

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ KASDI MERBAH OUARGLA

FACULTÉ DE MATHÉMATIQUES ET SCIENCES DE LA MATIÈRE

DÉPARTEMENT DE MATHÉMATIQUES



Mémoire présenté en vue de l'obtention du Diplôme :

Master en Mathématiques

Option : **Probabilités et statistiques**

Par :

DEBBAKH Ali

Titre :

Estimation non paramétrique des quantiles

Membres du Comité d'Examen :

Dr. MANSOUL Brahim	UKMO	President
Dr. BOUREDJI Hind	UMKB	Encadreur
Dr. BENBRAHIM Radhia	UKMO	Examineur

juin 2024

Dedication

À mes parents,

À mes frères et sœurs,

et à toute ma famille,

À mes chers amis,

À tous les membres de ma promotion et à tous mes professeurs,

Enfin, à tous ceux qui m'ont aidée de proche ou de loin.

Remerciements

*Je tiens à exprimer ma profonde gratitude envers le tout-Puissant pour nous avoir donné la force et la capacité d'accomplir ce travail. Je suis extrêmement reconnaissant envers mes parents pour avoir inculqué en nous la valeur du travail acharné, pour leurs sacrifices, leur affection et leur soutien moral. Je tiens à remercier sincèrement **Dr. Bouredji Hind** d'avoir accepté de superviser ce mémoire, pour sa présence, ses conseils précieux et ses orientations. Son dévouement, son éthique professionnelle et sa vaste connaissance pratique parlent d'elles-mêmes. Je remercie également tous les membres du jury pour avoir accepté d'évaluer notre travail. Ma reconnaissance la plus sincère va également aux enseignants qui ont contribué à mon éducation. À mes chers amis qui ont toujours été fidèles et solidaires.*

Enfin, à toutes les personnes qui ont contribué, directement ou indirectement, à l'achèvement de ce mémoire. Veuillez accepter toute ma gratitude sincère.

Liste des tableaux

1.1	Quelques noyaux symetriques	14
1.2	Efficacit�e des noyaux	16
1.3	Quelques noyaux asymetriques	16
2.1	<i>MSE</i> associe � la loi <i>Exp</i> (2) avec l'utilisation de la noyau Epanechnikov	33
2.2	<i>MSE</i> associe � la loi <i>Beta</i> (2, 3) avec l'utilisation de la noyau Epanechnikov	33
2.3	<i>MSE</i> associe � la loi <i>Gamma</i> (1, 2) avec l'utilisation de la noyau Epanechnikov	34
2.4	<i>MSE</i> associe � la loi <i>Exp</i> (2) avec l'utilisation de la noyau Biweight	34
2.5	<i>MSE</i> associe � la loi <i>Beta</i> (2, 3) avec l'utilisation de la noyau Biweight	35
2.6	<i>MSE</i> associe � la loi <i>Gamma</i> (1, 2) avec l'utilisation de la noyau Biweight	35
2.7	<i>MSE</i> associe � la loi <i>Normale</i> (0, 1) par les deux methodes h_{amisc} et h_{ucv}	36

Table des figures

1.1	La fonction de répartition empirique	9
1.2	Noyaux symétriques	15
1.3	Noyaux asymétriques	17
2.1	l'estimateur empirique et l'estimateur à noyau d'une loi exponentielle	36
2.2	l'estimateur empirique et l'estimateur à noyau d'une loi beta	37
2.3	l'estimateur empirique et l'estimateur à noyau d'une loi gamma	37
2.4	La fonction de quantile Exp(2) et les estimateurs empirique (vert), à noyau (rouge) .	38
2.5	La fonction de quantile Beta(2,3) et les estimateurs empirique (vert), à noyau (rouge)	39
2.6	La fonction de quantile Gamma(1,2) et les estimateurs empirique (vert), à noyau (rouge)	39

Table des matières

Dedication	i
Remerciements	ii
Liste des tableaux	iii
Table des figures	iv
Introduction	2
1 La statistique d'ordre et les quantiles	3
1.1 Statistique d'ordre	3
1.1.1 Définition des statistiques d'ordre	3
1.1.2 Distribution des statistiques d'ordre	4
1.1.3 Fonction de répartition empirique	8
1.1.4 Inverse Generalisé	9
1.2 les quantiles	9
1.2.1 Quantile d'ordre p	10
1.2.2 La fonction quantile	10
1.3 Approche de l'estimation paramétrique et non paramétrique	12
1.3.1 Estimation ponctuelle	12
1.3.2 Estimation par intervalle de confiance	13
1.3.3 Méthode du maximum de vraisemblance	13
1.3.4 Estimation par la méthode de noyau	13
2 L'estimation à noyau de la fonction quantile	21
2.1 La fonction de quantile empirique	21
2.1.1 Propriétés asymptotiques de l'estimateur	21
2.2 L'estimateur à noyau	22
2.2.1 Propriétés asymptotiques de l'estimateur à noyau	22
2.2.2 Sélection du paramètre de lissage	26
2.2.3 Normalité asymptotique	28
2.3 Application numérique	28
2.3.1 Données de simulation	28
2.3.2 Code R	30
2.3.3 Resultats	33
2.3.4 Discussion	37
Conclusion	40

Bibliographie	41
Annexe A : Logiciel <i>R</i>	43
Annexe B : Critères importants	44
Annexe C : Table de notations et abréviation	46

Introduction

L'estimation est un outil essentiel pour les statisticiens. On a deux méthodes d'estimations : la méthode paramétrique et la méthode non-paramétrique.

Les quantiles sont fréquemment utilisés en Statistique. Par exemple, la médiane est un indicateur robuste de la tendance centrale d'une population, l'intervalle interquartile est un bon indicateur de sa dispersion. Dans la pratique, les domaines d'utilisation des quantiles sont assez variés. En biologie utilisent les quantiles pour estimer des courbes de référence permettant d'analyser certaines propriétés biophysiques de la peau. Et ils sont utilisé dans plusieurs domaines comme climatologie, hydrologie, assurance et finance, etc.

L'estimation de quantiles ou fonction quantile est un problème de base en statistique. Il a été étudiée par : Nadaraya (1964), Yamato(1973), Azzalini (1981), Yang (1985), Parzen (1979), Falk (1984), Padgett (1986), Sheather et Marron (1990), Ralescu et Sun (1993), Harrell et Davis (1982) et Park(2006).

Les deux techniques statistiques d'estimation du quantile utilisées dans la littérature sont : méthode paramétrique et méthode non paramétrique. Nous intéressons ici plutôt aux approches non paramétriques, nous serons plus précis (estimation a noyau), comme nous savons que la fonction quantile vous donne de nombreuses informations sur vos données, trouver un moyen d'estimer une fonction quantile facilite la tâche du statisticien.

Ce mémoire suivra une approche similaire comme suit :

Premier chapitre est réservé aux quelques notations et définitions essentielle. Il traite des statistiques d'ordre qui servent de base au chapitre suivant, nous discutons d'une définition, distribution et propriétés générales des statistiques d'ordre, plus important la fonction de répartition empirique et les quantile parce que nous devons les connaître avant de passer par ce processus d'estimation. Puis on présente les deux différentes approches d'estimations de la fonction quantile utilisées : "l'approche paramétrique et l'approche non paramétrique", mais on concentre sur l'estimation non paramétrique. Deuxième chapitre est focalisé sur l'estimation non paramétrique de la fonction quantile plus précisément la méthode du noyau, ensuite on a étudié les propriétés asymptotique de l'estimateur. A la fin on fait une application numérique, on simule l'estimateur empirique de la fonction de quantile et l'estimateur à noyau, puis on fait la comparaison. Tous les résultats sont effectués à l'aide du logiciel R.

Chapitre 1

La statistique d'ordre et les quantiles

La statistique d'ordre permet de déterminer le rang d'une observation dans un ensemble de données triées, tandis que les quantiles permettent de diviser les données en parties égales en termes de pourcentage des observations. Ces concepts sont essentiels pour résumer et analyser des ensembles de données dans divers domaines, y compris la statistique descriptive et l'analyse exploratoire des données.

1.1 Statistique d'ordre

La Statistique d'ordre représente le rang ou la position qu'occupe une valeur spécifique parmi toutes les valeurs de l'échantillon, lorsque celui-ci est arrangé dans un ordre particulier, tel que croissant ou décroissant. Par exemple, si nous avons une liste de nombres triés par ordre croissant, la statistique d'ordre d'une observation est simplement sa position dans cette liste. Cette mesure est utile pour identifier les valeurs les plus petites ou les plus grandes dans un ensemble de données, ainsi que pour comparer les observations entre elles en termes de leur position relative dans la distribution.

1.1.1 Définition des statistiques d'ordre

Définition 1.1.1. Soit X_1, X_2, \dots, X_n un échantillon de variables aléatoires indépendantes et identiquement distribuées (iid). Les statistiques d'ordre de cet échantillon sont les valeurs ordonnées telles que [1]

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

permutation aléatoire $\hat{\sigma} \in \mathfrak{S}_n$ telle que

$$(X_{1,n}, \dots, X_{n,n}) = (X_{\hat{\sigma}(1)}, \dots, X_{\hat{\sigma}(n)})$$

Proposition 1.1.1. On a,

1. Deux cas importants des statistiques d'ordre sont les statistiques du minimum et du maximum, représentées respectivement par les valeurs extrêmes :

$$X_{1,n} = \min(X_1, \dots, X_n), X_{n,n} = \max(X_1, \dots, X_n)$$

Pour tout $i = 1, \dots, n$, la variable aléatoire $X_{i,n}$ s'appelle la i -ème statistique d'ordre de l'échantillon.

2. Les variables aléatoires des statistiques d'ordre ne sont plus indépendantes et ont des distributions différentes.
3. Soit \mathfrak{S}_n l'ensemble des permutation de $\{1, \dots, n\}$
 On suppose F continue, la loi de $\hat{\sigma}$ est la loi uniforme sur \mathfrak{S}_n . De plus la permutation $\hat{\sigma}$ est indépendante de la statistique d'ordre.

1.1.2 Distribution des statistiques d'ordre

La détection des points boiteux dans un échantillon est l'un des rôles plus importants des statistiques d'ordre en statistique [1], on s'intéresse aux lois suivantes :

Loi de $X_{1,n}$

$$F_{1,n}(x) = P\{X_{1,n} \leq x\} = 1 - (1 - F(x))^n,$$

d'où

$$f_{1,n}(x) = nf(x)(1 - F(x))^{n-1},$$

Loi de $X_{n,n}$

$$F_{n,n}(x) = P\{X_{n,n} \leq x\} = (F(x))^n,$$

d'où

$$f_{n,n}(x) = nf(x)(F(x))^{n-1}.$$

Les expressions de $F_{1,n}$ et $F_{n,n}$ peuvent s'obtenir très facilement en considérant les relations

$$\{X_{1,n} \geq x\} \Leftrightarrow \min(X_1, \dots, X_n) \geq x \Leftrightarrow \bigcap_{i=1}^n \{X_i \geq x\}$$

et

$$\{X_{n,n} \leq x\} \Leftrightarrow \max(X_1, \dots, X_n) \leq x \Leftrightarrow \bigcap_{i=1}^n \{X_i \leq x\}.$$

En utilisant la propriété d'indépendance des variables aléatoires X_1, \dots, X_n , nous en déduisons que

$$\begin{aligned} F_{1,n}(x) &= P\{X_{1,n} \leq x\} \\ &= 1 - P\{X_{1,n} \geq x\} \\ &= 1 - P\left\{\bigcap_{i=1}^n \{X_i \geq x\}\right\} \\ &= 1 - \prod_{i=1}^n P\{X_i \geq x\} \\ &= 1 - \prod_{i=1}^n [1 - P\{X_i \leq x\}] \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

et pour $X_{n,n}$, on a

$$\begin{aligned} F_{n,n}(x) &= P\{X_{n,n} \leq x\} \\ &= P\left\{\bigcap_{i=1}^n \{X_i \leq x\}\right\} \\ &= \prod_{i=1}^n P\{X_i \leq x\} \\ &= [F(x)]^n. \end{aligned}$$

Exemple 2.1.1.

Nous considérons un échantillon d'une v.a. X qui suit une loi de Gumbel de paramètres $\lambda \in \mathbb{R}, \delta \in \mathbb{R}_+^*$, comme $F_X(x) = \exp(-\exp(-(\frac{x-\lambda}{\delta})))$, $x \in \mathbb{R}$, nous constatons que :

$$F_{X_{n,n}}(x) = F_X(x)^n = \exp(-\exp(\frac{x - \lambda - \delta \ln(n)}{\delta})), \quad \forall x \in \mathbb{R};$$

le maximum d'un échantillon d'une loi de Gumbel suit la loi de Gumbel de paramètres $\lambda + \delta \ln(n)$ et δ .

Distribution de la k-ème statistique d'ordre

L'expression de la distribution de $X_{k,n}$ est :

$$\begin{aligned} F_{k,n}(x) &= P\{X_{k,n} \leq x\} \\ &= \sum_{r=k}^n \binom{n}{r} (F(x))^r (1 - F(x))^{n-r}. \end{aligned}$$

Nous en déduisons que la fonction de densité est :

$$f_{k,n}(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x),$$

Propriétés générales

Soit (X_1, \dots, X_n) est une suite de variables iid :

1. La densité conjointe de n statistiques d'ordre est,

$$f_{X_{1,n}, \dots, X_{n,n}}(x_1, \dots, x_n) = \begin{cases} n! \prod_{i=1}^n f(x_i) & \text{si } -\infty < x_1 < \dots < x_n < +\infty. \\ 0 & \text{sinon.} \end{cases}$$

2. La fonction de répartition de $X_{k,n}$ est donnée par,

$$F_{X_{k,n}}(x) = \sum_{i=k}^n \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i}, \quad \forall x \in \mathbb{R}$$

3. La fonction de densité de $X_{k,n}$ est donnée par,

$$f_{k,n}(x) = n \binom{n-1}{k-1} [F(x)]^{k-1} [1-F(x)]^{n-k}$$

4. La fonction de répartition conjointe de $X_{r,n}$ et $X_{s,n}$ est donnée par, $\forall x, y \in \mathbb{R}$,

$$F_{X_{r,n}, X_{s,n}}(x, y) = \sum_{s=j}^n \sum_{r=i}^s \frac{n!}{r!(s-r)!(n-s)!} [F(x)]^r [F(y) - F(x)]^{s-r} [1-F(y)]^{n-s},$$

5. La fonction de densité conjointe de $X_{r,n}$ et $X_{s,n}$ est donnée par, $\forall x, y \in \mathbb{R}$,

$$f_{X_{r,n}, X_{s,n}}(x, y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} [F(x)]^{r-1} f(x) [F(y) - F(x)]^{s-r-1} f(y) [1-F(y)]^{n-s},$$

Où $\binom{n}{i} = \frac{n!}{i!(n-i)!}$ est la combinaison de i objets parmi n objets sans remise.

