République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Seddik BenYahia - Jijel
Faculté des Sciences Exactes et Informatique
Département de Mathématiques

Mémoire de fin d'études

Présenté pour l'obtention du diplôme de

Master

Spécialité : Mathématiques.

Option : Probabilités et Statistique.

Thème

l'estimation de la fonction de quantile par la méthode de noyau

Présenté par :

Benamira Asma

Devant le jury :

Président :	**H. Cheraitia**	M.C.A Université de Jijel
Encadreur :	**M. Madi**	M.A.A Université de Jijel
Examinateur :	**Z. Djeridi**	M.C.B Université de Jijel

Promotion **2021/2022**

Remerciements

Mes remerciements vont tout premièrement à **Allah** tout puissant pour la volonté, la santé, et la patience qu'il m'a donné durant ces longue années d'étude et le courage pour terminer ce mémoire.

Ensuite, j'offre mes remerciements à mon encadreur Mm **Madi Meriem**, qui n'a pas été avare de toutes les connaissances que Allah lui a données, Et les instructions que vous nous avez données pour faire ce travail.

Je remercie les éminents enseignants **Mr Cheraitia Hassen** et **Mm Djeridi Zohra** du jury qui ont suivi mon travail et qui ont pris leur temps et leurs efforts pour le lire malgré leurs préoccupations et les circonstances, afin de le corriger et de le faire réussir.

Un grand merci à tous les enseignants du département de mathématiques de l'université de Jijel qui nous ont suivis pendant cinq années d'études à l'université, Je remercie tout particulièrement l'enseignante **Mm Boutana Imen** et enfin je tiens à remercier tous ceux qui ont contribué d'une manière ou d'une autre à la réalisation de ce travail.

DÉDICACE

À l'abondance de donner sans rien attendre en retour, celle qui m'a comblé de sa tendresse et de son amour après que j'ai détesté ma vie dans les affres de ce travail, **Ma mère**, qui ne remplira pas son droit quoi que je fasse.

À **mon père**, que Allah prolonge sa vie.

À mes merveilleuses amies **Afaf** et **Karima**.

À la joie de la maison **mes frères** chacun en son nom et lieu.

À ma chère enseignante **Mm Sellami Nawel.**

TABLE DES MATIÈRES

LISTE DES TABLEAUX

TABLE DES FIGURES

Notations

Nous utiliserons les notations suivantes tout au long de ce travail.

k	:	Un noyau.
h_n	:	La fenêtre, ou paramètre de lissage.
θ	:	Le paramètre de la distribution.
$\bar{x}$	:	La moyenne arithmétique de l'échantillon.
σ^2	:	La variance de la population.
μ	:	La moyenne de la population.
S^2	:	La variance de l'échantillon.
f	:	La densité de probabilité.
$\hat{f_n}$	:	L'estimateur par la méthode de noyau de la densité de probabilité.
F	:	Fonction de répartition.
F_n	:	La fonction de répartition empirique.
$\mathbb{1}_A$	:	La fonction indicatrice de l'ensemble A.
$AMSE$	:	L'erreur quadratique moyenne asymptotique.
h_{opt}	:	La fenêtre optimale.
f''	:	La deuxième dérivé de f.
$AMSE_{opt}$	:	L'erreur quadratique moyenne asymptotique optimale.
eff	:	L'efficacité d'un noyau par rapport au noyau d'Epanechnecov.
K_{as}	:	Noyau asymétrique.
Q_X	:	La fonction de quantile.
$\mathcal{G}(p)$	:	La loi géométrique.
Q_n	:	Estimateur empirique de la fonction quantile.
$\tilde{Q}_n$	:	Estimateur de la fonction quantile par la méthode de noyau.
F_θ	:	Une distribution de paramètre θ.
$\mathcal{F}$	:	Une famille de distributions.
F_θ^{-1}	:	L'inverse de la distribution de paramètre θ.
$\hat{\theta}$	:	Estimation de paramètre θ.
$\mathrm{F}_{\hat{\theta}}$	:	L'estimation de la distribution de paramètre θ.
$X_{(i)}$	:	Un statistique d'ordre.
$i.i.d.$	:	Indépendant identiquementes distribuées.
$\hat{F}_n$	:	Estimateur de la fonction de répartition empirique.

$IMSE$	:	L'erreur quadratique moyenne intégrée
$IMSE_{opt}$	:	L'erreur quadratique moyenne intégrée optimale.
ξ_{np}	:	Le quantile empirique d'ordre p.
$\hat{\xi}_{np}$	:	Le quantile empirique d'ordre p de l'échantillon.
k_*	:	C'est un noyau avec un rang m symétrique autour de zéro.
$Q^{'}, Q^{''}$	:	Le premier et la deuxième dérivé de la fonction quantile.
b_{opt}	:	La fenêtre optimale asymptotique de $Q^{''}_m(p)$.
a_{opt}	:	La fenêtre optimale asymptotique de $Q^{'}_m(p)$.
$k^{'}$	:	La dérivé de noyau k.
ps	:	Presque surement.
mq	:	Moyenne quadratique.
pr	:	Probabilité.
MSE	:	L'erreur quadratique moyenne.
$\hat{h}_{opt}$	:	L'estimateur de la fenêtre optimale de l'estimateur $\hat{Q}_n$.

INTRODUCTION GÉNÉRALE

L'inférence statistique est définie comme l'utilisation d'un échantillon aléatoire de la population d'étude pour arriver à des généralisations ou des caractéristiques inconnues dans la population. L'une des méthodes les plus importantes d'inférence statistique est l'estimation statistique. Lorsque nous étudions un phénomène à partir des données communautaires, nous obtenons des caractéristiques souvent inconnues. Pour les calculer, nous devons disposer de données sur tout le vocabulaire de la population, ce qui demande des efforts, du temps et un coût élevé. Nous effectuons le processus d'estimation à partir des données de l'échantillon sélectionné au hasard pour représenter la population en termes réels. La distribution de la population peut être connue, et donc le calcul dépend des méthodes paramétriques usuelles connues. Mais dans le cas où il est difficile de déterminer la distribution de probabilité, cela nécessite l'adoption de méthodes plus souples que celles paramétriques pour l'analyse des données, on applique les méthodes non paramétrique. Ces méthodes statistiques peuvent être utilisée pour obtenir des conclusions sur la population étudiée de l'échantillon, quelle que soit la distribution théorique de cette population. L'estimation non paramétrique ne nécessitent aucune hypothèse ou information sur les caractéristiques de distribution de la population, de plus le temps nécessaire pour analyser les données, est inférieur au temps d'analyse pour l'estimation paramétrique [17].

L'estimation de quantiles ou fonction quantile est un problème de base en statistique. Cette estimation a plusieurs domaines d'application, par exemple, en climatologie, en hydrologie, en assurance et en finance, etc.

Dans ce mémoire, on s'intéresse à l'estimation à noyau de la fonction de quantile. L'estimateur à noyau de la fonction de quantile est basé sur l'estimateur de la fonction de répartition introduit par *Nadaraya* (1964) [22]. De nombreux chercheurs ont étudie cet estimateur et ces propriétés, comme *Falk*(1984) [9] qui a montré que la performance asymptotique de l'estimateur à noyau est meilleure par rapport à celle de l'estimateur empirique de la même fonction. *Yang*(1985) [29] a établi la normalité asymptotique et la consistance en moyenne quadratique du même estimateur. *Sheather* et *Marron*(1990) [28] ont donné l'expression de l'erreur moyenne quadratique (MSE) et d'autres comme *Yamato*(1973), *Parzen*(1979) [23], *Azzalini*(1981) [3], *Harrell* et *Davis*(1982) [12], *Padgett*(1986), *Ralescu* et *Sun*(1993) et *Park*(2006).

Ce mémoire contient une introduction, deux chapitres et une conclusion.

Dans le premier chapitre, on a étudié la méthode d'estimation non paramétrique dite à noyau. On a donné une description de cette méthode et des estimateurs à noyau de la fonction de densité et de répartition. Ainsi on a parler des propriétés des noyaux les plus utilisé dans ce type d'estimation.

Dans le deuxième chapitre, on a abordé la définition de la fonction de quantile, puis on a donné l'expression de l'estimateur à noyau de cette fonction, ensuite on a étudié les propriétés asymptotique de l'estimateur.

CHAPITRE 1

MÉTHODE DE NOYAU

1.1 Introduction

Dans ce premier chapitre, on s'intéresse à la description et l'étude de la méthode d'estimation non paramétrique, dite à noyau ou encore méthode de *Parzen – Rosenblatt*, introduite par *Rosenblatt*(1956) [25] puis amélioré par *Parzen*(1962) [23]. Elle se base sur un échantillon d'une population statistique, et permet d'estimer la fonction de densité, de répartition et celle de quantile..., en tout points de support. Les estimateurs à noyau sont fonction de deux paramètres le noyau k et le paramètre de lissage h. L'estimation par noyau est la plus populaire parmi les méthodes d'estimation non paramétrique car elle semble commode, robuste et ne nécessite pas un choix multiple de paramètres k et h.

1.2 Estimation paramétrique et non paramétrique

Nous avons un modèle statistique $(\mathbb{E}, \mathbb{A}, \mathbb{P})$ où $\mathbb{E}$ est l'espace fondamental ou bien l'ensemble des observations possibles, $\mathbb{A}$ est une tribu sur $\mathbb{E}$, $\mathbb{P}$ est l'ensemble de probabilités ou de distribuions. Soit P une sous famille de $\mathbb{P}$, et considérons $X : \Omega \to \mathbb{E}$ une application mesurable. On peut toujours noter P par $(P_\theta, \theta \in \Theta)$, où Θ est l'ensemble qui définit les paramètres du modèle.

Soit T une application de P dans Θ' l'ensemble des estimateurs. Estimer $T(P)$ c'est-à-dire essayer de l'évaluer au vu de l'observation d'un échantillon de la variable aléatoire X qui est à valeurs dans $\mathbb{E}$. Donc, le paramètre à estimer est l'application :

$$\begin{cases} T : P \to \Theta' \\ P_\theta \mapsto T(P_\theta). \end{cases}$$

L'estimation de h est une fonction $T_n : x \mapsto T_n(\theta, X_1, X_2, ..., X_n)$ mesurable en termes d'observations $X_1, X_2, ..., X_n$.

1.2.1 L'estimation paramétrique (Θ est de dimension finie)

Si nous savons que T appartient à une famille avec un paramètre $\{T(x, \theta), \theta \in \Theta\}$ $\Theta \subset \mathbb{R}^S$ avec s est le nombre de paramètres de la distribuions P, et $T(.,.)$ une fonction bien connue. On parle d'estimation paramétrique, car estimation de T signifie l'estimation d'un paramètre appartenant à un espace fini Θ. Il existe deux façons d'estimer le paramètre de population inconnu : Estimation ponctuelle et estimation par intervalle.

Estimation ponctuelle :

Ce type est destiné à estimer le paramètre de la population inconnu avec une seule valeur. Il est calculé à partir des données de l'échantillon tiré. (Par exemple, lorsqu'un échantillon est tiré de la population et que la moyenne arithmétique $\bar{x}$ de l'échantillon

est calculée, elle est prise comme une estimation ponctuelle de la moyen de la population μ, la variance de l'échantillon S^2 est prise comme une estimation ponctuelle de la variance de la population σ^2, et la proportion du phénomène dans la population est déduite de la proportion de ce phénomène dans cet échantillon) La précision de cette estimation dépend de la nature et de la taille de l'échantillon prélevé dans la population.

Il est préférable de s'appuyer d'abord sur le deuxième type d'estimation statistique, qui est l'estimation de période, car l'estimation ponctuelle est rarement égale au paramètre à estimer.

Estimation par intervalle (estimation de période)

C'est un intervalle de valeurs réelles qui contient le paramètre inconnu, avec une borne supérieure et inférieure et avec une certaine probabilité appelée le niveau de confiance, symbolisé par $(1 - \alpha)$%qui donne notre confiance que ce paramètre va se trouver entre les limites de cet intervalle, avec α est la probabilité que se paramètre n'appartient pas à cet intervalle.

