

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider, Biskra
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie
Département de Mathématiques



Mémoire présenté pour obtenir le diplôme de
Master en “**Mathématiques Appliquées**”

Option : statistique

Par :

ADJAIMI Khadidja

Simulation des estimateurs de distribution non paramétrique

Devant le Jury :

Pr. BENATIA Fatah	UMKB	Président
Pr. BRAHIMI Brahim	UMKB	Encadreur
Dr. KHEIREDDINE Souraya	UMKB	Examinatrice

Soutenu Publiquement le 18 Juin 2023

Dédicace

Je dédie cet humble travail

*À ma chère **maman**, ma vie et mon bonheur pour son amour,
ses encouragements, et pour tous ses sacrifices afin que je cible ce travail.*

*À l'homme de ma vie qui a souffert sans me laisser souffrir, qui était toujours à
mes côtés, mon cher **papa**.*

À mes chers frères et sœurs, ma source de joie et de bonheur.

À tous mes amis.

À tous mes proches.

À tous mes professeurs et tous ceux qui ont contribué à mon éducation.

À mes camarades de promotion.

À tous ceux qui m'aiment.

À tous ceux que j'aime.

Remerciements

Nous remercions tout d'abord ALLAH de nous avoir donné la santé, la volonté, la force et le courage pour pouvoir surmonter les moments difficiles, et atteindre nos objectifs.

Mes remerciements à mon cher père, ma mère et frères pour leurs sacrifices et leur patience avec moi pour réaliser mon succès et mes rêves, qu'ALLAH les protège tous.

*Nous témoignons une reconnaissance particulière à notre promoteur monsieur **BRAHIMI Brahim**. Pour ses commentaires, remarques et suggestions qui ont donné une autre dimension à notre travail.*

Nous tenons aussi à le remercier pour tous ses efforts pour bien mener à notre formation durant notre cursus.

*Mes remerciements aux membres de Jury : Pr. **BENATIA Fatah**, et Dr. **KHEIREDDINE Souraya**, pour avoir accepté d'étudier et d'évaluer ce travail, nous sommes très reconnaissant à leurs remarques et commentaires qui nous aident beaucoup pour mieux présenter ce document.*

Nous avons beaucoup de reconnaissances envers tous les enseignants du département de Mathématique de la Faculté des Sciences exactes de Biskra.

Que tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail, trouvent ici nos sincères reconnaissances.

Table des matières

Remerciements	ii
Table des matières	iii
Table des figures	v
Liste des tables	vi
Notations et symbols	vii
Introduction	1
1 Introduction la statistique non paramétrique	3
1.1 Notions préliminaires	3
1.1.1 Rapapels d'inégalités classiques :	7
1.2 Estimateur et propriétés	9
1.2.1 Estimation	9
1.2.2 Estimation non-paramétrique de la foction répartition	9
1.2.3 Propriétés de la fonction de répartition empirique	10
2 Fonction des répartition et distribution	13

2.1 Estimation paramétrique	13
2.1.1 Estimation par la méthode du maximum de vraisemblance	
(MVS)	14
2.2 Estimation non paramétrique	20
2.2.1 Un estimateur simple de la densité : l'histogramme	20
2.2.2 Propriétés statistiques de l'histogramme	22
2.2.3 Estimateurs à noyaux	25
2.2.4 Propriétés asymptotiques de l'estimateur	32
2.2.5 Erreur quadratique moyenne	35
2.2.6 Erreur quadratique moyenne intégrée(MISE)	36
2.2.7 Choix du paramètre de lissage	36
2.2.8 Méthode de la validation croisée	40
2.2.9 Noyaux asymptotiquement optimales	47
Conclusion	48
Bibliographie	48

Table des figures

2.1	Fonction puissance du test δ	19
2.2	Estimation par histogramme de densité de khi-deux	22
2.3	Fonction de répartition empirique $n = 10$	26
2.4	Courbes des noyaux usuels	31
2.5	La densité théorique et la densité estimée pour $n = 200, Bandwith =$ $0.7803.$	37
2.6	La densité théorique $N(0, 1)$ et la densité estimée avec noyau Gaussien pour $n = 100, h = 0.05, h = 0.2, h = 0.4$ et $h = 0.8.$	38
2.7	La représentation graphique de UCV sur $[0.1, 1]$	45
2.8	L'estimateur de la densité e par validation croisée non biaisée	46

Liste des tableaux

2.1 modèles paramétriques	14
2.2 Noyaux usuelles	30
2.3 Efficacite des noyaux	47

Notations et symboles

Les abréviations et les symboles qui utilisés tout au long de ce mémoire sont expliqués

ci-dessous :

$\mathcal{N}(0, \hat{\sigma}^2)$: La loi normale centré à variance $\hat{\sigma}^2$

\xrightarrow{P} : Converge en probabilité

\mathbb{I} : La fonction indicatrice.

