler la valeur d'un ou de plusieurs paramètres associés à la distribution de la caractéristique d'intérêt dans la population. On construit pour cela un estimateur qui est une variable aléatoire, définie comme une fonction des variables de l'échantillon. On étudie ensuite ses propriétés liées à sa distribution,

comme son son espérance, sa variance, etc. L'idée générale est de vérifier théoriquement si les réalisations de cette variable aléatoire ont de grandes chances d'être "proches" de la vraie valeur du paramètre que l'on souhaite estimer.

On peut aussi comparer différents estimateurs afin de choisir le plus performant : on introduit pour cela les notions d'estimateur optimal et d'estimateur efficace. Une fois que l'on dispose d'un "bon" estimateur, on l'utilise pour obtenir une estimation.

Une estimation ponctuelle n'est rien d'autre que la réalisation de l'estimateur obtenue à partir de la réalisation de l'échantillon, c'est-à-dire à partir des données statistiques ou des observations. Pour obtenir une estimation, il suffit donc d'appliquer sur les données la "formule" qui définit l'estimateur en fonction des variables de l'échantillon.

Dans ce chapitre, nous commencerons par définir ce qu'est un estimateur et les propriétés intéressantes qu'il doit vérifier pour être performant. Nous étudierons ensuite deux techniques d'estimation ponctuelle : la méthode des moments et celle du maximum de vraisemblance.

5.2.1 Estimateur : Définition et Propriétés

L'objectif d'une procédure d'estimation est de révéler de l'information sur le (ou les) paramètre(s) d'intérêt de la population à partir d'un échantillon aléatoire. Le problème général est le suivant. On suppose que la caractéristique d'intérêt dans la population, notée X, est une variable aléatoire définie sur un univers probabilisé $(X(\Omega), \mathcal{F}, \mathbb{P})$. La loi de probabilité de cette variable aléatoire est représentée

— soit par une fonction de densité si X est une variable continue,

— soit par une fonction de masse si X est une variable discrète.

On suppose que cette fonction de densité ou de masse dépend d'un *paramètre* θ, qui est a priori *inconnu* et que l'on cherche à estimer. On note alors $f_X(x, \theta)$, $x \in X(\Omega)$, la fonction de densité ou de masse de la variable X.

Pour estimer le paramètre θ, on dispose d'un n-échantillon $(X_1, \ldots, X_n)$ dans lequel toutes les variables aléatoires X_i, pour $i = 1, \ldots, n$, sont supposées indépendantes et identiquement distribuées (i.i.d.), de même loi que X. On note $(x_1, \ldots, x_n)$ la réalisation de cet échantillon : cette réalisation correspond aux données (fichier Excel, tableau de valeurs, etc.) utilisées pour l'estimation.

Exemple 5.7. On suppose que la durée de vie d'un équipement, notée D, peut être représentée par une variable aléatoire positive, admettant une distribution exponentielle de paramètre $A > 0$ inconnu. Afin de l'estimer, on dispose de six relevés pour lesquels on a pu observer la durée écoulée (exprimée en heures) avant la rupture de l'équipement : (100, 102,95 , 78, 135,98). Ces six valeurs correspondent à la réalisation d'un échantillon aléatoire de taille $n = 6$, noté $(D_1, \ldots, D_6)$, où les variables D_i, pour $i = 1, \ldots, 6$, sont i.i.d. de même loi que D.

5.2.1.a Définition

La théorie générale de l'estimation repose sur la notion d'estimateur.

Définition 5.4. *Un* estimateur du paramètre θ *est une fonction des variables aléatoires* $X_1, \ldots, X_n$ *de l'échantillon. Cet estimateur, noté* $\widehat{\theta}_n$*, est défini par*

$$\widehat{\theta}_n = g(X_1, \ldots, X_n).$$

Un estimateur est une variable aléatoire*, puisque c'est une fonction des variables aléatoires de l'échantillon.*

Bien évidemment, cette fonction ou cette "formule" $g(\cdot)$ n'est pas choisie au hasard. L'idée est de trouver une fonction qui combine les réalisations de l'échantillon de sorte à révéler de l'information sur le paramètre d'intérêt θ.

Exemple 5.8 (Moyenne Empirique)**.** Supposons que les variables aléatoires $(X_1, \ldots, X_n)$ soient i.i.d. de même loi que X. La moyenne empirique (statistique descriptive)

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

est un estimateur de l'espérance. En effet, $\overline{X}_n$ est une fonction des variables $X_1, \ldots, X_n$ telle que

$$\overline{X}_n = \frac{1}{n} (X_1 + \cdots + X_n) = g(X_1, \ldots, X_n).$$

Définition 5.5. *Une* réalisation de l'estimateur $\widehat{\theta}_n$ *associée à une réalisation* $(x_1, \ldots, x_n)$ *de l'échantillon correspond à une* estimation (ponctuelle) *du paramètre θ. L'estimation est généralement notée $\widehat{\theta}_n(x)$ pour la différencier de la variable aléatoire (estimateur) $\widehat{\theta}_n$*

$$\widehat{\theta}_n(x) = g(x_1, \ldots, x_n).$$

Une estimation n'est donc rien d'autre que l'application de la "formule" $g(X_1, \ldots, X_n)$ aux données, c'est-à-dire aux réalisations de l'échantillon $(x_1, \ldots, x_n)$.

Exemple 5.9. Soit un échantillon (X_1, X_2) de variables i.i.d. de même loi qu'une variable X. On admet que $\overline{X}_2 = (X_1 + X_2)/2$ est un estimateur du paramètre $\theta = \mathbb{E}[X]$. Pour une réalisation $(x_1, x_2) = (10, 4)$ de l'échantillon, on obtient une estimation (ponctuelle) du paramètre θ égale à $\widehat{\theta}_2(x) = \frac{10+4}{2} = 7$.

Ainsi à ce stade du chapitre, il convient de bien distinguer la notion d'estimateur de la notion d'estimation (réalisation)

— Estimateur (variable aléatoire) : $\widehat{\theta}_n$;
— Estimation (constante) : $\widehat{\theta}_n(x)$.

5.2.1.b Méthodes d'estimation

On peut concevoir une méthode d'estimation comme une sorte de recette de cuisine qui permet d'obtenir un estimateur $\widehat{\theta}_n$ à partir des ingrédients $X_1, \ldots, X_n$. Plus formellement, on définit une méthode d'estimation de la façon suivante.

Définition 5.6. *Une méthode d'estimation est une méthode mathématique qui permet de dériver la forme fonctionnelle d'un estimateur $\widehat{\theta}_n = g(X_1, \ldots, X_n)$ à partir des variables aléatoires de l'échantillon $X_1, \ldots, X_n$.*

Pour un même problème, on peut parfois appliquer plusieurs méthodes d'estimation. À chaque méthode d'estimation correspond un estimateur particulier. Si l'on se restreint aux seules méthodes d'estimation paramétriques, il existe de nombreuses méthodes suivant le problème étudié et les hypothèses retenues. Citons par exemple

— la méthode des moindres carrés ordinaires ;
— la méthode des moindres carrés généralisés ;
— la méthode des moments généralisés (voir Section 2) ;
— la méthode du maximum de vraisemblance (voir Section 3) ;
— la méthode des variables instrumentales ;
— la méthode des doubles moindres carrés ordinaires.