1.2.2 L'estimation non paramétrique (Θ est de dimension infinie)

D'autre part, si nous savons seulement que T appartient à l'ensemble P de distributions de probabilité, qui est un espace de dimension infinie, alors nous disons qu'il s'agit d'une estimation non paramétrique ou fonctionnelle.

Les méthodes non paramétrique permet d'éviter quelques problèmes reliés aux méthodes paramétrique, par exemple, les méthode non paramétrique ne nécessitent pas l'hypothèse sur la distribution de la population. Ainsi ces méthodes d'estimation de la densité, par exemple, consistent a considérer une certaine fonction pour chacune des observations d'un échantillon de données, contrairement à la méthode paramétrique permettant d'ajuster une seule distribution sur l'ensemble des observation.

Récemment, l'intérêt pour les méthodes non paramétriques a augmenté en raison de leur popularité. Parmi les plus importantes de ces méthodes non paramétriques se trouve la méthode dont nous discuterons dans notre mémoire, qui est la méthode du noyau.

Remarque 1.2.1.
Il existe un autre type d'estimation appelé estimation semi paramétrique. Un modèle statistique est semi paramétrique s'il possède à la fois des paramètres de dimension finie et de dimension infinie. Formellement, si m est la dimension de Θ *est n est la taille de l'échantillon, les modèles semi paramétriques et non paramétriques ont* $m \to \infty$ *lorsque* $n \to \infty$ *si* $\frac{m}{n} \to 0$ *alors le modèle est semi paramétrique, si non le modèle est non paramétrique.*

1.3 Estimation par la méthode du noyau

Le concept de noyau a été introduit pour la première fois par *Parzen*(1962) [23] et *Rosenblatt*(1956) [25], mais c'est *Cacoulos*(1966) [4]qui a été le premier à utiliser le terme noyau (kernel) pour définir la fonction utilisée dans les méthodes non paramétriques. En hydrologie statistique, *Yakowitz* et *Adamowski*(1963) et *Flush*(1983) ont présenté indépendamment la méthode de noyau lors d'une conférence L'AGE à l'automne 1983.
C'est la plus utilisée parmi les méthodes d'estimation de densité non paramétriques, les fonctions d'agrégation, la fonction de fiabilité, la régression et autres. La popularité de l'estimateur par noyau peut s'expliquer par au moins trois raisons, la simplicité de sa forme, ses multiples motifs d'affinité, et sa flexibilité, qui s'explique par la liberté de l'utilisateur de choisir : le noyau et la fenêtre. L'estimateur à noyau est une fonction des deux paramètres, du noyau k (la densité de la loi statistique en général) et du coefficient de lissage h (la fenêtre ou bande de fréquence).

La méthode du noyau est une généralisation de la méthode de l'histogramme. Il a donc l'inconvénient de la non-uniformité de la fonction de densité résultante $\hat{f}_n$ l'estimateur de la densité (une fonction constante multi-définie).
Dans le graphique, la densité en un point x est estimée par le rapport des observations

$x_1, ..., x_n$, qui se rapproche de x. Par conséquent, nous dessinons un carré sa largeur est régie par un coefficient de lissage h, puis calculez le nombre d'observations qui appartiennent à ce carré. Cette estimation a de bonnes propriétés statistiques mais elle n'est pas continue. L'estimation par noyau vise à remédier à ce dilemme "trouver la continuité" en remplaçant le carré centré en x et la largeur h par gausse centré en x. Plus l'observation est proche du point x d'appui, plus la courbe en cloche est élevée, ce qui lui donne une grande valeur numérique. Au contraire une valeur numérique négligeable est attribuée aux observations très éloignées, l'estimateur est formé par la somme ou plutôt la moyenne des courbes en cloche. Comme le montre l'image suivante, il est clair qu'il continue. De bonnes références sur la méthode du noyau sont *Lall*(1995) [14] et *Izenman*(1991) [13].

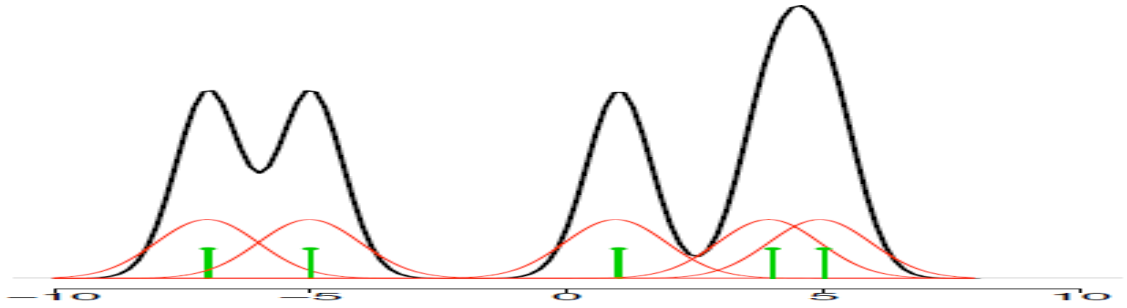

FIGURE 1.1 – Estimateur de densité par noyau $\hat{f}$. [16]

1.3.1 Estimateur de Parsen-Rozenblatt

Soit $X_1, X_2, ..., X_n$ des variable aléatoire identiquement distribuées et indépendantes i.i.d, copies de la variable aléatoire X de fonction de densité de probabilité f et de répartition F inconnues.

L'estimateur à noyau de la fonction de densité de probabilité f, noté $\hat{f}_n$, proposé par *Parzen*(1962) [23], *Rosenblatt*(1956) [25] est donné par :

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h_n}\right),$$

où $(h_n)_{n\geq 1}$ est une suite de nombres réels positifs, vérifiant $h_n \xrightarrow{n\to\infty} 0$, appelé fenêtre ou paramètre de lissage de l'estimateur, et k est un noyau ; une fonction borélienne, positive et intégrable telle que $\int_{\mathbb{R}} k(u)du = 1$. Une première justification concernant la forme de l'estimateur de *Parzen – Rozenblatt*, a été donnée précédemment. Une seconde est basée sur la fonction de répartition empirique associée à $(X_1, X_2, ..., X_n)$.

Fonction de répartition empirique :

La fonction de répartition empirique F_n est un estimateur simple de F [11]. Cette fonction est un très bon estimateur de F, elle est définie pour tout $x \in \mathbb{R}$ dans $]0, 1[$ par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{]-\infty,x[}(x_i).$$

On sait que :

$$\begin{aligned} f(x) &= \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h} \\ &\simeq \frac{F(x+h) - F(x-h)}{2h}. \end{aligned}$$

En remplaçant F par son estimateur F_n, on obtient :

$$\hat{f_n}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h},$$

alors :

$$\begin{aligned} \hat{f_n}(x) &= \sum_{i=1}^{n} \frac{\mathbb{1}_{x-h<X_i\leq x+h}}{2nh} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{1}_{-1<\frac{x-X_i}{h}\leq 1} \\ &= \frac{1}{hn}\sum_{i=1}^{n} k\Big(\frac{x-X_i}{h}\Big). \end{aligned}$$

Posons $u = \frac{x-X_i}{h}$, alors $k(u) = \frac{1}{2}\mathbb{1}_{]-1,1]}(u)$ tel que k est le noyau uniforme.

Basons sur $\hat{f_n}$, *Nadaraya*(1964) [22] a proposé un estimateur de la fonction de répartition F, noté $\hat{F}_n$, défini comme suit :

$$\begin{aligned}\hat{F}_n(x) &= \int_{-\infty}^{x} \hat{f_n}(t)dt \\ &= \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),\end{aligned}$$

où K_h est le noyau intégré défini par :

$$K_h(x) = \frac{1}{h_n}\int_{-\infty}^{x} k(\frac{t}{h})dt.$$

1.3.2 La fenêtre h et le noyau k

La fenêtre h :

Ce paramètre est positif, son but ou le but de son estimation est de modifier les données de manière à obtenir des estimateurs dont les caractéristiques convergent avec les propriétés des paramètres réels. On so concentre sur chaque observation et on détermine le degré d'homogénéité de l'estimation de la fonction de densité. Plus h est grand, plus l'estimateur est régulier, comme dans le cas de l'estimateur d'histogramme, contrairement au cas où h est petit, l'estimateur est irrégulier. Donc ce paramètre détermine le degré d'influence des observations sur l'estimation. Le sur-lissage peut masquer la plupart des propriétés de la vraie fonction de densité telles que l'asymétrie ou la multimodalité, tandis que le sous-lissage fait apparaître des détails artificiels sur le graphique de l'estimateur. Ce paramètre vérifie la condition suivante $h = h_n \xrightarrow{n\to\infty} 0$. Une manière courante d'obtenir la valeur de h consiste à supposer que l'échantillon est distribué selon une certaine loi paramétrique, par exemple selon la loi normal $N(\mu, \sigma^2)$ donc $h = 1,06\hat{\sigma}n^{-\frac{1}{5}}$. Malheureusement, l'estimation gaussienne n'est pas toujours efficace. Par exemple, lorsque n est petit, une autre façon d'opérer est de chercher à fixer h de manière optimale. En général, h_n est obtenu par des techniques de validation croisée.

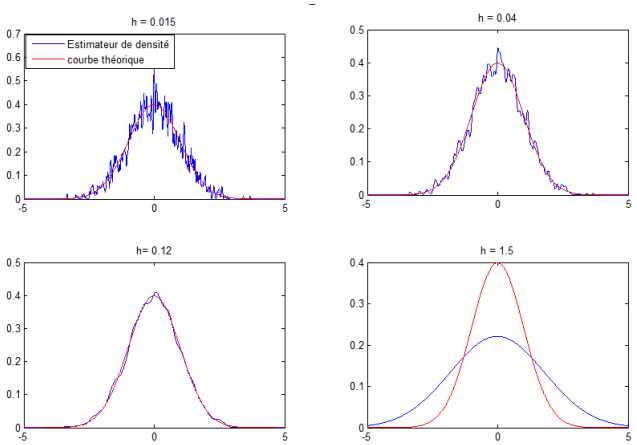

FIGURE 1.2 – Influence du paramètre de lissage sur l'estimation de la fonction de densité. [11]

Le noyau k

Le noyau est l'élément qui détermine la forme des "bosses" qui constituent l'estimation de la densité. En général le choix du noyau n'a pas vraiment d'importance. Plusieurs travaux ont été effectuer pour trouver un noyau optimale, qui serait préférable d'utilisation à tous les autres. *Epanechnikov*(1969) [6] a proposé sous certains critères, un noyau optimal qui porte son nom. *Rao*(1983) est arrivé à la conclusion que le choix d'un noyau autre que le noyau optimal n'entraînait qu'une légère perte de précision. *Lall* et al (1993) [15] considéraient que la sélection du noyau avait une certaine importance, mais que son effet sur l'ensemble de l'estimation était relativement faible.