Preuve.

1. On rappelle que, puisque $f(x)$ est la dérivée de $F(x)$, alors

$$\begin{aligned} f(x) &= \lim_{\delta x \rightarrow 0} \frac{F(x + \delta x) - F(x)}{\delta x} \\ &= \lim_{\delta x \rightarrow 0} \frac{P(x \leq X < x + \delta x)}{\delta x} \\ &= \frac{P(x_1 \leq X_{1,n} < x_1 + \delta x_1, \dots, x_n \leq X_{n,n} < x_n + \delta x_n)}{\delta x_1 \dots \delta x_n} \\ &= \frac{n! P(x_1 \leq X_{1,n} < x_1 + \delta x_1) \dots P(x_n \leq X_{n,n} < x_n + \delta x_n)}{\delta x_1 \dots \delta x_n} \\ &= n! \frac{P(x_1 \leq X_{1,n} < x_1 + \delta x_1)}{\delta x_1} \dots \frac{P(x_n \leq X_{n,n} < x_n + \delta x_n)}{\delta x_n} \\ f_{X_{1,n}, \dots, X_{n,n}}(x_1, \dots, x_n) &= n! \prod_{i=1}^n f(x_i), \text{ avec } -\infty < x_1 \leq \dots \leq x_n < +\infty. \end{aligned}$$

2. Pour $X_{k,n}$ on a,

$$\begin{aligned} F_k(x) &= P(X_{k,n} \leq x) = P(\text{au moins } k \text{ des } X_i \text{ sont inférieurs à } x) \\ &= \sum_{i=k}^n p(\text{exactement } i \text{ des } X_1, \dots, X_n \text{ sont inférieurs à } x) \\ &= \sum_{i=k}^n \binom{n}{i} [F(x)]^i [1-F(x)]^{n-i}, \forall x \in \mathbb{R}. \end{aligned}$$

3. La fonction de répartition $F_k(x)$ de $X_{k,n}$ est donnée par

$$\sum_{i=k}^n \binom{n}{i} [F(x)]^i [1-F(x)]^{n-i}, \forall x \in \mathbb{R}.$$

La densité de $X_{k,n}$ est donnée par $n \binom{n-1}{k-1} [F(x)]^{k-1} [1 - F(x)]^{n-k}$ (Preuve 2). Ça peut être montré par le fait que,

$$f_{X_{i,n}}(x) = \frac{d}{dp} \sum_{i=k}^n \binom{n}{i} [p]^i [1-p]^{n-i} = n \binom{n-1}{k-1} [p]^{k-1} [1-p]^{n-k}.$$

4. La fonction de distribution jointe de $(X_{r,n}, X_{s,n})$ est :

$$F_{X_{r,n}, X_{s,n}}(x, y) = P(\{X_{r,n} \leq x\} \cap \{X_{s,n} \leq y\}), (x, y) \in \mathbb{R}.$$

1^{re} cas : $x \geq y$

$$\begin{aligned} F_{X_{r,n}, X_{s,n}}(x, y) &= P(X_{r,n} \leq x, X_{s,n} \leq y) \\ &= P(X_{s,n} \leq y) F_{X_{r,n}, X_{s,n}}(x, y) \\ &= F_{X_{r,n}, X_{s,n}}(y) \end{aligned}$$

2^{me} cas : $x < y$

$$\begin{aligned} F_{X_{r,n}, X_{s,n}}(x, y) &= P(X_{r,n} \leq x, X_{s,n} \leq y) \\ &= P(\text{au moins } r \text{ de } X_1, \dots, X_n \text{ sont inférieurs à } x \\ &\quad \text{et au moins } s \text{ de } X_1, \dots, X_n \text{ sont inférieurs à } y) \\ &= \sum_{s=j}^n \sum_{r=i}^s \frac{n!}{r!(s-r)!(n-s)!} [F(x)]^r [F(y) - F(x)]^{s-r} [1 - F(y)]^{n-s}. \end{aligned}$$

5. La fonction de répartition $F_{X_{r,n}, X_{s,n}}(x, y)$ de $(X_{r,n}, X_{s,n})$ est donnée par, $\forall (x, y) \in \mathbb{R}$

$$\sum_{s=j}^n \sum_{r=i}^s \frac{n!}{r!(s-r)!(n-s)!} [F(x)]^r [F(y) - F(x)]^{s-r} [1 - F(y)]^{n-s},$$

donc la densité de $(X_{r,n}, X_{s,n})$ est :

$$\begin{aligned} f_{X_{r,n}, X_{s,n}}(x, y) &= \frac{d}{dp} \sum_{s=j}^n \sum_{r=i}^s \frac{n!}{r!(s-r)!(n-s)!} [F(x)]^r [F(y) - F(x)]^{s-r} [1 - F(y)]^{n-s} \\ &= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \\ &\quad \times [F(x)]^{r-1} f(x) [F(y) - F(x)]^{s-r-1} f(y) [1 - F(y)]^{n-s} \end{aligned}$$

□

Définition 1.1.2. (Point terminal)

Le point terminal d'une distribution F_X est défini par :

$$x_F = \sup\{x \in \mathbb{R} : F_X(x) < 1\}$$

Définition 1.1.3. (Fonction de queue)

Soit X une variables aléatoire de densité f_X et de fonction de répartition F_X . On appelle fonction de queue la fonction définie par :

$$\bar{F}_X(x) = 1 - F_X(x)$$

Remarque 1.1.1. On a,

- Une loi de probabilité est dite à queue lourde ou à queue épaisse si sa fonction de répartition vérifie :

$$\int_{-\infty}^{+\infty} e^{\lambda x} d\bar{F}_X(x) = \infty \text{ pour tout } \lambda > 0.$$

Sinon la loi est dite à queue légère ou à queue fine.

- Une loi de probabilité est dite à queue longue ou possédant une longue traîne si le support de sa fonction de répartition n'est pas majoré et si pour tout $y > 0$,

$$\frac{\bar{F}_X(x+y)}{\bar{F}_X(x)} \rightarrow 1 \text{ lorsque } x \rightarrow +\infty.$$

Les lois à queue longue sont des lois à queue lourde.

1.1.3 Fonction de répartition empirique

Définition 1.1.4. La fonction de répartition empirique de l'échantillon (X_1, \dots, X_n) notée F_n est donnée par :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(X_i \leq x)} = \frac{1}{n} \text{Card}\{i \in \{1, \dots, n\} : X_i \leq x\}, \quad x \in \mathbb{R}.$$

Donc pour $i = 1, \dots, n$ on a,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(X_i \leq x)} = \begin{cases} 0 & \text{si } x < x_1 \\ \frac{1}{n} & \text{si } x_1 \leq x < x_2 \\ \vdots & \\ \frac{i-1}{n} & \text{si } x_{i-1} \leq x < x_i \\ \vdots & \\ 1 & \text{si } x \geq x_n \end{cases}$$

Où $\mathbb{1}_A$ est la fonction l'indicatrice :

$$\mathbb{1}_A = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{sinon} \end{cases}$$

Théorème 1.1.1. (Glivenko-Canteli)

Soit X_n avec $n \in \mathbb{N}$ une suite de variables aléatoires réelles indépendantes et de même loi F . Alors :

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} [F_n(x) - F(x)] = 0$$

donc $F_n(x) \xrightarrow{P.s.} F(x)$.

Preuve. Comme $nF_n(x)$ représente tout bonnement le nombre de points de l'échantillon qui tombent à gauche de x :

$$nF_n(x) = \sum_{i=1}^n \mathbb{1}_{]-\infty, x]}(X_i) = \sum_{i=1}^n Y_i,$$

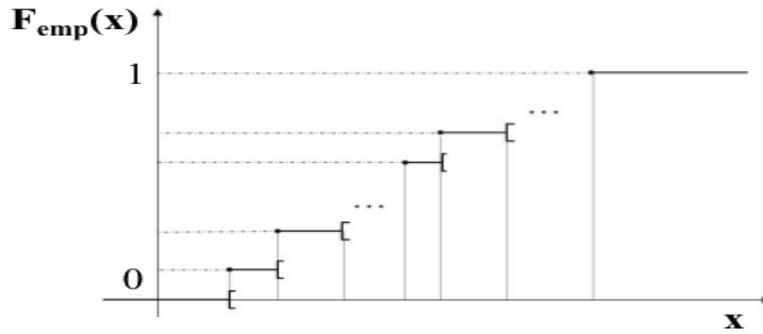


FIGURE 1.1 – La fonction de répartition empirique

où les Y_i sont i.i.d. selon une loi de Bernoulli de paramètre

$$p = P(Y_1 = 1) = P(X_i \leq x) = F(x),$$

d'où la loi binomiale pour leur somme. De la même façon, la loi des grands nombres appliquée aux variables Y_i assure que

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow[n \rightarrow \infty]{p.s.} \mathbb{E}[Y_1] = F(x).$$

□

1.1.4 Inverse Généralisé

Définition 1.1.5. On appelle inverse généralisé d'une fonction de distribution F définie de \mathbb{R} dans $[0, 1]$ la fonction notée $F^{\leftarrow}(x)$ définie de $[0, 1]$ dans \mathbb{R} par :

$$F^{\leftarrow}(x) = \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad 0 < p < 1$$

avec les conventions $\inf \mathbb{R} = -\infty$ et $\inf \emptyset = +\infty$.

Remarque 1.1.2. Si F est une fonction de répartition continue et strictement croissante alors la fonction inverse généralisé F^{\leftarrow} est équivalente à l'application réciproque F^{-1} .

Proposition 1.1.2. Soit F une fonction de répartition et F^{-1} son inverse généralisée. Alors :

- $F^{-1}(0) = -\infty$.
- F^{-1} est croissante.
- F^{-1} est continue à gauche.
- $\forall p \in [0, 1], F(x) \geq p \iff x \geq F^{-1}(p)$.

1.2 les quantiles

Les quantiles sont des mesures statistiques utilisées pour diviser une distribution de données en intervalles égaux en termes de proportion. Le p -ème quantile, noté $Q(p)$, est la valeur en dessous de laquelle une proportion p des observations se situe dans la distribution. Ils sont essentiels pour caractériser la répartition et la dispersion des données dans un ensemble statistique.

1.2.1 Quantile d'ordre p

Définition 1.2.1. Soit X est une variable aléatoire d'une distribution F continue, alors

$$F(x_p) = P(X \leq x_p) = p,$$

x_p : est appelé le quantile d'ordre p . Où $p \in]0, 1[$.

Remarque 1.2.1. Pour certain valeur de p , il y a des noms spéciaux au quantile :

- **La médiane** est le quantile d'ordre $p = \frac{1}{2}$. 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.
- **Le premier quartile** est le quantile d'ordre $p = \frac{1}{4}$. Ils servent à séparer les séries statistiques en quatre groupes de même taille.
- **Le troisième quartile** est le quantile d'ordre $p = \frac{3}{4}$.
- Le i -^{me} **quantile**, $1 \leq i \leq 4$, est le quantile d'ordre $p = \frac{i}{5}$.
- Le i -^{me} **décimale**, $1 \leq i \leq 9$, est le quantile d'ordre $p = \frac{i}{10}$.
- Le i -^{me} **vingtile**, $1 \leq i \leq 19$, est le quantile d'ordre $p = \frac{i}{20}$.
- Le i -^{me} **centile**, $1 \leq i \leq 99$, est le quantile d'ordre $p = \frac{i}{100}$.

Remarque 1.2.2. Comme remarque, on a,

1. Les quantiles sont appelé aussi fractiles.
2. Les quantiles sont souvent exprimés en pourcents, et donc on dira que x_p est le $100.p\%$ quantile de la distribution.
3. Les quartiles et centiles sont les échelles de centralisation sont particulièrement nécessaires lors du calcul de l'étendu.

1.2.2 La fonction quantile

Les quantiles sont des quantités qui peuvent être très utiles en statistique. Un quantile est obtenu à partir de la distribution de probabilité cumulée associée à une variable quantitative.

Définition 1.2.2. (fonction quantile)

Soit X une variable aléatoire qui a pour fonction de répartition : $F(x) = P(X \leq x)$, pour tout $0 < p < 1$. Dans le cas où $F(x)$ est continue et strictement croissante, le quantile d'ordre p de X est défini par :

$$Q_X(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

Cas des lois discrètes

La fonction de quantile d'une variable aléatoire discrète est une fonction en escalier, comme la fonction de répartition. Si les valeurs de X rangées par ordre croissant, la fonction de répartition est égale à,

$$F_n := F(X \leq x_n) = \sum_{i=1}^n P[X = x_i].$$

La fonction de quantile vaut :

$$Q(p) = x_n \mathbb{1}_{[F_n, F_{n+1}]}(p),$$

Exemple 2.3.1

Si X suit la loi géométrique de paramètre $q = 1 - p$, on a la fonction de répartition est défini comme suit,

$$G_q(x_n) = q(1 - q)^{x_n - 1}, x_n \in \mathbb{N}^*.$$

La fonction quantile Q est la fonction donnée par,

$$Q(p) = x_n, \text{ sur l'intervalle }]1 - (1 - q)^{x_n - 1}, 1 - (1 - q)^{x_n}].$$

Cas des lois continues

Dans le cas des lois usuelles, qui admettent une densité strictement positive sur un intervalle et nulle ailleurs, il existe un intervalle ouvert I tel que la fonction de répartition de la loi considérée réalise une bijection de I sur l'intervalle ouvert $]0, 1[$. Autrement dit, pour tout p strictement compris entre 0 et 1, il existe un unique élément x de I pour lequel $F_X(x) = p$. Ce nombre x est alors appelé le p -quantile de X ou de la loi de X . On peut le calculer par inversion de la fonction de répartition.