$\hat{f}_n(\cdot)$: Fonction distribution empirique

h : Paramètre de lissage

\mathbb{K} : Noyau.

$\mathbb{E}(X)$: Moyenne de X

$Var(X)$: Variance de X

\exists : Existe

\mathbb{R} : Ensemble de nombres réels

\mathbb{N} : Ensemble d'entiers non négatifs

(X_1, X_2, \dots, X_n) : Échantillons de taille n de X

$\#$: cardinale

v.a : Variable aléatoire

c-à-d : C'est à dire

TCL : Théorème centrale limite

iid : indépendantes et identiquement distribuées

Introduction

L'estimation non paramétrique est une approche statistique qui vise à estimer les caractéristiques d'une population ou d'une distribution de probabilité sans faire d'hypothèses sur sa forme fonctionnelle spécifique. Contrairement aux méthodes paramétriques qui supposent une forme de distribution prédéfinie avec un nombre fixe de paramètres, l'estimation non paramétrique permet une plus grande flexibilité en adaptant le modèle aux données observées.

L'estimation non paramétrique est particulièrement utile dans les situations où la structure sous-jacente de la population est inconnue ou complexe, ou lorsque les données présentent des caractéristiques inhabituelles qui ne correspondent pas à un modèle paramétrique prédéfini.

L'idée principale derrière l'estimation non paramétrique est de s'appuyer sur les informations contenues dans les données elles-mêmes pour construire une estimation de la distribution ou de la fonction de densité de probabilité. Au lieu de supposer une distribution spécifique, l'estimation non paramétrique utilise des techniques telles que l'ajustement de courbes, les méthodes par noyaux, les méthodes de lissage et les approches basées sur des histogrammes pour obtenir une approximation de la distribution à partir des données.

L'estimation non paramétrique permet de prendre en compte la complexité des données et peut fournir des estimations plus précises et robustes dans certaines situa-

tions. Cependant, elle peut également nécessiter des échantillons plus importants et peut être plus sujette aux problèmes de surajustement si elle est mal utilisée.

Dans cette approche, l'objectif est d'obtenir une estimation aussi proche que possible de la véritable distribution sous-jacente sans faire d'hypothèses restrictives sur sa forme. Cela permet d'obtenir une vision plus complète des données et d'explorer des modèles plus flexibles et adaptatifs.

Chapitre 1

Introduction la statistique non paramétrique

1.1 Notions préliminaires

Variable aléatoire réelle

On appelle variable aléatoire réelle (abréviation : v.a.r) toute application de dans \mathbb{R} :

$$\begin{aligned} X : \Omega &\longrightarrow \mathbb{R} \\ \omega &\longrightarrow X(\omega) \end{aligned}$$

– \mathbb{I} un sous-ensemble de \mathbb{R} . L'ensemble $X^{-1}(\mathbb{I}) = \{\omega \in \Omega / X(\omega) \in \mathbb{I}\}$, qui est un sous-ensemble de Ω , est un événement aléatoire : on le note $(X \in \mathbb{I})$.

On dit que la variable aléatoire X est :

1. Discrète finie si l'ensemble $X(\Omega)$ est fini, discrète infinie si l'ensemble $X(\Omega)$ est infini dénombrable.
2. Continue si l'ensemble $X(\Omega)$ est un intervalle de \mathbb{R} non réduit à un point (ou une réunion d'intervalles de \mathbb{R}).

Fonction de répartition

La fonction de distribution est un outil de référence pour définir uniformément la loi de probabilité de savoir si une variable aléatoire est discrète ou continue. Si cette fonction est connue, la probabilité peut être calculée pour n'importe quel intervalle, et donc pour pratiquement n'importe quel événement. C'est pourquoi le tableau de la loi de probabilité lui est donné.