1.3.3 Noyaux usuels

Noyau triangulaire :

L'avantage de ce noyau par rapport au autres est sa continuité partout, ce qui conduit à un estimateur $\hat{f}_n$ continue. Ce noyau s'écrit sous la forme :

$$k(u) = (1 - |u|)\mathbb{1}_{[-1,1]}(u).$$

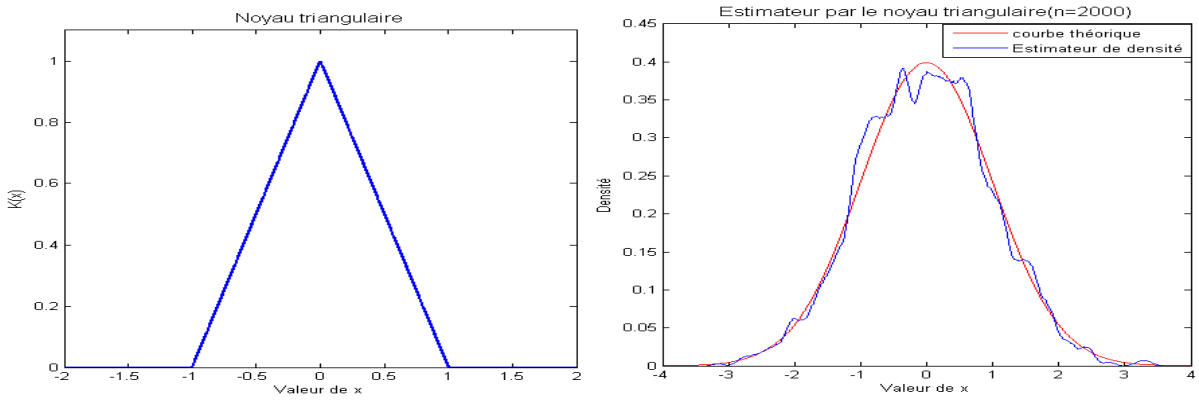

FIGURE 1.3 – Noyau triangulaire. [11]

Noyau d'Epanechnikov ou parabolique :

En 1969 [6], *Epanechnikov*, a donné la forme du noyau k défini par :

$$k(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right)\mathbb{1}_{[-\sqrt{5},\sqrt{5}]}(u)$$
$$= \frac{3}{4}(1-t^2)1_{[-1,1]}(u).$$

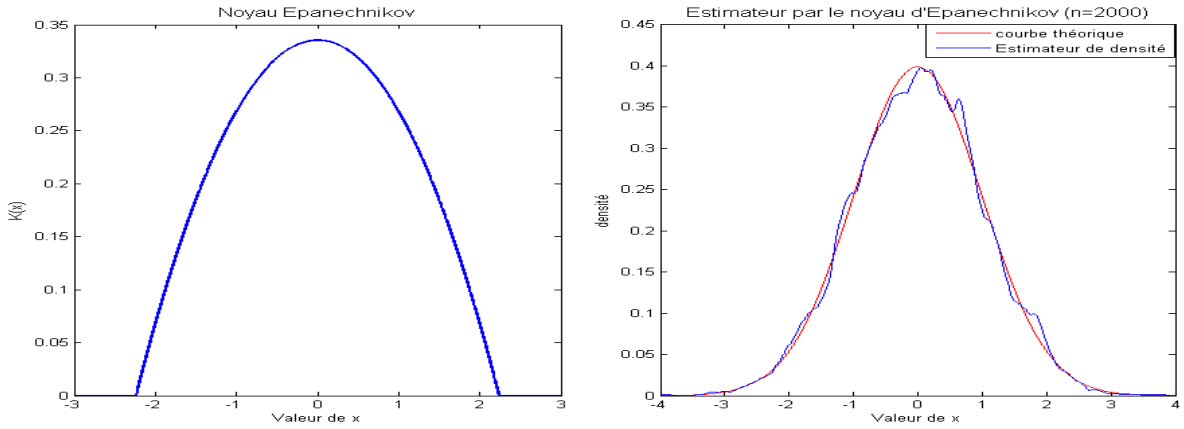

FIGURE 1.4 – Noyau d'Epanechnikov ou parabolique. [11]

Noyau quadratique ou biweight :

Le noyau de biweight est très intéressant car il donne un estimateur dérivable partout. En fait, il s'agit du noyau le plus simple parmi les noyaux de forme polynômial dérivable

partout. Ainsi, il assure le lissage locale de la fonction $\hat{f}_n$. Ce noyau est d'une forme très proche du noyau gaussien, il est donc préférable de l'utiliser. Il s'écrit sous la forme :

$$k(u) = \frac{15}{16}(1-u^2)^2 \mathbb{1}_{[-1,1]}.$$

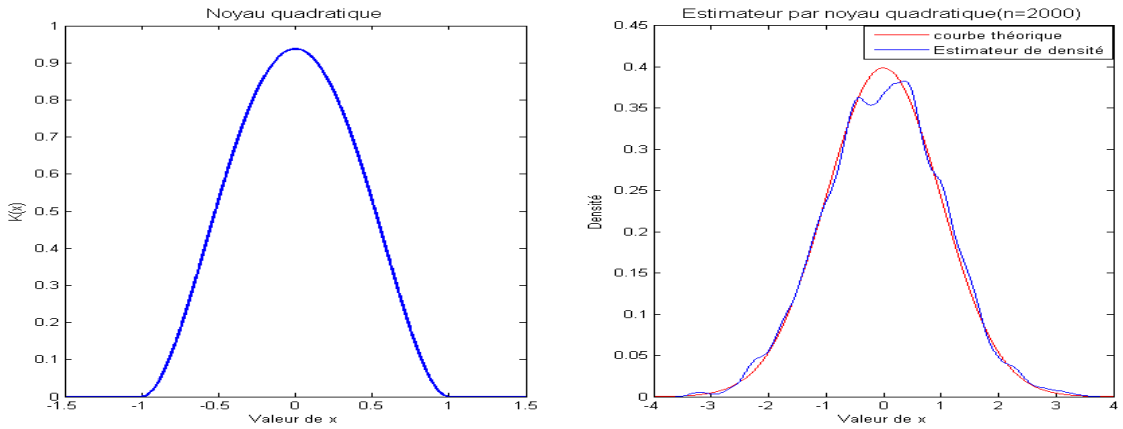

FIGURE 1.5 – Noyau quadratique ou biweight. [11]

Noyau Goussien :

L'avantage du noyau gaussien est que plus la valeur de n est élevée plus on élargit la fenêtre, ce qui a un effet de lissage globale important, mais le coût de calcul dans le cas de ce noyau est très élevé du fait de son support infini. Ce noyau s'écrit sous la forme :

$$k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right), u \in \mathbb{R}.$$

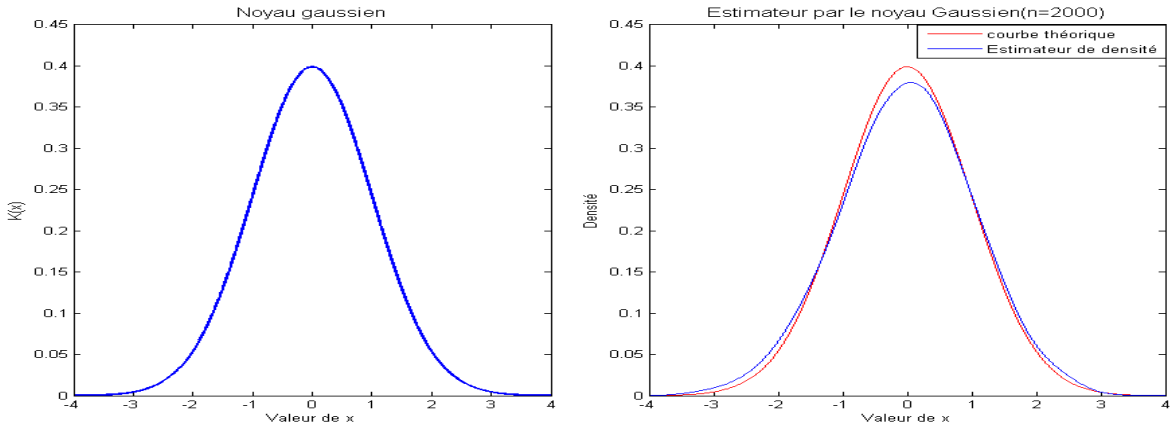

FIGURE 1.6 – Noyau Goussien. [11]

Noyau uniforme ou de Rozenblatt :

Ce noyau est de la forme :

$$k(u) = \frac{1}{2}\mathbb{1}_{[-1,1]}(u).$$

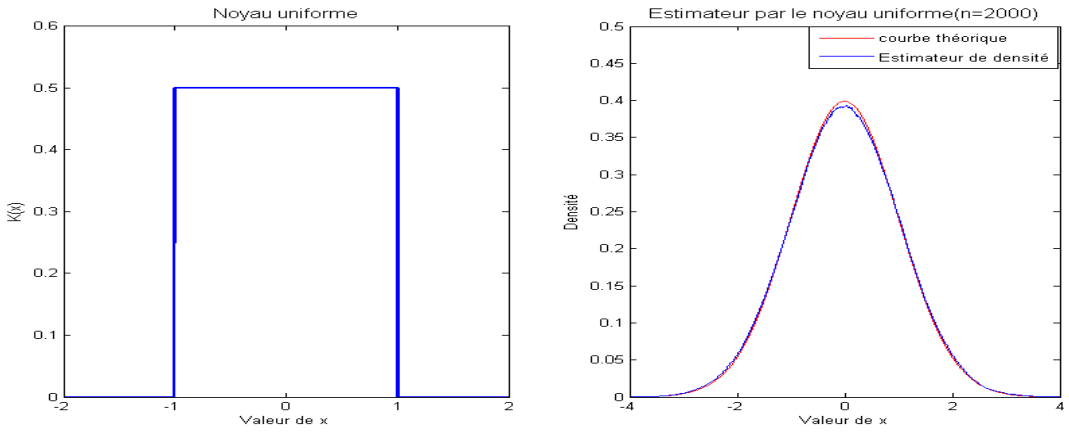

FIGURE 1.7 – Noyau uniforme ou de Rozenblatt. [11]

Noyau sinus :

La forme de se noyau est la suivant :

$$K(u) = \frac{1}{2\pi}\left(\frac{\sin\left(\frac{u}{2}\right)}{\frac{u}{2}}\right)^2, \quad u \neq 0.$$

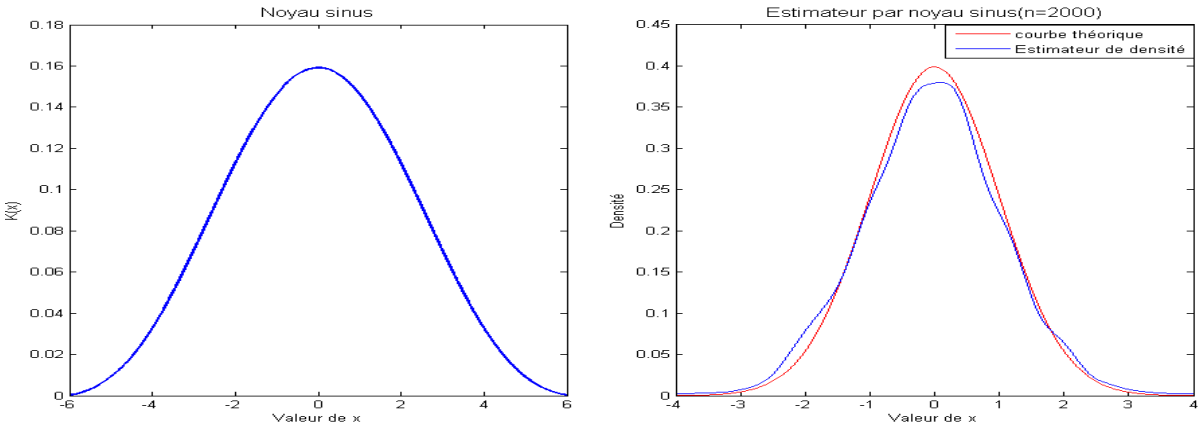

FIGURE 1.8 – Noyau sinus. [11]

Noyau cosinus :

Ce noyau est de la forme :

$$k(u) = \frac{\pi}{4}\cos\left(\frac{\pi u}{2}\right)\mathbb{1}_{[-1,1]}.$$

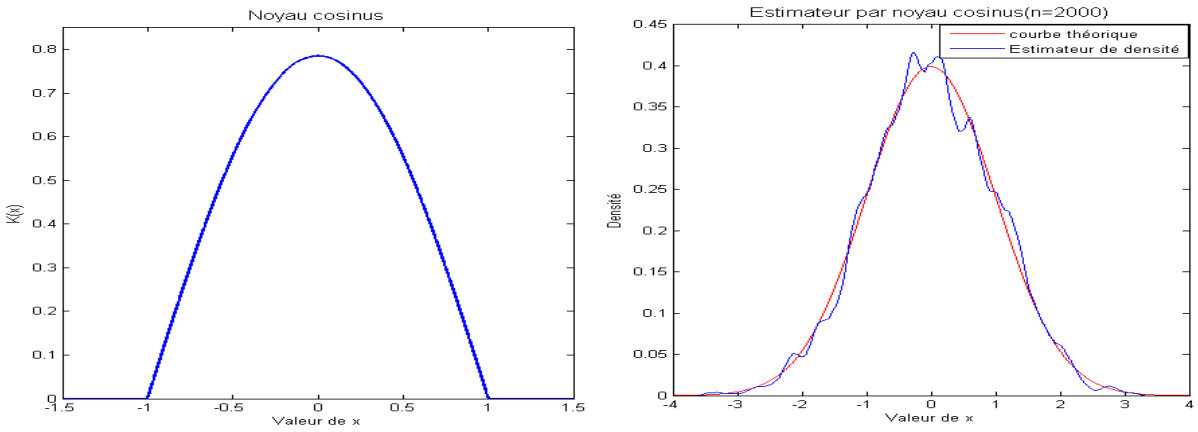

FIGURE 1.9 – Noyau cosinus. [11]

