République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en **Mathématiques**

Option : **Statistique**

Par

Choukal Lamia

Titre :

Estimation de la fonction de distribution

Membres du Comité d'Examen :

P.r	BENATIA Fateh	UMKB	Président
P.r	ABDALLAH Sayah	UMKB	Encadreur
P.r	ABDELLI Jihane	UMKB	Examinatrice

Juin 2024

Dédicace

.

Je dédie ce modeste travail tout d'abord à mes trés chers parents qui nous ont soutenu tout le long de mon parcour d'études ;

A mes trés chers parents ;

A ma trés chère amie : Almi Nassima ;

A mes soeurs et leurs enfants : Anfal ; Malak ; Ismail ;

A toute ma famille ;

REMERCIEMENTS

Tout d'abord je tiens à remercier dieu qui m'a donné le courage la patience et la volenté pour mener à bien ce travail

J'ai l'honneur et le plaisir d'exprimer mes profondes gratitudes à mon encadreur **Pr. Sayah Abdallah** pour ces apports scientifiques, ces orientations, ces remarques et aussi pour l'intéret porté a mon travialle .

Mes remerciments les plus vifs s'adressent également aux memberes de jury qui ont accepté s'examiner et d'évaluer ce travail **Pr. BENATIA Fateh** et **Pr ABDELLI Jihane**

Tout particulièrement je remercie profondément Almi NASSIMA pour son aide.

Finalement je tiens à remercier toute ma famille , mes amies et tous ceux qui ont contribué à la réalisation de ce memoire.

Merci...

Table des matières

Table des figures

Liste des tableaux

Introduction

Les méthodes non paramétriques deviennent progressivement populaires dans l'analyse satistique des de nombreux problèmes de domaines, tels que l'économie, la biologie et l'actuariat, par exemple en raison du réchauffement climatique. Le secteur de l'assurance est de plus en plus exposé à des événements extrêmes tels que des tempêtes de grêle, un volcan, etc. De tels événements entraînent des pertes catastrophiques. Il est nécessaire d'estimer la probabilité de tels événements et la probabilité que le paiement dépasse certains montants (par exemple 1, 000, 000 DA) pour que les compagnies d'assurance puissent déterminer des primes appropriées. Notons X le montant de l'indemnité d'un accident, la quantité d'intérêt est $P(X > x)$, où x est un montant de paiement prédéfini.

La connaissance de la fonction de distribution, ou leur estimateurs, permet de caractériser la variable aléatoire de manière plus complète. En particulier, nous pouvons dériver d'autres caractéristiques des variables aléatoires à partir de là, telles que les quantiles, la fonction de survie, le taux de risque, etc.

Il existe un certain nombre de méthodes d'estimation non paramétrique qui permettent d'estimer la fonction distribution.à savoir la methode empirique et la méthode la plus rencontrée, dans la littérature, est la méthode du noyau proposée par Nadaraya (1964). C'est une méthode très utilisée vu sa souplesse d'utilisation et ses propriétés. L'estimateurs à noyau est une fonction de deux paramètres k, appelé noyau, et h dit paramètre de lissage

Le mémoire est structuré de la maniére suivante

– **Premier chapitre** : "Préliminaire et notions de probabilités" : Il contient un rappel

sur les variables aléatoires (discrètes et continues) et l'estimation paramétrique de la fonction de distribution

- **Deuxième chapitre** : "Estimation non paramétrique de la fonction de distribution" : Elle sera consacrée à l'étude de l'estimation non paramétrique de la fonction de distribution où nous abordons les propriétés statistiques de chaque estimateur.
- **Troisième chapitre : "**Simulation " où nous donnons des exemples de simulation par le logiciel **R** qui expriment l'importance du choix de paramètre de lissage et du noyau.

Nous terminerons ce mémoire par une conclusion générale .

Chapitre 1

Préliminaire et notions de probabilités

L'objectif de ce chapitre est d'introduire certaines notions et concepts élémentaires nécessaires dont nous avons besoin autour de ce travail.

1.1 Variable aléatoire

Une des notions fondamentales des statistiques est celle de variable aléatoire.

Définition 1.1 (Variable aléatoire réelle) *Etant donné un ensemble fondamental Ω, une variable aléatoire réelle (v.a.r) X est une fonction de Ω dans :*

$$X : \omega \in \Omega \longmapsto X(\omega) \in \mathbb{R}.$$

Remarque 1.1 *Il y a deux types des variables aléatoires discretes et continues*

1.1.1 Variable aléatoire discrète

Soit $X : \omega \in \Omega \longmapsto X(\omega)$ une application. On dit que X est une variable aléatoire discrète (v.a.d) si :

– $X(\omega)$ est fini ou infini dénombrable

Exemple 1.1 (Cas discret) *On jette deux dés distincts et on s'intéresse au plus grand chiffre obtenu. Alors :* X $\{1, 2, 3, 4, 5, 6\}$

1.1.2 Variable aléatoire continue

On dit que X est une variable aléatoire continue (v.a.c) si elle prend des valeurs dans $\mathbb{R}$ (où éventuellement d'un intervalle).

Exemple 1.2 *La taille des individus d'une population*

$$X\,\{[1.60; 1.65[\,,\ \ [1.65; 1.70[\,,\ \ [1.75; 1.80[\}$$

1.1.3 Loi de probabilité d'une variable aléatoire

Définition 1.2 (Loi de probabilité) *La loi de probabilité est déterminée par la fonction de distribution, qui est définie pour tout réel par :*

$$F_X(x) = P(X \leq x),\ x \in \mathbb{R}.$$

Remarque 1.2 *La loi de probabilité d'une v.a.réelle discrète est définie par les probabilités individuelles :*

$$f(x) = P(X = x),\ x \in \mathbb{R}$$

Les probabilités f doivent satisfaire les deux conditions suivantes :
– $\sum_{i=1}^{n} f(x_i) = 1,$
– *Toutes les $f(x_i)$ sont positives.*

Exemple 1.4

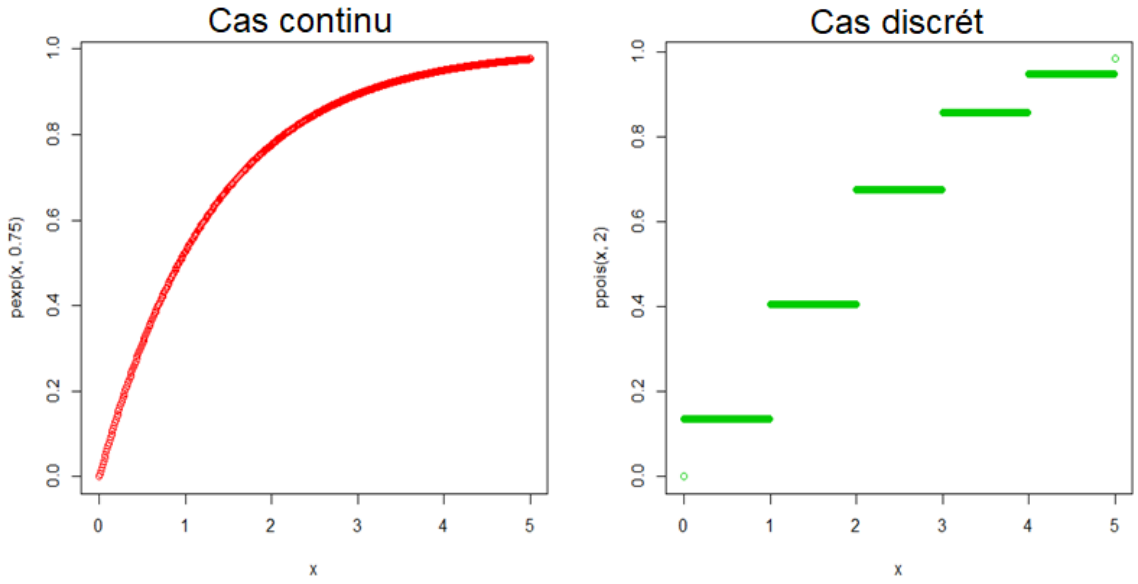

FIG. 1.1 – Représentation graphique de distribution

Exemple 1.3 *La fonction dans le tableau ci-dessous repésente une loi de probabilité*

x_i	*0*	*1*	*2*	*3*
$P(X = x_i)$	*0.17*	*0.43*	*0.13*	*0.27*

Remarque 1.3 *La loi d'une variable aléatoire dans le cas continu donne par la fonction de densité f qui vérifie :*

- $\forall x \in \mathbb{R},\ f(x) > 0$;
- $\int_{\mathbb{R}} f(x)dx = 1.$
- $F_X(x) = \int_{-\infty}^{x} f(t)dt$ *avec* F_X *est la fonction de distribution.*

Propriétés de la fonction de distribution :

1. $0 \leq F_X(x) \leq 1.$
2. $\lim_{x \to -\infty} F_X(x) = 0$ et $\lim_{x \to +\infty} F_X(x) = 1.$
3. F_X : est croissante et continue à droite.

