REMERCIEMENT

Avant tout, permettez nous de remercier nos professeurs qui nous ont soutenus le long de toutes nos études universitaires, en nous orientont et en nous motivant.

D'autre part, nous tenons à remercier notre professeur et encadreur qui nous a suivi le long de la prescription de notre mémoire en nous encourageant et en nous mettent dans la bonne voix.

Nous remercions également nos parents qui nous ont beaucoup aidé et nous ont procuré touts les moyens moraux et matériaux nous guidant à la réussite. Sans oublier nos collègues et toutes personnes nous ayant aidé de prés ou de loin dans nos recherches et dans l'élaboration de notre projet.

Dédicace

Je dédie ce modèste travail à tout les mombres de ma famille, en particulier à ma defunte mère qui m'a soutenue durant toute ma vie.

Bekki Fouzia

Table des matières

Introduction

La théorie de l'estimation est une des préoccupations majeures des statisticiens. Ainsi l'estimation non paramétrique a reçu un intérêt croissant tant sur le plan théorique que pratique. Cette branche de la statistique ne se résume pas à l'estimation d'un nombre fini de paramètres réels associés à la loi de l'échantillon (comme c'est le cas pour la théorie de l'estimation paramétrique), elle consiste généralement à estimer à partir des observations une fonction inconnue, élément d'une certaine classe fonctionnelle, telle que la fonction de répartition ou la fonction de densité à titre d'exemples.
L'objet central de ce mémoire est d'étudier l'estimation non paramétrique de ces dernières. Le lien évident qui existe entre fonction de répartition et densité, nous a conduits, à nous intéresser aux méthodes d'estimation de la densité qui peuvent être répertoriées dans deux classes principales : l'estimation par histogrammes et l'estimation par noyau qui peut être considérée comme une extension de l'estimateur par la méthode de l'histogramme.
Ce travail est divisé en trois chapitres, le premier chapitre englobe les différents outils statistiques qui seront utils dans la suite. Le deuxième chapitre est consacré à l'estimation non paramétrique de la fonction de répartition par la fonction de répartition empirique et l'estimateur à noyau. L'estimation de la densité par l'histogramme et l'estimateur à noyau fera l'objet du troisième chapitre.
Nous finissons par une conclusion générale pour ce travail de mémoire.
Nous considérerons dans tout le reste de ce travail que les variables sont $i.i.d.$ (indépendantes et identiquement distribuées).

Chapitre 1

Généralités

1.1 Estimateur et propriétés d'un estimateur

1.1.1 Estimateur et estimation

Définition 1.1.1 *Si $(X_1, \ldots, X_n)$ est un échantillon aléatoire d'effectif n de loi parente la loi de X, alors nous appelons estimateur du paramètre θ toute fonction h_n de l'échantillon aléatoire $(X_1, \ldots, X_n)$, notée $\widehat{\theta}_n$: $\widehat{\theta}_n = h_n(X_1, \ldots, X_n)$.*

Remarque 1.1.1

1. *À priori l'estimateur $\widehat{\theta}_n$ est à valeurs dans un ensemble Θ, contenant l'ensemble des valeurs possibles du paramètre θ.*
2. *$\widehat{\theta}_n$ est une variable aléatoire de loi de probabilité qui dépend du paramètre θ.*
3. *$\widehat{\theta}_n$ peut-être univarié ou multivarié.*

Définition 1.1.2 *Une fois l'échantillon prélevé, nous disposons de n valeurs observées $x_1, \ldots, x_n$, ce qui nous fournit une valeur $h_n(x_1, \ldots, x_n)$ qui est une réalisation de $\widehat{\theta}_n$ et que nous appelons estimation.*

Remarque 1.1.2

1. *Nous distinguons la variable aléatoire $\widehat{\theta}_n$ de sa valeur observé, notée $\widehat{\theta}_n(x_1, \ldots, x_n)$.*
2. *Nous utilisons les notations suivantes :*
   (i) $(X_1, \ldots, X_n)$ désigne l'échantillon aléatoire de taille n, et les n observations ne sont pas encore à disposition.

(ii) $(x_1, \ldots, x_n)$ *désigne une réalisation de l'échantillon aléatoire et les* n *observations sont à disposition.*

3. *Il faut systématiquement se demander : "suis-je entrain de manipuler une variable aléatoire ou l'une de ses réalisations ?"*

1.1.2 Propriétés d'un estimateur

Le chois d'un estimateur va reposer sur ses qualités, le premier défaut possible concerne la possibilité de comporter un biais.

Biais d'un estimateur

Définition 1.1.3 *Le biais de* $\widehat{\theta}_n$ *se définit par :* $B(\widehat{\theta}_n) = \mathbb{E}(\widehat{\theta}_n) - \theta$.

$\widehat{\theta}_n$ est un estimateur sans biais (ou non biaisé) du paramètre θ Si : $B(\widehat{\theta}_n) = 0$, c'est-à-dire Si $\mathbb{E}(\widehat{\theta}_n) = \theta$ (voir figure 1.1)

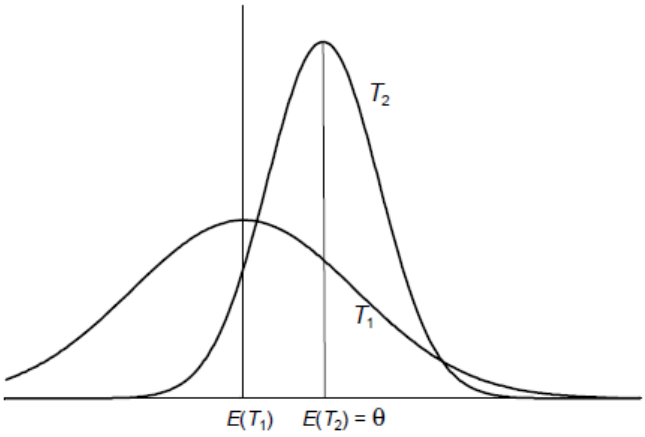

FIGURE 1.1 – Comparaison d'estimateur avec $\mathbb{E}(T_1) \neq \theta$ et $\mathbb{E}(T_2) = \theta$

Estimateur asymptotiquement sans biais

un estimateur $\widehat{\theta}_n$ est asymptotiquement sans biais pour θ Si $\lim\limits_{n \longrightarrow \infty} \mathbb{E}(\widehat{\theta}_n) = \theta$.