Noyau de Silverman :

Le noyau de Silverman est de la forme :

$$k(u) = \frac{1}{2}\exp\left(\frac{-|u|}{\sqrt{2}}\right)\sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right), \quad u \in \mathbb{R}.$$

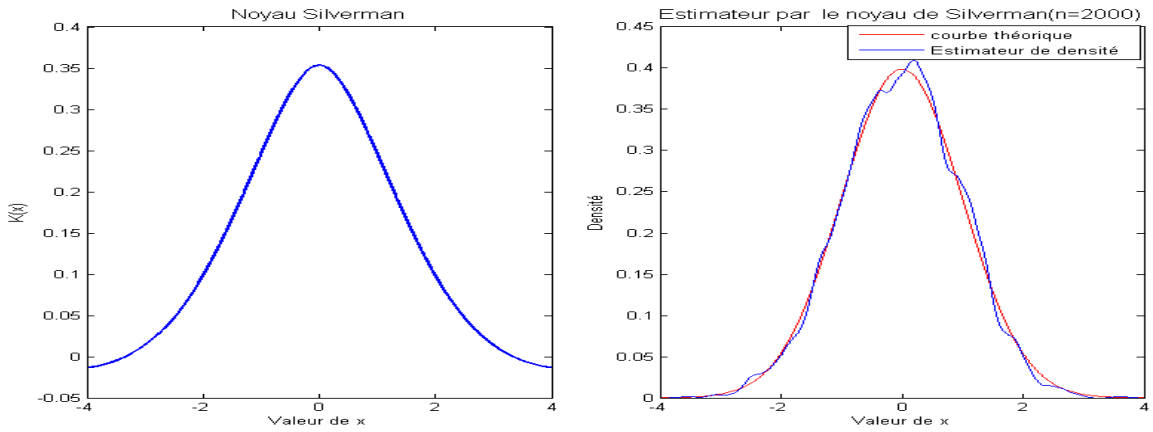

FIGURE 1.10 – Noyau de Silverman. [11]

1.4 Propriétés des noyaux

En partant de la définition même d'une fonction de densité, le noyau doit être positif en tout point et d'intégrale égale à 1. Lorsque l'on utilise un noyau qui répond aux exigences des fonctions de densité, on en déduit que $\hat{f}_n$, l'estimateur à noyau de la densité, est aussi une fonction de densité. De plus, $\hat{f}_n$ hérite de toutes les propriétés de continuité et de différentiabilité de sa fonction noyau (*Silverman*(1986)) [25].

1.4.1 Noyau optimal basée sur le critère de l'*IMSE*

Dans ses travaux, *Epanechnikov* [6] a étudié les propriétés d'une fonction de densité empirique à laquelle des noyaux de forme arbitraire ont été travaillés par d'autres chercheurs. Il a d'abord étudié les propriétés asymptotiques de la fonction de densité empirique, puis a étudié l'erreur causée par l'estimation $\hat{f}_n$ dans la fonction de densité réelle f. En fonction de l'*IMSE*, l'erreur moyenne quadratique intégrée, il a essayé de déterminer la forme de noyau optimale qui peut être utilisée à la place du noyau aléatoire. Il a d'abord imposé des contraintes sur le noyau utilisée uniquement pour simplifier les calculs d'optimisation du noyau (*Silverman*(1986)) [27] c'est être :
1) Symétrique et positif sur sa domaine de définition : $k(u) = k(-u) \geq 0$;

2) Son intégrale sur son domaine est égale à un : $\int_D k(u) = 1$;

3) Centré : $\int_D uk(u)du = 0$;

4) Une variance finie : $\int_D u^2 k(u)du < \infty$;

5) Des moments donnés par : $\int_D u^m k(u)du < \infty \quad 0 \leq m < \infty$.

Après avoir obtenu une approximation de *IMSE* :

$$IMSE(\hat{f_n}) = \frac{1}{4}h^4 k_2^2 \int (f''(x))^2 dx + \frac{1}{nh} \int (k(u))^2 du,$$

et en dérivant par rapport à h, puis l'égalisant à zéro, on obtient la valeur de h_{opt} qui minimise *IMSE* suivante :

$$h_{opt,IMSE} = \left\{ \frac{\int k^2(u)du}{nk_2^2 \int (f''(x))^2 dx} \right\}.$$

Nous ne pouvons pas le calculer directement car il dépend de la fonction de densité théorique inconnue.

En examinant l'expression de $IMSE_{opt}$:

$$IMSE_{opt} \simeq \frac{5}{4n^{\frac{-4}{5}}} k_2^{\frac{2}{5}} \left\{ \int [k(t)]^2 dt \right\}^{\frac{4}{5}} \left\{ \int [f''(x)]^2 dx \right\}^{\frac{1}{5}}.$$

On aperçoit que la seule façon de la réduire est de choisir k(kernel) qui réduit $\int k^2(u)du$ puisqu'on ne peut pas contrôler le terme $f''(x)$ et que le paramètre h a été déjà optimisé. Donc le problème de choisir le noyau optimal en fonction du critère *IMSE* se résume à réduire $\int k^2(u)du$ dans les conditions mentionnées précédemment. En considérant la formule du polynôme d'Euler [6] la fonction résultant de cette optimisation est appelée le noyau d'*Epanechnecov*. Il est tout à fait possible d'évaluer l'efficacité des autres noyaux par rapport au noyau optimal, en comparant la valeur $\int k^2 du$ des deux noyaux. *Silverman*(1986) [27] a déterminé l'efficacité de n'importe quel noyau par rapport au noyau d'*Epanechnecov* comme suit :

$$eff(k) = \frac{3}{5\sqrt{5}} \left\{ \int u^2 k(u)du \right\}^{-\frac{1}{2}} \left\{ \int k^2(u)du \right\}^{-1}.$$

La constante $\frac{3}{5\sqrt{5}}$ est la valeur de $\int k^2(u)du$ pour le noyau de *Epanechnicov*. Plus la valeur de $eff(k)$ est proche de 1, plus le noyau est comparable au noyau optimal. Et il a été montré que la plupart des noyaux peuvent être comparés aux noyaux d'*Epanechnicov*, par exemple, l'efficacité relative du noyau uniforme est 0.930, pour le noyau normal c'est 0.946, pour le noyau triangulaire c'est 0.986, pour le noyau de poids c'est 0.994, pour le noyau cosinus c'est 0.999, et pour le noyau de *Epanechnecov* c'est 1. Il n'est donc pas nécessaire de baser la sélection du noyau sur la base de la minimisation *IMSE* car très peu de précision est perdue en utilisant un noyau non optimal, il est donc préférable de choisir un noyau qui correspond au type d'estimation à effectuer à la place [5].

1.4.2 Autres critères d'optimisation

Nous avons vu que le noyau de *Epanechnecov* est le noyau qui sous-estime la valeur *IMSE* qui est calculée en fonction de l'estimation de la fonction de densité, en hydrologie par exemple on s'intéresse à l'estimation des grandeurs de période de retour plutôt qu'à la fonction de densité elle-même, c'est pourquoi il n'est pas certain que le noyau de *Epanechnecov* soit bien le noyau optimal à utiliser dans ce contexte. Il s'agira d'améliorer la fonction d'erreur calculée à partir de l'estimation des quantile. Dans la littérature, il semble y avoir peu d'études autres que celles d'*Epanechnecov* qui tentent de déterminer le noyau optimal, peut-être parce que la recherche s'est concentrée sur les techniques de calcul du facteur de lissage optimal plutôt que sur le noyau. Étant donné que la plupart des chercheurs dans ce domaine montrent qu'il y a peu d'avantages à sélectionner un noyau idéal si un autre noyau peut être utilisé sans perte significative de fidélité, comme mentionné précédemment, le choix du noyau peut avoir une certaine importance selon le contexte.

1.5 Types de noyaux

La méthode des noyaux a jusqu'à maintenant surtout été utilisée dans un contexte d'interpolation, ce qui pourrait expliquer le fait que dans la littérature on stipule généralement que le choix du noyau est sans importance [5]. Dans les prochaines paragraphes, on a distingué les noyaux par deux facteurs, le support et la symétrie, et on discute des situations où l'on a avantage à utiliser chacun des types.

1.5.1 Noyaux à support fini et à support non-fini

Un noyau à support fini est un noyau borné égal à zéro en dehors du domaine de la définition, par contre un noyau à support infini est asymptotique et donc non nul sur l'ensemble $\mathbb{R}$. Le tableau suivant contient les noyaux déjà mentionnés dans cette étude, classés par type de support [5].

Noyaux à support fini	**Noyaux à support non-fini**
Epanechnikov	Goussien
Cosinus	EVl
Biweight	Silverman
Triweight	Sinus
Triangulaire	Cauchy
Tniforme	

TABLE 1.1 – Noyaux classés selon le support.

Que le noyau soit fini ou non peut être important dans l'estimation, en particulier dans la région des "queues de la fonction de densité" les extrémité à gauche et à droite. Par exemple, lorsque l'on s'intéresse à l'estimation unilatérale à droite d'une fonction dépassant les notes disponibles ,c'est-à-dire que l'on est en extrapolation, le noyau à support fini peut être inefficace. La capacité de extrapolation étant fortement limitée, en revanche, les noyaux d'affinité permettent de sortir plus loin de l'échantillon puisque c'est non nul qu'il dépasse la largeur du paramètre de lissage.

1.5.2 Noyaux symétriques et asymétriques

Lall et al(1993) [15] ont affirmé que les noyaux asymétriques aident à réduire le biais dans l'estimation des observations récentes dans l'échantillon. Cependant, il a été montré que l'utilisation d'un noyau symétrique dans le cas où la densité théorique est symétrique, conduit à un biais nul dans l'original. Il est donc préférable d'utiliser des noyau symétriques dans ce cas.

Deux types d'asymétrie peuvent être identifiés, l'asymétrie positive et l'asymétrie négative. Une distinction est faite entre les deux types par le coefficient d'asymétrie suivante :

$$\int_{-a}^{a} \{k_{as}(t) - E(k_{as}(t))\}^3 dt,$$

tel que k_{as} noyau asymétrique et où [-a, a] est sont domaine de variation s'il est borné. Si ce coefficient est positif, alors la fonction k est asymétrique positive sinon elle est asymétrique négative, bien sûr s'il est nul, alors le noyau est symétrique.

Comme on n'est pas certain de la symétrie de la distribution théorique, il semble y avoir moins de risques à utiliser un noyau asymétrique. Parmi les noyau mentionnés précédemment, seul le noyau EV1 est asymétrique autour de zéro, comme le montre sa figure selon laquelle l'extrémité droite est relativement plus lourde que l'extrémité gauche.

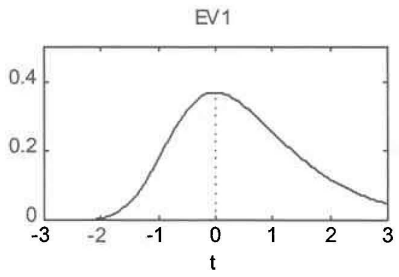

FIGURE 1.11 – Représentation graphique de noyau EV1. [5]

1.6 Conclusion

Dans ce chapitre, nous avons parlé des méthodes d'estimation statistique en général, et nous avons spécifiquement présenté la méthode non paramétrique du noyau. Nous avons conclu, des études déjà menées dans ce domaine, que le choix du noyau k n'a pas de grande d'importance, et même s'il est important, son effet sur la qualité de l'ensemble de l'estimateur est relativement faible, mais quant au paramètre de lissage h, il a un effet significatif, puisque la forme de cet estimateur change significativement à chaque petite différence dans la valeur de ce paramètre.