Exemple 2.3.2

Si X suit la loi exponentielle, $F_X : x \mapsto 1 - \exp(-\lambda x)$ réalise une bijection de $I =]0, +\infty[$ sur $]0, 1[$, le p -quantile de X est :

$$Q(p) = (F)^{-1}(p) = \frac{-\ln(1 - p)}{\lambda}.$$

Propriétés d'une fonction de quantile

La fonction de quantile posséder les propriété suivant :

- Elle est croissante.
- Elle est continue à gauche sur $]0, 1[$.
- La limite à droite de la fonction quantile 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.
- La fonction de quantile est la réciproque de la fonction de répartition, lorsque celle-ci réalise une bijection.
- Elle est continue sur un intervalle ouvert I sur $]0, 1[$, ce qui est le cas pour les lois usuelles admettant une densité.

Définition 1.2.3. (fonction quantile de queue)

La fonction quantile de queue est la fonction notée U définie par,

$$U(t) = F^{\leftarrow}\left(1 - \frac{1}{t}\right) \quad \text{avec } 1 < t < \infty$$

où F^{\leftarrow} étant l'inverse généralisé de F .

Définition 1.2.4. (fonction quantile extrême) la fonction quantile extrême ou quantile d'ordre $1 - p$, est définie comme suit :

$$x_{1-p} = Q(1 - p) = F^{-1}(1 - p) = \inf\{x \in \mathbb{R} : F(x) \geq 1 - p\}$$

où p proche de zéro.

1.3 Approche de l'estimation paramétrique et non paramétrique

Pour estimer la fonction de quantile nous recensons dans la littérature deux types d'approches que sont : "l'approche paramétrique et l'approche non paramétrique".

- L'approche paramétrique : On cherche à sa faire une idée de la valeur inconnue d'un paramètre, qui détermine la loi de probabilité d'une variable aléatoire. Il y a plusieurs méthodes pour estimer le paramètre d'une population inconnue.
- L'approche non paramétrique : Les méthodes non paramétriques ont été développées pour surmonter les problèmes rencontrés avec les méthodes paramétriques. Il est à noter que les méthodes non paramétriques ne nécessitent pas d'hypothèses sur la distribution de la population. Au lieu de cela, ces méthodes impliquent de considérer une distribution spécifique pour chaque observation individuelle dans l'ensemble de données, similaire à l'approche paramétrique qui consiste à ajuster une seule distribution à l'ensemble des observations.

Il y a deux propriétés fondamentales des estimateurs qui sont importantes : il faut obtenir une approximation de la bonne chose en moyenne, c'est l'absence de biais et il faut que l'augmentation de l'information disponible se traduise par une meilleure approximation, c'est la convergence. Mathématiquement cela se traduit par

- On dit qu'un estimateur est non biaisé si $E(\hat{R}) = R$ c'est-à-dire l'approximation mesure la bonne chose en moyenne.
- On dit qu'un estimateur est convergent si $Var(\hat{R}) \rightarrow 0$ lorsque n converge vers l'infini.

1.3.1 Estimation ponctuelle

L'estimation ponctuelle permet d'obtenir une approximation d'un paramètre de la population. Il est évidemment illusoire de vouloir connaître exactement un caractère aléatoire puisque sa principale propriété est justement de ne pas être connu avant la réalisation de l'expérience aléatoire et que les seules fenêtres que nous ayons sur ces phénomènes sont certaines de ses réalisations. D'une façon plus générale, posons θ un paramètre quelconque de la population qui peut être un indice de centre, de dispersion, de probabilité ou autre. Un estimateur de ce paramètre, que nous notons $\hat{\theta}$ est une fonction de l'échantillon qui donne une approximation de θ selon les observations disponibles, on note $\hat{\theta}(X_1, X_2, \dots, X_n)$. C'est une v. a. puisqu'elle dépend de l'échantillon particulier qui a été choisi mais c'est aussi une approximation du paramètre d'intérêt θ . On s'intéresse dans un premier temps aux paramètres principaux de la variable aléatoire : moyenne, variance (μ, σ^2). D'autres paramètres peuvent aussi être d'intérêt mais c'est plus rare. La précision de cette estimation dépend de la nature et de la taille de l'échantillon prélevé dans la population. Voici quelques estimateurs :

- Un estimateur ponctuel de la moyenne $E(X)$ est donné par

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

- Un estimateur ponctuel de la variance σ^2 est donné par

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

— Un estimateur ponctuel de l'écart type, σ , est donné par $S = \sqrt{S^2}$.

1.3.2 Estimation par intervalle de confiance

Soit $\alpha \in]0, 1[$. S'il existe des v.a $\theta_{min}(X_1, \dots, X_n)$ et $\theta_{max}(X_1, \dots, X_n)$ telles que

$$P(\theta \in [\theta_{min}(X_1, \dots, X_n), \theta_{max}(X_1, \dots, X_n)]) = 1 - \alpha,$$

on dit alors que $[\theta_{min}(X_1, \dots, X_n), \theta_{max}(X_1, \dots, X_n)]$ est un intervalle de confiance pour θ , avec coefficient de sécurité $1 - \alpha$. On le note $IC_{1-\alpha}(\theta)$.

Dans la pratique, on peut prendre par exemple $\alpha = 5\%$, ce qui nous donne un IC à 95% . Cela signifie qu'il y a 95% de chance que la valeur inconnue θ soit comprise entre $\theta_{min}(x_1, \dots, x_n)$ et $\theta_{max}(x_1, \dots, x_n)$.

1.3.3 Méthode du maximum de vraisemblance

La vraisemblance d'un échantillon (X_1, \dots, X_n) est définie par

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

si la loi de X est de densité f , et par

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n P(X_i = x_i | \theta)$$

si la loi de X est discrète. La log-vraisemblance est égale à $\ln L(X_1, \dots, X_n; \theta)$.

L'estimateur T_n du maximum de vraisemblance est la valeur de θ qui maximise L (ou $\ln(L)$).

Si L est deux fois dérivable, c'est la valeur θ telle que

$$\frac{\partial \ln(L)}{\partial \theta} = 0 \quad \text{et} \quad \frac{\partial^2 \ln(L)}{\partial \theta^2} < 0$$

Si Γ_n est un estimateur du maximum de vraisemblance du paramètre θ , alors $g(\Gamma_n)$ est aussi un estimateur du maximum de vraisemblance de $g(\theta)$ (mais il a nécessairement un biais si Γ_n n'en a pas).

1.3.4 Estimation par la méthode de noyau

La notion de noyau a été introduit pour la première fois par Parzen(1962)[14] et Rosenblatt(1956), mais Cacoulos(1966)[3] a été le premier qui utilise le terme noyau pour définir la fonction utilisée dans les méthodes non paramétriques. La méthode du noyau est une généralisation de la méthode de l'histogramme. C'est la plus utilisée parmi les méthodes d'estimation de densité non paramétriques, les fonctions d'agrégation,...

Définition 1.3.1. Une fonction $k : \mathbb{R} \rightarrow \mathbb{R}$ intégrable et telle que $\int_{-\infty}^{+\infty} k(y) dy = 1$ est appelée noyau.

Proposition 1.3.1. Il y a des contraintes sur le noyau utilisée uniquement pour simplifier les calculs d'optimisation du noyau :

- *Symétrique et positif sur sa domaine de définition : $k(x) = k(-x) \geq 0$.*
- $\int_{-\infty}^{+\infty} k(x) dx = 1$ et $k(x) \geq 0$. le noyau k est une densité de probabilité.
- $\int_{-\infty}^{+\infty} xk(x) dx = 0$ le noyau k est un noyau symétrique.
- $\int_{-\infty}^{+\infty} x^2k(x) dx = \mu_2$ une variance finie.

Noyaux symétriques

Les noyaux les plus couramment utilisés en pratique sont :

Noyau	Fonction
Epanechnikov	$k(u) = \frac{3}{4} (1 - u^2) \mathbb{1}_{[-1;1]}(u)$
Biweight	$k(u) = \frac{15}{16} (1 - u^2)^2 \mathbb{1}_{[-1;1]}(u)$
Triangulaire	$k(u) = (1 - u) \mathbb{1}_{[-1;1]}(u)$
Gaussien	$k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right)$
Uniforme	$k(u) = \mathbb{1}_{[-1;1]} \frac{(u)}{2}$
Cosinus	$k(u) = \frac{\pi}{4} \cos\left(\frac{u\pi}{2}\right) \mathbb{1}_{[-1;1]}(u)$

TABLE 1.1 – Quelques noyaux symétriques

La figure des ces noyaux sont présentées ci-dessous :

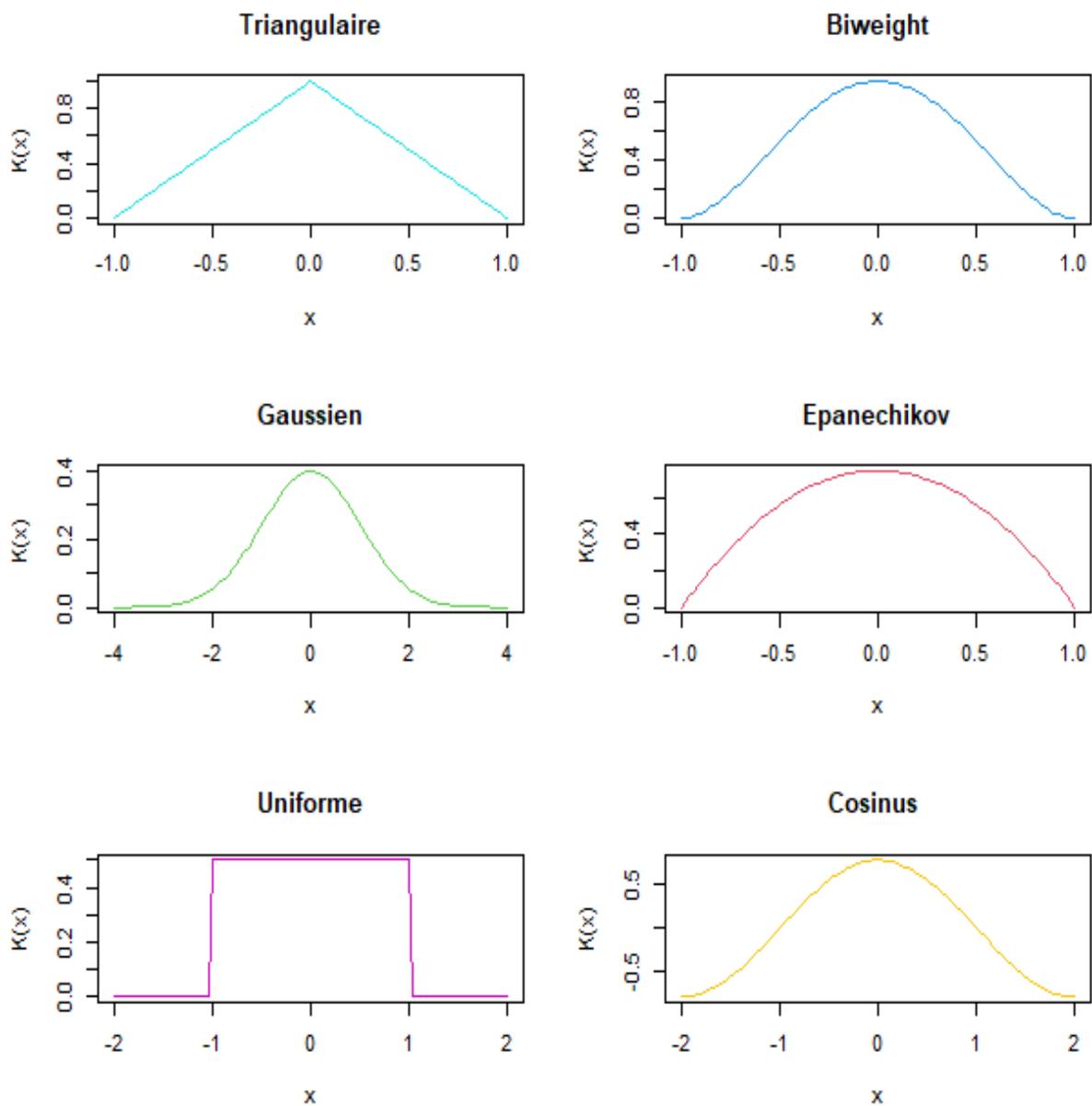


FIGURE 1.2 – Noyaux symétriques

Définition 1.3.2. L'efficacité d'un noyau k par rapporte noyau optimale est donnée par la relation [6]

$$Eff(k) = \left(\frac{R^4(k_{opt})\mu_2^2(k_{opt})}{R^4(k)\mu_2^2(k)} \right)^{\frac{1}{5}}.$$

avec

$$R(k) = \int_{-\infty}^{+\infty} k^2(x)dx.$$

Le noyau d'Epanechnikov est souvent appelé le noyau optimal k_{opt} .

Les efficacités des différents noyaux sont présentées dans le tableau 1.2 :

Noyau	μ_2	$R(k)$	Eff
Epanechnikov	1/5	3/5	1
Biweight	1/7	5/7	0.994
Triangulaire	1/6	2/3	0.986
Gaussien	1	$1/2\sqrt{\pi}$	0.951
Uniforme	1/2	1/2	0.930
Cosinus	$1 - (8/\pi^2)$	$\pi^2/16$	0.999

TABLE 1.2 – Efficacité des noyaux

Noyaux asymetriques

On résume les noyaux asymétriques dans le tableau 1.3 et le figure 1.3.