Écart quadratique moyen

Si $\widehat{\theta_n}$ est un estimateur de θ, nous mesurons la précision de $\widehat{\theta_n}$ par l'écart quadratique moyen $\mathbb{E}[(\widehat{\theta_n} - \theta)^2]$, noté EQM :

$$\begin{aligned} EQM &= \mathbb{E}[(\widehat{\theta_n} - \mathbb{E}(\widehat{\theta_n}) + \mathbb{E}(\widehat{\theta_n}) - \theta)^2] \\ &= \mathbb{E}[(\widehat{\theta_n} - \mathbb{E}(\widehat{\theta_n}))^2] + 2\mathbb{E}[(\widehat{\theta_n} - \mathbb{E}(\widehat{\theta_n}))(\mathbb{E}(\widehat{\theta_n}) - \theta)] + \mathbb{E}[(\mathbb{E}(\widehat{\theta_n}) - \theta)^2] \end{aligned}$$

Comme $\mathbb{E}(\widehat{\theta_n}) - \theta$ est une constante et que $\mathbb{E}[\widehat{\theta_n} - \mathbb{E}(\widehat{\theta_n})] = 0$, il vient

$$EQM = Var(\widehat{\theta_n}) + [\mathbb{E}(\widehat{\theta_n}) - \theta]^2$$

Donc, l'absence de biais n'est pas une garantie absolue de "bon estimateur". Il faut aussi tenir compte de sa variance ; par suit, deux estimateurs sans biais, le plus précis est donc celui de variance minimale (voir figure 1.2).

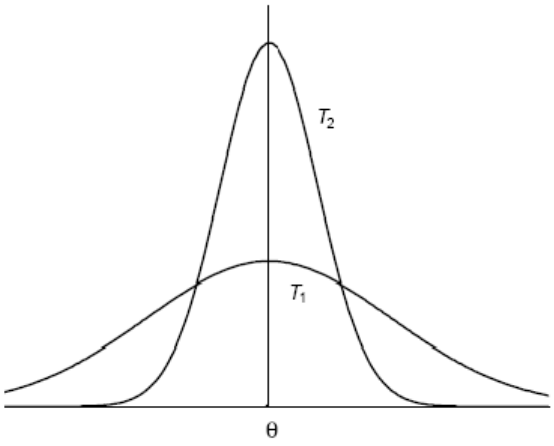

FIGURE 1.2 – Comparaison d'estimateur avec $\mathbb{E}(T_1) = \mathbb{E}(T_2)$ et $Var(T_1) > Var(T_2)$

Estimateur convergent

Définition 1.1.4 *un estimateur $\widehat{\theta}_n$ est convergent s'il converge en probabilité vers θ quand $n \longrightarrow \infty$.*

Propriété : Si un estimateur est sans biais et que sa variance tend vers 0 quand n tend vers l'infini, alors cet estimateur est convergent.

Remarque 1.1.3 *deux estimateurs convergents peuvent ne pas converger à la même vitesse.*

1.2 Lois des grands nombres

Théorème 1.2.1 *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires deux à deux indépendantes, de même loi ayant un moment d'ordre 2. Alors :*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mathbb{E}(X_1)$$

Théorème 1.2.2 *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes, de même loi dans L^1. Alors :*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s} \mathbb{E}(X_1)$$

1.3 Convergence presque complète

Notation

les petits "o" et les grands "O", que l'on trouve couramment dans la littérature sont rappelés de manière précise ici. Ces symboles ont été introduit par *Landau* pour simplifier les relations entre quantités (Stochastique ou non) de même ordre de grandeur, ou d'un ordre de grandeur inférieure asymptotiquement. Plus précisément, des relations liant les ordres de grandeur de quantités stochastiques sont expimés par les célèbres petit "o_p" et grands "O_p" définis comme suit :

– Si a_n est une suite de variables aléatoires et g est une fonction réelle de la variable entière n, alors la notation $a_n = o_p(g(n))$ signifie que :

$$\mathbb{P}(\lim_{n\to\infty} \frac{a_n}{f(n)}) = 0$$

– De manière similaire, la notation $a_n = O_p(g(n))$ signifie qu'il existe une constante $K > 0$ telle que : $\forall \varepsilon > 0 \quad \exists N_\varepsilon \in \mathbb{N}^*$ tel que :

$$\forall n > N_\varepsilon \qquad \mathbb{P}(|\frac{a_n}{g(n)}| > K) < \varepsilon$$

Définition 1.3.1 *On dit que la suite de variables aléatoires réelles $(X_n)_{n\in\mathbb{N}}$ converge presque complètement vers une variable aléatoire X lorsque $n \longrightarrow +\infty$ (et on note $\lim_{n\to+\infty} X_n = X \quad p.co$), si et seulement si :*

$$\forall \varepsilon > 0, \quad \sum_{n=0}^{\infty} \mathbb{P}[|X_n - X| > \varepsilon] < \infty$$

Définition 1.3.2 *On dit que la vitesse de convergence presque complète de la suite de variables aléatoire réelles $(X_n)_{n\in\mathbb{N}}$ vers X est d'ordre (U_n) ((U_n) étant une suite numérique déterministe), et on note $X_n = O_{p.co}(U_n)$, si et seulement si :*

$$\forall \varepsilon_0 > 0, \quad \sum_{n=0}^{\infty} \mathbb{P}[|X_n - X| > \varepsilon_0 U_n] < \infty$$

1.4 Statistique non paramétrique

Qu'est ce que la statistique non parametrique ?