5.2.1.c Propriétés

La question est de savoir ce qu'est un "bon" estimateur. Quelles propriétés doit satisfaire un estimateur pour être considéré comme performant ?

• Biais d'un estimateur

Définition 5.7. *On appelle* biais de l'estimateur $\widehat{\theta}_n$ *d'un paramètre* θ *la différence entre son espérance et le paramètre* θ

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

Un estimateur $\widehat{\theta}_n$ *d'un paramètre* θ *est dit* non biaisé *ou* sans biais *si l'espérance de sa distribution est égale à* θ

$$\mathbb{E}\left[\widehat{\theta}_n\right] = \theta$$

ou encore si son biais est nul.

Remarque 5.5. L'absence de biais n'est toutefois pas un critère suffisant pour discriminer des estimateurs alternatifs. Pour un même problème, on peut facilement trouver plusieurs estimateurs sans biais.

Un estimateur peut être biaisé, mais son biais peut diminuer avec la taille de l'échantillon, voire même tendre vers 0 : on parle alors d'estimateur asymptotiquement sans biais.

Définition 5.8. *Un estimateur* $\widehat{\theta}_n$ *d'un paramètre* θ *est dit* asymptotiquement sans biais *si son biais temps vers 0 quand la taille de l'échantillon tend vers l'infini*

$$\lim_{n\to+\infty} \text{biais}\left(\widehat{\theta}_n\right) = 0.$$

• **Précision d'un estimateur**

Définition 5.9 (Erreur quadratique moyenne)**.** *L'*erreur quadratique moyenne (EQM) *(Mean Squared Error (MSE) en aglais) d'un estimateur $\widehat{\theta}_n$ d'un paramètre inconnu θ est définie par*

$$\begin{aligned}\mathrm{EQM}\left(\widehat{\theta}_n\right) &= \mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \\ &= \mathbb{E}\left[\left(\widehat{\theta}_n - \mathbb{E}\left[\widehat{\theta}_n\right]\right)^2\right] + \mathbb{E}\left[\widehat{\theta}_n - \theta\right]^2 \\ &= \underbrace{\mathbb{V}\left(\widehat{\theta}_n\right)}_{\substack{\textit{fluctuations}\\ \textit{aléatoires}\\ \textit{autour de } \mathbb{E}\left[\widehat{\theta}_n\right]}} + \underbrace{\mathbb{E}\left[\widehat{\theta}_n - \theta\right]^2}_{\substack{\textit{carré du biais}\\ \textit{de } \widehat{\theta}_n}} .\end{aligned}$$

Remarque 5.6. Ainsi, pour un *estimateur sans biais*, son erreur quadratique moyenne est égale à sa *variance*.

Comment comparer deux estimateurs non biaisés ? Cette comparaison se fait sur la base de leur variance.

Propriété 5.4 (Comparaison d'estimateurs sans biais)**.** *Soient deux estimateurs sans biais $\widehat{\theta}_n^1$ et $\widehat{\theta}_n^2$. L'estimateur $\widehat{\theta}_n^1$ domine l'estimateur $\widehat{\theta}_n^2$, i.e. $\widehat{\theta}_n^1 \geq \widehat{\theta}_n^2$, si*

$$\mathbb{V}\left(\widehat{\theta}_n^1\right) \leq \mathbb{V}\left(\widehat{\theta}_n^2\right).$$

Remarque 5.7. Seuls des estimateurs non biaisés peuvent être comparés sur la base de leur variance.

• **Estimateur convergent**

Soit un estimateur $\widehat{\theta}_n = g(X_1, \ldots, X_n)$ d'un paramètre (ou d'un vecteur de paramètres) θ associé à un n-échantillon $(X_1, \ldots, X_n)$. Notons θ_0 la vraie valeur du paramètre θ.

Définition 5.10. *Un estimateur $\widehat{\theta}_n$ est* convergent au sens fort *s'il converge presque sûrement vers la vraie valeur du paramètre*

$$\widehat{\theta}_n \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \theta_0.$$

Un estimateur $\widehat{\theta}_n$ est convergent au sens faible *s'il converge en probabilité vers la vraie valeur du paramètre*

$$\widehat{\theta}_n \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} \theta_0.$$

Remarque 5.8. Lorsqu'un estimateur est qualifié de convergent sans plus de précision (*consistent* en anglais), cela signifie qu'il est convergent au sens faible.

La convergence est une des propriétés les plus importantes pour un estimateur. Elle signifie que si l'on applique l'estimateur à un très grand échantillon, les estimations (*i.e.* les réalisations de $\widehat{\theta}_n$) seront extrêmement concentrées autour de la vraie valeur du paramètre.

Propriété 5.5 (Convergence au sens faible)**.** *Soit un estimateur $\widehat{\theta}_n$ d'un paramètre (ou d'un vecteur de paramètres) θ tel que*

$$\lim_{n\to+\infty} \mathbb{E}\left[\widehat{\theta}_n\right] = \theta_0, \quad \lim_{n\to+\infty} \mathbb{V}\left(\widehat{\theta}_n\right) = 0,$$

où θ_0 est la vraie valeur du paramètre, alors cet estimateur est convergent au sens faible

$$\widehat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta_0.$$

Ainsi un estimateur sans biais dont la variance tend vers 0 quand la taille de l'échantillon tend vers l'infini est convergent.

• **Normalité asymptotique**

Définition 5.11. *Un estimateur en est* asymptotiquement normalement distribué *dès lors que*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0;\Sigma).$$

5.2.1.d Exemples d'estimateurs

• **Moyenne empirique**

Soit $(X_1,\ldots,X_n)$ un n-échantillon de variables aléatoires i.i.d. telles que $\mathbb{E}[X_i] = \mu$ et $\mathbb{V}(X_i) = \sigma^2$ finies, où le paramètre μ est inconnu. On définit la *moyenne empirique* par

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

qui est un estimateur de l'espérance μ inconnue. Étudions les propriétés de cet estimateur.

▷ *Biais de $\overline{X}_n$.* Puisque les variables X_i sont i.i.d. avec $\mathbb{E}[X_i] = \mu$, on a

$$\mathbb{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{n\times\mu}{n} = \mu.$$

$\overline{X}_n$ est donc un estimateur *sans biais* de μ.