Noyau	Fonction
Gamma	$k(u) = \frac{h^{1-1/xh}}{\Gamma(\frac{1}{xh}-1)} u^{-1/xh} \exp(-\frac{1}{hu}), u \in \mathbb{R}^+$
Inverse Gamma	$k(u) = \frac{h^{1-1/xh}}{\Gamma(\frac{1}{xh}-1)} u^{-1/xh} \exp(-\frac{1}{hu}), u \in \mathbb{R}^+$
Béta	$k(u) = \frac{u^{x/h}(1-u)^{(1-x)/h}}{\beta(\frac{x}{h}+1, \frac{(1-x)}{h}+1)}, u \in [0, 1]$
Log normal	$k(u) = \frac{1}{b\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log(u)-a}{b}\right)^2\right\}, u \in \mathbb{R}^+$
Gaussienne-inverse(a, b)	$k(u) = \frac{\sqrt{b}}{\sqrt{2\pi}u^3} \exp\left\{-\frac{b}{2a}\left(\frac{u}{a} - 2 + \frac{a}{u}\right)\right\}, u \in [0, +\infty[$

TABLE 1.3 – Quelques noyaux asymetriques

Proposition 1.3.2. (Estimateur de Parzen-Rozenblatt) L’estimateur à noyau de la fonction de densité de probabilité f , noté $f_n(x)$, proposé par Parzen(1962) et Rosenblatt(1956) est donné par :

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} k\left(\frac{X_i - x}{h}\right),$$

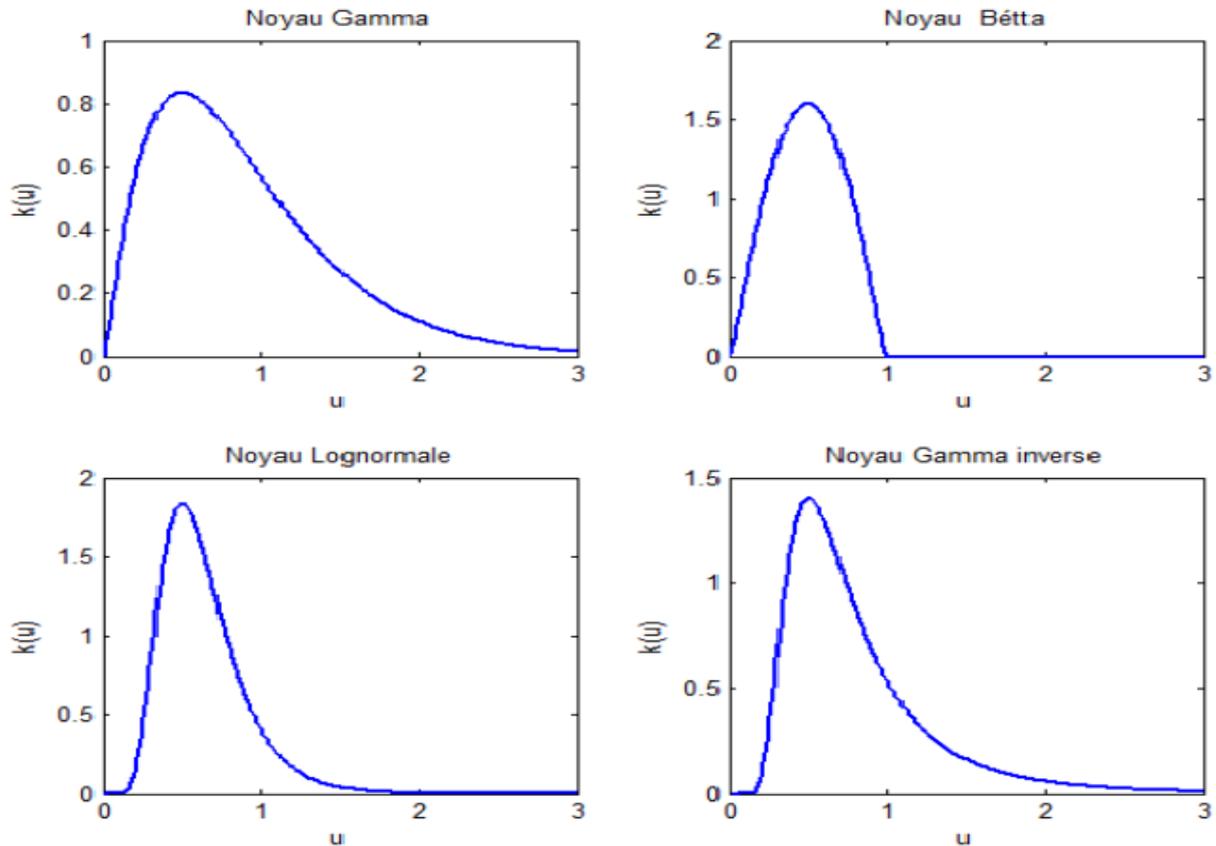


FIGURE 1.3 – Noyaux asymétriques

Preuve. La fonction de répartition empirique F_n est un bon estimateur de F . On sait que :

$$\begin{aligned} f(x) &= F^{(1)}(x) \\ &= \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} \\ &= \lim_{h \rightarrow 0} \frac{F(x+h) - F(x-h)}{2h}. \\ f(x) &\approx \frac{F(x+h) - F(x-h)}{2h}. \end{aligned}$$

En remplaçant F par son estimateur F_n , on obtient :

$$f_n(x) = \frac{F_n(x)(x+h) - F_n(x-h)}{2h} = \frac{1}{n} \sum_{i=1}^n \frac{1}{2h} \sum_{i=1}^n \mathbb{1}_{X_i \in]x-h, x+h]}(x)$$

Si on pose $k(x) = \frac{1}{2} \mathbb{1}_{]-1,1]}(x)$ alors on a,

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} k\left(\frac{X_i - x}{h}\right).$$

La fonction k est appelée le noyau de Rosenblatt (Uniforme)

□

Propriétés de l'estimateur à noyau f_n

Nous donnons les différentes propriétés d'un estimateur d'une densité de probabilité à noyau,

Le biais

$$\text{Biais}(f_n(x)) \approx \frac{h^2}{2} f^{(2)}(x) \int u^2 k(u) du + o(h^2)$$

Preuve. On a,

$$\begin{aligned} E[f_n(x)] &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h} E \left[k \left(\frac{x - X_i}{h} \right) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h} \int k \left(\frac{x - t}{h} \right) f(t) dt \\ &= \frac{1}{h} \int k \left(\frac{x - t}{h} \right) f(t) dt \end{aligned}$$

La transformation $z = \frac{x-t}{h}$, i. e. $t = -hz + x$, $\left| \frac{dz}{dt} \right| = \frac{1}{h}$

$$E[f_n(x)] = \int k(z) f(x - hz) dz$$

Un développement de Taylor de $f(x - hz)$ nous donne :

$$f(x - hz) = f(x) - hz f^{(1)}(x) + \frac{1}{2} (hz)^2 f^{(2)}(x) + o(h^2)$$

$$\begin{aligned} E[f_n(x)] &= \int k(z) f(x) dz - \int k(z) hz f^{(1)}(x) dz + \int k(z) \frac{(hz)^2}{2} f^{(2)}(x) dz + o(h^2) \\ &= f(x) \int k(z) dz - h f^{(1)}(x) \int z k(z) dz + \frac{h^2}{2} f^{(2)}(x) \int z^2 k(z) dz + o(h^2) \\ &= f(x) + \frac{h^2}{2} f^{(2)}(x) \int u^2 k(u) du + o(h^2) \end{aligned}$$

□

la variance

La variance de $f_n(x)$ est donnée par :

$$\begin{aligned} \text{Var}(f_n(x)) &= \text{Var} \left(\frac{1}{nh} \sum_{i=1}^n k \left(\frac{x - X_i}{h} \right) \right) \\ &= \frac{1}{n^2 h^2} \sum_{i=1}^n \text{Var} \left(k \left(\frac{x - X_i}{h} \right) \right) \end{aligned}$$

Preuve. Car les $X_i, i = 1, \dots, n$, sont i. i. d

$$\begin{aligned} \text{Var} \left(k \left(\frac{x - X_i}{h} \right) \right) &= E \left[k \left(\frac{x - X_i}{h} \right)^2 \right] - \left(E \left[k \left(\frac{x - X_i}{h} \right) \right] \right)^2 \\ &= \int k \left(\frac{x - X_i}{h} \right)^2 f(t) dt - \left(\int k \left(\frac{x - t}{h} \right) f(t) dt \right)^2 \\ \text{Var} (f_n(x)) &= \frac{1}{n} \int \frac{1}{h^2} k \left(\frac{x - t}{h} \right)^2 f(t) dt - \frac{1}{n} \left(\frac{1}{h} \int k \left(\frac{x - t}{h} \right) f(t) dt \right)^2 \\ &= \frac{1}{n} \int \frac{1}{h^2} k \left(\frac{x - t}{h} \right)^2 f(t) dt - \frac{1}{n} (f(x) + \text{Biais} (f_n(x)))^2 \end{aligned}$$

En effectuant le changement de variable suivant $z = \frac{x-t}{h}$, on obtient :

$$\text{Var} (f_n(x)) = \frac{1}{nh} \int k(x)^2 f(x - hz) dz - \frac{1}{n} (f(x) + o(h^2))^2$$

Et en effectuant un développement limité à l'ordre 2, il vient :

$$\begin{aligned} \text{Var} (f_n(x)) &= \frac{1}{nh} \int k(z)^2 (f(x) - hz f^{(1)}(x) + o(h)) dz - \frac{1}{n} (f(x) + o(h^2))^2 . \\ \text{Var} (f_n(x)) &= \frac{1}{nh} f(x) \int k^2(z) dz + o \left(\frac{1}{nh} \right) \end{aligned}$$

d'où :

$$\text{Var} (f_n(x)) \approx \frac{1}{nh} f(x) \int k^2(z) dz$$

□

Erreur quadratique moyenne (Mean-Squared Error MSE)

Le risque quadratique moyenne de l'estimateur de $f(\hat{x})$ de $f(x)$ est donne par

$$\begin{aligned} E\{\hat{f}_n(x) - f(x)\}^2 &= E \left\{ \hat{f}_n(x) - E\{\hat{f}_n(x)\} + E\{\hat{f}_n(x)\} - f(x) \right\}^2 \\ &= \text{Var}\{\hat{f}_n(x)\} + \left[E\{\hat{f}_n(x)\} - f(x) \right]^2 \\ &= \text{Var}\{\hat{f}_n(x)\} + \left[\text{Biais}\{\hat{f}_n(x)\} \right]^2 . \end{aligned}$$

a partir de la bais et variance asymptotiques on peut obtenir la MSE et la MISE

$$\begin{aligned} \text{MSE}\{\hat{f}_n(x)\} &= \frac{1}{4} h^4 \mu_2^2 \{f^{(2)}(x)\}^2 + \frac{1}{nh} f(x) R(k) + o \left(h^4 + \frac{1}{nh} \right) \\ \text{MISE}\{\hat{f}_n(x)\} &= \frac{1}{4} h^4 \mu_2^2 \int \{f^{(2)}(x)\}^2 dx + \frac{1}{nh} R(k) + o \left(h^4 + \frac{1}{nh} \right) , \end{aligned}$$

Donc, si $h \rightarrow 0$ et $nh \rightarrow \infty$ quand $n \rightarrow \infty$, on a que

$$E\{\hat{f}_n(x) - f(x)\}^2 \rightarrow 0$$

pour tout point x . L'estimateur simple $\hat{f}_n(x)$ est alors un estimateur consistant de $f(x)$

Remarque 1.3.1. *On a :*

- *Si h décroît, alors $(\text{Biais})^2$ décroît et la variance croît.*
- *Si h croît, alors $(\text{Biais})^2$ croît et la variance décroît.*

Chapitre 2

L'estimation à noyau de la fonction quantile

La méthode du noyau est l'une des méthodes d'estimation non paramétrique la plus utilisée. 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 et du coefficient de lissage. Ce chapitre est principalement consacré à l'étude d'estimation par la méthode à noyau des quantiles, nous mentionnons la général forme de cette estimateur, comment choisir le noyau et la fenêtre optimaux, et mentionner quelques détails connexes sur cette méthode et les propriétés asymptotiques de l'estimateur.

2.1 La fonction de quantile empirique

Définition 2.1.1. La fonction de quantile empirique de l'échantillon X_1, \dots, X_n , est donnée par [12] :

$$Q_n(p) = F_n^{-1}(p) = \inf\{x \in \mathbb{R} : F_n(x) \geq p\}, \quad 0 < p < 1$$

Définition 2.1.2. Soit (X_1, \dots, X_n) un n échantillon issu d'une distribution F et $X_{1,n}, \dots, X_{n,n}$ les statistiques d'ordre associées. Soit $p \in]0, 1[$. La statistique d'ordre $X_{([np]+1),n}$ s'appelle le quantile empirique d'ordre p de l'échantillon.

$$X_{([np]+1),n} = Q_n(p) = \inf\{x \in \mathbb{R} : F_n(x) \geq p\}$$

où $[np]$ désigne la partie entière de np .

Remarque 2.1.1. le p^{me} quantile empirique peut être aussi définir en fonction des statistiques d'ordre par

$$Q_n(p) = X_{i,n} \quad \text{ou} \quad \frac{i-1}{n} \leq p < \frac{i}{n}.$$

2.1.1 Propriétés asymptotiques de l'estimateur

Soit $p_m = \frac{m}{(n+1)}$ et $q_m = 1 - p_m$, si nous utilisons $X_{m,n}$ pour estimer le quantile d'ordre p_m , alors le biais asymptotique et la variance sont :

$$\text{Biais}\{X_{m,n}\} \approx \frac{p_m q_m Q^{(2)}(p_m)}{2(n+2)} + \frac{p_m q_m}{(n+2)^2} \left\{ \frac{1}{3}(q_m - p_m)Q^{(3)}(p_m) + \frac{1}{8}Q^{(4)}(p_m) \right\},$$

et

$$\begin{aligned} \text{Var}\{X_{m,n}\} &\approx \frac{p_m q_m}{(n+2)} (Q^{(1)}(p_m))^2 \\ &+ \frac{p_m q_m}{(n+2)^2} \left\{ 2(q_m - p_m) Q^{(1)}(p_m) Q_m^{(3)} + p_m q_m (Q^{(1)}(p_m) Q^{(3)}(p_m) + \frac{1}{2} Q^{(2)}(p_m)) \right\}. \end{aligned}$$

L'erreur quadratique moyenne asymptotique de $X_{m,n}$ doit être :

$$AMSE\{X_{m,n}\} = (ABiais\{X_{m,n}\})^2 + Avar\{X_{m,n}\}.$$

2.2 L'estimateur à noyau

Définition 2.2.1. *Estimateur de noyau pour un quantile est donné par Yang (1985) et par Parzen (1979) [14]. Il est défini par*

$$\tilde{Q}_n(p) = \frac{1}{h} \sum_{i=1}^n X_{i,n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} k\left(\frac{x-p}{h}\right) dx = \sum_{i=1}^n X_{i,n} \left[K\left(\frac{\frac{i}{n}-p}{h}\right) - K\left(\frac{\frac{i-1}{n}-p}{h}\right) \right]$$

où $X_{i,n}$ est la i -ième statistique d'ordre et h le paramètre de lissage qui vérifie $h \rightarrow 0$ quand $n \rightarrow \infty$

2.2.1 Propriétés asymptotiques de l'estimateur à noyau

Nous présentons dans cette partie les propriétés statistiques de l'estimateur à noyau de la fonction de quantile. Supposant que le noyau k , la fenêtre h et Q satisfaisaient les hypothèses suivantes :

Hypothèse 1. Q est dérivable.