La statistique paramétrique est le cadre classique de la statistique. Le modèle statistique est décrit par un nombre fini de paramètres.Typiquement $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \mathbb{R}^p\}$ est le modèle statistique qui d'écrit la distribution des variables observées.

Exemple

– $\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^{+*}\}$, modèle Gaussien.
– $\mathcal{M} = \{f(x, \theta) = h(x)\exp(\eta(\theta)T(x) - A(\theta)), \theta \in \mathbb{R}^p\}$ modèle des familles exponentielles.
– $\mathcal{M} = \{\Gamma(\alpha, \beta); (\alpha, \beta) \in \mathbb{R}^{+*}\}$ modèle loi Gamma.

Par opposition, en statistique non paramétrique, le modèle n'est pas d'écrit par un nombre fini de paramètres. Divers cas de figure peuvent se présenter, comme par exemple

– On s'autorise toutes les distributions possibles on ne fait aucune hypothèse sur la forme, la nature ou le type de la distribution des variables aléatoires.
– Le nombre de paramètres du modèle n'est pas fixé, et varie (augmente) avec le nombre d'observations.

Lemme 1.4.1 *(Inégalité de Brenstein-Frechet)*
Soit $X_1, \ldots, X_n$ une suite des variables aléatoires réelles indépendantes définies sur l'espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$
Si $\alpha_i \leq X_i \leq \beta_i, \quad \forall i \leq n$ où les α_i et les β_i sont des constantes réelles, alors :

$$(\forall t > 0) \quad \mathbb{P}(|\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))| \geq t) \leq 2\exp(\frac{-2t^2}{\sum_{i=1}^{n}(\beta_i - \alpha_i)^2})$$

Lemme 1.4.2 *(Inégalité de Hoeffding)*
Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes centrés de même loi telle que : $|X_1| \leq C_1, \quad \mathbb{E}(X_1^2) \leq C_2$ alors :

$$\mathbb{P}(\frac{1}{n}|\sum_{i=1}^{n} X_i| \geq \varepsilon) \leq 2\exp(\frac{-n\varepsilon^2}{4C_2}) \quad \forall \varepsilon \in]0, \frac{C_1}{C_2}[$$

Chapitre 2

Estimation non paramétrique de la fonction de répartition

Un problème récurrent en statistique est celui de l'estimation d'une fonction de répartition F à partir d'un échantillon de variables aléatoires réelles $X_1, X_2, ..., X_n$ indépendantes et de même loi inconnue. L'objectif de ce chapitre et l'estimation non paramétrique de la fonction de répartition. Pour ce but, ce chapitre est divisé en deux sections. La première section est consacrée à l'estimation par la fonction de répartition empirique. L'estimateur à noyau fera l'objet de la deuxième section.

2.1 La fonction de répartition empirique

Soit $X_1, X_2, \ldots, X_n$ un échantillon *i.i.d.* (indépendantes et identiquement distribuées) de fonction de répartition $F : x \longrightarrow F(x) = \mathbb{P}(X_1 \leq x)$
et $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ les observations ordonnées.
Supposons que F soit complètement inconnue. Comment estimer F, en se basant sur les observations $X_1, \ldots, X_n$?
Un bon estimateur pour F est la fonction de répartition empirique, notée $\widehat{F}_n$ définie par :

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$$

d'une manière plus précise, pour un n-échantillon $X_1, X_2, \ldots, X_n$ de X.

$$\forall i \in \{1, 2, \ldots, n\} \qquad \begin{array}{rcl} X_i : (\Omega, \mathcal{A}, \mathbb{P}) & \longrightarrow & (\mathbb{R}, \mathbb{B}_R) \\ \omega & \longrightarrow & X_i(\omega) = x_i \end{array}$$

$$\forall \omega \in \Omega \quad \text{on pose} \quad x_{(1)} = X_1(\omega) \leq x_{(2)} = X_2(\omega) \leq \ldots \leq x_{(n)} = X_n(\omega)$$

$$\widehat{F}_n(x) = \begin{cases} 0 & \text{si} \quad x < x_{(1)} \\ \frac{i}{n} & \text{si} \quad x_{(i)} \leq x < x_{(i+1)} \quad i \in \{1, \ldots, n-1\} \\ 1 & \text{si} \quad x \geq x_{(n)} \end{cases} \tag{2.1}$$

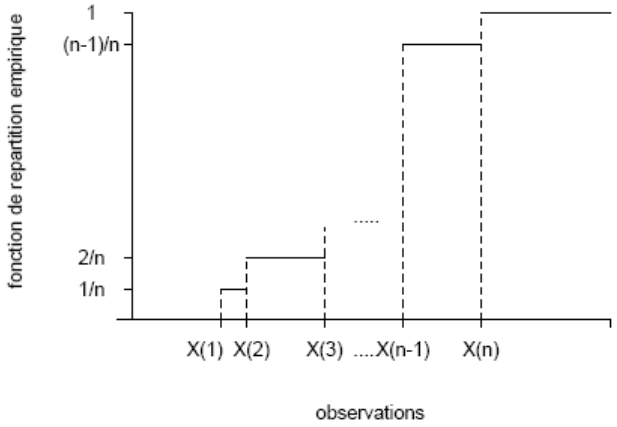

FIGURE 2.1 – Fonction de répartition empirique.