Définition 1.1.1 Soit X une variable aléatoire, on appelle fonction de répartition de X , que l'on not F_X , la fonction définit sur \mathbb{R} par :

$$\forall x \in \mathbb{R}, F(x) = P(X \leq x) = P_x(]-\infty, x]) = \int_{-\infty}^x f(t)dt. \quad (1.1)$$

Remarque 1.1.1 Une fonction de répartition doit vérifier un certain nombre de propriétés suivantes :

1. $F(x) \in [0, 1] \quad \forall x \in \mathbb{R}$.
2. La limite de $F(x)$ quand x tend vers $-\infty$ égale à 0 ($\lim_{x \rightarrow -\infty} F(x) = 0$).
3. La limite de $F(x)$ quand x tend vers $+\infty$ égale à 1 ($\lim_{x \rightarrow +\infty} F(x) = 1$).
4. La fonction F est un fonction croissante, si $x_1 \leq x_2$ alors $F(x_1) \leq F(x_2)$.
5. Pour tout nombre $a, b \in \mathbb{R}$ tels que $a < b$, $P(a < X \leq b) = F(b) - F(a)$.
6. $\forall x \in \mathbb{R}, F(X > a) = 1 - F(X \leq a) = 1 - F(a)$.

La densité de probabilité

Définition 1.1.2 Une fonction $f : \mathbb{R} \rightarrow \mathbb{R}$ est appelée densité de probabilité si elle est positive (en tout $x \in \mathbb{R}$ ou elle est définie, $f(x) \geq 0$). Intégrable sur \mathbb{R} et si :

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Variabes aléatoires discrètes et continues

Une v.a X est dite discrète (v.a.d) ssi elle est valeur dans un ensemble E fini ou dénombrable on note $E = \{x_1, x_2, \dots, x_n\}$, la loi de probabilité d'une v.a.d est :

$$F(x) = P(X \leq x) = \sum_{x_i \in E} P(X = x_i).$$

Dans ce cas $F(x)$ est une fonction en escalier présentant des sauts.

Une v.a X est dite continue (v.a.c) ssi sa fonction de répartition est continue et presque partout dérivable, la loi de probabilité d'une v.a.c est :

$$\forall x \in \mathbb{R}, F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt \text{ et } f(x) = \frac{dF}{dx}(x).$$

Dans ce cas $F(x)$ est une fonction continue, de plus :

$$\forall B \in \mathbb{R}, P(X \in B) = \int_B f(x)dx,$$

où $f_X(x)$ est une densité de probabilité de X .

Espérance mathématique

– Si X est une v.a.d son espérance mathématique est définie par :

$$\mathbb{E}(X) = \sum_{x_i \in E} x_i P(X = x_i).$$

– Si X est une v.a.c son espérance mathématique est définie par :

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f_X(x)dx.$$

Le espérance d'une fonction de la variable aléatoire

Proposition 1.1.1 *Soit X est une v.a et h une fonction définie sur \mathbb{R} ; Alors :*

– Si X est une v.a.d on a :

$$\mathbb{E}(h(X)) = \sum_{x_i \in E} h(x_i) P(X = x_i).$$

– Si X est une v.a.c on a :

$$\mathbb{E}(h(X)) = \int_{-\infty}^{+\infty} h(x) f_X(x) dx.$$

Le moment d'un variable aléatoire Soit k un entier naturel, le moment d'ordre k de X est $\mathbb{E}(X^k)$ Défini par :

$$\mathbb{E}(X^k) = \int_{-\infty}^{+\infty} x^k f_X(x) dx.$$

Et le moment centré d'ordre k est :

$$\mathbb{E} [X - \mathbb{E}(X^k)].$$

Définition 1.1.3 *Le moment absolu d'ordre k par rapport au point a , est égal, sous réserve de l'existence de l'intégrale*

$$\mathbb{E}(X - a)^k = \int_{-\infty}^{+\infty} (x - a)^k f_X(x) dx.$$

Variance et écart-type

La variance d'une v.a X est :

$$Var(x) = \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X)^2 - \mathbb{E}(X^2).$$

L'écart-type est la racine de la variance $\sigma(X) = \sqrt{Var(x)}$.

1.1.1 Rapapels d'inégalités classiques :

Inégalité de Markov

Soit X une v.a.réel. positive telle que $\mathbb{E}(X) < \infty$. Alors :

$$\forall t > 0, P(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Inégalité de Bienaymé-Tchebycheff (B-T)

Soit X une v.a.réel. telle que $E(X^2) < \infty$. Alors pour tout :

$$P(|X - \mathbb{E}(X)| > t) \leq \frac{Var(X)}{t^2}.$$

LGN La loi forte des grands nombres (LGN) s'applique et $F_n(x)$ convergent presque sûrement vers $F(x)$.

$$\hat{F}_n(x) \xrightarrow{p.s} F(x), \quad \text{quand } n \longrightarrow \infty.$$

TCL Le théorème central limite (TCL) est un résultat clé qui établit la convergence vers une distribution normale lorsque la taille de l'échantillon augmente, ce qui permet d'effectuer des inférences statistiques robustes dans de nombreux domaines.