1.1.4 Caractéristiques d'une variable aléatoire

Soit X une variable aléatoire

– **Moment d'ordre** r

1. **Moment non centré d'ordre** r

Le moment non centré d'ordre r de la v.a.r X, notée $E(X^r)$ est définie par

Cas discret

$$E(X^r) = \sum_i x_i^r P(X = x_i).$$

Cas continu

$$E(X^r) = \int_{\mathbb{R}} x^r f(x)dx.$$

Remarque 1.4 *L'espérance mathématique est le moment d'ordre* 1

2. **Moment centré d'ordre** r

Définition 1.3

Le moment centré d'ordre r de la v.a.r X, notée $E(X_c^r)$ est définie par

– **Cas discret**

$$E(X_c^r) = \sum_i (x_i - E(X))^r P(X = x_i).$$

Cas continu

$$E(X_c^r) = \int_{\mathbb{R}} (x - E(X))^r f(x)dx.$$

Remarque 1.5 *Le moment centré d'ordre 1 n'a aucun intérêt puisque est égal* 0*,*

Le moment centré d'ordre 2 *est la variance.*

– **Variance** Si le moment d'ordre 2 existe, Alors la variance (Var) est définie par :

$$Var(X) = E[(X - E(X))^2],$$

– **L'écart type** (σ) de X

$$\sigma(X) = \sqrt{Var(X)}.$$

Exemple 1.5 (Quelques lois discrètes et ses caractéristiques) – ***Loi de Bernoulli***

On dit qu'une variable aléatoire X suit une loi de Bernoulli de paramètre $p \in [0,1]$ ce que l'on note $X \rightarrow \mathbb{B}(p)$

$$P(X = x) \begin{cases} p & si \quad x = 1 \\ 1-p & si \quad x = 0 \end{cases}; \quad 0 \leq p \leq 1$$

$$E(X) = p \; ; \qquad Var(X) = (1-p)p$$

– ***Loi binomiale***

Si on renouvelle n fois de manière indépendante une épreuve de Bernoulli de paramètre p

On dit que la variable aléatoire X suit une loi binomiale de paramètres (n,p) on la note par $\mathbb{B}(n,p)$

$$P(X = x) = \begin{cases} C_n^x p^x (1-p)^{n-x}, & si \qquad x \in 0,1,2...,n \\ 0 & sinon \end{cases}; \quad n \in \mathbb{N}$$

$$E(X) = np \; ; \qquad Var(X) = n(1-p)p$$

– ***Loi de Poisson***

la loi de Poisson est une loi qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé , alors :

On dit que la variable aléatoire X suit une loi de poisson de paramètre λ on la note par $\mathbb{P}(\lambda)$

$$P(X = x) = \lambda^x \frac{e^{-\lambda}}{x!}; \quad x \in \mathbb{R}$$

$$E(X) = \lambda \ ; \qquad Var(X) = \lambda$$

– **Loi uniforme**

La loi uniforme est la loi de probabilité continue la plus simple, définie sur un ensemble borné $[a; b]$,Elle est utilisée pour modéliser une variable répartie uniformément sur l'ntervalle $[a; b]$ on note par $\mathbb{U}([a; b])$

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si} \quad x \in [a; b] \\ 0 & \text{sinon} \end{cases}$$

$$E(X) = \frac{a+b}{2} \ ; \qquad Var(X) = \frac{(b-a)^2}{12}$$

– Loi normale ou loi de Laplace-Gauss de parametres (μ, σ)

La loi normale est une loi qui dépend de deux paramètres : son espérance, un nombre réel noté μ, et son écart type, un nombre réel positif noté ■..on la note par $\mathbb{N}(\mu, \sigma)$

La densité de probabilité de la loi normale d'espérance μ et d'écart type ■ est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad ; x \in \mathbb{R}$$

$$E(X) = \mu \ ; \qquad Var(X) = \sigma^2$$

– Loi exponentielle de parametre λ on note par $\mathcal{E}(\lambda)$

$$f(x) = \lambda e^{-\lambda x} \quad ; x \in \mathbb{R}^+$$

$$E(X) = \frac{1}{\lambda} \; ; \qquad\qquad Var(X) = \frac{1}{\lambda^2}$$

1.1.5 Convergence des suites de variables aléatoires

Une suite (X_n) de variables aléatoires étant une suite de fonctions de Ω dans $\mathbb{N}$, il existe diverses façons de définir la convergence de (X_n) dont certaines jouent un grand rôle en calcul des probabilités.

1. **Convergence en probabilité**

   La suite (X_n) converge en probabilité vers une v.a.r X et on note $X_n \xrightarrow{P} X$ si :

   $$\forall \varepsilon > 0, \text{ on a } \lim_{n \to +\infty} P(\{|X_n - X| > \varepsilon\}) = 0.$$

2. **Convergence presque sûre ou convergence forte**

   Définissons d'abord l'égalité presque sûre de deux variables aléatoires :

   Définition 1.4 *X et Y sont égales presque sûrement si $P(\{w : X(w) \neq Y(w)\}) = 0$.*

   La convergence presque sûre se définit alors par :

   Définition 1.5 *La suite (X_n) converge presque sûrement vers X et on note $X_n \xrightarrow{p.s} X$ si*

   $$P\left(\left\{w : \lim_{n \to +\infty} X_n(w) \neq X(w)\right\}\right) = 0$$

2. **Convergence en loi**

   Bien qu'elle est la plus faible, elle est très utilisée en pratique car elle permet d'approximer la fonction de distribution de X_n par celle de X.

   Définition 1.6 *Soient (X_n) une suite de variables aléatoires et X une variable aléatoire sur un même espace probabilisé (Ω, P), de fonctions de distribution respective-*

ment F_n et F ; on dit que (X_n) convergent vers X en loi (et on note $X_n \xrightarrow{L} X$) si en tout point x où F est continue, les $F_n(x)$ convergent vers $F(x)$.

3. **Convergence en moyenne quadratique**

Définition 1.7 *On dit que la suite de v.a.r $(X_n)_{n\in N}$ converge en moyenne quadratique vers une v.a.r X (et on note $X_n \xrightarrow{m.q} X$) si :*

$$\lim_{n\to+\infty} E[(X_n - X)^2] = 0.$$

– **Lois des Grands Nombres**

Ces lois décrivent le comportement asymptotique de la moyenne de l'échantillon. Elles sont de deux types : lois faibles mettant en jeu la convergence en probabilité et lois fortes relatives à la convergence presque sûre.

Théorème 1.1 *Si $(X_1, ..., X_n)$ est un échantillon d'une v.a.r X tel que $E|X| < \infty$, alors*

$$\textit{Loi faible} \qquad \overline{X}_n \xrightarrow{P} \mu \quad \textit{quand } n \to \infty.$$

$$\textit{Loi forte} \qquad \overline{X}_n \xrightarrow{p.s} \mu \quad \textit{quand } n \to \infty,$$

où $\mu := E(X)$ et $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

– **Théoréme Central Limite**

L'étude de somme $S_n = n\ \overline{X}_n$ de variables indépendantes et de même loi joue un rôle capital en statistique. Le théorème suivant connu sous le nom de théorème central limite établit la convergence vers la loi normale.