Hypothèse 2. $Q^{(i)}$ (la dérivée i ème de Q) sont bornées.

Hypothèse 3. k est de support compact.

Hypothèse 4. $h \rightarrow 0$ quand $n \rightarrow \infty$

On peut écrire \tilde{Q}_n sous la forme,

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

où F_n est la fonction de répartition empirique.

On définit la fonction $\varphi(k)$ comme suit :

$$\varphi(k) = 2 \int uk(u)K(u)du.$$

où K désigne la primitive de k .

Proposition 2.2.1. *Sous certaines hypothèses sur le noyau k et le paramètre de lissage h , on a*

1. Biais de $\tilde{Q}_n(p)$

$$\text{Biais}(\tilde{Q}_n(p)) = \frac{h^2}{2!} Q^{(2)}(p) \int y^2 k(y) dy + o(h^2).$$

L'estimateur \tilde{Q}_n de la fonction Q est asymptotiquement non biaisé.

2. Variance de $\tilde{Q}_n(p)$

$$\text{Var}(\tilde{Q}_n(p)) = \frac{p(1-p)}{n} (Q^{(1)}(p))^2 - \frac{h}{n} (Q^{(1)}(p))^2 \int xk(x)K(x)dx + o(n^{-1}h).$$

Preuve.

On a

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

Soit le changement de variable $y = \frac{x-p}{h}$ donc $x = yh + p$ et $dx = hdy$, par conséquent le

Biais de \tilde{Q}_n sera

$$\text{Biais}(\tilde{Q}_n(p)) = \int_{-\frac{p}{h}}^{\frac{1-p}{h}} k(y)Q(yh+p)hdy - Q(p).$$

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

$$Q(yh+p) = Q(p) - \frac{yh}{1!}Q_X^{(1)}(p) + \frac{y^2h^2}{2!}Q^{(2)}(p) + o(y^2h^3).$$

$$\begin{aligned} \text{Biais}(\tilde{Q}_n(p)) &= \int_{-\frac{p}{h}}^{\frac{1-p}{h}} k(y)[Q(p) - \frac{yh}{1!}Q^{(1)}(p) + \frac{y^2h^2}{2!}Q^{(2)}(p) + o(y^3h^3)] - Q(p) \\ &= \int_{-\frac{p}{h}}^{\frac{1-p}{h}} k(y)Q(p)dy - \int_{-\frac{p}{h}}^{\frac{1-p}{h}} yhQ_X^{(1)}(p)k(y)dy + \int_{-\frac{p}{h}}^{\frac{1-p}{h}} k(y)\frac{y^2h^2}{2!}Q^{(2)}(p)dy \\ &\quad + o(y^2h^3) \int_{-\frac{p}{h}}^{\frac{1-p}{h}} k(y)dy - Q(p) \\ &= Q(p) \int_{-\frac{p}{h}}^{\frac{1-p}{h}} k(y)dy - hQ^{(1)}(p) \int_{-\frac{p}{h}}^{\frac{1-p}{h}} yk(y)dy + \frac{h^2}{2!}Q^{(2)}(p) \int_{-\frac{p}{h}}^{\frac{1-p}{h}} y^2k(y)dy \\ &\quad + o(y^2h^3) - Q(p). \end{aligned}$$

Lorsque n tend vers l'infini, on obtient :

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{Biais}(\tilde{Q}_n(p)) &= Q(p) \int k(y)dy - hQ^{(1)}(p) \int yk(y)dy + \\ &\quad \frac{h^2}{2!}Q^{(2)}(p) \int y^2k(y)dy + o(y^2h^3) - Q(p) \\ &= \frac{h^2}{2!}Q^{(2)}(p) \int y^2k(y)dy + o(y^2h^3). \end{aligned}$$

$\lim_{n \rightarrow \infty} \text{Biais}(\tilde{Q}_n(p)) = 0$, alors \tilde{Q}_n est asymptotiquement non biaisé.

□

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

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

Théorème 2.2.1. *En supposant que $Q^{(2)}$ est continue au voisinage de p et k est la densité d'un noyau, symétrique autour de 0 . Pour tout $p \in (0, 1)$.*

— Quand F est symétrique et $p \neq 0.5$, on a :

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

— Quand F est symétrique, pour $p = 0.5$, on a :

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

Preuve.

Premièrement on considère $p \in (0, 1)$ avec $p \neq 0.5$ et F symétrique. $Q^{(2)}$ est continue au voisinage de p . On a :

$$\begin{aligned} \text{Biais}(\tilde{Q}_n(p)) &= \sum_{i=1}^n \left[\int_{\frac{i-1}{n}}^{\frac{i}{n}} K\left(\frac{t-p}{h}\right) dt \right] \left\{ Q\left(\frac{i}{n+1}\right) - Q(p) \right\} + o\left(\frac{1}{n}\right) \\ &= \int_0^1 K\left(\frac{t-p}{h}\right) \{Q(t) - Q(p)\} dt + o\left(\frac{1}{n}\right). \\ &= \frac{1}{2h^2} \left[\int u^2k(u)du \right] Q^{(2)}(p) + o(h^2) + o\left(\frac{1}{n}\right) \end{aligned}$$

avec

$$\text{Var}(\tilde{Q}_n(p)) = \frac{p(1-p)}{n} [Q^{(1)}(p)]^2 - \frac{h}{n} [Q^{(1)}(p)]^2 \int k(u)k(u)du + o\left(\frac{h}{n}\right)$$

or

$$MSE(\tilde{Q}_n(p)) = \text{Var}(\tilde{Q}_n(p)) + \text{Biais}^2(\tilde{Q}_n(p))$$

d'où

$$\begin{aligned} \text{MSE}(\tilde{Q}_n(p)) &= \frac{p(1-p)}{n} [Q^{(1)}(p)]^2 - 2\frac{h}{n} [Q^{(1)}(p)]^2 \int uk(u)k(u)^{-1}du \\ &= \frac{1}{4h^4} [Q^{(2)}(p)]^2 \left[\int u^2k(u)du \right]^2 + o\left(\frac{h}{n}\right) + o(h^4) \end{aligned}$$

□

Corollaire 2.2.1. *En s'appuyant sur Falk (1984) [5], Sheader et Marron (1990) ont donné l'erreur quadratique moyenne asymptotique (AMSE) de $\tilde{Q}_n(p)$ comme suit :*

- Lorsque F est symétrique et $p \neq 0.5$

L'erreur quadratique moyenne asymptotique de $\tilde{Q}_n(p)$ est

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

si $\tilde{Q}_n^{(1)}(p) > 0$, la fenêtre optimale asymptotique pour $AMSE(\tilde{Q}_n(p))$ est

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

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

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

- Lorsque F est symétrique et $p = 0.5$

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

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

$R(k)$: est précédemment connu.

Corollaire 2.2.2. *Pour tout p , avec $p \neq 0.5$ et F symétrique, le paramètre de lissage optimal asymptotique est donné par*

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

où

$$\alpha(k) = \left[\frac{\varphi(k)}{\left\{ \int u^2k(u)du \right\}^2} \right]^{\frac{1}{3}}$$

et

$$\beta(Q) = \left[\frac{Q^{(1)}(p)}{Q^{(2)}(p)} \right]^{\frac{2}{3}}$$

Avec $h = h_{\text{opt}}$

$$\text{MSE} \left(\tilde{Q}_n(p) \right) = \frac{p(1-p)}{n} [Q^{(1)}(p)]^2 + o\left(n^{-\frac{4}{3}}\right)$$

Si F est symétrique et $p = 0.5$ prendre $h = o(n^{-\frac{1}{2}})$. Cela rend les deux premiers termes dans h de MSE pour $\tilde{Q}(0.5)$ dans le même ordre

$$\text{MSE}\{\tilde{Q}_n(0.5)\} = \frac{0.25}{n} (Q^{(1)}(0.5))^2 + o(n^{-\frac{3}{2}}).$$

La convergence

Corollaire 2.2.3. *L'estimation $\tilde{Q}_n(p)$ converge vers $Q(p)$ en moyen quadratique c'est à dire :*

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

Preuve. Comme est asymptotiquement sans biais et sa variance tends vers 0, alors

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

c-a-d :

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

□

2.2.2 Sélection du paramètre de lissage

Le paramètre de lissage joue un rôle essentiel dans la mise en œuvre d'une estimation pratique. Il détermine le compromis entre le degré de régularité obtenu et la proximité de l'estimation avec la vraie distribution.

Le paramètre de lissage optimal

Pour un choix optimal de h pour $Q_n(p)$, avec $p \neq 0.5$ quand F est symétrique. Nous voyons que pour un choix de k la valeur asymptotiquement optimale de h dépend des première et deuxième dérivées du quantile. Les estimateurs de $Q^{(1)}(p)$ et $Q^{(2)}(p)$ sont nécessaires pour le choix du h [15] [2].

$$\hat{Q}^{(1)}(p) = \sum_{i=1}^n \left[\int_{\frac{i-1}{n}}^{\frac{i}{n}} a^{-2} k^{(1)}(a^{-1}(t-p)) dt \right] X_{i,n}$$

et

$$\hat{Q}^{(2)}(p) = \sum_{i=1}^n \left[\int_{\frac{i-1}{n}}^{\frac{i}{n}} b^{-3} k^{(1)}(b^{-1}(t-p)) dt \right] X_{i,n}$$

tel que k c'est un noyau d'ordre t -ème symétrique autour de zéro(c'est à dire

$$\int k(u) du = 1, \int u^J k(u) = 0, J = 1, 2, \dots, m-1, \int u^t k(u) du < \infty, \quad k \in L^2(-\infty, +\infty), k^t \in \text{lip}(\alpha).$$

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

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

tel que :

$$\hat{\beta} = \left\{ \frac{\hat{Q}^{(1)}(p)}{\hat{Q}^{(2)}(p)} \right\}^{\frac{2}{3}},$$

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 $\hat{\beta}$.

Méthodes validation croisée (cross validation)

la Méthode validation croisée le type des moindres carrés a été introduit par Rudemo (1982) et Bowman (1984) [2]. L'idée de base des méthodes de validation croisée est de trouver une fonction de score CV (h) ayant la même structure que $MISE(h)$ et dont le calcul est plus simple, lors de la mesure de la proximité entre une fonction estimée et une fonction réelle, la MISE (Moyenne Intégrée des Carrés de l'Erreur) est définie comme suit :

$$MISE(h) = E \int_0^1 \left\{ \tilde{Q}_n(p) - Q(p) \right\}^2 dp$$

La valeur qui minimise la MISE, notée h est le paramètre de lissage optimal, mais en pratique, cette valeur est inconnue. Pour approximer la MISE, on utilise souvent

$$ASE(h) = \frac{1}{n} \sum_{i=1}^n \left\{ \tilde{Q} \left(\frac{i}{n} \right) - Q \left(\frac{i}{n} \right) \right\}^2.$$

Le $Q(p)$ inconnu est remplacé par Q_n et une fonction de procédure de validation croisée est créée comme suit :

$$ASE(h) = \frac{1}{n} \sum_{i=1}^n \left\{ \tilde{Q}_{-i} \left(\frac{i}{n} \right) - Q_n \left(\frac{i}{n} \right) \right\}^2.$$

où \tilde{Q}_{-i} désigne l'estimateur à noyau évalué à l'observation x_i , mais construit à partir des données avec l'observation x_i omise.

Le principe de cette méthode représentant chaque observation par une fonction delta de Dirac $\delta(x - x_i)$, dont l'espérance est $f(x)$, et la comparant avec une estimation de densité basée sur le reste des données.

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n k \left(\frac{x - x_i}{h} \right),$$

quand $h \rightarrow 0$, $k(x - x_i) \rightarrow \delta(x - x_i)$.

La méthode du noyau pour la fonction de répartition

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n K \left(\frac{x - x_i}{h} \right),$$

Lorsque $h \rightarrow 0$, la fonction de distribution K est une fonction de distribution, h est la largeur de bande qui contrôle le degré de lissage.

$$K \left(\frac{x - X_i}{h} \right) \rightarrow \mathbb{1}(x - x_i),$$

avec $\mathbb{1}(x - x_i)$ est la fonction indicatrice.

$$\mathbb{1}(x - x_i) = \begin{cases} 1, & \text{if } x - x_i \geq 0, \\ 0, & \text{sinon.} \end{cases}$$

avec $\text{MISE}(h)$ formule quand $h \rightarrow 0$

$$\tilde{Q}(p) \longrightarrow \delta\left(\frac{i}{n} - p\right) X_{i,n},$$

et donc une fonction de validation croisée peut être écrite comme

$$\text{CV}(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 \left\{ \delta\left(\frac{i}{n} - p\right) X_{i,n} - \tilde{Q}_{-i}\left(\frac{i}{n}\right) \right\}^2 dp.$$

Le paramètre de lissage h est ensuite choisi pour minimiser cette fonction. En soustrayant un terme qui caractérise les performances du vrai p , nous avons

$$H(h) = \text{CV}(h) - \frac{1}{n} \sum_{i=1}^n \int_0^1 \left\{ \delta\left(\frac{i}{n} - p\right) X_{i,n} - Q\left(\frac{i}{n}\right) \right\}^2 dp$$

2.2.3 Normalité asymptotique

Théorème 2.2.2. *Si $Q_X^{(1)} > 0$, et si Q_X a une dérivée seconde bornée au voisinage de $q \in (0, 1)$, avec k a un support borné, tel que $\int k(x)dx = 1$ et $h \xrightarrow{n \rightarrow 0} 0$, on a,*

$$\frac{n^{\frac{1}{2}} \left[\tilde{Q}_n(p) - E\left(\tilde{Q}_n(p)\right) \right]}{\sqrt{\text{Var}\left[\tilde{Q}_n(p)\right]}} \longrightarrow N(0, 1).$$

Preuve. Voir Falk (1985). □

2.3 Application numérique

On définit simulation statistique toute méthode qui utilise des séquences des nombres aléatoires. Les nombres aléatoires peuvent être générés par les observations d'expériences de nature aléatoire ou par l'ordinateur : ceux ci sont appelés nombres pseudo-aléatoires. Dans ce partie, en appliquant les deux méthodes non paramétriques de les chapitres précédent pour estimé la fonction quantile et puis en calculant l'erreur quadratique moyenne (MSE). Cette étude de simulation est réalisée avec le logiciel R.