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{\mathbb{L}} \mathcal{N}(0, F(x)(1 - F(x))), \quad \text{quand } n \longrightarrow \infty.$$

Le TCL a des applications importantes en statistique inferentielle. Par exemple, il permet de justifier l'utilisation de tests d'hypothèse basés sur la distribution normale. Il permet également de construire des intervalles de confiance pour les estimations

basées sur les échantillons.

Convergence en probabilité

Nous rappelons ici les notions de convergence des suites des variables aléatoires, aussi deux théorèmes fondamentaux en statistique on va les données.

On dit que $\{X_n\}_{n \geq 1}$ converge en probabilité vers une v.a

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

Théorème 1.1.1 (*Loi des grands nombres*) : Soit $\{X_n\}_{n \geq 1}$ une suite $m = \mathbb{E}(X_i)$ et $\sigma^2 = \text{Var}(X_i)$ finies, notons $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. alors :

$$\bar{X}_n \xrightarrow{p.s} m, \text{ quand } n \rightarrow \infty.$$

$$\bar{X}_n \xrightarrow{p} m, \text{ quand } n \rightarrow \infty.$$

Théorème 1.1.2 (*Central limite*) : Soit $\{X_n\}_{n \geq 1}$ une suite m et σ^2 . Alors :

$$\frac{U_n - nm}{\sigma\sqrt{n}} \xrightarrow{loi} N(0,1), \text{ quand } n \rightarrow \infty,$$

où $U_n = \sum_{i=1}^n X_i$ Cette Théorie prouve que quelle que soit la loi des v.a considérées leur moyenne se comporte asymptotiquement comme une loi normale, ce qui montre l'importance de la loi normale dans la modélisation statistique. La démonstration de ce théorème est basée sur la fonction caractéristique. Pour la preuve on réfère à [Devroye et Györfi \(1985\)](#), page 66].

1.2 Estimateur et propriétés

1.2.1 Estimation

Définition 1.2.1 Soient (X_1, \dots, X_n) un n -échantillon aléatoire simple issu d'une variable aléatoire X (discrète ou continue) et θ un paramètre associé à la loi probabilité P_θ de θ . Un estimateur de θ est une variable aléatoire T fonction des X_i

$$T = f(X_1, X_2, \dots, X_n).$$

Si on considère n observations x_1, x_2, \dots, x_n l'estimateur T fournira une estimation de notée :

$$\hat{\theta} = f(x_1, x_2, \dots, x_n).$$

Exemple 1.2.1 \bar{X} et S^2 , sont les estimateurs de μ et σ^2 (resp).

Comparaison d'estimateurs

Définition 1.2.2 On dit que l'estimateur T_n^1 domine l'estimateur T_n^2 si pour tout θ .

Définition 1.2.3 soit T_n^1 et T_n^2 deux estimateurs de θ , T_n^1 est dit plus efficace que T_n^2 si :

$$\text{Var}(T_n^1) < \text{Var}(T_n^2).$$

1.2.2 Estimation non-paramétrique de la foction répartition

On considère $X = (X_1, X_2, \dots, X_n)$ un n -échantillon iid de fonction de répartition $F(x) = P(X_i \leq x)$ tels que :

- F est une fonction croissante.

- F continue à droite, $\forall \varepsilon > 0, \lim_{\varepsilon \rightarrow 0} F_X(x + \varepsilon) = F_X(x)$.
- $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow +\infty} F(x) = 1$.

Définition 1.2.4 *la fonction de répartition empirique associée à $X_n = (X_1, \dots, X_n)$ est la fonction aléatoire définie par :*

$$\begin{aligned} \hat{F}_n : \mathbb{R} &\rightarrow [0, 1] \\ x &\rightarrow \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(X_i \leq x)} \end{aligned}$$

D'autre terme X_1, X_2, \dots, X_n v.a.r, iid prenant une valeur inférieure ou égale à x .