Théorème 1.2 (TCL) *Si $X_1, X_2, ...$ est une suite des v.a.r indépendantes et identiquement distribuées (i.i.d) de moyenne μ et de variance finie σ^2, alors*

$$\frac{(S_n - n\mu)}{\sigma\sqrt{n}} \xrightarrow{L} \mathcal{N}(0,1) \qquad \textit{quand } n \to \infty.$$

1.2 Estimateur

Définition 1.8 *On appelle un estimateur de θ toute fonction mesurable T_n de $X = (X_1, X_2, ...X_n)$ dans Θ autrement dit*

$$T : X \to \Theta$$

$$X \to T(X)$$

$T(X)$ s'appelle l'estimateur de θ

1.2.1 Qualité d'un estimateur :

1. **Estimateur avec biais** :Un estimateur T_n de θ est dit avec biais ou biasé si pour tout $\theta \in \Theta$ (Θ ouvert de $\mathbb{R}$) et tout entier positif n

$$E(T_n) = \theta + b(n, \theta).$$

La quantité $b(n, \theta)$ est le biais de l'estimateur T_n.

Exemple 1.6

$$Y = \frac{1}{n+1} \sum_{i=1}^{n} X_i$$

est un estimateur biaisé de $E(X)$ car :

$$\begin{aligned} E(Y) &= \frac{1}{n+1} \sum_{i=1}^{n} E(X_i) \\ &= \frac{n}{n+1} E(X). \end{aligned}$$

2. **Estimateur sans biais :** Un estimateur T_n de θ et $\theta \in \Theta$ est dit sans biais si :

$$b(n, \theta) = 0 \text{ alors } :E(T_n) = \theta$$

Exemple 1.7 *la moyenne empirique*

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

est un estimateur sans biais de la moyenne μ

3. **Estimateur asymptotiquement sans biais :** Un estimateur T_n de θ est dit asymptotiquement sans biais si :

$$\lim_{n\to\infty} b(n,\theta) = 0 \text{ alors } \lim_{n\to\infty} E(T_n) = \theta.$$

Exemple 1.8

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

est un estimateur asymptotiquement sans biais de σ^2.

Définition 1.9 (Erreur quadratique moyenne) *Soit* T_n *un estimateur de* θ*, l'erreur quadratique moyenne de* T_n *notée* (MSE) *est :*

$$MSE(T_n) = E[(T_n - \theta)^2].$$

Remarque 1.6 *L'erreur quadratique moyenne peut être écrite comme un somme de la variance et du carré du biais de l'estimateur*

$$MSE(T_n) = Var(T_n) + [E(T_n) - \theta]^2.$$

1.3 Estimation paramétrique de la fonction de distribution

L'estimation paramétrique est souvent perçue comme l'une des méthodes d'estimation les plus précises et les plus fiables, mais elle nécessite un effort important dès le départ.
Soit $X \sim F$, avec $F(x) = P(X \leq x)$ la fonction de distribution de X.

X est le vecteur formé par n-échantillon $X_1, ..., X_n$.de densité $f(X, \theta_1, ..., \theta_k)$ où : $\theta_1, ..., \theta_k$ sont les paramètres inconnus

1.3.1 Estimation par la méthode des moments

La méthode des moments consiste à estimer les paramètres $\theta_1, ..., \theta_k$, enégalisant les moments empiriques calculés à partir de l'échantillon avec les moments théoriques
de même ordre.où

– Moments d'ordre r de la population (théorique) est definie par

$$\mu_j = E(X_j), \quad j = 1, 2, ..., k$$

– Moments empirique d'ordre r de l'échantillon est definie par

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

La solution du système $\bar{X}_j = \mu_j, \qquad j = 1; k$ nous donne les estimateurs de $\theta_1, ..., \theta_k$

Remarque 1.7 *Dans la plupart des cas,les estimateurs obtenus par la méthode des moments sont consistants,convergents,asymptotiquement normaux mais en général ne sont pas éfficaces.*

Exemple 1.9 *Nous disposons d'un n-échantillon $(X_1, ..., X_n)$ de loi de Poisson de para-*

mètre $\lambda > 0$ inconnu. Rappelons que pour tout $n \in N$,

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

– *Le moment d'order* 1 *est definie par*

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} ke^{-\lambda}\frac{\lambda^{k-1+1}}{k(k-1)!} \\ &= \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \end{aligned}$$

– *Donc l'estimateur de λ par la méthode des moments est*

$$\widehat{\lambda} = \overline{X}$$

Application

Basé sur le logiciel R, on generate un échantillon suit la loi de poisson de taille 100 de parametre $\lambda = 2$ puis on calcule le moment d'orde 1 on trouve $\widehat{\lambda} = 2.10$

1.3.2 Estimation par la méthode du maximum de vraisemblance

L'estimation du maximum de vraisemblance est une méthode statistique courante utilisée pour inférer les paramètres de la distribution de probabilité d'un échantillon donné .Cette méthode consiste à recherche le paramètre $\widehat{\theta}$ qui maximise la fonction de vraisemblance $L(x, \theta)$.

– La fonction vraisemblance d'un échantillon $x_1, ..., x_n$ indépendent et identiquement distribué de densité f definie par

$$L(x, \theta) = \prod_{i=1}^{n} f\left(x_i, \theta\right)$$

Le principe de l'estimation par maximum de vraisemblance est de dire que plus la probabilité d'avoir obtenu les observations est forte, plus le modéle est proche de la réalité. Ainsi, on retient le modéle pour le quel la vraisemblance de notre échantillon est la plus élevée :

$$\widehat{\theta}^{MV} = arg\ max\ L(x, \theta)$$

En pratique, on va savoir un probléme de résoudre directement en raison de la présence du produit mais il suffit de prendre le logarithme de la vraisemblance

$$\widehat{\theta}^{MV} = arg\ max\ ln\ L(x, \theta)$$

Pour trouver le maximum, on résoud l'équation du premier ordre :

$$\left.\frac{\partial L(x,\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}^{MV}} = 0$$

On utilise le méme exemple précédent, on souhaite estimer le paramétre λ d'une loi de Poisson à partir d'un $n = 100$ échantillon

La fonction de vraisemblance sécrit

$$\begin{aligned} L(x_1, x_{2,\ldots,}x_n, \lambda) &= \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!} \end{aligned}$$

Il est plus simple d'utiliser le logarithme, la vraisemblance étant positive

$$\begin{aligned} \ln\left(L(x_1, x_{2,\ldots,}x_n; \lambda)\right) &= \ln(e^{-n\lambda}) + \sum_{i=1}^{n}\ln\left(\frac{\lambda^{x_i}}{x_i!}\right) \\ &= -n\lambda + \ln(\lambda)\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\ln(x_i!) \end{aligned}$$

La dériviée premiére

$$\frac{\partial \ln\left(L(x_1, x_{2,\ldots,}x_n; \lambda)\right)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i$$

S'annule pour

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La dériviée seconde :

$$\frac{\partial^2 \ln\left(L(x_1, x_{2,\ldots,}x_n; \lambda)\right)}{\partial \lambda^2} = \frac{-1}{\lambda^2}\sum_{i=1}^{n} x_i.$$

Est toujours négative ou nulle. ainsi $\widehat{\lambda} = \overline{X}$.

Application

Basé sur la méme échantillon précedent de loi de poisson, d'aprés simlpe calcule on trouve $\widehat{\lambda} = 2.01$.