▷ *Précision de $\overline{X}_n$.* Nous devons calculer la variance de la moyenne empirique $\overline{X}_n$. Comme les variables X_i sont i.i.d. avec $\mathbb{V}(X_i) = \sigma^2$, on a

$$\mathbb{V}\left(\overline{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left(X_i\right) = \frac{n\times\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Comme $\overline{X}_n$ est sans biais, on a

$$\text{EQM}\left(\overline{X}_n\right) = \mathbb{V}\left(\overline{X}_n\right) = \frac{\sigma^2}{n}.$$

$\triangleright$ $\overline{X}_n$ *est-il convergent ?* Puisque $\overline{X}_n$ est sans biais et que

$$\lim_{n\to+\infty} \mathbb{V}\left(\overline{X}_n\right) = \lim_{n\to+\infty} \frac{\sigma^2}{n} = 0,$$

alors d'après la Propriété 5.5, l'estimateur $\overline{X}_n$ est *convergent au sens faible*

$$\overline{X}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \mu.$$

$\triangleright$ $\overline{X}_n$ *est-il asymptotiquement normal ?* Puisque les X_i sont i.i.d. et que $\mathbb{E}[X_i] = \mu$ et $\mathbb{V}(X_i) = \sigma^2$ sont finies avec $\sigma^2 \neq 0$, alors d'après le TCL, on a

$$\frac{\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]}{\sqrt{\mathbb{V}\left(\overline{X}_n\right)}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}(0;1).$$

Ainsi $\overline{X}_n$ est *asymptotiquement normalement distribué* puisque

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}\left(0;\sigma^2\right).$$

• Variance empirique

Supposons que les variables aléatoires $(X_1,\ldots,X_n)$ soient i.i.d. de même loi que X, avec $\mathbb{E}[X] = \mu$ et $\mathbb{V}(X) = \sigma^2$. Comme $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$, en utilisant la moyenne empirique comme estimateur de $\mathbb{E}[X]$, on obtient la définition de la *variance empirique* suivante

$$V_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2,$$

qui est un estimateur de la variance σ^2.

Cet estimateur est-il sans biais ? On a

$$\begin{aligned}
\mathbb{E}[V_n] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \overline{X}_n\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2 - 2\overline{X}_n X_i + \overline{X}_n^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\overline{X}_n\right] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\overline{X}_n^2\right]
\end{aligned}$$

En utilisant le fait que les variables X_i sont i.i.d. de moyenne μ et de variance σ^2, on a $\mathbb{E}[X_i^2] = \mathbb{V}(X_i) + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2$, $\mathbb{E}\left[\overline{X}_n^2\right] = \mathbb{V}\left(\overline{X}_n\right) + \mathbb{E}\left[\overline{X}_n\right]^2 = \frac{\sigma^2}{n} + \mu^2$ (d'après ce qui précède) et

$$\mathbb{E}\left[X_i\overline{X}_n\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[X_iX_j\right] = \begin{cases} \dfrac{1}{n}\mathbb{E}[X_i^2] & \text{si } j = i \\ \dfrac{1}{n}\mathbb{E}[X_i]\mathbb{E}[X_j] & \text{si } j \neq i \end{cases} = \frac{\sigma^2+\mu^2}{n} + (n-1)\frac{\mu^2}{n} = \frac{\sigma^2}{n} + \mu^2.$$

D'où

$$\mathbb{E}[V_n] = \left(\sigma^2 + \mu^2\right) - \frac{2}{n}\sum_{i=1}^{n}\left(\frac{\sigma^2}{n} + \mu^2\right) + \left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2\frac{n-1}{n}.$$

Ainsi V_n n'est pas un estimateur sans biais de la variance. Par contre, $\lim_{n\to+\infty}\mathbb{E}[V_n] = \sigma^2$, donc V_n est un estimateur asymptotiquement sans biais de la variance.

• Variance empirique corrigée

Pour construire un estimateur sans biais de la variance à partir de l'expression de la variance empirique V_n, on voit qu'il faut la corriger par un facteur $\frac{n}{n-1}$ (par linéarité de l'espérance). On obtient alors la *variance empirique corrigée*

$$S_n^2 = \frac{n}{n-1}V_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

Alors

$$\mathbb{E}[S_n^2] = \frac{n}{n-1}\mathbb{E}[V_n] = \sigma^2,$$

et donc S_n^2 est un estimateur sans biais de la variance.

• Comparaison de deux estimateurs de la moyenne

Soit un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires i.i.d. telles que $\mathbb{E}[X_i] = \mu$. On considère deux estimateurs $\widehat{\mu}_n^1$ et $\widehat{\mu}_n^2$ de l'espérance μ. Le premier estimateur correspond à la moyenne empirique et le deuxième estimateur n'est rien d'autre que la première variable de l'échantillon

$$\widehat{\mu}_n^1 = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i, \quad \widehat{\mu}_n^2 = X_1.$$

Ces deux estimateurs sont des estimateurs sans biais de μ. En effet

$$\mathbb{E}\left[\widehat{\mu}_n^1\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{n\times\mu}{n} = \mu,$$

$$\mathbb{E}\left[\widehat{\mu}_n^2\right] = \mathbb{E}[X_1] = \mu.$$

puisque les variables X_i sont i.i.d. avec $\mathbb{E}[X_i] = \mu$.

Par ailleurs

$$\mathbb{V}\left(\widehat{\mu}_n^1\right) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{n\times\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

car les X_i sont indépendants et de même variance, et

$$\mathbb{V}\left(\widehat{\mu}_n^2\right) = \mathbb{V}(X_1) = \sigma^2.$$

On obtient $\mathbb{V}\left(\widehat{\mu}_n^1\right) \leq \mathbb{V}\left(\widehat{\mu}_n^2\right)$ dès lors que la taille d'échantillon n est supérieure ou égale à un. L'estimateur $\widehat{\mu}_n^1$ est préféré à $\widehat{\mu}_n^2$.

5.3 Méthode du maximum de vraisemblance

La procédure du maximum de vraisemblance est une méthode d'estimation ponctuelle dans laquelle on construit la probabilité jointe (cas discret) ou la densité jointe (cas continu) associée aux observations, appelée *vraisemblance de l'échantillon*. La vraisemblance est une fonction des observations et des paramètres inconnus de la distribution : elle mesure la plausibilité des données observées conditionnellement à une hypothèse de distribution sur la variable d'intérêt et à une valeur des paramètres. Le principe du maximum de vraisemblance consiste alors à déterminer la valeur des paramètres qui rend l'échantillon observé le plus vraisemblable. Dit autrement, la forme de l'estimateur du maximum de vraisemblance est déterminée par la *maximisation de la vraisemblance de l'échantillon*.

Cette méthode d'estimation est sans doute la plus utilisée en statistique et en économétrie. Les paramètres de la plupart des modèles non-linéaires considérés de nos jours en marketing, en finance, en gestion des risques (scoring bancaire), en assurance, etc., sont estimés par maximum de vraisemblance. Une des raisons de ce succès est que, sous des hypothèses relativement générales dites *hypothèses de régularité*, l'estimateur du maximum de vraisemblance présente de très bonnes propriétés (sans biais, efficace, convergent et asymptotiquement normalement distribué). C'est pourquoi cette méthode d'estimation est aujourd'hui disponible dans tous les logiciels d'économétrie et dans certains tableurs.