2.3.1 Données de simulation

Nous souhaitons tester les différentes méthodes d'estimation considérées dans notre étude sur quelque distributions de probabilité. Notre objectif est de :

- Vérifier les propriétés d'efficacité de ces méthodes d'estimation.

- Comparer les méthodes de sélection du paramètre de lissage.

Les noyau utilisé sont l'Epanechnikov et Biweight,

$$k(u) = \frac{3}{4} (1 - u^2) \mathbb{1}_{[-1;1]}(u)$$

et

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

respectivement.

On prend un échantillon suit une loi choisit de taille $n = 10$, $n = 50$, $n = 100$ et $n = 500$ et pour chaque distribution nous simulons $R = 100$.

Pour l'application numérique nous choisissons les distributions suivantes :

- Une loi exponentielle de paramètre $\lambda = 2$

$$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty[}(x);$$

- Une loi Beta de paramètre (α, β) :

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad 0 \leq x \leq 1, \alpha > 0, \beta > 0$$

- Une loi Gamma de paramètre (λ, a) :

$$f_X(x) = \frac{\lambda^a}{\Gamma(a)} e^{-\lambda x} x^{a-1} \mathbb{1}_{x \geq 0}(x)$$

- Une loi normale de paramètre (m, σ) :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Les étapes de simulation :

- Générer l'échantillon X_1, \dots, X_n issue d'une v.a d'une loi choisi.
- Ordonner l'échantillon simulé (Statistique d'ordre $X_{i,n}$)
- Calculer le paramètre de lissage h par les deux methodes.
- Calculer Q_n et \tilde{Q}_n l'estimateurs de la fonction quantile donnée par

$$Q_n(p) = \inf \{x : F_n(x) \geq p\}, 0 \leq p \leq 1$$

et

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

- Calculer le MSE de chaque estimateurs dans les cas : $p = 0.1$, $p = 0.5$ et $p = 0.7$.
- Représenter la fonction de quantile et leurs estimateurs (Empirique et à noyau), avec tous les tailles de l'échantillon.
- Tracer la boîte de moustache des estimateurs de la fonction quantile dans le cas au $p = 0.5$ ($Q(0.5)$).

2.3.2 Code R

Résultats numériques résumés dans les tableaux

```

1  ### R s u l t a t s n u m r i q u e s ###
2  rm ( list = ls () )
3  library ( kedd )
4  library ( MASS )
5  library ( Metrics )
6  n =500 # Taille
7  R =100 # Repetition
8  k = fonction ( t ) {(3 / 4) * (1 - t ^2) * ( abs ( t ) <=1) } # ##
    Noyau
9  p =0.1 # P r o b a b i l i t
10 qt = qbeta (p ,2 ,3) # Quantile theorique d      ordre p
11 W = V = numeric ( R )
12 for ( j in 1: R ) {
13 X = rbeta (n ,2 ,3) # chantillon
14 Y = sort ( X ) # Statistique d      ordre
15
16 # L      estimateur empirique de la fonction quantile #
17 q = fonction ( p ) {
18 m = floor ( p * n )
19 s = m +1
20 Y [ s ]
21 }
22 # ### Le parametre de lissage ###
23 ban = h . ucv (X , deriv . order =0 , lower =0.1 , upper =0.5 ,
    tol =0.001 , kernel = "
24 epanechnikov " )
25 h = ban $ h
26 # ### L      estimateur a noyau de la fonction quantile #####
27 qx = fonction ( p ) {
28 M = numeric ( n )
29 for ( i in 1: n ) {
30 M [ i ]= integrate (k ,((( i -1) / n ) -p ) /h ,(( i / n ) -p ) /
    h ) $ value
31 }
32 sum ( Y * M )
33 }
34 W [ j ]= q ( p )
35 V [ j ]= qx ( p )
36 }
37 # ## MSE ##
38 mse (W , qt )
39 mse (V , qt )

```

Graph des noyaux symétriques

```

1 K1=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}
2 K2=function(t){(15/16)*((1-t^2)^2)*ifelse(abs(t)<=1,1,0)}
3 K3=function(t){dnorm(t)}
4 K4=function(t){ifelse(abs(t)<=1,(3/4)*(1-t^2),0)}
5 K5=function(t){ifelse(abs(t)<=1,(1/2),0)}
6 K6=function(t){(pi/4)*(cos((pi*t)/2))}
7 op=par(mfrow=c(3,2))
8 curve(K1(x),-1,1,ylab="K(x)",main="Triangulaire",col="5")
9 curve(K2(x),-1,1,ylab="K(x)",main="Biweight",col="4")
10 curve(K3(x),-4,4,ylab="K(x)",main="Gaussien",col="3")
11 curve(K4(x),-1,1,ylab="K(x)",main="Epanechnikov",col="2")
12 curve(K5(x),-2,2,ylab="K(x)",main="Uniforme",col="6")
13 curve(K6(x),-2,2,ylab="K(x)",main="Cosinus",col="7")
14 par(op)

```

Graph de la fonction de quantile et leurs estimateurs

```

1 par(mfrow=c(2,2))
2 library(kedd)
3 n=10
4 R=10
5 k=function(t){(15/16)*((1-t^2)^2)*(abs(t)<=1)}
6 p=seq(0,0.9,length.out=10)
7 qu=qexp(p,2)
8 W=V=numeric(R)
9 for(j in 1 :R){
10 X=rexp(n,2)
11 ##### Empirique #####
12 Y=sort(X)
13 q=function(p){
14 m=floor(p*n)
15 s=m+1
16 Y[s]
17 }
18 ban=h.ucv(X,deriv.order=0,lower=0.1,upper=0.5,tol=0.001,kernel="
    biweight")
19 h=ban$h
20 ##### A noyau #####
21 qx=function(p)
22 {
23 M=numeric(n)
24 for (i in 1:n)
25 {
26 M[i]=integrate(k,(((i-1)/n)-p)/h,((i/n)-p)/h)$value
27 }
28 sum(Y*M)
29 }
30 W[j]=q(p[j])
31 V[j]=qx(p[j])
32 }
33 plot(p,qu,type="l",xlab="Probabilit",ylab="Quantile",main="n=10")
34 lines(p,W,col="3",lty=3,lwd=4)
35 lines(p,V,col="2",lty=5,lwd=2)

```

Résultats graphiques résumés dans les figures

```

1  rm(list=ls())
2  library(kedd)
3  library(MASS)
4  library(Metrics)
5  n=50
6  R=100
7  k=function(t){(3/4)*(1-t^2)*(abs(t)<=1)}
8  p=0.5
9  qt=qbeta(p,2,3)
10 W=V=numeric(R)
11 for(j in 1 :R){
12 X=rbeta(n,2,3)
13 ##### Empirique #####
14 Y=sort(X)
15 q=function(p){
16 m=floor(p*n)
17 s=m+1
18 Y[s]
19 }
20 ##### Param tre de lissage #####
21 ban=h.ucv(X,deriv.order =0,lower=0.1,upper=0.5,tol=0.001, kernel="
    epanechnikov")
22 h=ban$h
23 ##### A noyau #####
24 qx=function(p)
25 {
26 M=numeric(n)
27 for (i in 1:n)
28 {
29 M[i]=integrate(k,(((i-1)/n)-p)/h,((i/n)-p)/h)$value
30 }
31 sum(Y*M)
32 }
33 W[j]=q(p)
34 V[j]=qx(p)
35 }
36 ##### Graph #####
37 boxplot(W,V,names=c("Empirique","A noyau"),col=c("green", "blue"),
    main="Beta(2,3)",xlab="Estimateurs", ylab="Quantile")
38 abline(h =qt, col = "red")
39 legend("top", legend=c("Empirique", "A noyau"),col=c("red", "blue")
    ,cex=0.7, fill = c("green", "blue"))

```

2.3.3 Resultats

Les résultats sont résumés dans les tableaux suivants.

Exponentielle(2)				
Taille	Estimateurs	$p = 0.1$	$p = 0.5$	$p = 0.7$
10	Q_n	6.224077×10^{-03}	4.503244×10^{-02}	6.661868×10^{-02}
	\tilde{Q}_n	4.929086×10^{-03}	3.437501×10^{-02}	3.611019×10^{-02}
50	Q_n	8.088671×10^{-04}	5.294587×10^{-03}	1.415154×10^{-02}
	\tilde{Q}_n	8.060631×10^{-04}	4.647644×10^{-03}	1.312734×10^{-02}
100	Q_n	3.70823×10^{-04}	2.349422×10^{-03}	7.325803×10^{-03}
	\tilde{Q}_n	2.907593×10^{-04}	2.16352×10^{-03}	6.672089×10^{-03}
500	Q_n	4.854168×10^{-05}	4.064432×10^{-04}	1.183833×10^{-03}
	\tilde{Q}_n	3.578088×10^{-05}	3.804042×10^{-04}	1.105871×10^{-03}

TABLE 2.1 – MSE associe à la loi $Exp(2)$ avec l'utilisation de la noyau Epanechnikov

Beta(2, 3)				
Taille	Estimateurs	$p = 0.1$	$p = 0.5$	$p = 0.7$
10	Q_n	8.374795×10^{-3}	8.448417×10^{-3}	1.017245×10^{-2}
	\tilde{Q}_n	2.769762×10^{-3}	5.953935×10^{-3}	5.991416×10^{-3}
50	Q_n	1.088497×10^{-3}	1.593778×10^{-3}	1.971852×10^{-3}
	\tilde{Q}_n	5.330961×10^{-4}	1.288345×10^{-3}	1.485239×10^{-3}
100	Q_n	6.449444×10^{-4}	7.703698×10^{-4}	1.020885×10^{-3}
	\tilde{Q}_n	3.070057×10^{-4}	6.858172×10^{-4}	4.132497×10^{-3}
500	Q_n	1.371505×10^{-4}	1.393908×10^{-4}	1.861999×10^{-4}
	\tilde{Q}_n	1.259711×10^{-4}	1.176644×10^{-4}	1.667876×10^{-4}

TABLE 2.2 – MSE associe à la loi $Beta(2, 3)$ avec l'utilisation de la noyau Epanechnikov

Gamma(1, 2)				
Taille	Estimateurs	$p = 0.1$	$p = 0.5$	$p = 0.7$
10	Q_n	6.638256×10^{-3}	3.003331×10^{-2}	8.89889×10^{-2}
	\tilde{Q}_n	4.275602×10^{-3}	2.145518×10^{-2}	5.4863×10^{-2}
50	Q_n	1.03646×10^{-3}	4.631102×10^{-3}	1.430065×10^{-2}
	\tilde{Q}_n	9.567885×10^{-4}	4.583543×10^{-3}	1.416453×10^{-2}
100	Q_n	3.027614×10^{-4}	2.86526×10^{-3}	7.999032×10^{-3}
	\tilde{Q}_n	2.709697×10^{-4}	2.760871×10^{-3}	7.275338×10^{-3}
500	Q_n	5.620288×10^{-5}	4.901122×10^{-4}	1.497403×10^{-3}
	\tilde{Q}_n	3.879496×10^{-5}	4.364999×10^{-4}	1.360773×10^{-3}

TABLE 2.3 – MSE associe à la loi $Gamma(1, 2)$ avec l'utilisation de la noyau Epanechnikov

Exponentielle(2)				
Taille	Estimateurs	$p = 0.1$	$p = 0.5$	$p = 0.7$
10	Q_n	1.015296×10^{-2}	1.015296×10^{-2}	6.979066×10^{-2}
	\tilde{Q}_n	5.430142×10^{-3}	5.430142×10^{-3}	4.125249×10^{-2}
50	Q_n	5.710332×10^{-4}	5.710332×10^{-4}	1.30315×10^{-2}
	\tilde{Q}_n	5.254367×10^{-4}	5.254367×10^{-4}	1.265053×10^{-2}
100	Q_n	3.776917×10^{-4}	3.776917×10^{-4}	5.100647×10^{-3}
	\tilde{Q}_n	2.617827×10^{-4}	2.617827×10^{-4}	5.092308×10^{-3}
500	Q_n	7.124239×10^{-5}	7.124239×10^{-5}	1.284887×10^{-3}
	\tilde{Q}_n	5.735591×10^{-5}	5.735591×10^{-5}	1.238791×10^{-3}

TABLE 2.4 – MSE associe à la loi $Exp(2)$ avec l'utilisation de la noyau Biweight

L'exemple présenté dans le tableau 2.7, réalisé dans le cadre d'estimation de la fonction de quantile d'une loi normale, centrée réduite, à partir d'un échantillons de taille $n=50$ et $n = 100$, est une illustration de l'influence du choix du paramètre de lissage sur les caractéristiques de l'estimateur.

Beta(2, 3)				
Taille	Estimateurs	$p = 0.1$	$p = 0.5$	$p = 0.7$
10	Q_n	9.604419×10^{-3}	9.604419×10^{-3}	8.441122×10^{-3}
	\tilde{Q}_n	2.739162×10^{-3}	2.739162×10^{-3}	6.046021×10^{-3}
50	Q_n	1.145857×10^{-3}	1.145857×10^{-3}	2.007222×10^{-3}
	\tilde{Q}_n	7.027038×10^{-4}	7.027038×10^{-4}	1.428235×10^{-3}
100	Q_n	5.790959×10^{-4}	5.790959×10^{-4}	1.106927×10^{-3}
	\tilde{Q}_n	4.053582×10^{-4}	4.053582×10^{-4}	9.155626×10^{-4}
500	Q_n	1.051367×10^{-4}	1.051367×10^{-4}	2.15214×10^{-4}
	\tilde{Q}_n	1.021969×10^{-4}	1.021969×10^{-4}	1.785644×10^{-4}

TABLE 2.5 – MSE associe à la loi $Beta(2, 3)$ avec l'utilisation de la noyau Biweight

Gamma(1, 2)				
Taille	Estimateurs	$p = 0.1$	$p = 0.5$	$p = 0.7$
10	Q_n	6.752974×10^{-3}	6.752974×10^{-3}	9.241681×10^{-2}
	\tilde{Q}_n	3.55131×10^{-3}	3.55131×10^{-3}	5.087853×10^{-2}
50	Q_n	9.068209×10^{-4}	9.068209×10^{-4}	1.350777×10^{-2}
	\tilde{Q}_n	7.525022×10^{-4}	7.525022×10^{-4}	1.330856×10^{-2}
100	Q_n	3.84128×10^{-4}	3.84128×10^{-4}	5.654582×10^{-3}
	\tilde{Q}_n	2.702279×10^{-4}	2.702279×10^{-4}	5.334711×10^{-3}
500	Q_n	4.454219×10^{-5}	4.454219×10^{-5}	9.851274×10^{-4}
	\tilde{Q}_n	3.140671×10^{-5}	3.140671×10^{-5}	9.796221×10^{-4}

TABLE 2.6 – MSE associe à la loi $Gamma(1, 2)$ avec l'utilisation de la noyau Biweight

On prend un échantillon qui suit la loi exponentielle de paramètre 2 et un échantillon suit la loi beta de paramètre 2 et 3, de taille $n = 50$, on difinit la boîte à moustaches des valeurs de l'estimateur empirique et l'estimateur à noyau de la fonction quantile par taille d'échantillon. Les resultats graphiques sont resumés dans les graphes 2.1, 2.2 et 2.3.