Chaque X_i ayant une probabilité $F(x)$ d'être inférieure ou égale à x .

$$\begin{aligned} F_n(x) &= \frac{\text{nombre d'observation} \leq x}{n} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(X_i \leq x)} \quad (1.2) \\ &= \begin{cases} 0 & \text{si } x \leq X_1, \\ \frac{i}{n} & \text{si } X_i \leq x \leq X_{i+1}, i = 1, \dots, n-1, \\ 1 & \text{si } x \geq X_n. \end{cases} \end{aligned}$$

1.2.3 Propriétés de la fonction de répartition empirique

Une fonction de répartition empirique est une fonction qui décrit la répartition empirique d'un échantillon de données. Il est souvent utilisé pour estimer la fonction de répartition d'une population inconnue à partir d'un échantillon d'observations.

Espérance

Pour le terme de la espérance, on a :

$$\hat{F}_n(x) = \mathbb{E}(\hat{F}_n(x)) = F(x).$$

Biais

Pour le terme de la biais, on a :

$$Biais(\hat{F}_n) = \mathbb{E}(\hat{F}_n(x)) - F(x).$$

Donc :

$$Biais(F_n) = 0.$$

Variance

Pour le terme de la variance, on a :

$$Var(\hat{F}_n(x)) = \mathbb{E}(\hat{F}_n(x)^2) - \mathbb{E}(\hat{F}_n(x)).$$

On va calculer $\mathbb{E}(\hat{F}_n(x)^2)$, on obtient :

$$\begin{aligned} \mathbb{E}(\hat{F}_n^2(x)) &= \mathbb{E}\left(\frac{1}{n^2} \sum_{i=1}^n (\mathbb{I}_{(X_i \leq x)})^2\right) = \frac{1}{n^2} \mathbb{E}\left(\sum_{i=1}^n \mathbb{I}_{(X_i \leq x)}^2 + \sum_{i=1, j \neq i}^n \mathbb{I}_{(X_i \leq x)} \mathbb{I}_{(X_j \leq x)}\right) \\ &= \frac{1}{n} F(x) + \frac{n-1}{n} F(x)^2. \end{aligned}$$

D'ou

$$\begin{aligned} Var(\hat{F}_n(x)) &= \frac{1}{n} F(x) + \frac{n-1}{n} F(x)^2 - F(x)^2 \\ &= \frac{1}{n} F(x)(1 - F(x)). \end{aligned} \tag{1.3}$$

Critères d'erreurs

L'évaluation de la similarité entre l'estimateur \hat{f}_n et la vraie densité f estimé, nécessite des critères d'erreur. La mesure la plus naturelle utilisée est la moyenne intégrée des erreurs quadratiques. Ainsi, on définit d'abord la moyenne des erreurs quadratiques (Mean Squared Error *MSE*).

– L'erreur quadratique moyenne *MSE* :

$$MSE(f(x), \hat{f}_n(x)) = \mathbb{E} \left[f(x) - \hat{f}_n(x) \right]^2 = Var(\hat{f}_n(x)) + Biails(\hat{f}_n(x))^2.$$

– L'erreur quadratique moyenne intégrée *MISE* :

Une mesure globale de l'efficacité de l'estimateur \hat{f}_n est obtenue en intégrant *MSE*. Il s'agit de l'Erreur Quadratique Moyenne Intégrée (en anglais Mean Integrated Squared Error) (*MISE*), qui s'écrit comme suit :

$$MISE(f(x), \hat{f}_n(x)) = \int_{\mathbb{R}} \mathbb{E} \left[f(x) - \hat{f}_n(x) \right]^2 dx = \int_{\mathbb{R}} \left[Var(\hat{f}_n(x)) + Biails(\hat{f}_n(x))^2 \right] dx.$$

– L'erreur quadratique intégrée *ISE* :

Notons qu'il est plus pratique d'utiliser l'intégrale des erreurs quadratiques (*ISE*) pour l'évaluation du *MISE*. L'*ISE* est défini par :

$$ISE(f(x), \hat{f}_n(x)) = \int_{\mathbb{R}} \left[f(x) - \hat{f}_n(x) \right]^2 dx.$$

Chapitre 2

Fonction des répartition et distribution

2.1 Estimation paramétrique

La statistique paramétrique est une méthode d'analyse statistique qui suppose que les données suivent une distribution de probabilité connue ou spécifiée par un certain nombre de paramètres. Elle est souvent utilisée lorsque les données sont suffisamment bien décrites par une distribution normale ou une autre distribution connue, et que les paramètres de cette distribution peuvent être estimés à partir des données.