5.3.1 Fonction de vraisemblance

Soit X une variable aléatoire (discrète ou continue) définie sur un univers probabilisé $(X(\Omega), \mathcal{F}, \mathbb{P})$, dont la loi de probabilité est caractérisée par une fonction de densité ou une fonction de masse notée $f_X(x;\theta)$, $x \in X(\Omega)$. Cette fonction dépend d'un paramètre inconnu, noté θ, avec $\theta \in \Theta \subset \mathbb{R}$, où Θ désigne l'ensemble des valeurs possibles pour ce paramètre. Afin d'estimer θ, on dispose d'un n-échantillon $(X_1, \ldots, X_n)$ de variables i.i.d. de même loi que X. La réalisation de cet échantillon (observations) est notée $(x_1, \ldots, x_n)$ ou x en abrégé.

Nous pouvons alors déterminer la vraisemblance de l'échantillon, définie par la densité ou la probabilité jointe associée aux réalisations de l'échantillon. Si les variables $X_1, \ldots, X_n$ sont indépendantes, cette densité ou cette probabilité jointe peut s'écrire comme le produit des densités ou des probabilités marginales.

Définition 5.12. *La* fonction de vraisemblance *des réalisations de l'échantillon* $(x_1, \ldots, x_n)$ *est définie par*

$$\begin{array}{rcl} L : X(\Omega)^n \times \Theta & \to & \mathbb{R}_+ \\ (x_1, \ldots, x_n; \theta) & \mapsto & L(x_1, \ldots, x_n; \theta) = \displaystyle\prod_{i=1}^{n} f_X(x_i; \theta) \end{array}$$

La vraisemblance étant définie comme un produit de fonctions de densité ou de probabilités (fonction de masse), cette quantité est nécessairement positive.

Définition 5.13. *La* fonction de log-vraisemblance *des réalisations de l'échantillon* $(x_1, \ldots, x_n)$

est définie par

$$\begin{aligned} \ell : X(\Omega)^n \times \Theta &\to \mathbb{R} \\ (x_1, \ldots, x_n; \theta) &\mapsto \ell(x_1, \ldots, x_n; \theta) = \ln L(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \ln f_X(x_i; \theta) \end{aligned}$$

La fonction de log-vraisemblance, contrairement à la vraisemblance, peut être positive ou négative.

Exemple 5.10 (Loi exponentielle). On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires continues, positives et i.i.d. On suppose que les variables X_i admettent une distribution exponentielle $\mathcal{E}(1/\theta)$, où $\theta > 0$ est un paramètre inconnu. La fonction de densité des variables X_i est définie par

$$f_X(x_i; \theta) = \frac{1}{\theta} \exp\left(-\frac{x_i}{\theta}\right), \quad x_i \in \mathbb{R}_+.$$

Puisque les variables X_i sont indépendantes, la vraisemblance associée aux réalisations de l'échantillon $(x_1, \ldots, x_n)$ est définie par

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta) = \prod_{i=1}^{n} \frac{1}{\theta} \exp\left(-\frac{x_i}{\theta}\right) = \frac{1}{\theta^n} \exp\left(-\frac{1}{\theta} \sum_{i=1}^{n} x_i\right).$$

La log-vraisemblance de l'échantillon est définie par

$$\ell(x_1, \ldots, x_n; \theta) = \ln L(x_1, \ldots, x_n; \theta) = -n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^{n} x_i.$$

5.3.2 Estimateur du maximum de vraisemblance

Avant de définir l'estimateur du maximum de vraisemblance comme la quantité qui maximise la fonction de log-vraisemblance, il convient de s'assurer que le paramètre θ est identifiable à partir de cette fonction.

Définition 5.14. *Le paramètre θ est* identifiable *(ou estimable) pour l'échantillon $(x_1, \ldots, x_n)$, si pour toutes valeurs θ^* et θ telles que $\theta^* \neq \theta$, les lois jointes des variables $(x_1, \ldots, x_n)$ sont différentes.*

Tous les problèmes que nous considérerons dans ce cours sont identifiables. Sous cette hypothèse, on peut définir l'estimateur du maximum de vraisemblance comme suit.

Définition 5.15. *L'estimateur du maximum de vraisemblance $\widehat{\theta}_n^{MV}$ du paramètre $\theta \in \Theta$ est la solution du problème de maximisation suivant*

$$\widehat{\theta}_n^{MV}(x) = \arg\max_{\theta \in \Theta} \ell(x_1, \ldots, x_n; \theta) = \arg\max_{\theta \in \Theta} \ln L(x_1, \ldots, x_n; \theta).$$

De façon équivalente, on peut considérer le programme de maximisation de la vraisemblance $L(x_1, \ldots, x_n; \theta)$. Mais il est souvent plus simple de maximiser la log-vraisemblance que la vraisemblance d'un échantillon.

Remarque 5.9. La résolution de ce programme permet d'obtenir l'estimation $\widehat{\theta}_n^{MV}(x)$ associée aux données $x_1, \ldots, x_n$. De cette estimation, l'on déduit ensuite la forme fonctionnelle de l'estimateur $\widehat{\theta}_n^{MV}$ exprimée comme une fonction des variables aléatoires $X_1, \ldots, X_n$. Toutefois, afin de simplifier les notations, nous utiliserons $\widehat{\theta}_n$ à la place de $\widehat{\theta}_n(x)$ dans le programme d'optimisation et dans les conditions du premier et second ordres. Au-delà des notations, il convient de bien faire la différence entre les deux concepts.

La résolution du programme de maximisation de la log-vraisemblance, qui définit l'estimateur du maximum de vraisemblance, requiert de calculer la dérivée première et la dérivée seconde de cette fonction par rapport au paramètre θ. Ces dérivées correspondent respectivement au gradient et à la hessienne.

La condition nécessaire du programme de maximisation de la log-vraisemblance correspond à l'équation de log-vraisemblance.

Définition 5.16. *On appelle* équation de log-vraisemblance *l'équation associée à la* condition du premier ordre *(condition nécessaire) du programme de maximisation de la log-vraisemblance*

$$\text{C1} : \qquad \frac{\partial \ln L(x_1, \ldots, x_n; \theta)}{\partial \theta}\Big|_{\theta=\widehat{\theta}_n} = 0.$$

Ainsi, le gradient évalué au point $\widehat{\theta}_n$ (réalisation) doit être nul. La résolution de cette équation en $\widehat{\theta}_n$ permet d'obtenir l'estimation du maximum de vraisemblance en fonction des réalisations de l'échantillon (données) $x_1, \ldots, x_n$. De cette forme fonctionnelle, on déduira ensuite l'estimateur du maximum de vraisemblance. Mais avant cela, il convient de s'assurer que la solution $\widehat{\theta}_n$ est un maximum en vérifiant la condition suffisante du programme de maximisation.