CHAPITRE 2

ESTIMATION À NOYAU DE LA FONCTION DE QUANTILE

2.1 Introduction

Dans la statistique descriptive, les quantiles sont utilisés pour déterminer les valeurs quantitatives, qui sont les valeurs qui divisent la population étudiée en classes de nombres égaux. Dans la théorie des probabilités, l'idée de valeurs quantitatives s'applique généralement aux vraies données aléatoires, en divisant les données en périodes contenant la même quantité de données ou en les divisant en parties de probabilité égale. La fonction de quantile permet de définir une fonction de répartition inverse (fonction de quantile). La méthode la plus populaire pour simuler la loi d'une variable aléatoire est basé sur cette fonction. Elle représente un puissant moyen de prévision, et son estimation est l'un des principaux problèmes en statistique.

2.2 Fonction de quantile

La fonction quantile d'une variable aléatoire (ou loi de probabilité) est l'inverse de sa fonction de distribution. Lorsque cette fonction de distribution est strictement croissante, son inverse est déterminé sans ambiguïté. Mais la fonction de distribution reste constante sur tout intervalle dans lequel une variable aléatoire ne peut pas prendre de valeurs. C'est pourquoi nous donnons la définition suivante.

Définition 2.2.1.
Soit X une variable aléatoire à valeurs dans $\mathbb{R}$*, et* F_X *sa fonction de répartition. On appelle fonction de quantile de X la fonction, notée* Q_X*, de* $]0,1[$ *dans* $\mathbb{R}$*, qui à* $p \in]0,1[$ *associe :*

$$Q_X(p) = \inf\{x \in \mathbb{R} : \ F_X(x) \geq p\}.$$

Par convention, on peut décider que $Q_X(0)$ *est la plus petite des valeurs possibles pour X et* $Q_X(1)$ *est la plus grande (elles sont éventuellement infinies).*

2.2.1 Lois discrètes :

La fonction de quantile d'une variable aléatoire discrète est une fonction en escalier, comme la fonction de répartition. Si X prend les valeurs $x_1, x_2, ..., x_n$, rangées par ordre croissant, la fonction de répartition est égale à : $F_X(x) = P(X = x_1) + P(X = x_2) + ... + P(X = x_i)$, $x \in [x_{i-1}, x_i[$. La fonction de quantile vaut :

$$Q_X(p) = \begin{cases} x_1 & pour\ p \in]0, F_1] \\ \vdots & \\ x_i & pour\ p \in]F_i, F_{i+1}] \\ \vdots & \end{cases}$$

Par exemple, pour la loi géométrique $\mathcal{G}(p)$, la fonction quantile est la fonction qui, pour tout i=1,2,..., vaut i sur l'intervalle $]1-(1-p)^{i-1}, 1-(1-p)^i]$.

2.2.2 Lois continues

Plaçons-nous dans le cas le plus fréquent, où la densité f_X est strictement positive sur un intervalle de $\mathbb{R}$ (son support) et nulle ailleurs. Si l'intervalle est $[a, b]$, la fonction de répartition est nulle avant a si a est fini, elle est strictement croissante de 0 à 1 entre a et b, elle vaut 1 après b si b est fini. Toute valeur p strictement comprise entre 0 et 1 est prise une fois et une seule par F_X. La valeur de $Q_X(p)$ est le point x unique, compris entre a et b, tel que $F_X(x) = p$ [1].

Par exemple calculons la fonction quantile de la loi exponentielle $\mathcal{E}(\lambda)$, de fonction de répartition $1 - \exp(-\lambda x)\mathbb{1}_{\mathbb{R}^+}(x)$.

Pour tout $p \in]0, 1[$, $(1 - \exp(-\lambda x)) = p \iff x = Q_X(p) = -\frac{1}{\lambda}\log(1 - p)$.

Remarque 2.2.1.

Le continuité de la fonction de répartition F_X, garantie pour tout élément $p \in]0, 1[$ une valeur appelée quantité d'ordre p pour laquelle $P(X \leq x) = p$. Mais cette quantile n'est pas nécessairement unique. Lorsqu'une distribution de X a des zones de probabilité nulles.

Si le support de f_X soit un intervalle [a, b], là où il est susceptible d'être $a = -\infty$ et ou $b = +\infty$ où F_X strictement croissante sur l'ensemble des valeurs de x, de sorte que $0 < F(x) < 1$. Donc, F_X est bijective et le quantile d'ordre p est unique pour tout $p \in]0, 1[$.

Dans le cas d'une variable aléatoire X discrète, la fonction de distribution F_X est discontinue et l'existence de la valeur x pour laquelle $p(X \leq x) = p$ n'est pas garanti pour toutes les p-valeurs.

2.2.3 Propriétés d'une fonction de quantile Q_X

La fonction Q_X posséder les propriété suivant :

1) Q_X est croissante ;

En effet, comme Q_X est l'inverse de F_X et F_X est croissante, alors Q_X est croissance.

2) Q_X est continue à gauche sur $]0 ; 1[$;

3) La limite à droite de Q_X en 0 est la borne inférieure du support de la loi de X et sa limite à gauche en 1 est la borne supérieure de ce support ;

4) La fonction Q_X est la réciproque de la fonction de répartition F_X, lorsque celle-ci

réalise une bijection ;

5) Q_X est continue sur un intervalle ouvert I sur $]0;1[$, ce qui est le cas pour les lois usuelles admettant une densité ;

6) Pour tout x, $Q_X(F_X(x)) \leq x$, avec égalité lorsqu'il n'existe aucun y strictement inférieur à x tel que $F_X(y) = F_X(x)$;

7) Pour tout p strictement compris entre 0 et 1, $F_X(Q_X(p)) \geq p$, avec égalité lorsque p est une valeur prise par F_X ;

8) $Q_X \circ F_X \circ Q_X = Q_X$;

9) $F_X \circ Q_X \circ F_X = F_X$;

10) $Q1$ et $Q2$ sont deux fonctions de quantiles alors :

_ $Q1 + Q2$ est une fonction de quantiles ;

_ $Q1 \cdot Q2$ est une fonction de quantile si $Q1$ et $Q2$ sont à valeurs positives.

Remarque 2.2.2.

Pour certain valeur de p, il y a des noms spéciaux au quantile :

• *Lorsque* $p = \frac{1}{2}$*, le quantile est appelé médiane il sépare la série statistique en deux groupes de taille égale, l'un contenant les plus petites valeurs et l'autre les plus grandes valeurs. Son interprétation :*

Au moins la moitié des valeurs sont inférieures ou égales à la médiane(me).

Au moins la moitié des valeurs sont supérieures ou égales à la médiane(me).

• *On dit le* i^{ieme} *quartile* ($i \in [1,4[$) *ou la quantile d'ordre* $p = \frac{1}{4}$*, servant à séparer les séries statistiques en quatre groupes de même taille. Son interprétation :*

Au moins un quart des valeurs sont inférieures ou égales à Q_1.

Au moins les trois quarts des valeurs sont inférieures ou égales à Q_3.

• *On dit le* i^{ieme} *décimal* ($i \in [1,10[$) *ou la quantile d'ordre* $p = \frac{1}{10}$*. Qui est séparée la série statistique en dix groupes de même taille. Son interprétation :*

Au moins dixième des valeurs inférieures ou égales à D_1.

Au moins les neuf dixièmes des valeurs inférieures ou égales à D_9.

• *On dit Le* i^{ieme} *centile* ($i \in [1,100[$) *ou la quantile d'ordre* $p = \frac{1}{100}$*, séparant la série en cent groupes de même taille. Son Interprétation :*

Environ 1٪ des valeurs sont inférieures ou égales à C_1.

1% des valeurs sont supérieures ou égales à C_{99}.

Remarque 2.2.3.

$D_1 = C_{10} = q_{0.1}$;

$Q_1 = C_{25} = q_{0.25}$;

$Q_3 = C_{75} = q_{0.75}$;

$Q_2 = D_5 = C_{50} = me = q_{0.5}$.

Remarque 2.2.4.

Ces quartiles, décimales et centiles sont les échelles de centralisation sont particulièrement nécessaires lors du calcul de l'étendu, qui est l'une des mesures de dispersion les plus simples, lorsqu'il est nécessaire de se débarrasser des valeurs aberrantes qui se trouvent généralement au début et à la fin des valeurs après ils sont disposés en ordre croissant [2].

2.3 Estimation de la fonction de quantile

L'estimation des quantiles ou l'estimation de la fonction des quantiles représente un enjeu important pour les applications. L'estimation de la fonction quantile dépend de deux méthodes ou approches : L'approche paramétrique et non paramétrique.

2.3.1 L'approche paramétrique

Quant au paramétrique, on estimer les paramètres inconnus qui déterminent la loi de probabilité de la variable aléatoire. On suppose que F_X est continue et appartient à la famille : $\mathcal{F} = \{F_\theta, \theta \in \Theta \subset \mathbb{R}^K\}$.
Un estimateur de quantile naturel s'obtient en remplaçant θ par son estimation $\hat{\theta}$ et à partir de laquelle

$$\hat{Q}_X(p) = F^{-1}_{\hat{\theta}}(p) = \inf\{x \in \mathbb{R} : F_{\hat{\theta}} \geq p\}, \quad 0 < p < 1$$

L'estimateur $\hat{\theta}$ peut être obtenu de plusieurs manières, comme *moment*, *maximum de vraisemblance*,...etc.

2.3.2 L'approche non-paramétrique

La fonction de quantile Q_X d'une variable aléatoire est la fonction réciproque de sa fonction de répartition F_X, elle est définie par :

$$Q_X(p) = \inf\{x \in \mathbb{R} : \ F_X(x) \geq p\}, \ \ 0 < p < 1.$$

Le quantile d'ordre p de F_X est noté ξ_p est donné par $\xi_p = Q_X(p)$.
Quant au non paramétrique, en générale l'estimation de $Q_X(p)$ est :

$$\hat{Q}_X(p) = \inf\{x \in \mathbb{R} : \hat{F}_X(x) \geq p\}, \ \ 0 < p < 1.$$

Plusieurs méthodes ont été consacrées pour l'estimation non paramétrique de la fonction de quantile. Le type d'estimation le plus simple est une estimation empirique.

Définition 2.3.1. ***(Statistiques d'ordre)***
Soit $X_1, X_2, ..., X_n$ un échantillon de variables aléatoires, la statistique d'ordre de cette échantillons sont $X_* = (X_{(1)}, X_{(2)}, ..., X_{(n)})$ vérifient $X_{(1)} < X_{(2)} < ... < X_{(n)}$.

Définition 2.3.2. ***(Quantile empirique)***
La fonction de quantile empirique est la fonction correspondante à la fonction de distribution empirique F_n, définie par :

$$Q_n(p) = \inf\{x \in \mathbb{R} : F_n(x) \geq p\},$$

tel que

$$F_n = \frac{1}{n} \sum \mathbb{1}_{]-\infty,x]}(X_i).$$

Son estimateur correspondant est :

$$\hat{Q}_n(p) = \inf\{x \in \mathbb{R} : \hat{F}_n(x) \geq p\}.$$

Alors afin d'estimer la quantile dans le cas d'un échantillon i.i.d. $X_1, X_2, ..., X_n$, nous connaissons la fonction de quantile empirique d'ordre p à partir des statistiques d'ordre de l'échantillon.

Exemple 2.3.1.

1) Pour X discrète, soient $p \in]0,1[$*, la statistique d'ordre* $X_{([np]+1)}$ *est appelée le quantile empirique d'ordre p de l'échantillon nous le symbolisons avec* $\hat{\xi}_{np}$*, tel que : la fonction* $x \to [x]$ *est le plus petit entier supérieur ou égal à x.*

2) pour X continue, le quantile empirique est calculée comme suit :

$$\hat{\xi}_{np} = \varrho_{i-1} + \frac{p_n - n_{(i-1)c}}{n_i}(\varrho_i - \varrho_{i-1}),$$

où $[\varrho_{i-1}, \varrho_i[$ *est la classe contenant le quantile,* p_n *c'est l'ordre de quantile,* $n_{(i-1)c}$ *est l'effectif cumulé de la classe précédent de* $[\varrho_{i-1}, \varrho_i[$*,* n_i *l'effectif de* $[\varrho_{i-1}, \varrho_i[$ *et* $n = \sum n_i$*.*

Remarque 2.3.1.