2.1.1 Propriétés

Proposition 2.1.1 *$\widehat{F}_n(x)$ est un estimateur sans biais pour $F(x)$, et il converge en moyenne quadratique.*

Preuve
L'objectif est de vérifier que :

$$\mathbb{E}[\widehat{F}_n(x)] = F(x) \tag{2.2}$$

et

$$\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] \longrightarrow 0 \quad \text{quand} \quad n \to \infty \tag{2.3}$$

Pour démontrer (2.2), on peux remarquer que d'après (2.1), $\widehat{F}_n(x)$ est une variable aléatoire dont les valeurs sont dans $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$, ce qui implique que la variable aléatoire $n\widehat{F}_n(x)$ est a valeurs dans $\{0, 1, \ldots, n\}$, alors $n\widehat{F}_n(x)$ est une variable aléatoire Binomiale de paramètres $(n, F(x))$.
On en déduit que

$$\begin{array}{rcl} \mathbb{E}[n\widehat{F}_n(x)] & = & nF(x) \\ n\mathbb{E}[\widehat{F}_n(x)] & = & nF(x) \\ \mathbb{E}[\widehat{F}_n(x)] & = & F(x) \end{array}$$

ce qui montre que $\widehat{F}_n(x)$ est un estimateur sans biais.
Il nous reste maintenant de vérifier (2.3)
En effet

$$\begin{array}{rcl} Var(\widehat{F}_n(x)) & = & Var(\widehat{F}_n(x) - F(x)) \\ & = & \mathbb{E}[(\widehat{F}_n(x) - F(x))^2] - [\mathbb{E}(\widehat{F}_n(x) - F(x))]^2 \\ & = & \mathbb{E}[(\widehat{F}_n(x) - F(x))^2] \,(\text{car} \quad \widehat{F}_n(x) \,\text{est un estimateur sans biais}) \end{array}$$

Donc pour montrer que $\widehat{F}_n(x)$ converge en moyenne quadratique vers $F(x)$, il suffit de montrer que $\lim\limits_{n\to\infty} Var(\widehat{F}_n(x)) = 0$

En effet, comme la variable aléatoire $n\widehat{F}_n(x) \rightsquigarrow \mathcal{B}(n, F(x),$

$$Var(n\widehat{F}_n(x)) = nF(x)(1 - F(x))$$

donc

$$\begin{array}{l} n^2 Var(\widehat{F}_n(x)) = nF(x)(1 - F(x)) \\ Var(\widehat{F}_n(x)) = \frac{1}{n}F(x)(1 - F(x)) \longrightarrow 0 \text{ quand } n \to \infty \end{array}$$

■

Remarque

La loi forte des grands nombres nous montre que $\widehat{F}_n(x)$ est un estimateur fortement consistant c'est-à-dire

$$\forall x \in \mathbb{R} \qquad \widehat{F}_n(x) \xrightarrow{p.s} F(x)$$

Théorème 2.1.1 *Soit $X_1, X_2, \ldots, X_n$ un n-échantillon de X de fonction de répartition F, alors*

$$\widehat{F}_n(x) - F(x) = O(\sqrt{\frac{\log n}{n}}) \qquad p.co.$$

Preuve
Par définition, il suffit de montrer que :

$$\exists \varepsilon > 0, \quad \sum_{n=0}^{\infty} \mathbb{P}[|\widehat{F}_n(x) - F(x)| > \varepsilon \sqrt{\frac{\log n}{n}}] < \infty$$

Soit $\varepsilon > 0$, alors d'après (2.2), ona

$$\mathbb{P}[|\widehat{F}_n(x) - F(x)| > \varepsilon \sqrt{\frac{\log n}{n}}] = \mathbb{P}[|\widehat{F}_n(x) - \mathbb{E}(F(x))| > \varepsilon \sqrt{\frac{\log n}{n}}]$$

Par définition de l'estimateur de $\widehat{F}_n(x)$ on a

$$\begin{aligned}
&\mathbb{P}[|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x]}(X_i) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\mathbb{1}_{]-\infty,x]}(X_i))| > \varepsilon \sqrt{\frac{\log n}{n}}] = \\
&\mathbb{P}[[\frac{1}{n}\sum_{i=1}^{n} |\mathbb{1}_{]-\infty,x]}(X_i) - \mathbb{E}(\mathbb{1}_{]-\infty,x]}(X_i))| > \varepsilon \sqrt{\frac{\log n}{n}}] = \\
&\mathbb{P}[\sum_{i=1}^{n} |\mathbb{1}_{]-\infty,x]}(X_i) - \mathbb{P}(\mathbb{1}_{]-\infty,x]}(X_i))| > \varepsilon \sqrt{n \log n}]
\end{aligned}$$

En appliquant l'inégalité de *Brrenstein-Frechet* avec $\alpha_i = 0$, $\beta_i = 1$, $\quad X_i = \mathbb{1}_{]-\infty,x]}(X_i)$ et $t = \varepsilon\sqrt{n \log n}$
il vient

$$\begin{aligned}
\mathbb{P}[|\widehat{F}_n(x) - F(x)| > \varepsilon \sqrt{\tfrac{\log n}{n}}] &\leq 2\exp(\tfrac{-2\varepsilon^2 n \log n}{n}) \\
&\leq 2\exp(\log n^{-2\varepsilon^2}) = 2n^{-2\varepsilon^2}
\end{aligned}$$

Il suffit de prendre $2\varepsilon^2 > 1$ ∎

2.2 L'estimateur à noyau

Définition 2.2.1 *Soit $X_1, \ldots, X_n$ un n-échantillon de X de fonction de répartition F, l'estimateur*

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} H(\frac{x - X_i}{h_n})$$

s'appelle estimateur à noyau H pour la fonction de répartition.

La fonction H est une fonction de répartition et $(h_n)_n$ est une suite de nombres réels positifs tels que $h_n \longrightarrow 0 \quad \text{quand}\, n \longrightarrow 0$.