Normale(0,1)			
Estimateurs		$h_{amise} = 0.296445$	$h_{ucv} = 0.2896815$
$n = 50$	Q_n	2.959062×10^{-2}	2.920883×10^{-2}
	\tilde{Q}_n	2.273278×10^{-2}	2.058456×10^{-2}
Estimateurs		$h_{amise} = 0.2962181$	$h_{ucv} = 0.2796114$
$n = 100$	Q_n	1.518199×10^{-2}	1.192599×10^{-2}
	\tilde{Q}_n	1.084014×10^{-2}	1.001059×10^{-2}

TABLE 2.7 – MSE associée à la loi $Normale(0, 1)$ par les deux méthodes h_{amise} et h_{ucv}

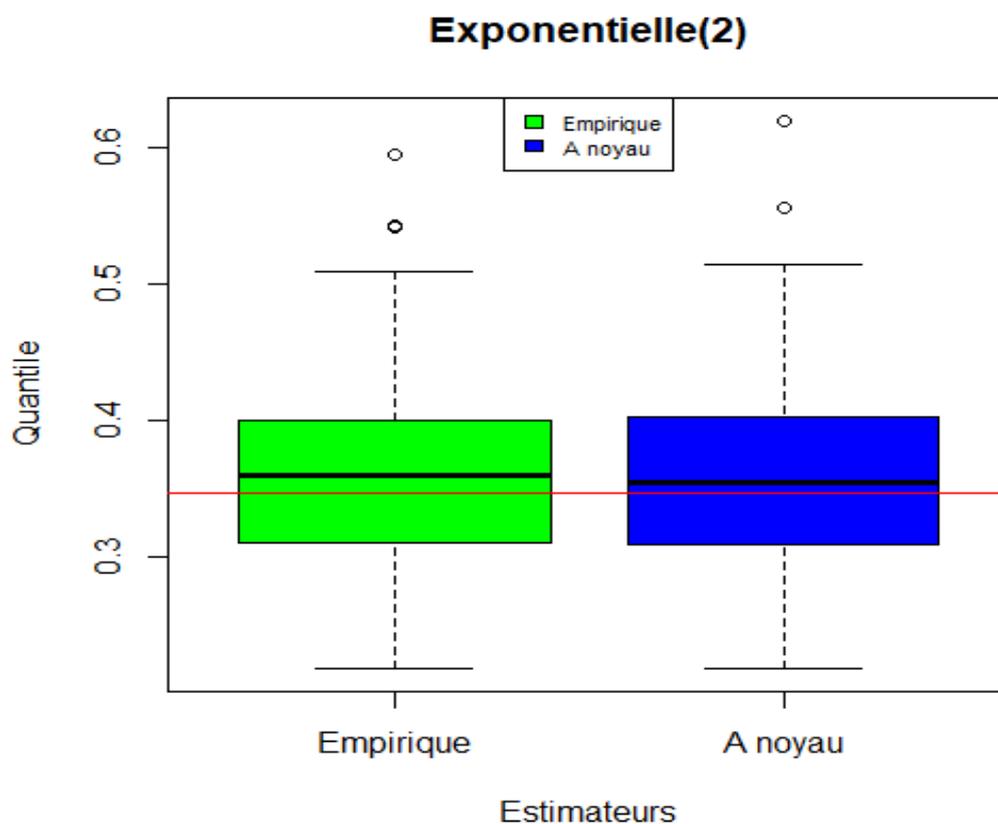


FIGURE 2.1 – l'estimateur empirique et l'estimateur à noyau d'une loi exponentielle

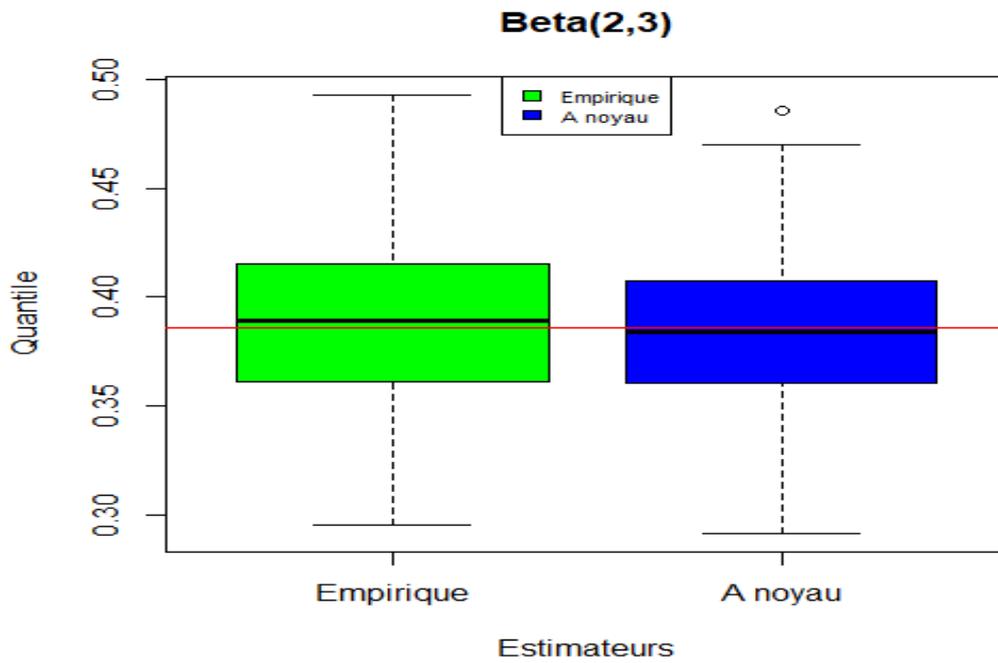


FIGURE 2.2 – l'estimateur empirique et l'estimateur à noyau d'une loi beta

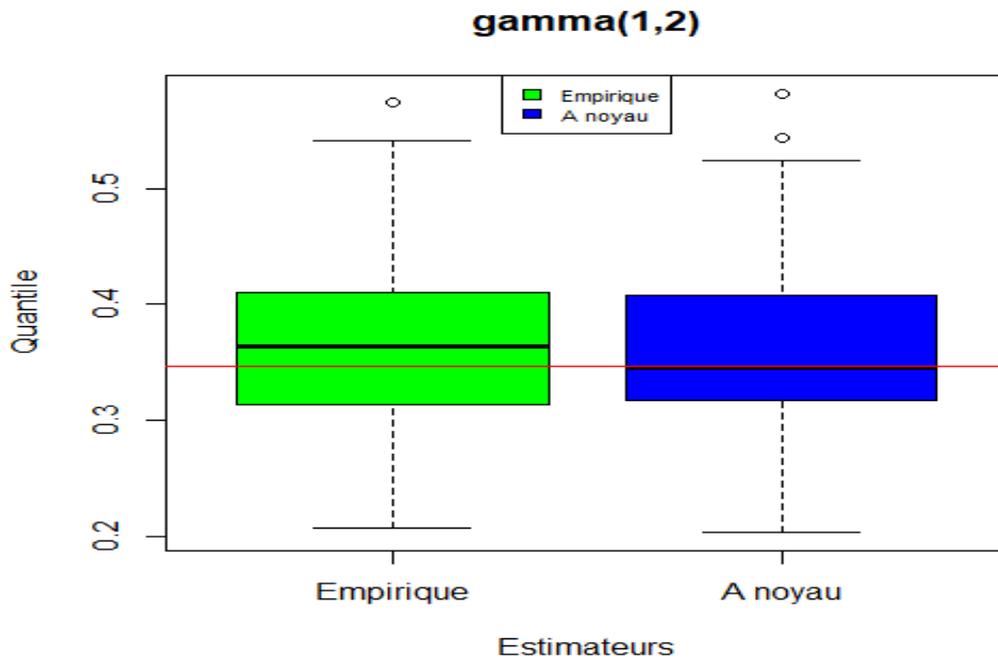


FIGURE 2.3 – l'estimateur empirique et l'estimateur à noyau d'une loi gamma

2.3.4 Discussion

Les résultats de simulations obtenus nous permettent de conclure que :

- L'augmentation de la taille de l'échantillon entraîne la décroissance du MSE , c-à-d la convergence du MSE vers zéro est indépendante du noyau mais dépend essentiellement de la taille de l'échantillon.
- Les valeurs obtenues pour le noyau Epanechnikov sont légèrement meilleurs aux valeurs obtenues pour la noyau Biweight.
- Le tableau 2.7 montre que le paramètre de lissage h_{ucv} a donné de bons résultats par rapport à h_{amise} .
- Généralement, l'estimateur à noyau \tilde{Q}_n performe mieux que l'estimateur empirique Q_n .
- D'après 2.1, 2.2 and 2.3, on peut conclure que, concernant l'estimateur à noyau, la médiane (centre de boîte) est proche de la valeur de quantile théorique (ligne en rouge) mieux que l'estimateur empirique dans tout les distributions.

Les graphes suivants nous confirment les résultats précédents :

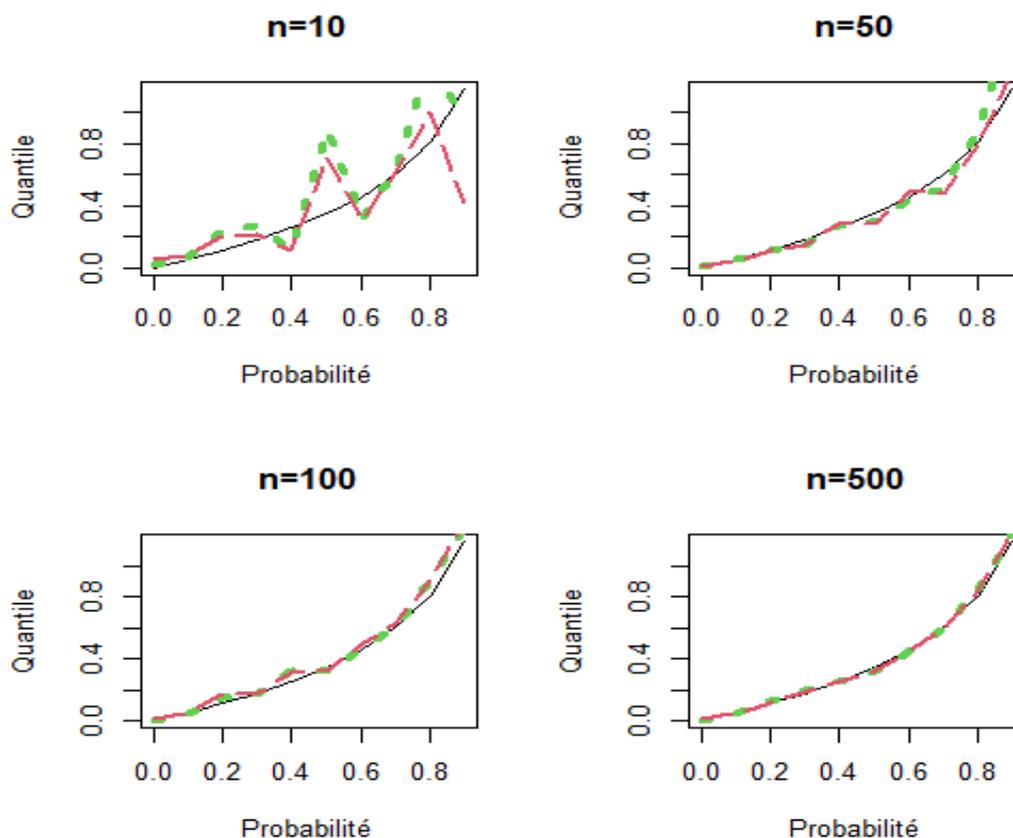


FIGURE 2.4 – La fonction de quantile $\text{Exp}(2)$ et les estimateurs empirique (vert), à noyau (rouge)

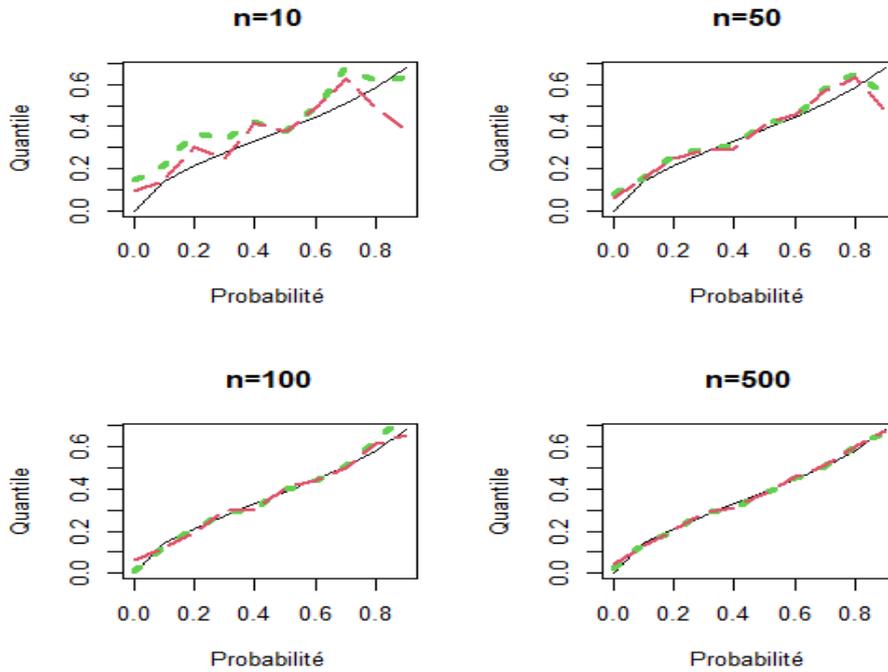


FIGURE 2.5 – La fonction de quantile Beta(2,3) et les estimateurs empirique (vert), à noyau (rouge)