Le quantile empirique est une valeur du l'ensemble $X_{(1)}, X_{(2)}, ..., X_{(n)}$*, donc pour étudier le comportement des quantile empirique, on étudie les statistiques ordonnées.*

Convergence des quantiles empiriques

La proposition suivante montre qu'un quantile empirique est un estimateur d'un quantile théorique.

Proposition 2.3.1. *[19]*

Soit $p \in]0,1[$*, supposons que F est continue et qu'il existe une seule solution* ξ_p *à l'équation* $F(x) = p$*. Soit* $(l(n), n \geq 1)$ *une suite d'entiers telle que :*

$$1 \leq l(n) \leq n \text{ et } \lim_{n\to\infty} \frac{l(n)}{n} = p.$$

Alors, la suite des quantiles empiriques $(X_{(k(n))}), n \geq 1$*, converge presque sûrement vers* ξ_p*, c-à-d,*

$$X_{(l(n))} \xrightarrow{ps} \xi_p \text{ quand } n \to \infty,$$

où ξ_p *est le quantile d'ordre p.*

Erreur quadratique moyenne de $\hat{Q}_n$

L'erreur moyenne quadratique de $\hat{Q}_n$ est minimale sur les hypothèses suivantes :

1-f est différentiable et sa dérivée est $f^{'}$;

2-$f^{'}$ est continu dans le voisinage de ξ_p et $f^{'}(\xi_p) \neq 0$;

3-$\int_{\mathbb{R}} xk(x)dx = 0$ et $\mu_2 = \int_{\mathbb{R}} x^2 k(x)dx < \infty$.

Théorème 2.3.2. *[20] [19]*

Sous les hypothèses 1-3 Nous avons pour chaque $p \in]0,1[$, l'erreur quadratique moyenne de $\hat{Q}_n(p)$ est :

$$MSE\{\hat{Q}_n(p)\} = \frac{p(1-p)}{nf^2(\xi_p)} + \frac{h^4}{4}\frac{(f^{'}(\xi_p))}{f^{''}(\xi_p)} - \frac{h}{nf(\xi_p)}\psi(k) + O\left(\frac{h}{4} + h\right),$$

tel que :

$$\psi(k) = 2\int_R uK_h(u)k(u)du.$$

L'erreur quadratique moyenne asymptotique est :

$$AMSE\{\hat{Q}_n(p)\} = \frac{p(1-p)}{nf(\xi_p)^2} + \frac{h^4}{4}\frac{f^{'}(\xi_p)}{f(\xi_p)^2}u_2(k)^2 - \frac{h}{n}\frac{1}{f(\xi_p)}\psi(k).$$

Approximativement le paramètre de lissage optimal pour $\hat{Q}_n(p)$ est donné par :

$$\hat{h}_{opt} = \left\{\frac{f(\xi_p)\psi(k)}{n(f^{'}(\xi_p))^2(\mu_2(k))^2}\right\}^{\frac{1}{3}}.$$

Corollaire 2.3.3.

Si en remplaçant la valeur h_{opt} dans l'expression AMSE de $\hat{Q}_n(p)$, nous obtenons :

$$AMSE_{opt}\{\hat{Q}_n(p)\} = \frac{p(1-p)}{f^2(\xi_p)} - \frac{3}{4}\left\{\frac{(\psi(k))^4}{nf^2(\xi_p)f^{'}(\xi_p)^2\mu_2^2(k)}\right\}^{\frac{1}{3}}.$$

h_{opt} approximatif pour $\hat{Q}_n(p)$ en utilisant l'estimateur de dérivé

Nous pouvons proposer une méthode alternative pour estimer $f^{'}(\xi_p)$ et $f(\xi_p)$ dans

$\hat{h}_{opt}$ qui utilisent des estimateurs de ces dérivées, notez que :

$$Q'_X(p) = \frac{1}{f(F^{-1}(p))} = \frac{1}{f(Q(p))} = \frac{1}{f(\xi_p)}.$$
$$Q''_X(p) = -\frac{f'(Q(p))}{f^3(Q(p))} = -\frac{f'(\xi_p)}{f^3(\xi p)}.$$

Alors l'équation précédente de $\hat{h}_{opt}$ devient :

$$\hat{h}_{opt} = \left\{ \frac{Q'_n(p)^5 \psi(k)}{n Q''_n(p)^2 \mu_2(k)} \right\}^{\frac{1}{3}}.$$

2.3.3 Estimation à noyau de la fonction de quantile

Soit $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ la statistique d'ordre des variables aléatoires i.i.d. $X_1, X_2, ..., X_n$ de fonction de répartition F absolument continue. Rappelons que la fonction de quantile Q est L'inverse à gauche de F définie par :

$$Q_X(p) = \inf\{x \in \mathbb{R} : F_X(x) \geq p\}, 0 < p < 1.$$

Étant donné que les statistiques de classement individuelles varient, les mesures de l'échantillon souffrent d'un manque d'efficacité. Afin de réduire cette variance, différentes méthodes ont été proposées pour estimer les quantités d'échantillons à l'aide de statistiques de classement pondérées. Une classe commune de ces estimateurs est appelée estimateurs à noyau, qui a suggérée *Parzen*(1962). Une version des estimateurs à noyau est comme suit :

$$\tilde{Q}_n(p) = \frac{1}{h_n} \sum_{i=1}^{n} X_{(i)} \int_{\frac{i-1}{n}}^{\frac{i}{n}} k\left(\frac{x-p}{h_n}\right) dx.$$

2.4 Propriété de l'estimateur à noyau de la fonction de quantile

Nous présentons dans cette partie les propriétés statistiques de l'estimateur $\tilde{Q}_n$.

2.4.1 Biais de $\tilde{Q}_n$

Supposant que le noyau k, la fenêtre h et Q satisfaisaient les hypothèses suivantes :

H1) $h_n \to 0$ *quand* $n \to \infty$;

H2) $\int_{-\infty}^{+\infty} y^2 k(y) dy < \infty$;

H3) $Q_X^{''}$ et $Q_X^{'}$ sont bornées ;

H4) $\int_{-\infty}^{+\infty} yk(y) dy < \infty$.

Proposition 2.4.1.

Sous les hypothèses H1-H4, l'estimateur $\tilde{Q}_n$ *de la fonction* Q_X *est asymptotiquement non biaisé. C'est-à-dire,* $\lim_{n\to\infty} biais(\tilde{Q}_n(p)) = 0$.

Démonstration.

On a :

$$\begin{aligned}
biais(\tilde{Q}_n(p)) &= E(\tilde{Q}_n(p)) - Q_X(p) \\
&= E\left(\sum_{i=1}^{n} X_{(i)} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{1}{h_n} k\left(\frac{x-p}{h_n}\right) dx\right) - Q_X(p) \\
&= \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{1}{h_n} k\left(\frac{x-p}{h_n}\right) dx E(X_{(i)}) - Q_X(p) \\
&= \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{1}{h_n} k\left(\frac{x-p}{h_n}\right) dx Q(\frac{i}{n+1}) - Q_X(p) \\
&= \int_0^1 \frac{1}{h_n} k\left(\frac{x-p}{h_n}\right) Q_X(x) dx - Q_X(p).
\end{aligned}$$

Soit le changement de variable $y = \frac{x-p}{h_n}$ donc $x = yh_n + p$ et $dx = h_n dy$, par conséquent le biais de $\tilde{Q}_n$ sera

$$biais(\tilde{Q}_n(p)) = \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} k(y) Q(yh_n + p) h_n dy - Q_X(p).$$

On applique un développement de Taylor d'ordre 2 pour la fonction $Q(yh_n + p)$, on obtient :

$$Q(yh_n + p) = Q_X(p) - \frac{yh_n}{1!} Q_X^{'}(p) + \frac{y^2 h_n^2}{2!} Q_X^{''}(p) + O(y^2 h_n^3).$$

$$\begin{aligned}
biais(\tilde{Q}_n(p)) &= \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} k(y)[Q_X(p) - \frac{yh_n}{1!} Q_X^{'}(p) + \frac{y^2 h_n^2}{2!} Q_X^{''}(p) + O(y^3 h_n^3)] - Q_X(p) \\
&= \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} k(y) Q_X(p) dy - \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} yh_n Q_X^{'}(p) k(y) dy + \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} k(y) \frac{y^2 h_n^2}{2!} Q_X^{''}(p) dy \\
&\quad + O(y^2 h_n^3) \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} k(y) dy - Q_X(p) \\
&= Q_X(p) \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} k(y) dy - h_n Q_X^{'}(p) \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} yk(y) dy + \frac{h_n^2}{2!} Q_X^{''}(p) \int_{\frac{-p}{h_n}}^{\frac{1-p}{h_n}} y^2 k(y) dy \\
&\quad + O(y^2 h_n^3) - Q_X(p).
\end{aligned}$$

Lorsque n tend vers l'infini, on obtient :

$$\begin{aligned}\lim_{n\to\infty} biais(\tilde{Q}_n(p)) &= Q_X(p)\int_{-\infty}^{+\infty} k(y)dy - h_n Q_X^{'}(p)\int_{-\infty}^{+\infty} yk(y)dy + \frac{h_n^2}{2!}Q_X^{''}(p)\int_{-\infty}^{+\infty} y^2k(y)dy \\ &\quad + O(y^2h_n^3) - Q_X(p) \\ &= \frac{h_n^2}{2!}Q_X^{''}(p)\int_{-\infty}^{+\infty} y^2k(y)dy + O(y^2h_n^3).\end{aligned}$$

$\lim_{n\to\infty} biais(\tilde{Q}_n(p)) = 0$, alors $\tilde{Q}_n$ est asymptotiquement non biaisé. □

2.4.2 Variance de $\tilde{Q}_n$

Falk (1984, p263) [9] a prouvé que la variance de $\tilde{Q}_n$ est minimale sous les hypothèses suivantes :

H5) k est de support compact ;

H6) $Q_X^{'}$ est dérivable ;

H7) $Q_X^{''}$ est bornée ;

H8) $\mu_2 = \int_{-\infty}^{+\infty} y^2k(y)dy < \infty$;

H9) $\int (q - h_n t)tk(t)j(t)dt < \infty$,

en écrivant $\tilde{Q}_X$ sous la forme

$$\tilde{Q}_n(p) = \int_0^1 F_n^{-1}(x)h_n^{-1}k\left(\frac{p-x}{h_n}\right)dx,$$

où F_n est la fonction de répartition empirique.

Proposition 2.4.2.

Sous les hypothèses H5-H9, l'estimateur $\tilde{Q}_n$ de la fonction Q_X est a variance minimal asymptotique. C'est à dire la variance de $\tilde{Q}_n(p)$ défini par :

$$Var(\tilde{Q}_n(p)) = \frac{p(1-p)}{n}(Q_X^{'}(p))^2 - \frac{h}{n}(Q_X^{'}(p))^2\int xk(x)K_h(x)dx + O(n^{-1}h).$$

est tend vers 0 lorsque n tend vers l'infini.

Corollaire 2.4.3.

L'estimation $\tilde{Q}_n(p)$ converge vers $Q_X(p)$ en moyen quadratique c'est à dire

$$\tilde{Q}_n(p) \xrightarrow{mq} Q_X(p).$$

Démonstration.

Comme $\tilde{Q}_n(p)$ est asymptotiquement sans bais et sa variance tend vers 0, alors

$$\lim_{n\to\infty} E|\tilde{Q}_n(p) - Q_n(p)|^2 = 0,$$

c-à-d :

$$\tilde{Q}_n(x) \xrightarrow{mq} Q_X(x).$$

□

2.4.3 Erreur quadratique moyenne (*MSE*) de $\tilde{Q}_n$

Le théorème suivant donne une expression de l'erreur quadratique moyenne asymptotique de $\tilde{Q}_n(p)$ basée sur l' expression de la variance calculée par Falk(1984).