2.2.1 Propriétés

Proposition 2.2.1 *Soit $X_1, \ldots, X_n$ un n-échantillon de X de fonction de répartition F et $\widehat{F}_n$ un estimateur à noyau H pour la fonction F tel que : $\int yH'(y)dy < \infty$.*
Alors : $\widehat{F}_n$ est asymptotiquement sans biais.

Preuve
L'objectif est de vérifier que $\mathbb{E}(\widehat{F}_n(x)) \longrightarrow F(x)$.

En effet : $\mathbb{E}[\widehat{F}_n] = \frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} H(\frac{x - X_i}{h_n})] = \mathbb{E}(H(\frac{x - X}{h_n}))$ (car $X_1, \ldots, X_n$ est un échantillon de X).

D'où $\mathbb{E}(\widehat{F}_n(x)) = \int_{-\infty}^{+\infty} H(\frac{x - z}{h_n}) f_X(z)dz$

Une intégration par parties, nous donne :

$$\begin{aligned}\mathbb{E}(\widehat{F}_n(x)) &= [H(\frac{x - z}{h_n})F_X(z)]_{-\infty}^{+\infty} + \frac{1}{n}\int_{-\infty}^{+\infty} H'(\frac{x - z}{h_n})F_X(z)dz \\ &= 0 + \frac{1}{n}\int_{-\infty}^{+\infty} H'(\frac{x - z}{h_n})F_X(z)dz\end{aligned}$$

Le changement de variable $y = \dfrac{x - z}{h_n}$, nous conduit à

$$\mathbb{E}[\widehat{F}_n] = \int_{-\infty}^{+\infty} H'(y)F_X(x - h_{ny})dy$$

Maintenant on va faire un développement de Taylor de la fonction F_X au point x, et à l'ordre 1

$$F_X(x - h_n y) = F_X(x) - h_n y f_X(\xi) + O(h_n^2) \quad \text{avec} \quad \xi \in [x - h_n y, x]$$

alors :

$$\begin{aligned} \mathbb{E}[\widehat{F}_n(x)] &= \int_{-\infty}^{+\infty} H'(y)[F_X(x) - h_n y f_X(\xi)]dy \\ &= F_X(x)\int_{-\infty}^{+\infty} H'(y)dy - h_n \int_{-\infty}^{+\infty} yH'(y)f_X(x)dy + O(h_n^2) \end{aligned}$$

Donc

$$\begin{aligned} \mathbb{E}[\widehat{F}_n(x)] &= F_X(x) - h_n \int_{-\infty}^{+\infty} yH'(y)f_X(x)dy \\ &= F_X(x) - h_n f_X(x)\int_{-\infty}^{+\infty} yH'(y)dy + O(h_n^2) \end{aligned}$$

au passage à la limite quand $n \to \infty$, $\quad h_n \to 0 \quad \mathbb{E}(\widehat{F}_n(x)) \to F(x)$

D'ou $\widehat{F}_n$ est un estimateur asymptotiquement sans biais. ■

Proposition 2.2.2 *Sous les mêmes conditions de la proposition précédente, l'estimateur à noyau $\widehat{F}_n(x)$ converge en moyenne quadratique vers $F(x)$.*

Preuve

L'objectif est de vérifier que : $\lim_{n\to\infty} \mathbb{E}[\widehat{F}_n(x) - F(x)]^2 = 0$. On a :

$$\begin{aligned} Var(\widehat{F}_n(x) &= Var(\widehat{F}_n(x) - F(x))) \\ &= \mathbb{E}[(\widehat{F}_n(x) - F(x))^2] - (\mathbb{E}(\widehat{F}_n(x) - F(x)))^2 \end{aligned}$$

Il suffit donc de vérifier $Var(\widehat{F}_n(x)) \to 0 \quad$ quand $n \to \infty$

En effet :

$$\begin{aligned} Var(\widehat{F}_n(x)) &= Var(\frac{1}{n}\sum_{i=1}^{n} H(\frac{x - X_i}{h_n})) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(H(\frac{x - X_i}{h_n})) \\ &= \frac{1}{n}[\mathbb{E}((H(\frac{x - X}{h_n}))^2) - (\mathbb{E}(H(\frac{x - X}{h_n})))^2] \end{aligned}$$

Donc, on peux poser

$$Var(\widehat{F}_n(x)) = \frac{1}{n}[I_1 - I_2]$$

pour $I_2 = (\mathbb{E}(\widehat{F}_n(x)))^2 \to (F(x))^2$
De même :

$$\begin{aligned} I_1 = \mathbb{E}(H^2(\frac{x-X}{h_n})) &= \int_{-\infty}^{+\infty} H^2(\frac{x-z}{h_n}) f_X(x) dz \\ &= \left[H^2(\frac{x-z}{h_n}) F_X(x) \right]_{-\infty}^{+\infty} + \frac{1}{h_n} \int_{-\infty}^{+\infty} (H^2)'(\frac{x-z}{h_n}) F_X(z) dz \\ &= F(x) \int_{-\infty}^{+\infty} (H^2)'(y) dy \end{aligned}$$

En suivant les mêmes étapes que pour la proposition précédente, on arrive à

$$\lim_{n\to\infty} Var(\widehat{F}_n(x)) = 0$$

■

Chapitre 3

Estimation non paramétrique de la densité

La fonction de répartition empirique joue un rôle crucial dans l'étude de la loi parente d'un échantillon. Cependant elle ne permet pas d'obtenir des résultats très précis sur la structure de cette loi. En revanche, lorsque les variables de l'échantillon admettent une densité f celle-ci donne beaucoup plus d'informations sur la loi parente : dispersion, mode, etc... on peux dire aussi que f donne une information visuelle importante sur la répartition des valeurs.
Considérons une quantité aléatoire continue X qui possède une fonction de densité de probabilité f. La spécification de la fonction f donne une description naturelle de la distribution de X, et permet de retrouver les probabilités associées à X par l'équation suivante :

$$P(a < X < b) = \int_a^b f(x)dx \quad \text{pour tout } a < b$$

Supposons maintenant que nous ayons n observations $X_1, \ldots, X_n$ provenant d'une fonction de densité inconnue. L'estimation de densité, que nous considérons dans notre travail, est la construction, à partir des données observées, d'une fonction estimée qui représente (approche) cette fonction de densité.
Le problème évoqué dans notre travail est alors "Comment estimer une densité à partir d'un ensemble de données ?"