Définition 5.17. *La* condition du second ordre *(condition suffisante) du programme de maximisation de la log-vraisemblance consiste à vérifier que la hessienne évaluée au point $\widehat{\theta}_n$ est négative*

$$\text{C2} : \qquad \frac{\partial^2 \ln L(x_1, \ldots, x_n; \theta)}{\partial \theta^2}\Big|_{\theta=\widehat{\theta}_n} < 0.$$

Exemple 5.11 (Loi exponentielle). On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires continues, positives et i.i.d. de distribution exponentielle $\mathcal{E}(1/\theta)$, où $\theta > 0$ est un paramètre inconnu. On a vu que la log-vraisemblance de l'échantillon est définie par

$$\ln L(x_1, \ldots, x_n; \theta) = -n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^{n} x_i.$$

Ainsi la condition du premier ordre du programme de maximisation de la log-vraisemblance s'écrit

$$\begin{aligned} \text{C1} : \qquad & \frac{\partial \ln L(x_1, \ldots, x_n; \theta)}{\partial \theta}\Big|_{\theta=\widehat{\theta}_n} = 0 \\ \iff \quad & -\frac{n}{\widehat{\theta}_n} + \frac{1}{\widehat{\theta}_n^2} \sum_{i=1}^{n} x_i = 0 \iff \widehat{\theta}_n(x) = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n \\ \iff \quad & \widehat{\theta}_n^{MV} = \overline{X}_n. \end{aligned}$$

On vérifie maintenant la condition du second ordre du programme de maximisation de la log-vraisemblance pour $\widehat{\theta}_n(x) = \bar{x}_n$.

$$\text{C2}: \quad \frac{\partial^2 \ln L(x_1, \ldots, x_n; \theta)}{\partial \theta^2}\Big|_{\theta=\bar{x}_n} = \frac{n}{\bar{x}_n^2} - \frac{2}{\bar{x}_n^3}\sum_{i=1}^n x_i = \frac{n}{\bar{x}_n^2} - \frac{2n\bar{x}_n}{\bar{x}_n^3} = -\frac{n}{\bar{x}_n^2} < 0$$

car $\bar{x}_n^2 > 0$. Nous avons donc bien un maximum. Par conséquent, l'estimateur du maximum de vraisemblance du paramètre θ correspond à la moyenne empirique.

Exemple 5.12 (**Loi de Poisson).** On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires entières, positives et i.i.d. de distribution de Poisson $\mathcal{P}(\lambda)$, où $\lambda > 0$ est un paramètre inconnu. On dispose de réalisations $(x_1, \ldots, x_n)$, où $x_i \in \mathbb{N}$, $i = 1, \ldots, n$. Alors la fonction de masse de la variable aléatoire X_i s'écrit

$$f_X(x_i; \lambda) = \mathbb{P}(X_i = x_i; \lambda) = \frac{\lambda^{x_i}}{x_i!}e^{-\lambda}, \quad x_i \in \mathbb{N}.$$

On en déduit la fonction de vraisemblance de l'échantillon

$$L(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^n \mathbb{P}(X_i = x_i; \lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}e^{-n\lambda},$$

et la log-vraisemblance de l'échantillon

$$\ln L(x_1, \ldots, x_n; \lambda) = -n\lambda + \ln\lambda \sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!).$$

Ainsi la condition du premier ordre du programme de maximisation de la log-vraisemblance s'écrit

$$\begin{aligned}
\text{C1}: \quad & \frac{\partial \ln L(x_1, \ldots, x_n; \lambda)}{\partial \lambda}\Big|_{\lambda=\widehat{\lambda}_n} = 0 \\
\iff \quad & -n + \frac{1}{\widehat{\lambda}_n}\sum_{i=1}^n x_i = 0 \iff \widehat{\lambda}_n(x) = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}_n \\
\iff \quad & \widehat{\lambda}_n^{MV} = \overline{X}_n.
\end{aligned}$$

On vérifie maintenant la condition du second ordre du programme de maximisation de la log-vraisemblance pour $\widehat{\lambda}_n(x) = \bar{x}_n$.

$$\text{C2}: \quad \frac{\partial^2 \ln L(x_1, \ldots, x_n; \lambda)}{\partial \lambda^2}\Big|_{\lambda=\bar{x}_n} = -\frac{1}{\bar{x}_n^2}\sum_{i=1}^n x_i < 0$$

car $\bar{x}_n^2 > 0$ et les x_i sont positifs. Nous avons donc bien un maximum. Par conséquent, l'estimateur du maximum de vraisemblance du paramètre λ correspond à la moyenne empirique.

Exemple 5.13 (**Loi normale).** $\triangleright$ On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires i.i.d. de distribution gaussienne $\mathcal{N}(\mu; \sigma^2)$, où $\mu \in \mathbb{R}$ est un paramètre inconnu et

$\sigma^2 > 0$ est connu. On dispose de réalisations $(x_1, \ldots, x_n)$, où $x_i \in \mathbb{R}$, $i = 1, \ldots, n$. Alors la densité de la variable aléatoire X_i s'écrit

$$f_X(x_i; \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right), \quad x_i \in \mathbb{R}.$$

On en déduit la fonction de vraisemblance de l'échantillon

$$\begin{aligned} L(x_1, \ldots, x_n; \mu) &= \prod_{i=1}^{n} f_X(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right), \end{aligned}$$

et la log-vraisemblance de l'échantillon

$$\ln L(x_1, \ldots, x_n; \mu) = -n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Ainsi la condition du premier ordre du programme de maximisation de la log-vraisemblance s'écrit

$$\begin{aligned} \text{C1} : \quad & \frac{\partial \ln L(x_1, \ldots, x_n; \mu)}{\partial \mu}\Big|_{\mu = \widehat{\mu}_n} = 0 \\ \iff & \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \widehat{\mu}_n) = 0 \iff \sum_{i=1}^{n} x_i = n\widehat{\mu}_n \iff \widehat{\mu}_n(x) = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n \\ \iff & \widehat{\mu}_n^{MV} = \overline{X}_n. \end{aligned}$$

On vérifie maintenant la condition du second ordre du programme de maximisation de la log-vraisemblance pour $\widehat{\mu}_n(x) = \bar{x}_n$.

$$\text{C2} : \quad \frac{\partial^2 \ln L(x_1, \ldots, x_n; \mu)}{\partial \mu^2}\Big|_{\mu = \bar{x}_n} = \frac{\partial}{\partial \mu} \left(\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)\right)\Big|_{\mu = \bar{x}_n} = -\frac{n}{\sigma^2} < 0.$$

Nous avons donc bien un maximum. Par conséquent, l'estimateur du maximum de vraisemblance du paramètre μ correspond à la moyenne empirique.