Par la représentation graphique, en comparant les deux méthodes, on remarque que la méthode du maximum de vraisemblence est plus précise que la méthode des moments, comme indiqué dans le graphe suivant

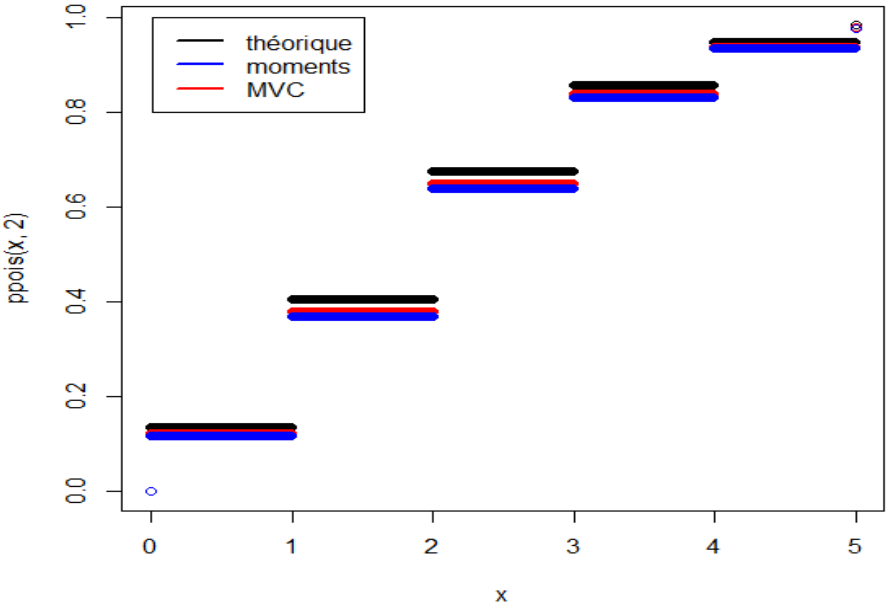

FIG. 1.2 – Comparaison entre les deux méthodes paramétriques

Chapitre 2

Estimation non-paramétrique de la distribution

L'estimation non paramétrique d'une fonction de distribution est un objectif fondamental dans de nombreux domaines dans lesquels les analystes s'intéressent à l'estimation du risque d'occurrence d'un événement particulier, par exemple en raison du réchauffement climatique. Le secteur de l'assurance est de plus en plus exposé à des événements extrêmes tels que des tempêtes de grêle, un volcan, etc. De tels événements entraînent des pertes catastrophiques. Il est nécessaire d'estimer la probabilité de tels événements et la probabilité que le paiement dépasse certains montants (par exemple 1, 000, 000 DA) pour que les compagnies d'assurance puissent déterminer des primes appropriées. Notons X le montant de l'indemnité d'un accident, la quantité d'intérêt est $P(X > x)$, où x est un montant de paiement prédéfini.

Dans ce chapitre, nous allons présenter une étude non paramétrique de la fonction de distribution tellque la méthode empirique et la méthode du noyau ainsi que ses propriétés statistiques.

2.1 Estimateur empirique de la fonction de distribution

Dans de nombreuses applications, la loi de F est inconnue et on dispose d'un n observations $x_1, x_2, ..., x_n$, de fonction de distribution F inconnue. Comment obtenir un estimateur de F à partir de la seule information contenue dans l'échantillon ?

Ce problème, que l'on désigne généralement par estimation non paramétrique de la fonction de distribution a fait l'objet de multiples travaux par des méthodes diverses, à savoir :

L'estimateur empirique et l'estimateur par la methode de noyau.que nous vous présentons dans ce qui suit.

Soit $X_1, X_2, ..., X_n$ suit des variables aléatoires indépendantes identiquement distribuées comme une variable aléatoire X, dont la fonction de distribution $F(x)$ est absolument continue

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

avec $f(x)$ fonction de densité de probabilité.

L'estimateur non paramétrique le plus couramment utilisée pour la fonction F est l'estimateur empirique (EDF) noté par F_n est donne par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$$

$$F_n(x) = \frac{card\left\{X_i \leq x, i = 1, .., n\right\}}{n}$$

$$= \begin{cases} 0 & \text{si} \quad x < X_{1,n} \\ \dfrac{i}{n} & \text{si } X_{i,n} \leq x < X_{i+1,n} \\ 1 & \text{si} \quad x \geq X_{n,n} \end{cases}$$

où $X_{i,n}$ la statistique d'order de X , $i = 1, 2, ...n$

$$X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n}$$

et $\mathbb{I}(x)$ est la fonction de l'indicatrice :

$$\mathbb{I}(x) = \begin{cases} 1 & \text{si } X_i \leq x \\ 0 & \text{Sinon} \end{cases}$$

2.1.1 Propriétés statistiques

1. Biais et Variance de l'estimateur F_n

$$E(Fn(x)) = F(x) \quad \text{et} \quad V(Fn(x)) = \frac{F(x)(1 - F(x))}{n}$$

Alors, pour tout point x, F_n est une estimateur sans biais de F.

2. La loi forte des grands nombres nous donne

$$\forall x \in \mathbb{R} : F_n(x) \overset{p.s}{\to} F(x) \text{ si } n \to +\infty.$$

3. La convergence presque sûrement uniforme de F_n vers F est définie par :

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \overset{p.s}{\to} 0, \text{ quand } n \to +\infty$$

Malgré les bonnes propriétés statistiques de F_n, le fait connu que le lissage peut perdre Figeur (2.1), on pourrait préférer une autre estimateur plutôt lisse

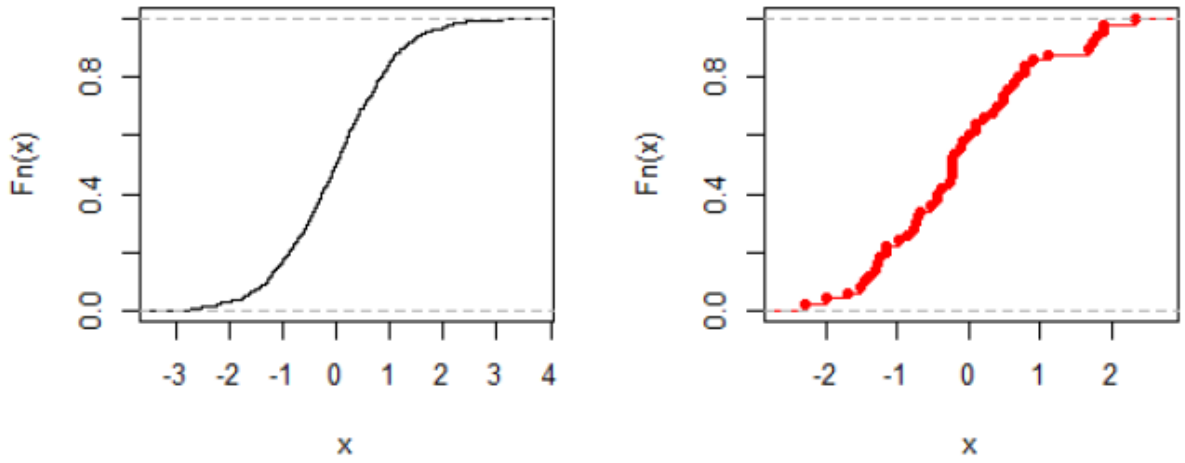

FIG. 2.1 – Performance de l'estimateur empirique dans les deux cas continu et discrét

2.2 Estimation de la fonction de distribution par la méthode du noyau

Nadaraya [9] a proposé une alternative non paramétrique estimateur plus lisse à l'estimateur EDF, à savoir l'estimateur de distribution du noyau (KDF) que nous avons noté $\widetilde{F_n}$. Cet estimateur est obtenu en intégrant l'estimateur de densité à noyau de Rosenblatt -Parzen ($1959 - 1962$), où qu'est noté par $\widetilde{f}_n$, que nous présentons brièvement dans la sous-section suivante.

2.2.1 L'estimateur à noyau de la densité

$$\widetilde{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right), \quad x \in \mathbb{R}$$

où $h := h_n$ est le paramètre de lissage, qui contrôle le lissage de l'estimateur et qui vérifie ($h \to 0, nh \to \infty$ lorsque $n \to +\infty$) et $k(.)$ est la fonction de noyau.