Les approches d'estimation de densité se divisent en deux catégories : les approches paramétriques et les approches non paramétriques.
Dans l'approche paramétrique, nous supposons que les données proviennent d'une famille

de distributions paramétriques connue. L'estimation de la densité est obtenue en estimant les paramètres de la distribution à partir des données et en substituant ces estimés dans la formule de densité pour cette distribution. Par exemple, supposons que les données proviennent d'une distribution normale de moyenne μ et de variance σ^2. L'estimation de la densité f relative à ces observations est obtenue en trouvant les estimés de μ et de σ^2 à partir de ces données et en substituant ces derniers dans la formule de densité pour une distribution normale.
L'approche non paramétrique prend son sens lorsqu'on ne possède aucune information précise sur la forme et la classe de la vraie densité. Dans cette approche, ce sont les observations qui vont nous permettre de déterminer l'estimation de la densité f. Cette approche sera utilisée puisque l'on ne peut supposer aucune forme de fonction préspécifiée pour f.

3.1 L'histogramme

L'estimateur non paramétrique de la densité le plus populaire est l'histogramme introduit par John Graunt au $XVII^{\text{ème}}$ siècle et est définit comme suit :
Supposons que l'on ait $x_1, \ldots, x_n$ observations issus d'une même loi de probabilité de densité f, où f est à support borné $[a, b[$; pour estimer f par la méthode de l'histogramme ce qui revient à approcher f par une fonction étagée, on découpe $[a, b[$ en k classes $[\alpha_i, \alpha_{i+1}[$, où $i = 1, \ldots, k$ avec $a = \alpha_1$ et $b = \alpha_{k+1}$.
l'estimateur histogramme s'écrit alors, $\forall t \in [a, b[$, $\exists i = 1, \ldots, k$ tel que $t \in [\alpha_i, \alpha_{i+1}[$ et

$$\widehat{f}_n(t) = \frac{f_i}{\alpha_{i+1} - \alpha_i}$$

où f_i est la fréquence du nombre de points de la classe correspondante. Ce que l'on peut encore écrire plus précisément : $\forall t \in [a, b[$

$$\widehat{f}_n(t) = \sum_{i=1}^{k} \frac{f_i}{\alpha_{i+1} - \alpha_i} \mathbb{1}_{[\alpha_i, \alpha_{i+1}[}(t)$$

où

$$f_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{[\alpha_i, \alpha_{i+1}[}(x_j)$$

donc

$$\widehat{f}_n(t) = \sum_{i=1}^{k} \frac{1}{n(\alpha_{i+1} - \alpha_i)} \sum_{j=1}^{n} \mathbb{1}_{[\alpha_i, \alpha_{i+1}[}(x_j)$$

La figure (3.1) montre quatre histogrammes basés sur le même ensemble de données. Ces données représentent le poids à la naissance de cinquante enfants ayant un syndrome respiratoire idiopathique sévère. Les deux premiers histogrammes sont basés sur une largeur de classe petite et grande ($h = 0,2$ et $h = 0,8$) respectivement. Les deux autres histogrammes sont basés sur la même largeur de classe ($h = 0.4$) mais avec les intervalles placés de façons différentes. On remarque que chacun des histogrammes donne une impression différente sur la forme de la densité. Une largeur de classe trop petite conduit à un histogramme plus découpé, tandis que d'une largeur de classe trop grande résulte un histogramme plus lissé comme le montrent les figures (3.1 (a) et (b)). Les figures (3.1 (c) et (d)) montrent que le placement des intervalles a aussi un impact sur la forme de la densité suggérée, puisque ces histogrammes sont très différents l'un de l'autre.

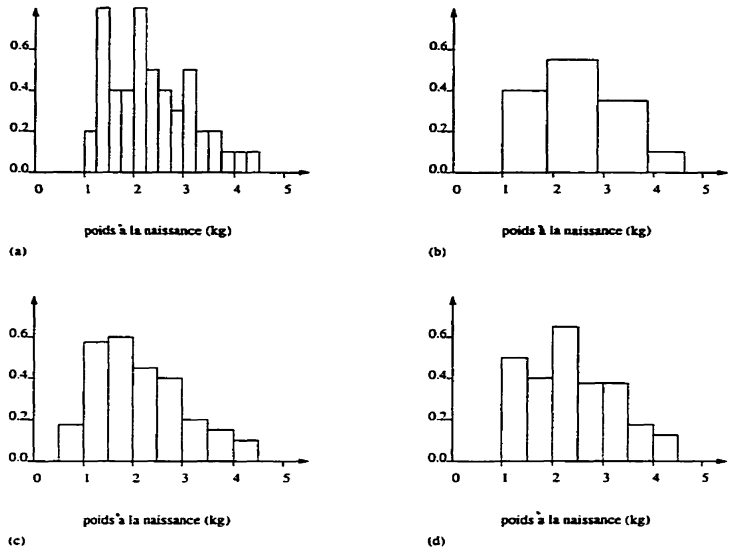

FIGURE 3.1 – l'influence de l'origine et la largeur de classe

L'histogramme souffre de défauts évidents car par construction, tous les points d'un intervalle ont la même densité estimée, ce qui n'est pas réaliste. Une autre difficulté que l'on retrouve avec l'histogramme est que celui-ci estime toutes les densités par une fonction

étagée ce qui n'est pas toujours le cas. De plus, cet estimateur est une fonction discontinue, il constitue donc une mauvaise approximation d'une fonction continue et par conséquent, l'application de certaines opérations sur l'estimé, comme par exemple une dérivée ou une intégration, devient impossible ou très difficile à effectuer.
Afin de résoudre ce problème, l'estimateur de la fenêtre mobile a été introduit, il généralise intuitivement la méthode d'estimation par histogramme, et il est très utilisé en estimation non paramétrique.