Théorème 2.4.4. *[28]*

Supposons que Q''_X est continue dans le voisinage p et k est une densité d'un noyau symétrique par rapport à zéro. Avec un support compact. l'erreur quadratique moyenne de $\tilde{Q}_n(p)$ est :

• *lorsque F_X est symétriques et $p \neq \frac{1}{2}$ (ou lorsque F_X est asymétriques) :*

$$MSE\{\tilde{Q}_n(p)\} = \frac{p(1-p)}{n}(Q'_X(p))^2 + \frac{h^4}{4}(Q''_X(p))^2\mu_2^2(k) - \frac{h}{n}(Q'_X(p))^2\psi(k) + O\left(\frac{h}{n} + h^4\right).$$

• *Lorsque F_X est symétrique et $p = \frac{1}{2}$:*

$$MSE\{\tilde{Q}_n(p)\} = \frac{1}{n}(Q'(0.5))^2(0.25 - 0.5h\psi(k) + (nh)^{-1}R(k)) + O(n^{-1}h + (nh)^{-2}),$$

tel que $R(k) = \int_{\mathbb{R}} k^2(x)dx$.

Démonstration.
Découle directement des deux propositions précédentes. □

Notez que pour une sélection raisonnable de h (c'est-à-dire tendant vers zéro plus rapidement que $n^{-\frac{1}{4}}$*), le terme dominant pour MSE est la variance asymptotique de la quantité d'échantillon.*

Corollaire 2.4.5.
En s'appuyant sur Falk (1984), Sheader et Marron (1990) ont donné l'erreur quadratique moyenne asymptotique (AMSE) de $\tilde{Q}_n(p)$ *comme suit :*
• *Lorsque* F_X *est symétrique et* $p \neq \frac{1}{2}$ *:*
L'erreur quadratique moyenne asymptotique de $\tilde{Q}_n$ *est :*

$$AMSE\{\tilde{Q}_n(p)\} = \frac{p(1-p)}{n}(Q^{'}(p))^2 + \frac{h^4}{4}(Q^{''}(p))^2\mu_2^2(k) - \frac{h}{n}(Q^{'}(p))^2\psi(k).$$

- Si $Q_X^{'}(p) > 0$*, la fenêtre optimale asymptotique pour* $AMSE\{\tilde{Q}_n(p)\}$ *est*

$$\tilde{h}_{opt} = \left\{\frac{(Q_X^{'}(p))^2\psi(k)}{n(Q_X^{''}(p))^2(\mu_2(k))^2}\right\}^{\frac{1}{3}},$$

et à partir de cette valeur, on obtient l'erreur quadratique moyenne asymptotique optimale suivante :

$$\begin{aligned} AMSE_{opt}\{\tilde{Q}_n(p)\} &= \frac{1}{n}\left\{p(1-p)(Q_X^{'}(p))^2 + \frac{3}{4}\left(\frac{(Q_X^{'}(p))^8(\psi(k))^4}{n(Q_X^{''}(p))^2\mu_2^2(k)}\right)^{\frac{1}{3}}\right\} \\ &= n^{-1}p(1-p)(Q_X^{'}(p))^2 + O(n^{-\frac{4}{3}}). \end{aligned}$$

Les estimateurs des dérivés de la fonction de quantile peuvent être utilisés pour estimer $\tilde{h}$ *comme suit :*

$$\tilde{h}_{opt} = \left\{\frac{f(\xi_p)^4\psi(k)}{nf^{'}(\xi_p)^2\mu_2(k)^2}\right\}^{\frac{1}{3}}.$$

• *Lorsque F_X est symétrique et $p = \frac{1}{2}$:*

L'erreur quadratique moyenne asymptotique de $\tilde{Q}_n$ est :

$$AMSE\{\tilde{Q}_n(p)\} = \frac{1}{n}(Q_X^{'}(0.5)^2\{0.25 - 0.5h\psi(k) + (nh)^{-1}R(k)\}.$$

$R(k)$: est précédemment connu.

Remarque 2.4.1.

Si elle était $Q_X^{'} = 0$ nous avons besoin de termes d'ordre supérieur peut afficher AMSE de $\tilde{Q}_n(p)$ par :

$$AMSE\{\tilde{Q}_n(p)\} = (\frac{1}{4} - \frac{1}{n})h^4Q_X^{''}(p)^2\mu_2^2(k) + 2n^{-1}h^2Q_X^{''}(p)^2\int(p - ht)tk(t)J(t)dt,$$

tel que $J(t) = \int_{-\infty}^{t} xk(x)dx$. La preuve est fournie en [20].

Azalin(1981) *[3] a considéré les estimateurs de quantités obtenus en inversant l'estimateur du noyau de la fonction de distribution et a obtenu un résultat lié à notre théorie précédente, la théorie précédente produit le corollaire suivant :*

Corollaire 2.4.6. *[28]*

Nous supposons que pour chaque $p \neq 0.5$ et F_X symétrique, le paramètre de lissage optimal est donné par :

$$h_{opt} = \alpha(k) \cdot \beta(Q) \cdot n^{-\frac{1}{3}},$$

tel que :

$$\alpha(k) = \left[2\int_R uk(u)K_h(u)du \Big/ \left\{\int_R u^2k(u)du\right\}^2\right]^{\frac{1}{3}} et\ \ \beta(Q) = [Q_X^{'}(p)/Q_X^{''}(p)]^{\frac{2}{3}}.$$

avec $h = h_{opt}$.

$$MSE(\hat{Q}(p)) = \frac{p(1-p)}{n}(Q_X^{'}(p))^2 + O(n^{-\frac{4}{3}}).$$

Si F est symétrique et $p = \frac{1}{2}$ prendre $h = 0(n^{-\frac{1}{2}})$ Cela rend les deux premiers termes dans

h de MSE pour $\tilde{Q}_n(0.5)$ *dans le même ordre et*

$$MSE\{\tilde{Q}_n(0.5)\} = \frac{0.25}{n}(Q'_X(0.5))^2 + O(n^{-\frac{3}{2}}).$$

Cependant, comme le terme dans hn^{-1} *est négatif et que le terme dans* $n^{-2}h^{-1}$ *est positif, il n'y a pas de fenêtre unique qui réduise l'erreur quadratique asymptotique moyenne de* $\tilde{Q}_n(0.5)$ *lorsque F est symétrique, c'est-à-dire h satisfaisant* $h = cte \cdot n^{-m}$ $(0 < m \leq \frac{1}{2})$ *pour les grandes valeurs de la constante, un estimateur produira une erreur quadratique asymptotique moyenne inférieure à* $\xi_{0.5}$.

Déterminer le paramètre de lissage optimal

Dans cette section, nous proposons un choix basé sur les données pour h le paramètre de lissage de $\tilde{Q}_n(p)$ *pour tout p sauf 0.5 si* F_X *est symétrique.*

On voit du corollaire précédent que pour un choix donné de k la valeur optimale asymptotique de h dépend des dérivées première et seconde de la quantile :

Considéré les estimateurs $Q'_X(p)$ *et* $Q''_X(p)$ *nécessaire de choisir de h si la dérivées première et seconde de k existent, alors nous pouvons estimer ces quantiles par les dérivées première et seconde de* $\tilde{Q}_n(p)$. *Depuis lors, l'intérêt pour le ratio* $\{Q'_X(p)/Q''_X(p)\}^{\frac{2}{3}}$ *il semble naturel de s'intéresser aux noyaux d'ordre supérieur pour tenter d'éliminer les problèmes associés à l'estimation du rapport résultat dans ce*

$$Q'_m(p) = \sum_1^n \left\{ \int_{\frac{i-1}{n}}^{\frac{i}{n}} a^{-2}k'_*(a^{-1}(t-p))dt \right\} X_i,$$

$$Q''_m(p) = \sum_1^n \left\{ \int_{\frac{i-1}{n}}^{\frac{i}{n}} b^{-3}k''_*(b^{-1}(t-p))dt \right\} X_i,$$

tel que : k_* *c'est un noyau d'ordre m symétrique autour de zéro (c'est,* $\int k_*(u)du = 1$, $\int u^J k_*(u) = 0,\ J = 1, 2, ..., m-1,$ $\int u^m k_*(u)du < \infty,$ $k_* \in L^2(-\infty, +\infty)$ *et* $k^m \in lip(\alpha)$.*).*

L'estimation du quantile nous donne asymptotiquement le paramètre de lissage optimal suivant :

$$\hat{h}_{opt} = \alpha(k) \cdot \hat{\beta} \cdot n^{-\frac{1}{3}},$$

tel que :

$$\hat{\beta} = \left\{ \frac{\hat{Q}'_m(p)}{\hat{Q}''_m(p)} \right\}^{\frac{2}{3}},$$

et $\alpha(t)$ précédemment connu.

Le problème est alors de choisir des valeurs pour la fenêtre a et b qui se traduisent par une efficacité approximative $\hat{\beta}$.

Théorème 2.4.7. *[28]*

Supposons que $Q_X^{(m+2)}$ est continue au voisinage de p et que k_* est un noyau de support compact, d'ordre m et symétrique autour de zéro. La fenêtre optimale asymptotique de $\hat{Q}'_m(p)$ est donnée par :

$$a_{opt} = \mu_m(k_*) \cdot \gamma_m(Q) \cdot n^{-\frac{1}{(2m+1)}},$$

tel que :

$$\mu_m(k_*) = \left\{ (m!)^2 \int_{\mathbb{R}} k_*^2(u)du \Big/ 2m \Big[\int_{\mathbb{R}} u^m k_*(u)du \Big]^2 \right\}^{\frac{1}{(2m+1)}},$$

$$\gamma_m(Q) = \{ Q'_X(p) / Q_X^{(m+1)}(p) \}^{\frac{2}{(2m+1)}}.$$

La fenêtre asymptotique optimale de $\hat{Q}''_m(p)$ est donnée par :

$$b_{opt} = \tau_m(k_*) \cdot \delta_m(Q) \cdot n^{-\frac{1}{(2m+3)}},$$

tel que :

$$\tau_m(k_*) = \left\{ 3(m!)^2 \int_{\mathbb{R}} (k'(u))^2 du / 2m \Big[\int u^m k_*(u)du \Big]^2 \right\}^{\frac{1}{(2m+3)}}$$

et

$$\delta_m(Q) = \{ Q'_X(p) / Q_X^{(m+2)}(p) \}^{\frac{2}{2m+3}}.$$

cette approche a été utilisée avec succès par Hall et Sheather(1988) *ont réussi à choisir la largeur de fenêtre d'un estimateur $Q'_X(0.5)$.*

2.4.4 Normalité asymptotique

Théorème 2.4.8. *[8]*

Si Q_X a une dérivée seconde bornée au voisinage de $q \in (0, 1)$ et si $Q_X^{'} > 0$, k a un support borné,tel que $\int k(x)dx = 1$ et $h_n \xrightarrow{n\to 0} 0$ on a :

$$\frac{n^{\frac{1}{2}}[\tilde{Q}_n(p) - E(\tilde{Q}_n(p))]}{var[\tilde{Q}_n(p)]} \to N(0,1).$$

Démonstration.

Voir.[8] □

2.5 Conclusion

Dans ce chapitre, on a étudié les propriétés asymptotiques de l'estimateur à noyau de la fonction de quantile, On a conclut que c'et un bon estimateur asymptotiquement non biais, de variance minimale et d'erreur moyenne quadratique plus petit que celle de l'estimateur empirique. On a aussi parlé de la dérivée de la fonction de quantile, qui fait l'objectif de plusieurs travaux comme Jones [10].

CONCLUSION GÉNÉRALE

Le travail présenté dans ce mémoire est l'estimation de la fonction des quantiles par la méthode du noyau.

Dans un premier temps, nous avons introduit l'idée de la méthode non paramétrique du noyau, que nous utiliserons pour estimer la fonction quantile. L'estimateur à noyau est une fonction de deux paramètres : la fonction K appelée noyau et h appelé paramètre de lissage ou fenêtre. Si le choix du noyau K n'est pas un problème dans l'estimation de la densité de probabilité, il n'en est pas de même pour le choix du paramètre de lissage h qui ne dépend que de la taille de l'échantillon.