$\triangleright$ Considérons maintenant que μ est connu et σ^2 inconnu. La fonction de vraisemblance de l'échantillon s'écrit

$$L(x_1, \ldots, x_n; \sigma^2) = \prod_{i=1}^{n} f_X(x_i; \sigma^2) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right),$$

et la log-vraisemblance de l'échantillon s'écrit

$$\ln L(x_1, \ldots, x_n; \sigma^2) = -\frac{n}{2} \ln(\sigma^2) - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Ainsi la condition du premier ordre du programme de maximisation de la log-vraisemblance s'écrit

$$\begin{aligned}
\text{C1}: \quad & \frac{\partial \ln L(x_1,\dots,x_n;\sigma^2)}{\partial \sigma^2}\Big|_{\sigma^2=\widehat{\sigma}_n^2} = 0 \\
\iff & -\frac{n}{2\widehat{\sigma}_n^2} + \frac{1}{2(\widehat{\sigma}_n^2)^2}\sum_{i=1}^n (x_i-\mu)^2 = 0 \iff -n + \frac{1}{\widehat{\sigma}_n^2}\sum_{i=1}^n (x_i-\mu)^2 = 0 \\
\iff & \widehat{\sigma}_n^2(x) = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2 \iff \widehat{\sigma}_n^{2,MV} = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2.
\end{aligned}$$

On vérifie maintenant la condition du second ordre du programme de maximisation de la log-vraisemblance pour $\widehat{\sigma}_n^2$.

$$\begin{aligned}
\text{C2}: \quad \frac{\partial^2 \ln L(x_1,\dots,x_n;\sigma^2)}{\partial (\sigma^2)^2}\Big|_{\sigma^2=\widehat{\sigma}_n^2} &= \frac{n}{2(\widehat{\sigma}_n^2)^2} - \frac{1}{(\widehat{\sigma}_n^2)^3}\sum_{i=1}^n (x_i-\mu)^2 \\
&= \frac{n}{2(\widehat{\sigma}_n^2)^2} - \frac{n\widehat{\sigma}_n^2}{(\widehat{\sigma}_n^2)^3} = -\frac{n}{2(\widehat{\sigma}_n^2)^2} < 0.
\end{aligned}$$

Nous avons donc bien un maximum. Par conséquent, l'estimateur du maximum de vraisemblance du paramètre σ^2 est $\widehat{\sigma}_n^{2,MV} = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$.

Remarque 5.10. Nous verrons, sur l'exemple de la loi normale, comment on peut estimer simultanément les deux paramètres μ et σ^2 (quand ils sont tous deux inconnus) par la méthode du maximum de vraisemblance. La définition de l'estimateur du maximum de vraisemblance demeure inchangée, mais les conditions nécessaire et suffisante font alors intervenir respectivement le gradient et la hessienne.

5.3.3 Extension au cas avec plusieurs paramètres

Lorsque l'on considère un vecteur de paramètres $\theta = (\theta_1,\dots,\theta_d)' \in \Theta \subset \mathbb{R}^d$, la définition de l'estimateur du maximum de vraisemblance demeure inchangée. La seule différence est que, dans ce cas, la condition du premier ordre fait intervenir le gradient qui est un vecteur de dimension d et la condition du second ordre fait intervenir la hessienne qui est une matrice de dimension $d \times d$.

On a alors

$$\text{C1}: \quad \nabla_\theta \ell(x_1,\dots,x_n;\theta)\big|_{\theta=\widehat{\theta}_n} = 0_{\mathbb{R}^d} \iff \forall j = 1,\dots,d, \quad \frac{\partial \ln L(x_1,\dots,x_n;\theta)}{\partial \theta_j}\Big|_{\theta=\widehat{\theta}_n} = 0.$$

On rappelle que la matrice hessienne est une matrice symétrique de dimension $d \times d$ définie, pour $x = (x_1,\dots,x_n)$, par

$$H(x;\theta) = (H_{ij}(x;\theta))_{1\le i,j\le d}, \quad \text{où} \quad H_{ij}(x;\theta) = \frac{\partial^2 \ln L(x_1,\dots,x_n;\theta)}{\partial\theta_i \partial\theta_j}.$$

La condition du second ordre du programme de maximisation de la log-vraisemblance est alors la suivante

$$\text{C2}: \quad H(x;\widehat{\theta}_n) \text{ est définie négative.}$$

Rappelons qu'une matrice est définie négative lorsque toutes ses valeurs propres sont négatives. Considérons un exemple avec deux paramètres ($d = 2$).

Exemple 5.14 (Loi normale). On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires i.i.d. de distribution gaussienne $\mathcal{N}(\mu; \sigma^2)$, où $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ sont des paramètres inconnus. On dispose de réalisations $(x_1, \ldots, x_n)$, où $x_i \in \mathbb{R}$, $i = 1, \ldots, n$. On a déjà vu que la log-vraisemblance de l'échantillon s'écrit

$$\ln L(x_1, \ldots, x_n; \mu) = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Ainsi la condition du premier ordre du programme de maximisation de la log-vraisemblance s'écrit, pour $\theta = (\mu, \sigma^2)'$ et $\widehat{\theta}_n = (\widehat{\mu}_n, \widehat{\sigma}_n^2)'$,

$$\text{C1 :} \qquad \nabla_\theta \ln L(x_1, \ldots, x_n; \theta)\big|_{\theta=\widehat{\theta}_n} = 0_{\mathbb{R}^2} \iff \begin{cases} \dfrac{\partial \ln L(x_1, \ldots, x_n; \mu, \sigma^2)}{\partial \mu}\Big|_{\mu=\widehat{\mu}_n, \sigma^2=\widehat{\sigma}_n^2} = 0 \\ \dfrac{\partial \ln L(x_1, \ldots, x_n; \mu, \sigma^2)}{\partial \sigma^2}\Big|_{\mu=\widehat{\mu}_n, \sigma^2=\widehat{\sigma}_n^2} = 0 \end{cases}$$

$$\iff \begin{cases} \dfrac{1}{\widehat{\sigma}_n^2}\displaystyle\sum_{i=1}^{n}(x_i - \widehat{\mu}_n) = 0 \\ -\dfrac{n}{2\widehat{\sigma}_n^2} + \dfrac{1}{2(\widehat{\sigma}_n^2)^2}\displaystyle\sum_{i=1}^{n}(x_i - \widehat{\mu}_n)^2 = 0 \end{cases} \iff \begin{cases} \displaystyle\sum_{i=1}^{n} x_i = n\widehat{\mu}_n \\ -n + \dfrac{1}{\widehat{\sigma}_n^2}\displaystyle\sum_{i=1}^{n}(x_i - \widehat{\mu}_n)^2 = 0 \end{cases}$$

$$\iff \begin{cases} \widehat{\mu}_n(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = \bar{x}_n \\ \widehat{\sigma}_n^2(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \widehat{\mu}_n(x))^2 \end{cases} \iff \begin{cases} \widehat{\mu}_n^{MV} = \overline{X}_n \\ \widehat{\sigma}_n^{2,MV} = V_n \end{cases}.$$

On vérifie maintenant la condition du second ordre du programme de maximisation de la log-vraisemblance. Pour cela, on doit construire la matrice hessienne

$$H(x; \mu, \sigma^2) = \begin{pmatrix} \dfrac{\partial^2 \ln L(x_1, \ldots, x_n; \mu, \sigma^2)}{\partial \mu^2} & \dfrac{\partial^2 \ln L(x_1, \ldots, x_n; \mu, \sigma^2)}{\partial \mu \partial \sigma^2} \\ \dfrac{\partial^2 \ln L(x_1, \ldots, x_n; \mu, \sigma^2)}{\partial \sigma^2 \partial \mu} & \dfrac{\partial^2 \ln L(x_1, \ldots, x_n; \mu, \sigma^2)}{\partial (\sigma^2)^2} \end{pmatrix}$$