Lemme 2.1 *Si k est une densité de probabilité, alors $\widetilde{f}_n$ est aussi une densité de probabilité, en effet*

on utilise un changement de variable $t = \dfrac{y-x}{h} \Longrightarrow dt = \dfrac{1}{h}dy$, et on obtient :

$$\begin{aligned}\int_{\mathbb{R}} \widetilde{f}_n(x)dx &= \frac{1}{h}\int_{\mathbb{R}} k\left(\frac{X-x}{h}\right)dx \\ &= \int_{\mathbb{R}} k\left(t\right)dt = 1.\end{aligned}$$

Définition 2.1 *Un noyau est dit symétrique si, pour tout t dans son ensemble de définition $k(t) = k(-t)$.*

Tout noyau symétrique vérifie les conditions suivantes

1. $k(t) \geq 0$, $\forall t \in \mathbb{R}$.
2. $\displaystyle\int_{-\infty}^{\infty} tk(t)dt = 0.$
3. $\displaystyle 0 < \int_{-\infty}^{\infty} t^2k(t)dt < \infty$
4. $\displaystyle\int_{-\infty}^{\infty} k^2(t)dt < \infty$
5. $\int_{\mathbb{R}} |k(t)|dt < \infty$, $\sup\limits_{-\infty<t<+\infty} |k(t)| < \infty$, $\lim\limits_{t\to\infty} |tk(t)| < \infty$

2.2.2 Exemples des noyaux (Noyaux symétriques)

Le tableau suivant présente quelques noyaux continus symétriques et leurs formes qui sont présentées (see Silverman,1986)

Kernel	$k(t)$
Uniform	$\frac{1}{2} I\left(\lvert t\rvert \leq 1\right)$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, for $t \in \mathbb{R}$
Epanechnikov	$\frac{3}{4}\left(1-t^2\right) I\left(\lvert t\rvert \leq 1\right)$
Quartic or Biweight	$\frac{15}{16}\left(1-t^2\right)^2 I(\lvert t\rvert \leq 1)$
Triangular or Triweight	$\frac{35}{32}\left(1-t^2\right)^3 I\left(\lvert t\rvert \leq 1\right)$

TAB. 2.1 – Quelques fonction de noyaux

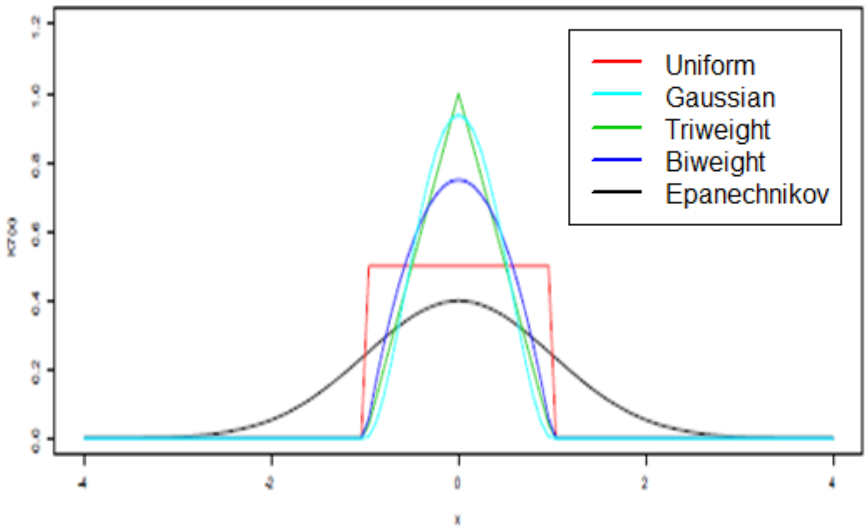

FIG. 2.2 – Représentation graphique des noyaux usuels

2.2.3 Estimateur à noyau de la fonction de distribution

Pour obtenir un estimateur non paramétrique de F, nous intégrons l'estimateur de la densité $\widetilde{f}_n$, qui est défini par :

$$\widetilde{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \quad x \in \mathbb{R}$$

où la fonction K est définie par :

$$K(x) = \int_{-\infty}^{x} k(t)dt.$$

Exemple 2.1 *Dans cet exemple, nous donnons la représentation graphique de l'estimateur empirique et l'estimateur à noyau pour* $k(t) = \frac{3}{4}(1-t^2)$ *et* $h = 0.12$ *d'une distribution de loi normale* $N(0;1)$*.Il est très clair que l'estimateur à noyau est plus lisse que l'estimateur empirique..*

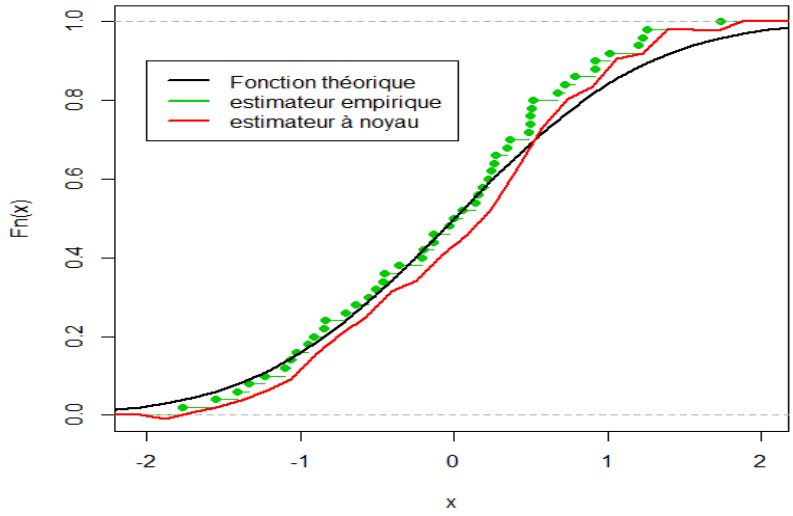

FIG. 2.3 – Comparaison le lissage de l'estimateur empirique et l'estimateur à noyau

Propriétés statistiques

– **Bias**

$$Bias\left(\widetilde{F_n}(x)\right) = \frac{1}{2}h^2 f'(x)\mu_2(k) + o(h^2), \tag{2.1}$$

Preuve. pour tout $x \in \mathbb{R}$

$$Bias\left(\widetilde{F_n}(x)\right) = E\left(\widetilde{F_n}(x)\right) - F(x)$$

$$\begin{aligned} E\left(\widetilde{F_n}(x)\right) &= E\left(K\left(\frac{x-X_i}{h}\right)\right) \\ &= \int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) dF(y) \\ &= \int_{-\infty}^{x-h} 1.f(y)dy + \int_{x-h}^{x+h} K\left(\tfrac{x-y}{h}\right) f(y)dy + \int_{x+h}^{+\infty} 0.f(y)dy \end{aligned}$$

On pose $t = \dfrac{x-y}{h}$ et nous utilisons une extension de Taylor de $f(x-ht)$, on trouve

$$\begin{aligned} E\left(\widetilde{F_n}(x)\right) &= F(x-h) + \int hK(t)f(x-ht)dt \\ &= F(x) - hf(x) + \frac{1}{2}h^2 f'(x) + o(h^2) \\ &+ \int hK(t)\left(f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + o(h^2)\right)dt \\ &= F(x) + \frac{1}{2}h^2 f'(x)\mu_2(k) + o(h^2). \\ E\left(\widetilde{F_n}(x)\right) - F(x) &= \frac{1}{2}h^2 f'(x)\mu_2(k) + o(h^2) \end{aligned}$$

où

$$\mu_j(k) = \int_{-\infty}^{\infty} x^j k(x)dx \text{ for } j \geq 1.$$

■

– **Variance**

$$Var\left(\widetilde{F_n}(x)\right) = \frac{1}{n}F(x)\left(1-F(x)\right) - \frac{h}{n}f(x)\varphi\left(k\right) + o\left(\frac{h}{n}\right).$$

Preuve.

$$Var\left(\widetilde{F_n}(x)\right) = E\left(\widetilde{F_n}(x)\right)^2 - \left(E\left(\widetilde{F_n}(x)\right)\right)^2$$

Nous ne traitons que du premier terme puisque la deuxième est donné

$$\begin{aligned} E\left(\widetilde{F_n}(x)\right)^2 &= \int_{-\infty}^{\infty}\left(K\left(\frac{x-y}{h}\right)\right)^2 f(y)dy \\ &= F(x-h) + \int hK^2(t)f(x-bt)dt \end{aligned}$$