3.2 L'Histogramme mobile

Une première amélioration due à *Rosenblatt* (1956) est la méthode de l'histogramme mobile :
Rappelons que la densité de probabilité f est égale à la dérivée de la fonction de répartition F. On peut donc écrire :

$$\begin{aligned} f(x) &= \lim_{h\to 0} \frac{F(x+h)-F(x-h)}{2h} \\ &= \lim_{h\to 0} \frac{\mathbb{P}(x-h<X\leq x+h)}{2h} \end{aligned}$$

Un estimateur de $f(x)$ est alors :

$$\begin{aligned} \widehat{f}_n(x) &= \frac{1}{2h}\#\{i, x-h<X_i\leq x+h\} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{1}_{\{x-h<X_i\leq x+h\}} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{1}_{\{-1\leq \frac{x-X_i}{h}<1\}} \end{aligned}$$

Notons que cet estimateur peut encore s'écrire comme :

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} W(\frac{x-X_i}{h})$$

où

$$W(y) = \begin{cases} 1/2 & \text{si } y \in [-1,1[\\ 0 & \text{sinon.} \end{cases}$$

3.2.1 Propriétés

Biais

$$\begin{aligned}\mathbb{E}(\widehat{f}_n(x)) - f(x) &= \frac{1}{2h}\mathbb{E}(\widehat{F}_n(x+h) - \widehat{F}_n(x-h)) - f(x) \\ &= \frac{1}{2h}(F(x+h) - F(x-h)) - f(x)\end{aligned}$$

Si $h \to 0$ et $nh \to \infty$ quand $n \to \infty$, on a :

$$\lim_{n\to\infty} \mathbb{E}(\widehat{f}_n(x)) = f(x)$$

Variance

$$\begin{aligned}Var(\widehat{f}_n(x)) &= \frac{1}{4n^2h^2}Var(2nh\widehat{f}_n(x)) \\ &= \frac{n(F(x+h) - F(x-h))(1 - F(x+h) + F(x-h))}{4n^2h^2} \\ &\leq \frac{1}{2nh}\frac{F(x+h) - F(x-h)}{2h}\end{aligned}$$

Il en résult que, si $h \to 0$ et $nh \to \infty$ quand $n \to \infty$, on a :

$$\lim_{n\to\infty} Var(\widehat{f}_n(x)) = 0$$

3.3 L'estimateur à noyau

Malgré que l'histogramme mobile est un estimateur consistent et qu'il n'a pas le problème du choix d'origine à o comme le cas de l'histogramme, il présente l'inconvénient d'être discontinu aux points $X_i \pm h$.

Ainsi une généralisation de cet estimateur a été introduite par *Parzen* (1962) en posant

$$\widehat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{X_i - x}{h_n})$$

où h_n est une suite de réels strictement positifs, tendant vers zéro quand $n \to \infty$ appelée paramètre de lissage et K une fonction mesurable appelée noyau.

L'estimateur de Parzen-Rosenblat a connu un très grand succès parmi les estimateurs non paramétriques, ceci est dû à sa simplicité et sa convergence vers la densité f.
Dans la section suivante, nous introduisons les propriétés statistiques élémentaires de l'estimateur à noyau.

3.3.1 Propriétés

La propriété ci-dessous exprime le fait que, lorsque h est petit, la convolution avec K_h perturbe peu une fonction de L^1

Lemme 3.3.1 *(Lemme de Bochner) Soit K un noyau de Parzen-Rosenblat et $g \in L^1$, alors en tout point x où g est continue.*

$$\lim_{h\to 0}(g * K_h)(x) = g(x)$$

avec $K_h(x) = \frac{1}{h}K(\frac{x}{h})$ *et* $(g * K_h)(x) = \int g(x-y)K_h(y)dy$

Proposition 3.3.1 *Soit $X_1, \ldots, X_n$ un n-échantillon de X de densité f et $\widehat{f}_n$ l'estimateur à noyau K, alors $f_n(x)$ converge en moyenne quadratique vers $f(x)$ avec les conditions $h_n \to \infty$ quand $n \to \infty$ et $nh_n \to \infty$ quand $n \to \infty$*

Preuve

L'objectif est de vérifier que :

$$\lim_{n\to\infty} \mathbb{E}[\widehat{f}_n(x) - f(x)]^2 = 0$$

On a

$$Var(\widehat{f}_n(x)) = Var(\widehat{f}_n(x) - f(x)) = \mathbb{E}([\widehat{f}_n(x) - f(x)]^2) - \mathbb{E}([\widehat{f}_n(x) - f(x)])^2$$

D'où

$$\mathbb{E}[\widehat{f}_n(x) - f(x)]^2 = Var(\widehat{f}_n(x)) + \mathbb{E}([\widehat{f}_n(x) - f(x)])^2$$

Il suffit de montrer que

$$\lim_{n\to\infty} \mathbb{E}[\widehat{f}_n(x) - f(x)] = 0 \tag{3.1}$$

et que

$$\lim_{n\to\infty} Var(\widehat{f}_n(x)) = 0 \tag{3.2}$$

En effet pour (3.1)

$$\begin{aligned}\mathbb{E}(\widehat{f}_n(x)) &= \mathbb{E}[\frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x-X_i}{h})] \\ &= \frac{1}{h}\int K(\frac{x-z}{h})f(z)dz\end{aligned}$$