Ainsi au point $\widehat{\theta}_n = (\widehat{\mu}_n, \widehat{\sigma}_n^2)'$, on a

$$\begin{aligned} H(x; \widehat{\mu}_n, \widehat{\sigma}_n^2) &= \begin{pmatrix} -\dfrac{n}{\widehat{\sigma}_n^2} & -\dfrac{1}{(\widehat{\sigma}_n^2)^2}\displaystyle\sum_{i=1}^{n}(x_i - \widehat{\mu}_n) \\ -\dfrac{1}{(\widehat{\sigma}_n^2)^2}\displaystyle\sum_{i=1}^{n}(x_i - \widehat{\mu}_n) & \dfrac{n}{2(\widehat{\sigma}_n^2)^2} - \dfrac{1}{(\widehat{\sigma}_n^2)^3}\displaystyle\sum_{i=1}^{n}(x_i - \widehat{\mu}_n)^2 \end{pmatrix} \\ &= \begin{pmatrix} -\dfrac{n}{\widehat{\sigma}_n^2} & -\dfrac{1}{(\widehat{\sigma}_n^2)^2}(n\widehat{\mu}_n - n\widehat{\mu}_n) \\ -\dfrac{1}{(\widehat{\sigma}_n^2)^2}(n\widehat{\mu}_n - n\widehat{\mu}_n) & \dfrac{n}{2(\widehat{\sigma}_n^2)^2} - \dfrac{n\widehat{\sigma}_n^2}{(\widehat{\sigma}_n^2)^3} \end{pmatrix} = \begin{pmatrix} -\dfrac{n}{\widehat{\sigma}_n^2} & 0 \\ 0 & -\dfrac{n}{2(\widehat{\sigma}_n^2)^2} \end{pmatrix} \end{aligned}$$

Cette matrice diagonale est définie négative car les éléments de sa diagonale sont tous négatifs. Nous avons bien un maximum. Les estimateurs du maximum de vraisemblance des paramètres μ et σ^2 sont définis par

$$\widehat{\mu}_n^{MV} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\sigma}_n^{2,MV} = V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

5.3.4 Score, hessienne et quantité d'information de Fisher

Définition 5.18. *Le* score *de l'échantillon $(X_1, \ldots, X_n)$ est une variable aléatoire définie par*

$$S_n(X;\theta) = \frac{\partial \ln L(X_1, \ldots, X_n;\theta)}{\partial \theta} = \frac{1}{L(X_1, \ldots, X_n;\theta)} \frac{\partial L(X_1, \ldots, X_n;\theta)}{\partial \theta}.$$

Propriété 5.6. *Pour toute valeur du paramètre $\theta \in \Theta$, le score de l'échantillon vérifie*

$$\mathbb{E}\left[S_n(X;\theta)\right] = 0.$$

Exemple 5.15 (Loi exponentielle). On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires continues, positives et i.i.d. On suppose que les variables X_i admettent une distribution exponentielle $\mathcal{E}(1/\theta)$, où $\theta > 0$ est un paramètre inconnu. La fonction de densité des variables X_i est définie par

$$f_X(x_i;\theta) = \frac{1}{\theta}\exp\left(-\frac{x_i}{\theta}\right), \quad x_i \in \mathbb{R}_+.$$

La log-vraisemblance de l'échantillon est définie par

$$\ell(x_1, \ldots, x_n;\theta) = \ln L(x_1, \ldots, x_n;\theta) = -n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i.$$

Le score de l'échantillon $(X_1, \ldots, X_n)$ s'écrit alors

$$S_n(X;\theta) = \frac{\partial \ln L(X_1, \ldots, X_n;\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left(-n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} X_i\right) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i.$$

On vérifie que son espérance est nulle

$$\mathbb{E}\left[S_n(X;\theta)\right] = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}\mathbb{E}[X_i] = -\frac{n}{\theta} + \frac{n\theta}{\theta^2} = 0.$$

Définition 5.19. *La* quantité d'information de Fisher *associée à l'échantillon est une constante définie par la variance du score*

$$I_n(\theta) = \mathbb{V}\left(S_n(X;\theta)\right) = \mathbb{E}\left[S_n(X;\theta)^2\right],$$

car le score est centré.

Cette quantité mesure l'information apportée par l'échantillon sur le paramètre θ. Elle est croissante avec la taille de l'échantillon (plus la taille de l'échantillon est grande, plus l'information sur le paramètre est importante).

Propriété 5.7. *Si le domaine de définition des X_i ne dépend pas de θ et que la quantité d'information de Fisher existe, alors*

$$I_n(\theta) = \mathbb{E}\left[-\frac{\partial^2 \ln L(X_1, \ldots, X_n; \theta)}{\partial \theta^2}\right].$$

De plus, les quantités d'information de Fisher de chaque observation x_i, $i = 1, \ldots, n$, sont identiques et on a alors

$$I_n(\theta) = nI_1(\theta) = n\mathbb{E}\left[-\frac{\partial^2 \ln L(X_i; \theta)}{\partial \theta^2}\right],$$

où $I_1(\theta)$ est la quantité d'information de Fisher apportée par une observation.

Exemple 5.16 (Loi exponentielle). On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires continues, positives et i.i.d. On suppose que les variables X_i admettent une distribution exponentielle $\mathcal{E}(1/\theta)$, où $\theta > 0$ est un paramètre inconnu. On a vu que le score de l'échantillon $(X_1, \ldots, X_n)$ s'écrit

$$S_n(X; \theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i.$$

La quantité d'information de Fisher associée à l'échantillon est alors

$$I_n(\theta) = \mathbb{V}\left(S_n(X; \theta)\right) = \frac{1}{\theta^4}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{n\theta^2}{\theta^4} = \frac{n}{\theta^2},$$

puisque les X_i sont i.i.d. La quantité d'information de Fisher apportée par une observation est alors $I_1(\theta) = I_n(\theta)/n = \theta^{-2}$.

5.3.5 Estimateur efficace

Nous avons vu comment comparer deux estimateurs non baisés à l'aide de leurs variances. Mais existe-t-il un estimateur sans biais qui soit plus efficace que tous les autres ? C'est la notion d'*estimateur optimal*.

Définition 5.20. *Un* estimateur optimal *au sens du critère de la variance (ou de l'erreur quadratique) est l'estimateur sans biais qui possède la* variance la plus faible *parmi tous les estimateurs sans biais.*

Il est souvent difficile, voire impossible, de montrer qu'un estimateur est optimal. Une alternative consiste à montrer que la variance d'un estimateur atteint une certaine borne en deçà de laquelle les variances des estimateurs sans biais ne peuvent pas descendre. C'est le concept de borne de Cramer-Rao ou de borne FDCR (Frechet- Darmois - Cramer - Rao).

Il est important de noter que la borne FDCR ne peut être établie que sous un certain nombre d'hypothèses, c'est pourquoi le concept d'efficacité au sens FDCR est plus restrictif que le concept d'optimalité, même si l'idée est similaire.