Utilisation de la propriété $K(t) = 1 - K(-t)$, et l'expansion de la série Taylor que nous obtenons

$$\begin{aligned}
E\left(\left(K\left(\frac{x-X_i}{h}\right)\right)^2\right) &= F(x-h) + \int h\left(1-K(-t)\right)^2 f(x-ht)dt \\
&= F(x-h) + \int hf(x-ht)dt + \int hK^2(-t)f(x-ht)dt \\
&- 2\int_{-1}^{1} hK(-t)f(x-ht)dt \\
&= F(x-h) - F(x-h) + F(x+h) + \int hK^2(t)\left\{f(x)+o(1)\right\}dt \\
&- 2\int hK(t)\left\{f(x)+o(1)\right\}dt \\
&= F(x) + hf(x) + hf(x)\int K^2(t)dt - 2hf(x)\int K(t)dt + o(h) \\
&= F(x) + hf(x) - 2hf(x) + hf(x)\int K^2(t)dt + o(h) \\
&= F(x) - hf(x) + hf(x)\int K^2(t)dt + o(h)
\end{aligned}$$

Expression pour $Var\left(\widetilde{F_n}(x)\right)$ peut être calculé comme suit :

$$\begin{aligned}
Var\left(\widetilde{F_n}(x)\right) &= \frac{1}{n}\Big[F(x) - hf(x) + f(x)h\int K^2(t)dt + o(h) \\
&- \left(F(x) + \tfrac{1}{2}h^2f'(x)\mu_2(k) + o(h^2)\right)^2\Big] \\
&= \frac{1}{n}F(x)\left(1-F(x)\right) + \frac{h}{n}f(x)\left(\int K^2(t)dt - 1\right) + o\left(\frac{h}{n}\right) \\
&= \frac{1}{n}F(x)\left(1-F(x)\right) - \frac{h}{n}f(x)\varphi(k) + o\left(\frac{h}{n}\right).
\end{aligned}$$

où

$$\varphi(k) = 2\int xK(x)\,k(x)dx.$$

■

Le résultat précédent montre que la variance asymptotique de $\widetilde{F_n}$ est inférieur à la variance du EDF. Il est évident que pour des valeurs plus élevées de h, la quantité $hf(x)\varphi(k)$ Augmente ce qui donne une expression de variance plus faible mais un biais plus important.

Cette observation a des implications importantes pour le choix de paramètre de h.

Plusieurs autres propriétés de l'estimateur $\widetilde{F_n}$ ont fait l'objet d'une enquête. Nadaraya (1964), winter(1973) et Yamato (1973) a prouvé une convergence presque uniforme de$\widetilde{F_n}$ àF, Watson et Leadbetter(1964) normalité asymptotique établie pour $\widetilde{F_n}$, et Winter (1979) a montré que $\widetilde{F_n}$ a la propriété Chung-Smirnov, qui

$$\limsup_{n\to\infty}\left\{\left(\frac{2n}{\log\log n}\right)^{1/2}\sup_{x\in\mathbb{R}}\left|\widetilde{F_n}(x)-F(x)\right|\right\}\leq 1,$$

avec probabilité 1. Reiss (1981) a souligné que la perte de partialité à l'égard de F_n est compensé par un gain de variance. Ce résultat est appelé la déficience de F_n en ce qui concerne : $\widetilde{F_n}$ Falk (1983) a fourni une solution complète à la question de savoir lequel des F_n ou $\widetilde{F_n}$ est le meilleur estimateur de F. En utilisant le concept de déficience relative, les conditions(comme $n\to\infty$) sur K et $h=h_n$ sont dérivés, ce qui permet à l'utilisateur de décider exactement si un estimateur de fonction de distribution de noyau donné doit être préféré à l'EDF.

Azzalini (1981) dérivé également d'une expression asymptotique de l'erreur quadratique moyenne MSE de $\widetilde{F_n}\ (x)$ et déterminé le paramètre de lissage asymptotiquement optimal, pour avoir un MSE plus bas pour F_n, , et il a obtenu les expressions asymptotiques de l'erreur quadratique intégrée moyenne $MISE$ de $\widetilde{F_n}\ (x)$, fou plus de détails voir (Mack, 1984, et Hill, 1985).

Afin de proposer des méthodes d'estimation de la bande passante, des mesures d'écart qui quantifient la qualité de la bande passante $\widetilde{F_n}$ comme estimateur pour$F$ doit être introduit. L'une de ces mesures est l'erreur quadratique moyenne , qui, dans le cas de l'estimateur de la fonction de distribution du noyau, est définie comme suit :

$$\begin{aligned} MSE\left(\widetilde{F_n}(x)\right) &= E\left\{\left(\widetilde{F_n}(x) - F(x)\right)^2\right\} \\ &= Bias^2\left(\widetilde{F_n}(x)\right) + Var\left(\widetilde{F_n}(x)\right) \\ &= \frac{1}{4}f'^2(x)h^4\mu_2^2(k) + \frac{1}{n}F(x)\left(1 - F(x)\right) - \frac{h}{n}f(x)\varphi\left(k\right) + o\left(h^4 + \frac{h}{n}\right), \end{aligned} \tag{2.2}$$

et l'expression asymptotique de la $MSE\left(\widetilde{F_n}(x)\right)$ est :

$$AMSE\left(\widetilde{F_n}(x)\right) = \frac{1}{4}f'^2(x)h^4\mu_2^2(k) + \frac{1}{n}F(x)\left(1 - F(x)\right) - \frac{h}{n}f(x)\varphi\left(k\right).$$

L'erreur quadratique intégrée moyenne asymptotique($AMISE$) est trouvé en intégrant le $AMSE\left(\widetilde{F_n}(x)\right)$ qui est

$$AMISE\left(\widetilde{F_n}(x)\right) = \int\left(\frac{1}{4}f'^2(x)h^4\mu_2^2(k) + \frac{1}{n}F(x)\left(1 - F(x)\right) - \frac{h}{n}f(x)\varphi\left(k\right)\right)dx.$$

– La valeur de h qui minimise le $AMSE\left(\widetilde{F_n}(x)\right)$ est :

$$\widetilde{h} = \left(\frac{f(x)\varphi(k)}{nf'^2(x)\mu_2^2(k)}\right)^{1/3}.$$

– La valeur de h qui minimise la $AMISE$ peut être calculé en dérivant l'expression de la $AMISE\left(\widetilde{F_n}(x)\right)$, définir l'équation sur 0 et le résoudre pour h. Le résultat est appelé

$$\widetilde{h}_{opt} = \left(\frac{\varphi(k)}{n\mu_2(k)^2\int f'(x)^2dx}\right)^{1/3}. \tag{2.3}$$

1. L'amélioration de $\widetilde{F_n}\ (x)$ est inversement proportionnelle à $\rho\left(f'\right) = \int\acute{f}(x)^2dx$. Nous nous attendons donc à ce que les gains soient minimes lorsque la densité est forte.

2. Le choix du noyau k n'affecte que le $AMISE$ par $\varphi(k)$ (des valeurs plus élevées réduisent la valeur $AMISE$).
3. L'estimateur $\widetilde{F_n}(x)$ est asymptotiquement plus efficace que le $F_n(x)$ voir (Swanapoel, 1988).

2.2.4 Choix du paramétre de lissage

Le paramètre de lissage est le second élément de la méthode d'estimation à noyau. Ce paramètre est indispensable pour la convergence de l'estimateur à noyau et donc l'efficacité du lissage et la qualité de l'estimation (voir figure (2.4)). En pratique, pour évaluer une paramétre de lissage globale optimale (2.5), nous devons développer une méthode pour remplacer la distribution réelle par son estimateur. Plusieurs méthodes existent déjà d'obtenir différents sélecteurs de bande passante en fonction des détails de la procédure développé pour minimiser (2.4) sans avoir besoin d'une estimation supplémentaire de la distribution Dérivés.