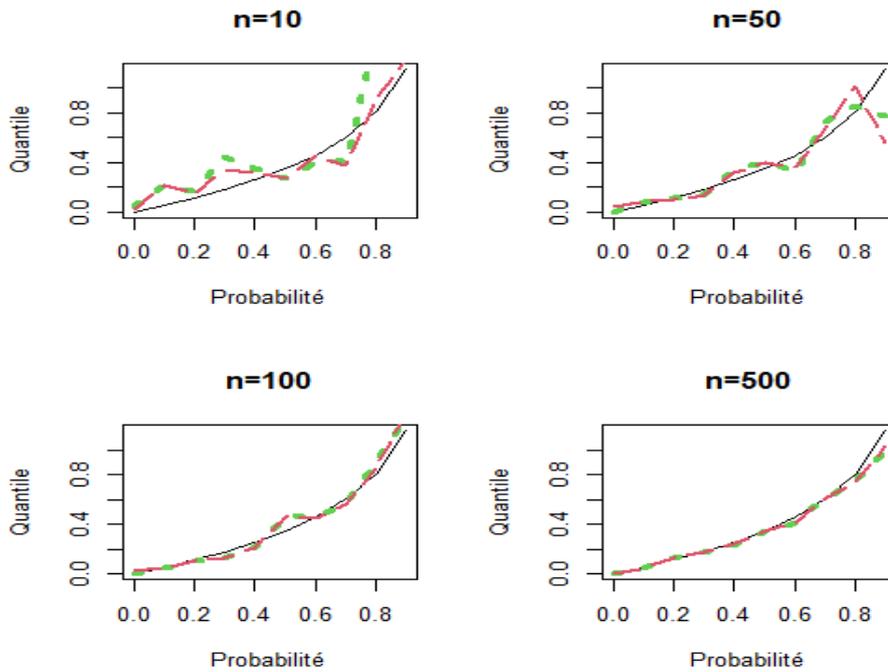


FIGURE 2.6 – La fonction de quantile Gamma(1,2) et les estimateurs empirique (vert), à noyau (rouge)

Conclusion

Dans un premier temps, nous parlons des définitions et des propriétés de la statistique d'ordre comme approche de l'estimation, et nous expliquons qu'est-ce que la fonction répartition empirique et son rôle pour trouver la fonction quantile, ainsi que quelques propriétés de la quantile et sa dénomination populaire.

Dans un second temps, nous donnons un résumé de la méthode paramétrique et non paramétrique, ensuite, nous avons trouvé différentes définitions et critères d'estimation non paramétrique d'une fonction quantile. nous avons présenté l'estimateur de Rosenblatt et Parzen (1962), mentionné des noyau symétrique qui est habituellement utilisé dans cette méthode comme (Epanechnikov, Gaussien, Biweight,..., etc), et expliqué comment sélectionner le bon paramètre de lissage h en détaillant le choix théorique et pratique optimal de h , ainsi que la convergence en probabilité de cet estimateur et ses propriétés asymptotiques de Biais et de variance.

A la fin, pour confirmer notre étude, nous simulons une densité d'échantillon en utilisant la méthode des noyaux. Ensuite, nous définissons la fonction quantile à partir de cette densité. Le dernier partie valide les notions introduites précédemment.

Perspectives :

Parmi les perspectives de ce travail, il serait intéressant de :

- Utiliser d'autres méthodes pour déterminer le paramètre de lissage comme l'approche bayésienne, ...
- Effectuer des simulations sur des données réelles.
- Effectuer des simulations sur d'autre distributions.
- Comparer entre les estimateurs par l'utilisation d'autre critère.

Bibliographie

- [1] Arnold, B. C. , Balakrishnan, N. , & Nagaraja, H. N. (1992). *A First Course in Order Statistics*. John Wiley & Sons, Inc. , Massachusetts, 2nd ed.
- [2] Al-Kenani, A. , & Yu, K. (2012). New Bandwidth Selection for Kernel Quantile Estimators. *Journal of Probability and Statistics*, Volume 2012.
- [3] Cacoullos, T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18 : 178-189.
- [4] David, H. A. , & Nagaraja, H. N. (2003). *Order Statistics*, 3rd Edition. John Wiley & Sons, Inc. , New York.
- [5] Falk, M. (1984). Relative Deficiency of Kernel Type Estimators of Quantiles, *The Annals of Statistics*,
- [6] Ferhat, L. (2012). Memoire estimation d'une fonction de densite par la methode des noyaux et application a la VaR ,Tizi-Ouzou
- [7] Introduction à la statistique non paramétrique. CNRS, Laboratoire Statistique & Génome, Evry, ENSIIE - (2013/2014).
- [8] Jones, M. C. (1992). Estimating densities, quantiles, quantile densities, and density quantiles. *Annals of the Institute of Statistical Mathematics*, 44, 721-727.
- [9] Lanju, Z., B. S., M. S. (2002). Bandwidth selections for a class of smooth quantile estimators. Dean of the Graduate School.
- [10] MY Cheng, S Sun. (2006). Bandwidth Selection for Kernel Quantile Estimation. *China Statistical Journal*, 44 page 271–295.
- [11] Michel Lejeune (2010). *Statistique. Statistique et probabilités appliquées. La théorie et ses applications*, Deuxième édition Springer
- [12] Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory Proba. Appl.*, 9, 497-500

-
- [13] Probabilités et statistique avec R [https://ljk.imag.fr/membres/Bernard Ycart/mel/dr/node11.html](https://ljk.imag.fr/membres/Bernard_Ycart/mel/dr/node11.html).
- [14] Parzen, E. (1979). Non parametric Statistical Data Modelling. *Journal of the American Statistical Association*.
- [15] Sheather, S. J. , & Marron, J. S. (1990). Quantile Estimators. *Journal of the American Statistical Association*, 85(410), 410-416.
- [16] SILVERMAN, B. W. (1986). Density estimation for statistics and data analysis, Chapman and Hall, London , 175, 0–412.
- [17] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions technip.
- [18] Soni, P., Dewan, I., and Jain, K. (2012). Nonparametric estimation of quantile density function. *Computational Statistics and Data Analysis*, 56(12), 3876-3886.
- [19] Tsybakov, A. B. (2003). Introduction à l'estimation non paramétrique (Vol. 41). Springer Science and Business Media.

Annexe A : Logiciel R

R est un système, communément appelé langage et logiciel, qui permet de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possible la manipulation des données, les calculs et les représentations graphiques. R a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand nombre de procédures statistiques appelées paquets. Ces derniers permettent de traiter assez rapidement des sujets aussi variés que les modèles linéaires (simples et généralisés), la régression (linéaire et non linéaire), les séries chronologiques, les tests paramétriques et non paramétriques classiques, les différentes méthodes d'analyse des données,...Plusieurs paquets, tels `ade4`, `FactoMineR`, `MASS`, `multivariate`, `scatterplot3d` et `rgl` entre autres sont destinés à l'analyse des données statistiques multidimensionnelles. Il a été initialement créé, en 1996, par Robert Gentleman et Ross Ihaka du département de statistique de l'Université d'Auckland en Nouvelle Zélande. Depuis 1997, il s'est formé une équipe "R Core Team" qui développe R. Il est conçu pour pouvoir être utilisé avec les systèmes d'exploitation Unix, Linux, Windows et MacOS. Un élément clé dans la mission de développement de R est le Comprehensive R Archive Network (CRAN) qui est un ensemble de sites qui fournit tout ce qui est nécessaire à la distribution de R, ses extensions, sa documentation, ses fichiers sources et ses fichiers binaires. Le site maître du CRAN est situé en Autriche à Vienne, on peut y accéder par l'URL : "<http://cran.r-project.org/>".



Les autres sites du CRAN, appelés sites miroirs, sont répartis partout dans le monde. R est un logiciel libre distribué sous les termes de la "GNU Public Licence". Il fait partie intégrante du projet GNU et possède un site officiel à l'adresse "<http://www.R-project.org/>". Il est souvent présenté comme un clone de S qui est un langage de haut niveau développé par les AT and T Bell Laboratories et plus particulièrement par Rick Becker, John Chambers et Allan Wilks. S est utilisable à travers le logiciel S-Plus qui est commercialisé par la société Insightful (<http://www.splus.com/>).

Annexe B : Critères importants

A. Critères de convergence

La convergence en moyenne quadratique

La suite de variables aléatoires $(X_n)_{n \geq 1}$ converge vers X une autre variable aléatoire, en moyenne d'ordre $p > 1$ si :

$$\lim_{n \rightarrow \infty} E(|X_n - X|^p) = 0$$

pour $p = 1$ convergence en moyenne vers une variable aléatoire X , si :

$$\lim_{n \rightarrow \infty} E(|X_n - X|) = 0.$$

On note $X_n \rightarrow X$ en moyenne.

pour $p = 2$ de convergence en moyenne quadratique vers une variable aléatoire X , si :

$$\lim_{n \rightarrow \infty} E(|X_n - X|^2) = 0.$$

On note $X_n \rightarrow X$ en moyenne quadratique.

La convergence presque sûre

La suite de variables aléatoires $(X_n)_{n \geq 1}$ converge vers X une autre variable aléatoire, en presque sûre, si $P(\Omega') = 1$

$$\Omega' := \{\omega \in \Omega; \lim_{n \rightarrow +\infty} X_n = X\}$$

B. Critères d'erreurs

Pour mesurer les performances théoriques des estimateurs et identifier le meilleur, il est nécessaire de spécifier un critère d'erreur. Nous considérons la fonction de quantile Q et son estimateur Q_n .

L'erreur quadratique intégrée ISE

$$ISE(Q(p), Q_n(p)) = \int [Q(p) - Q_n(p)]^2 dp.$$

L'erreur quadratique moyenne MSE

$$MSE(Q(p), Q_n(p)) = \mathbb{E}(Q(p) - Q_n(p))^2$$

donc

$$MSE(Q(p), Q_n(p)) = [\text{Biais}Q_n(p)]^2 + \text{Var}Q_n(p).$$

L'erreur quadratique moyenne intégrée MISE

$$\begin{aligned} MISE(Q(p), Q_n(p)) &= \int MSE(Q(p), Q_n(p)) dp \\ &= \int \mathbb{E}[Q(p) - Q_n(p)]^2 dp \end{aligned}$$

Annexe C : Table de notations et abréviation

- i. i. d. : Indépendant identiquement distribuées.
- v.a : Variable aléatoire.
- ps : Presque sûrement.
- mq : Moyenne quadratique.
- P : Probabilité.
- E : Espérance.
- Var : Variance.
- MSE : Erreur quadratique moyenne.
- $AMSE$: L'erreur quadratique moyenne asymptotique.
- MISE : L'erreur quadratique moyenne intégrée.
- k : La noyau.
- h : La fenêtre, ou paramètre de lissage.
- θ : Le paramètre de la distribution.
- μ : La moyenne de la population.
- S^2 : La variance de l'échantillon.
- $\mathbb{1}_A$: La fonction indicatrice de l'ensemble A .
- h_{opt} : La fenêtre optimale.
- $f^{(2)}$: La deuxième dérivée de f .
- $AMSE_{\text{opt}}$: L'erreur quadratique moyenne asymptotique optimale.
- eff : L'efficacité d'un noyau par rapport au noyau d'Epanechnecov.
- Q_n : Estimateur empirique de la fonction quantile.
- \tilde{Q}_n : Estimateur à noyau de Q_X .
- $\hat{\theta}$: Estimation de paramètre θ .
- \hat{F}_θ : L'estimation de la distribution de paramètre θ .
- \hat{F}_n : Estimateur de la fonction de répartition empirique.
- $Q^{(1)}, Q^{(2)}$: Le premier et la deuxième dérivée de la fonction quantile.
- b_{opt} : La fenêtre optimale asymptotique de $Q_m^{(2)}$.
- a_{opt} : La fenêtre optimale asymptotique de $Q_m^{(1)}$.
- \hat{h}_{opt} : L'estimateur de la fenêtre optimale de l'estimateur \hat{Q}_n .

Abstract

Nonparametric estimation in statistics is one of the common and valid methods that lead to satisfying results.

The main goal of our dissertation is to apply the non-parametric method to estimate the quantile function using the kernel method which is characterized by the choice of the kernel and the smoothing parameter, also called smoothing window. The choice of the smoothing parameter is crucial for the local precision as well as for the global precision of the estimator. At last, we enriched our work with a numerical application part, where we made a comparison between the empirical estimator and the kernel estimator of the quantile function.

Keywords: non-parametric estimation, quantile function, kernel method, simulations.

Résumé

L'estimation non paramétrique en statistique est l'une des méthodes courantes et valides qui mènent à des résultats satisfaisants.

Le but principal de notre mémoire est d'appliquer la méthode non paramétrique pour estimer la fonction quantile en utilisant la méthode du noyau qui est caractérisée par le choix du noyau et du paramètre de lissage appelé aussi fenêtre de lissage. Le choix du paramètre de lissage est crucial pour la précision locale ainsi que pour la précision globale de l'estimateur. En fin de compte, nous avons enrichi notre travail avec une partie d'application numérique, où nous avons fait une comparaison entre l'estimateur empirique et l'estimateur à noyau de la fonction quantile.

Mots clés: estimation non paramétrique, fonction quantile, méthode du noyau, simulations.

ملخص

التقدير غير المعلمي في الإحصاء هو إحدى الطرق الشائعة التي تؤدي إلى نتائج مرضية. الهدف الرئيسي من مذكرتنا هو تطبيق الطريقة غير المعلمية لتقدير دالة التوزيع الكمي باستخدام طريقة النواة التي تتميز باختيار النواة ومعلمة التنعيم التي تسمى أيضًا نافذة التنعيم. يعتبر اختيار معلمة التنعيم أمرًا بالغ الأهمية لكل من الدقة المحلية والدقة العامة للمقدر. في النهاية، قمنا بإثراء عملنا بجزء من التطبيق العددي، حيث قمنا بمقارنة بين المقدر التجريبي والمقدر بالنواة لدالة التوزيع الكمي.

الكلمات الرئيسية: التقدير غير المعلمي، دالة التوزيع الكمي، طريقة النواة، المحاكاة.