Propriété 5.8 (Borne FDCR)**.** *Soit $(X_1, \ldots, X_n)$ un n-échantillon i.i.d. où X_i admet une fonction de densité (ou de masse) $f_X(x;\theta)$ dépendant d'un paramètre θ. Soit $\widehat{\theta}_n$ un estimateur sans biais de θ, i.e., $\mathbb{E}\left[\widehat{\theta}_n\right] = \theta$. Si la fonction $f_X(x;\theta)$ est régulière alors*

$$\mathbb{V}\left(\widehat{\theta}_n\right) \geq I_n^{-1}(\theta_0) = \textit{borne FDCR ou borne de Cramer-Rao},$$

où $I_n(\theta_0)$ correspond à la quantité d'information de Fisher associée à l'échantillon et évaluée en θ_0, vraie valeur du paramètre θ.

Définition 5.21. *Un estimateur est* efficace *au sens de la borne FDCR (Frechet - Darmois - Cramer - Rao) ou de la borne Cramer-Rao, si*

$$\mathbb{V}\left(\widehat{\theta}_n\right) = I_n^{-1}(\theta_0),$$

où $I_n(\theta_0)$ correspond à la quantité d'information de Fisher associée à l'échantillon et évaluée en θ_0, vraie valeur du paramètre θ.

On dit aussi qu'un estimateur efficace est BUE (*Best Unbiased Estimator*). Cela traduit le fait que c'est le meilleur des estimateurs sans biais en termes de variance.

5.3.6 Propriétés du maximum de vraisemblance

La question qui se pose à présent est de savoir si l'estimateur du maximum de vraisemblance est un « bon » estimateur (sans biais, efficace, convergent, asymptotiquement normal). Afin d'étudier ces propriétés, nous allons poser des hypothèses sur la distribution de la variable d'intérêt X, qualifiées d'*hypothèses de régularité*. Les voici :

- **Hypothèse 1 :** la fonction $\ln f_X(x_i;\theta)$ est trois fois différentiable par rapport à θ. Ses dérivées sont continues et finies pour toute valeur de x_i et de θ.
- **Hypothèse 2 :** Les espérances des dérivées première et seconde de $\ln f_X(X_i;\theta)$ par rapport à θ existent.
- **Hypothèse 3 :** la vraie valeur de θ, notée θ_0, appartient à un ensemble compacte.

Sous ces hypothèses de régularité, on peut montrer que l'estimateur du maximum de vraisemblance présente de bonnes propriétés.

1. L'estimateur du maximum de vraisemblance est *convergent* :

   $$\widehat{\theta}_n^{MV} \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} \theta_0,$$

   où θ_0 désigne la vraie valeur du paramètre θ.
2. L' estimateur du maximum de vraisemblance est *efficace* :

   $$\mathbb{V}(\widehat{\theta}_n^{MV}) = I_n(\theta_0)^{-1},$$

   où $I_n(\theta_0)$ désigne la quantité d'information de Fisher associée à l'échantillon et évaluée au point θ_0, vraie valeur du paramètre.

3. L'estimateur du maximum de vraisemblance est *asymptotiquement normalement distribué* :
$$\sqrt{n}\left(\widehat{\theta}_n^{MV} - \theta_0\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}\left(0; I_1(\theta_0)^{-1}\right),$$
où θ_0 désigne la vraie valeur du paramètre et $I_1(\theta_0) = I_n(\theta_0)/n$ correspond à la quantité d'information moyenne de Fisher évaluée au point θ_0.

Exemple 5.17 (Loi de Poisson). On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires entières, positives et i.i.d. de distribution de Poisson $\mathcal{P}(\theta)$, où $\theta > 0$ est un paramètre inconnu. Soit θ_0 la vraie valeur du paramètre. On a vu que $\widehat{\theta}_n^{MV} = \overline{X}_n$. Cet estimateur est sans biais et convergent puisque

$$\mathbb{E}[\widehat{\theta}_n^{MV}] = \mathbb{E}[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \theta_0, \quad \mathbb{V}(\widehat{\theta}_n^{MV}) = \mathbb{V}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{\theta_0}{n} \xrightarrow[n\to+\infty]{} 0.$$

Déterminons la borne FDCR. Le score et la dérivée seconde (stochastique) de la log-vraisemblance associés à l'échantillon sont respectivement définis par

$$S_n(X;\theta) = \frac{\partial \ln L(X;\theta)}{\partial\theta} = -n + \frac{1}{\theta}\sum_{i=1}^{n} X_i, \quad \frac{\partial^2 \ln L(X;\theta)}{\partial\theta^2} = -\frac{1}{\theta^2}\sum_{i=1}^{n} X_i.$$

La quantité d' information de Fisher associée à l'échantillon s'écrit

$$I_n(\theta_0) = \mathbb{E}\left[-\frac{\partial^2 \ln L(X;\theta)}{\partial\theta^2}\Big|_{\theta=\theta_0}\right] = \frac{1}{\theta_0^2}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{n\theta_0}{\theta_0^2} = \frac{n}{\theta_0}.$$

Ainsi $\mathbb{V}(\widehat{\theta}_n^{MV}) = I_n(\theta_0)^{-1}$ et, par conséquent, l'estimateur $\widehat{\theta}_n^{MV}$ est efficace.

On a vu par ailleurs que la moyenne empirique est un estimateur asymptotiquement normalement distribué du paramètre θ.

Exemple 5.18 (Loi exponentielle). On considère un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires continues, positives et i.i.d. On suppose que les variables X_i admettent une distribution exponentielle $\mathcal{E}(1/\theta)$, où $\theta > 0$ est un paramètre inconnu. Soit θ_0 la vraie valeur du paramètre. On a vu que $\widehat{\theta}_n^{MV} = \overline{X}_n$. Cet estimateur est sans biais

$$\mathbb{E}[\widehat{\theta}_n^{MV}] = \mathbb{E}[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \theta_0,$$

et de plus,

$$\mathbb{V}(\widehat{\theta}_n^{MV}) = \mathbb{V}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \frac{\theta_0}{n} \xrightarrow[n\to+\infty]{} 0.$$

Ainsi l'estimateur $\widehat{\theta}_n^{MV}$ est convergent.

Nous avons vu que dans ce cas, la quantité d'information de Fisher moyenne et la quantité d'information de Fisher associée à l'échantillon sont respectivement définies par

$$I_1(\theta_0) = \frac{1}{\theta_0^2}, \quad I_n(\theta_0) = \frac{n}{\theta_0^2}.$$

Puisque le problème est régulier, l'estimateur $\widehat{\theta}_n^{MV} = \overline{X}_n$ est asymptotiquement normalement distribué

$$\sqrt{n}\left(\widehat{\theta}_n^{MV} - \theta_0\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}\left(0; I_1(\theta_0)^{-1}\right),$$

soit dans notre cas

$$\sqrt{n}\left(\widehat{\theta}_n^{MV} - \theta_0\right) \xrightarrow[n\to+\infty]{\text{loi}} \mathcal{N}\left(0; \theta_0^2\right).$$