Méthode de validation croisée par moindres carrés (see Bowman1984)

Cette méthode appellée aussi méthode de validation croisée non biaisée (Unbiaised Cross-Validation "UCV") proposée par Rudemo (1982) [42] et Bowman (1984). Le principe de cette méthode est la minimisation d'un estimateur de l'erreur quadratique moyenne intégrée MISE par rapport à hn. En effet, le MISE dépend de la fonction inconnue F . On va remplacer MISE (2.2) par une fonction de h_n, mesurable par rapport à l'échantillon et dont la valeur , pour chaque $h_n > 0$, est un estimateur sans biais de MISE(h_n), pour cela,on a

$$CV_B(h) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\left(\mathbb{1}_{[0,+\infty[}(x-x_i) - \widehat{F}_{-i}(x)\right)^2 dx,$$

Méthode Plug-in

L'ideé de base de la procédure de Plug-in pour le choix du paramètre h, est d'estimer dans l'expression de h_{opt} théorique (2.3), la quantité inconnue.$\int f^{'}(x)^2 dx$ En effet on suppose que $f(x)$ appartient à une famille de distributions normales $N(\mu, \sigma^2)$, de moyenne μ et variance σ^2 inconnues. Sous cette hypothèse

$$\hat{\Psi}_{\mathrm{r}} = \frac{(-1)^{r/2} r!}{(2\hat{\sigma}(x_i))^{r+1} (r/2)! \pi^{1/2}},$$

ou

$$\hat{\sigma}(x_i) = min \left(\hat{s}, \frac{Q_3 - Q1}{1.349} \right)$$

Exemple 2.2 *Le paramètre de lissage donneé par la méthodes de plug-in pour un noyau gaussien est défini par*

$$h_n^{opt} = 1.06\, \widehat{\sigma}\, n^{-1/5}$$

Chapitre 3

Simulation

Dans ce chapitre, nous présentons les résultats d'une application numérique qui a été menée pour étudier l'influence des deux choix paramètre de lissage h et de la fonction de noyau k sur le comportement de l'estimateur à noyau de la distribution, afin d'étudier divers cas nous générons un échantillon aléatoire de taille $n = 200$ à partir de deux distributions exponentielle de parametre 2 et Normal de parametre $(0, 1)$, on utilise deux noyaux différents Biweight et Triweight pour la valeur de h on calcule h_{opt} et h_{cv} à l'aide de package "kerdiest".

3.1 L'influence de choix du paramètres à le performance de l'estimateur à noyau de distribution

– **L'influence de choix de noyau**

Dans ce cas on fixé la valeur de $h = 0.1253$ et on change le noyaux k, d'après la figure (3.1), on remarque que l'estimateur est bien performant pour chaque noyau utilisé, ce qui signifie que le choix de noyau a une faible influence.

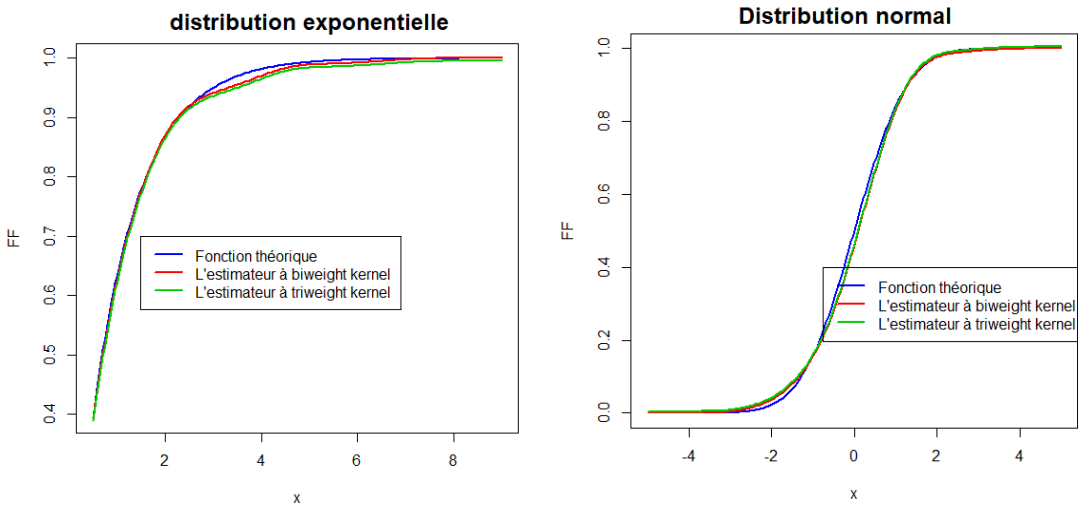

FIG. 3.1 – L'influence de choix de noyau

– **L'influence de choix de paramètre de lissage**

Dans ce cas on fixé on fixe le noyaux $k(t) = \frac{35}{32}\left(1-t^2\right)^3 I\left(|t| \leq 1\right)$, et on calcule h_{opt} et h_{Al}, d'apres la figure (3.2), on remarque que le choix de h a une forte influence sur la qualité de l'estimateur.

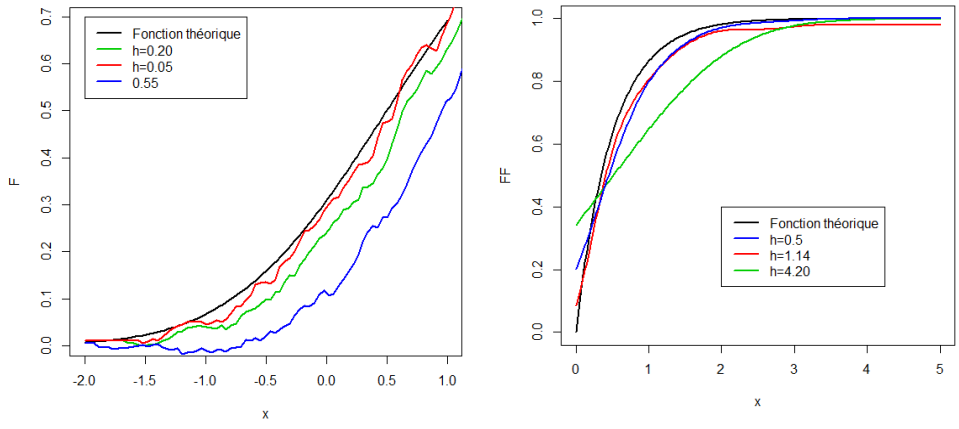

FIG. 3.2 – L'influence de choix de paramètre de lissage

3.2 Performance des différents estimateurs non parametriques

Les résultats de la simulation mesure la performance des différents estimateurs non parametriques dans la signification du biais et de la Mse, sont résumées dans les tableaux suivante pour deux distribution exponentielle (2) et distribution Normal $(0, 1)$, on utilise le noyau triweight. Les résultats sont redimensionnés par le facteur 0,001. Après avoir examiné les tableaux, nous pouvons voir que l'estimateur à noyau de la distribution a un biais et Mse plus faible que l'estimateur empirique.

Estimateur\ h	hopt=0.2681	hcv=0.2895	hal=0.1445
$\widetilde{F_n}(x)$	1.495(0.2618)	1.387(0.3077)	1.577(0.5910)
$F_n(x)$	2.015(0.6617)	2.015(0.6617)	2.015(0.6617)

TAB. 3.1 – Bias (Mse) de distribution exponentielle (2) pour le noyau triweight

Estimateur\ h	hopt=0.7562	hcv=0.8176	hal=0.3445
$\widetilde{F_n}(x)$	1.694(1.111)	1.740(1.336)	1.846(1.415)
$F_n(x)$	3.468(1.257)	3.468(1.257)	3.468(1.257)

TAB. 3.2 – Bias (Mse) de distribution Normal (0,1) pour le noyau triweight

3.3 Application sur les donnée réelles

Les données décrivent une étude de rémission (en mois) d'un échantillon aléatoire de 126 patients atteints de cancer rapportée dans Lee et Wang [14] resumée dans le tableau suivant Nous comparons les performances de l'estimateur à noyau et de l'estimateur empirique par la représentation graphique (), les resultats montrent que les deux estimateurs sont bien performant mais l'estimateur à noyau il est plus lisse.