Dans un second temps, nous définissons la fonction quantile et nous donnons quelques propriété. que ce soit dans l'approche paramétrique ou non paramétrique, nous nous sommes intéressés à la méthode du noyau parce qu'elle est "populaire" vu sa souplesse d'utilisation et elle présente de bonnes propriétés asymptotiques. ce qui le montre l'étude des propriété de l'estimateur à noyau de la fonction de quantile.

ANNEXE

Caractéristiques d'un bon estimateur

Dans de nombreux aspects appliqués, lors de la recherche d'une estimation pour l'un des paramètres de la population d'étude, le chercheur est confronté a de nombreux critères comme estimateurs de ce paramètre, et afin de déterminer le meilleur de ces capacités. Les caractéristiques d'un bon estimateur doivent être connues comme critère de choix d'une de ces métriques. Un bon estimateur se caractérise les quatre caractéristiques :

sans biais

soit $X_1, X_2, ..., X_n$ *un échantillon aléatoire d'une communauté à densité de probabilité* $f(x, \theta)$ *tel que* θ *inconnu et soit* $\hat{\theta}$ *estimateur de* θ*. On dit qu'il s'agit d'un estimateur sans biais lorsque sa valeur attendue est égale à la vraie valeur d'un paramètre de la population c-à-d* $E(\hat{\theta}) = \theta$*, ou c'est celui qui fera vraisemblablement référence au vrai paramètre lorsque la taille de l'échantillon augmente,*

$$\lim_{n\to\infty}(E|\hat{\theta} - \theta|) = 0.$$

Consistance

L'estimateur est caractérisé par la consistance s'il est sans biais, et sa variance devient nulle lorsque la taille de l'échantillon augmente, c'est-a-dire si ces deux conditions sont remplies :

$$\lim_{n\to\infty}(E|\hat{\theta}-\theta|)=0 \; et \; \lim(\sigma_{\theta}^{2})=0.$$

Efficacité relative

Un estimateur efficace est celui qui a la plus petite variance par rapport aux variances des autres estimateurs.

Efficacité

Soit $\hat{\theta}_1$ et $\hat{\theta}_2$ deux estimateur non biaisés pour le paramètre θ. Pour déterminer l'efficacité, nous comparons var($\hat{\theta}_1$) avec var($\hat{\theta}_2$) Si le rapport est inférieur à 1, le premier estimateur est le plus efficace, et vice versa.

Erreur quadratique moyenne MSE

Soit $\hat{\theta}$ un estimateur pour le paramètre inconnu θ L'erreur quadratique moyenne est définie comme suit :

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= E(\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2) \\ &= E(\hat{\theta}^2 - 2\theta E(\hat{\theta}) + \theta^2) \\ &= E(\hat{\theta}^2 - E(\hat{\theta})^2 + (E(\hat{\theta}))^2 - 2\theta E(\hat{\theta}) + \theta^2 \\ &= var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 \\ &= var(\hat{\theta}) + biaise(\theta). \end{aligned}$$

Remarque 2.5.1.
si $\hat{\theta}$ est un estimateur sans biais pour le paramètre θ. On a $biais(\theta) = 0$ alors $MSE(\hat{\theta}) = var(\hat{\theta})$.

Définition 2.5.1.
On dit que l'estimateur $\hat{\theta}_n(x)$ du paramètre $\theta(x)$
1) est faiblement consistant si

$$\forall x \in \mathbb{R} \quad \hat{\theta}_n(x) \xrightarrow{pr} \theta(x), \qquad quandn \to \infty.$$

2) est faiblement et uniformément consistant si

$$\sup_{x\in\mathbb{R}} |\hat{\theta}_n(x) - \theta(x)| \xrightarrow{pr} 0, \qquad quandn \to \infty.$$

3) est fortement consistant si

$$\forall x \in \mathbb{R} \quad \hat{\theta}_n(x) \xrightarrow{p.s} \theta(x), \qquad quandn \to \infty.$$

4) est fortement et uniformément consistant si

$$\sup_{x\in\mathbb{R}} |\hat{\theta}_n(x) - \theta(x)| \xrightarrow{p.s} 0 \qquad \textit{quand } n \to \infty.$$

5) est asymptotiquement sans biais si

$$\forall x \in \mathbb{R}, \qquad \lim_{n\to\infty} E(\hat{\theta}_n(x)) = \theta(x).$$

6) est asymptotiquement et uniformément sans biais si

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} E(\hat{\theta}_n(x) - \theta(x)) = 0.$$

Critères de convergence

Parmi toutes les qualités que peut avoir un estimateur, on s'intéresse souvent à sa consistance, c'est à dire, au fait qu'un estimateur f_h converge ou non vers f. La convergence d'un estimateur peut être faible (en probabilité) ou forte (presque sûrement ou en moyenne quadratique).

On donne quelques résultats de convergence des estimateurs à noyaux de la littérature. Soit X une variable aléatoire et (X_n) suite de variable aléatoire définit sur le même espace de probabilité $(\omega, \mathbb{A}, \mathbb{P})$.

Convergence en probabilité

La suite (X_n) *converge en probabilité vers* X *si :*

$$\forall \epsilon > 0 \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

La loi faible des grand nombres

Si les variables aléatoires X_n *sont deux à deux non covariées, de même loi, d'espérance* μ*, de variance* σ^2*, alors* $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{pr} \mu$.

Convergence presque sure

La suit (X_n) *converge presque surement vers* X *si* $P(w/\lim_{n\to\infty} X_n(w) = X(w)) = 1$.

Loi fort des grand nombres

Si les variable aléatoire X_n *sont mutuellement indépendantes de même loi, d'espérance* μ *alors* $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{ps} \mu$.

Théorème 2.5.1.
Soit $(X_n)_n$, $(Y_n)_n$ *sont des vecteurs aléatoire.*
i) $X_n \xrightarrow{ps} X \Longrightarrow X_n \xrightarrow{pr} X$.

Démonstration.
Soit $B = \{w, X_n(x) \to X(w)\}$. Soit $\epsilon > 0$ fini. Soit $A_n = U_{m\geq n}\{d(X_m, X) \geq \epsilon\}$. Pour tout $\epsilon > 0$, la suit A_n est décroissant.
Si $w \in B$, il existe n tel que pour tout $m \geq n d(X_n, X) \leq \epsilon$. C'est à dire que $w \in A_n^c$. Par suite on en déduit que $\mathbb{P}(A_n) \to 0$.
On conclut en remarquant que $\{w, d(X_n(w), X(w)) \geq \epsilon\} \subset A_n$. □

RÉSUMÉ

Dans ce mémoire on a intéressé à la méthode d'estimation non paramétrique à noyau. On a étudiée les propriétés de l'estimateur à noyau de la fonction de quantile, inverse de la fonction de répartition. Cet estimateur est asymptotiquement non biaisé, de variance minimale et asymptotiquement normale lorsque n tend vers l'infini.

Mots clés : Fonction de quantile, Estimation à noyau, Estimation non paramétrique, Estimation à noyau de la fonction de quantile.

BIBLIOGRAPHIE

[1] ***Adrian** et **autre.** (2022), quantile.density : Quantiles of a Density Estimate, web site : https ://rdrr.io/cran/spatstat.core/man/quantile.density.html, consulte le 14 jean 2022.*

[2] ***Al Baldawi,A. A.**,(2009) Méthodes statistiques pour les sciences économiques et l'administration des affaires avec l'utilisation d'un programme SPSS, Maison d'édition Wael, première édition 2009.*

[3] ***Azzalini, A**. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. Bio metrika. 68,326-328.*

[4] ***Cacoullos, T.** (1966). Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics, 18 : 178-189.*

[5] ***Dany Faucher**. (1999), Estimation non paramétrique des quantiles de crue par la méthode des noyaux, Mémoire présenté pour l'obtention du grade de Maître ès science (M.Sc.), Baccalauréat en statistique. Université du Québec INRS-Eau.*

[6] ***Epanechnikov, V. A**. (1969). Non parametric Estimation of a Multivariate Probability Density. Theory of Probability and its Applications. 14(1) :153-158.*

[7] ***Faucher, D. , Rasmussen, P. F. et Bobée, B.**,(2021) Estimation non paramétrique des quantiles de crue par la méthode des noyaux Nonparametric estimation of quantiles by the kernel method.*

[8] **Falk, M.** *(1985). Asymptotic normality of the kernel quantile estimator. The Annals of Statistics, 13(1), 428-433.*

[9] **Falk, M.** *(1984), "Relative Deficiency of Kernel Type Estimators of Quantiles," The Annals ofStatistics, 12,261-268.*

[10] **Jones, M.C.** *(1992), estimating densities, quantiles, quantile densities and density quantiles. Ann. Inst. Statist. Math. Vol. 44, No. 4, 721-727*

[11] **Hadj Amar, K.** *et* **Khalfi, N**. *(2015-2016), Etude comparative des méthodes de sélection du paramètre de lissage dans l'estimation de la densité de probabilité par la méthode du noyau. Mémoire Master, Département des Mathématiques, Faculté des Sciences, Université M'hamed Bougara Boumerdes.*

[12] **Harrell, F.E.** *et* **Davis, C.E.** *(1982), A New Distribution-Free Quantile Estimator, Biometrika, 69, 635-640.*

[13] **Izenman, A. J.** *(1991). Recent developments in nonparametric density estimation. J. Am. Stat. Assoc., 86(413) : 205-224.*

[14] **Lall, U.** *(1995). Recent advances in nonparametric function estimation : Hydrologie applications. Reviews of geophysics, supplement, 1093-1102, U.S. National Report to International Union of Geodesy and Geophysics 1991-1994.*

[15] **Lall, U., Y.-1. Moon** *et* **K. Bosworth** *(1993). Kernel flood frequency estimators : bandwitdh selection and keme1 choice. Water Resour. Res., 29(4) : 1003-1015.*

[16] **Larry Wasserman**. *All of Statistics : A Concise Course in Statistical Inference, Springer Berlin Heidelberg New York Barcelona Hong Kong London Milan Paris Tokyo.*

[17] **Madjid Abdel Aziz, A. A.** *(2018), Some nonparametric estimators for right censored survival data, Journall of Economiics and Admiiniisttrattiive Sciiences 2019; Voll. 25, No.113 Pages : 475-498.*

[18] **Maesono Y.** *et* **Penev S.**, *(2013), Edgeworth expansion for the kernel quantile estimator, Ann Inst Stat Math, DOI 10.1007/s10463-012-0369-6.*

[19] **Menaceur Sabrina**, *(2020), Estimation à noyau de la fonction des quantiles, MASTER en Mathématiques, Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie, Université Mohamed Khider, Biskra.*

[20] ***Ming-Yen Cheng1*** *et* ***Shan Sun2****, Bandwidth Selection for Kernel Quantile Estimation, JEL subject classification : C14, C13.*

[21] ***Malet, J.****,CEA 2017 Première année, Quantiles et simulation,cours on ligne https ://www.institutdesactuaires.com/global/gene/link.php*

[22] ***Nadaraya, E. A.*** *(1964). Some new estimates for distribution function. Theory of Probab. Appl. 9, 497-500.*

[23] ***Parzen, E.*** *(1979). Non parametric Statistical Data Modelling. Journal of the American Statistical Association. 74,105-131.*

[24] ***Rochet, P.****, Statistique : compléments de cours, Préparation à l'Agrégation, université de nantes.*

[25] ***Rosenblatt, M.*** *(1956). Remarks on some non parametric estimates of a density function. Ann. Math. Statist. 27, 832-837.*

[26] ***Sayah, A.*** *(2015). Kernel quantile estimation for heavy-tailed distributions. Thèse de doctorat, Université MohamedKhider, Biskra. 18, 178.189. 1966.*

[27] ***Silverman, B.W.****, (1986). Density estimation for statistics and data analysis. Chapman Hall, London.*

[28] ***Sheather, S. J. et J. S. Marron*** *(1990). Kernel quantile estimator. J. Am. Stat. Assoc., 85(410) : 410-416.*

[29] ***Yang, S-S.*** *(1985), A Smooth Nonparametric Estimator of a Quantile Function, Journal ofthe American Statistical Association, 80, 1004-1011.*