En posant $t = x - z$, on arrive à : $\mathbb{E}(\widehat{f}_n(x)) = \frac{1}{h_n}\int K(\frac{t}{h_n})f(x-t)dt$

d'après le lemme de *Bochner* et le fait que $\lim_{n\to\infty} h_n = 0$ et $\frac{1}{h_n}K(\frac{t}{h_n}) = K_{h_n}(t)$, on obtient

$$\lim_{n\to\infty}\frac{1}{h_n}\int K(\frac{t}{h_n})f(x-t)dt = f(x)\int K(t)dt$$

D'où

$$\lim_{n\to\infty}\mathbb{E}(\widehat{f}_n(x)) = f(x)$$

pour (3.2), on a

$$\begin{aligned}Var(\widehat{f}_n(x)) &= Var[\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h})] \\ &= \frac{1}{nh^2}Var(K(\frac{x-X}{h})) \\ &= \frac{1}{nh}[\frac{1}{h}\mathbb{E}(K^2(\frac{x-X}{h}))] - \frac{1}{n}[\frac{1}{h}\mathbb{E}(K(\frac{x-X}{h}))]^2 \\ &= \frac{1}{nh}[K_h^2 * f - (K_h * f)^2]\end{aligned}$$

Il suffit d'appliquer le lemme *Bochner* sur le noyau $K' = \dfrac{K^2}{\int K^2(z)dz}$ ∎

Maintenant, nous introduisons les hypothèses de base permettant de donner un théorème général sur la convergence presque complète.

1. f est continue au voisinage de x, un point fixé de $\mathbb{R}$
2. le paramètre de lissage h_n est tel que : $\lim_{n\to\infty} h_n = 0$ et $\lim_{n\to\infty}\dfrac{\log n}{nh_n} = 0$

3. K est borné, intégrable et à support compact.

Théorème 3.3.1 *Si les conditions* $1, 2$ *et* 3 *sont vérifiées, alors :*

$$\lim_{n\to\infty} \widehat{f}_n(x) = f(x) \qquad p.co$$

Preuve

La démonstration de ce théorème est basée sur la décomposition suivante :

$$\widehat{f}_n(x) - f(x) = (\widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x))) - (f(x) - \mathbb{E}(\widehat{f}_n(x)))$$

Il suffit donc de montrer que

$$\lim_{n\to\infty} \mathbb{E}(\widehat{f}_n(x)) = f(x) \tag{3.3}$$

et

$$\widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x)) = O(\sqrt{\frac{logn}{nh_n}}) \qquad p.co. \tag{3.4}$$

Pour (3.3), Nous avons :

$$\begin{aligned} \mathbb{E}(\widehat{f}_n(x)) &= \frac{1}{h_n}\int_{-\infty}^{+\infty} K(\frac{x-t}{h_n}) f(t)dt \\ &= \int K(z) f(x - zh_n)dz \end{aligned}$$

La continuité uniforme de f sur le support compact de K entraine $f(x - zh_{_n}) \to f(x)$, uniformément en z. D'où $\lim_{n\to\infty} \mathbb{E}(\widehat{f}_n(x)) = f(x)$

Pour (3.4), on a

$$\begin{aligned} \widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x)) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n}[K(\frac{x-X_i}{h_n}) - \mathbb{E}(K(\frac{x-X_i}{h_h}))] \\ &= \frac{1}{n}\sum_{i=1}^{n} M_i \end{aligned}$$

où $M_i = \dfrac{1}{h_n}[K(\dfrac{x-X_i}{h_n}) - \mathbb{E}(K(\dfrac{x-X_i}{h_n}))]$

En utilisant l'hypothèse 3, on a : $|M_i| < \dfrac{C}{h_n}$

d'autre part le changement de variable $z = \frac{x-t}{h_n}$, nous donne

$$\begin{aligned} \frac{1}{h_n}\mathbb{E}[\frac{1}{h_n}K^2(\frac{x-X}{h_n})] &= \frac{1}{h_n^2}\int K^2(\frac{x-t}{h_n})f(t)dt \\ &= \frac{1}{h_n}\int K^2(z)f(x-zh_n)dz \end{aligned}$$

Comme K est bornée et f est continue sur le support compact de K, on a l'existence d'une constante C, telle que :

$$\mathbb{E}(M_i^2) < \frac{C}{h_n}$$

On obtient alors, en appliquant l'inégalité de Hoeffding,

$$\widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x)) = O(\sqrt{\frac{logn}{nh_n}}) \qquad p.co.$$

■

3.4 Conclusion

La problématique posée dans ce mémoire est l'estimation non paramétrique de la fonction de répartition et de la densité.
Deux méthodes pour estimer la fonction de répartition ont été utilisées : la fonction de répartition empirique, et l'estimateur à noyau. Notons que L'estimation par la fonction de répartition empirique est très utile car de nombreuses statistiques peuvent s'exprimer comme des fonctionnelles de celle-ci.
Pour ce qui concerne la densité, à part l'histogramme, l'estimateur à noyau est probablement l'estimateur le plus utilisé et certainement le plus étudié mathématiquement. Il possède des propriétés qui le rendent fort intéressant. Car, si le noyau K est une fonction de densité alors l'estimateur à noyau $\widehat{f}_n$ est lui aussi une fonction de densité. De plus, ce dernier possède les propriétés de continuité et de différentiabilité du noyau K. De sorte que si, par exemple, K est la densité normale dors $\widehat{f}_n$ possède des dérivées de tout ordre.

Bibliographie

[1] Geoffrey, J. (1974). Sur l'estimation d'une densité dans un espace métrique. C.R. Acad. Paris Sér. A, 278 :1449-1452.

[2] MATIAS, C. (2012). Introduction à la statistique non paramétrique. CNRS, Laboratoire Statistique, Génome, Évry. SFDS.

[3] Tsybakov, A.(2003). Introduction à l'estimation non-paramétrique. Springer science and business media, New York.