0.08	2.09	3.48	4.87	6.94	3.52	6.97	13.29	2.26	5.06	9.22	25.74	2.46
8.66	13.11	23.63	0.20	2.23	4.98	9.02	0.40	3.57	7.09	13.80	0.50	3.64
3.70	9.74	0.81	5.32	14.77	3.88	10.34	0.90	5.34	15.96	2.69	7.62	2.75
5.17	14.76	2.62	7.32	32.15	5.32	14.83	2.69	7.59	36.66	4.23	10.75	1.19
7.28	26.31	3.82	10.06	2.64	7.39	34.26	4.18	10.66	1.05	5.41	16.62	43.01
79.05	2.87	7.87	17.36	3.02	1.46	18.10	4.34	5.85	11.98	1.76	3.31	12.63
1.35	5.62	11.64	1.40	11.79	7.93	5.71	4.40	8.26	19.13	4.51	21.73	12.07
6.54	12.03	37	46.12	1.26	14.24	2.83	5.49	12.02	17.14	11.25	2.54	8.37
8.53	9.47	17.12	6.93	6.76	25.82	4.33	3.36	2.02	2.07	7.66	0.51	6.25
5.09	7.26	4.26	5.41	8.65	7.63	22.69	20.28	3.36				

TAB. 3.3 – Tableau des données réelles

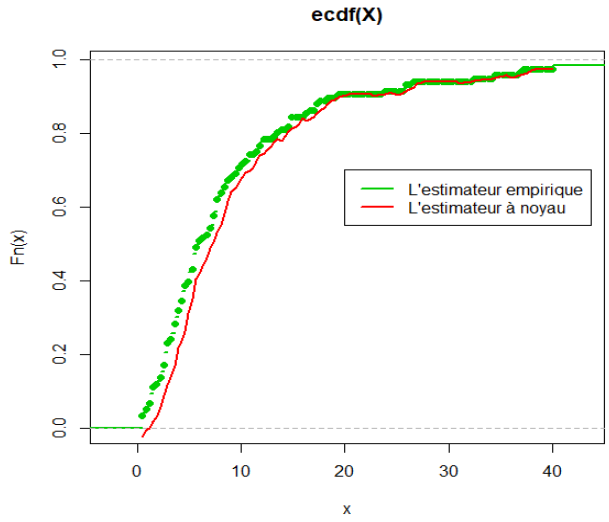

FIG. 3.3 – Comportement de l'estimateur empirique et l'estimateur à noyau pour les donneé réelles

Conclusion

Nous avons présenté les différentes méthodes d'estimation de la fonction de distribution à savoir l'estimateur empirique et l'estimateur à noyau. L'estimateur à noyau était le meilleur estimateur par rapport à l'estimateur empirique. On a vu que l'estimateur à noyau de la distribution dépend de deux paramétres la fenêtre h et le noyau k, une étude numérique réalisé pour étudier l'influence de ses parametres. Une comparaison à l'aide de Bias et Mse montre que l'estimateur à noyau de la fonction de distribution est le meilleur que l'estimateur empirique

En conclusion, nous disons que le choix du noyau n'a pas d'influence majeure sur la qualité de l'estimateur par contre le choix de la fenêtre h est crucial, ceci est illustré par les résultats de simulations.

Bibliographie

[1] Azzalini, A., A note on the estimation of a distribution function and quantiles by a kernel method, Biometrika, vol. 68, pp. 326 − 328, 1981.

[2] Cowling, A., and Hall, P., "On Pseudodata Methods for Removing Boundary Effects in Kernel Density Estimation", Journal of the Royal Statistical Society, Ser. B, vol. 58, pp. 551 − 563, 1996.

[3] Koláček, J. and Karunamuni, R.J., On boundary correction in kernel estimation of ROC curves, Austrian Journal of Statistics, vol. 38, pp. 17 − 32, 2009.

[4] Nadaraya, E.A., Some new estimates for distribution functions, Theory of Probability and its Application, vol. 15, pp. 497 − 500, 1964.

[5] Parzen, E., On estimation of a probability density function and mode. Annals of Mathematical Statistics, vol. 33, pp. 1065 − 1076, 1962.

[6] Reiss, R.D., Nonparametric estimation of smooth distribution functions, Scandinavian Journal of Statistics, vol. 8, pp. 116 − 119, 1981.

[7] Rosenblatt, M., Remarks on Some Nonparametric Estimates of a Density Function, Annals of Mathematical Statistics, vol. 27, pp. 832 − 837, 1956.

[8] Silverman, B.W., Density Estimation for Statistics and Data Analysis, London : Chapman and Hall, 1986.

[9] Tenreiro, C., Boundary kernels for distribution function estimation, Statistical Journal, CMUC, vol. 11, pp. 169 − 190, 2013.

[10] Watson, G.S. and Leadbetter, M.R., Hazard analysis II, Sankhyā Ser. A, vol. 26, pp. $101 - 116$, 1964.

[11] Winter, B.B., Convergence rate of perturbed empirical distribution functions, Journal of Applied Probability, vol. 16, pp. $163 - 173$. 1979.

[12] Zhang, S. and Karunamuni, R.J. On Nonparametric Density Estimation at the Boundary, Nonparametric Statistics, vol. 12, pp. $197 - 221$, 2000.

[13] Yamato, H., Uniform convergence of an estimator of a distribution function, Bulletin of Mathematical Statistics, vol. 15, pp. $69 - 78$, 1973.

Annexe : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

Notation	**Signification**
Ω	un ensemble fondamental
$\mathcal{F}$	ensemble desévénements
P	loi de probabilité
$(\Omega, \mathcal{F}, P)$	espace probabilisable
v.a.r	variable aléatoire réelle
i.i.d	indépendant et identiquement distribuées
f	densité de probabilité
F	fonction de répartition
$E(X)$ ou μ	espérance mathématique ou moyenne de X
Var ou σ^2	variance
$\mathbb{R}$	ensemble des nombres réels
$\mathbb{N}$	ensemble des nombres naturel
MSE	erreur quadratique moyenne
$MISE$	erreur quadratique moyenne intégrée
$AMISE$	erreur quadratique moyenne intégrée asymptotique
k	noyau
h_{opt}	paramétre de lissage (h) optimale
eff	effecacité
$\widehat{f}_n$	estimateur de la densité par noyau

Notation	**Signification**
$\mathcal{N}(\mu, \sigma^2)$	loi deloi normale (ou de Gauss) à deux paramètre $\mu \in \mathbb{R}$ et $\sigma^2 > 0$
$\mathcal{N}(0, 1)$	oi normale standard (centrée réduite)
$\xi(\lambda)$	loi exponentielle
$\xrightarrow{P}$	convergence en probabilité
$\xrightarrow{L}$	Convergence en loi
$\xrightarrow{Mq}$	Convergence en moyenne quadratique
$\xrightarrow{p.s}$	convergence presque sûre
θ	paramètre inconnu
Θ	ensemble des valeurs possibles du paramètre θ
T	un estimateur
$\overline{X}$	estimateur de la moyenne
S^2	estimateur de la variance
$\widetilde{S}^2$	estimateur sans biais de la variance

Résumé

-Dans ce travail, nous étudions l'estimation de la fonction de distribution avec un échantillon indépendant identiquement distribué par la méthode du noyau, où nous discutons d'abord des principales caractéristiques de l'estimation paramétrique et non paramétrique et nous nous concentrons ensuite sur l'estimation non paramétrique en identifiant l'effet de l'estimateur du noyau et ses principales caractéristiques.

Enfin, nous avons effectué une simulation à l'aide du logiciel R, qui nous permet d'observer l'effet de la fenêtre et le noyau.

Mots clé: Estimation non paramétrique, Fonction de distribution, Estimateur à noyau.

Abstract

-In this work, we study the estimation of the distribution function with an independent identically distributed sample in the kernel method, where we first discuss the main concepts of parametric and non-parametric estimation and then focus on non-parametric estimation by identifying the effect of the kernel function and its main characteristics.

Finally, we performed a simulation using R software, which allows us to observe the effect of the window and the kernel

Key words: Non-parametric estimation, Distribution function, Kernel estimateur

ملخص

-في هذا العمل ، سندرس تقدير دالة التوزيع من خلال عينة مستقلة موزعة بشكل متطابق بطريقة النواة ، حيث بداية تطرقنا لاهم المفاهيم الاساسية في التقدير البارامتري و غير البارامتري ثم تم التركيز على التقدير غير البارامتري من خلال تحديد تاثير دالة النواة و خصائصها الاساسية.

في الاخير قمنا بعملية محاكاة بواسطة برنامج R ، الذي يسمح لنا بمراقبة تاثير النافذة و النواة

الكلمات المفتاحية : تقديرغيرالبارامتري، دالة التوزيع، تقدير النواة