Définition 2.1.1 *Un modèle paramétrique est un modèle dans lequel le type de loi de X est supposé connu, mais il dépend d'un paramètre inconnu θ de dimension n . La famille des lois de probabilités possibles pour X peut alors s'écrire $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^n\}$, par exemple :*

Modèle Gaussien	$\{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$
Modèle de Poisson	$\{P(\lambda), \lambda > 0\}$
Modèle Exponentiel	$\{\mathcal{E}(\lambda), \lambda > 0\}$

TAB. 2.1: modèles paramétriques

Sont des modèles paramétriques.

Dans ce cadre paramétrique, le problème est celui de l'estimation des paramètres par lesquels on obtiendra une estimation complète de la loi X . Il existe plusieurs façons de construire un estimateur pour un paramètre donné. Les plus populaires sont la méthode des moments et la méthode du maximum de vraisemblance.

2.1.1 Estimation par la méthode du maximum de vraisemblance (MVS)

Soit X une variable aléatoire réelle de loi paramétrique (discrète ou continue), dont on veut estimer le paramètre. Alors on définit une fonction f telle que :

$$f(\theta, x) = \begin{cases} f_\theta(x) & \text{si } X \text{ est une v.a continue de densité } f. \\ P_\theta(X = x) & \text{si } X \text{ est une v.a discrète de probabilité ponctuelle } P. \end{cases}$$

La méthode de maximisation de la vraisemblance est une approche pour estimer les paramètres d'un modèle statistique en se basant sur les données observées. L'objectif est d'identifier les valeurs des paramètres qui maximisent la probabilité d'observer les données observées. La vraisemblance est une fonction de densité de probabilité qui décrit la probabilité d'observer les données en fonction des valeurs des paramètres.

Définition 2.1.2 *On appelle fonction de vraisemblance de pour une réalisation*

(x_1, x_2, \dots, x_n) d'un échantillon, la fonction de θ :

$$\mathcal{L}(x_1, x_2, \dots, x_n; \theta) = f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i, \theta).$$

Définition 2.1.3 La méthode consistant estimer θ par la valeur qui maximise \mathcal{L} (vraisemblance) s'appelle méthode du maximum de vraisemblance

$$\hat{\theta} = \left\{ \theta / \mathcal{L}(\hat{\theta}) = \sup_{\theta} \mathcal{L}(\theta) \right\}.$$

La méthode de maximisation de la vraisemblance consiste à trouver les valeurs des paramètres θ qui maximisent cette fonction de vraisemblance $\mathcal{L}(x_1, x_2, \dots, x_n; \theta)$. Cela revient à résoudre l'équation suivante :

$$\hat{\theta} = \arg \max_{\theta} (\mathcal{L}(x_1, x_2, \dots, x_n; \theta)),$$

où $\hat{\theta}$ est l'estimateur des paramètres.

Ceci est un problème d'optimisation. On utilise généralement le fait que si \mathcal{L} est dérivable et si \mathcal{L} admet un maximum global en une valeur, alors la dérivée première s'annule en et que la dérivée seconde est négative.

Réciproquement, si la dérivée première s'annule en $\theta = \hat{\theta}$ et que la dérivée seconde est négative en $\theta = \hat{\theta}$; alors est θ un maximum local (et non global) de $\mathcal{L}(x_1, x_2, \dots, x_n; \theta)$. Il est alors nécessaire de vérifier qu'il s'agit bien d'un maximum global. La vraisemblance étant positive et le logarithme népérien de la vraisemblance (le produit se transforme en somme, ce qui est plus simple dérivé). Ainsi en pratique :

1. La condition nécessaire :

$$\frac{\partial(\mathcal{L}(x_1, x_2, \dots, x_n; \theta))}{\partial \theta} = 0, \text{ ou } \frac{\partial \ln(\mathcal{L}(x_1, x_2, \dots, x_n; \theta))}{\partial \theta} = 0. \quad (2.1)$$

permet de trouver la valeur $\hat{\theta}$.

2. $\theta = \hat{\theta}$ est un maximum local si la condition suffisante est remplie au point critique :

$$\frac{\partial^2(\mathcal{L}(x_1, x_2, \dots, x_n; \theta))}{\partial \theta^2}(\hat{\theta}) \leq 0, \text{ ou } \frac{\partial^2 \ln(\mathcal{L}(x_1, x_2, \dots, x_n; \theta))}{\partial \theta^2}(\hat{\theta}) \leq 0. \quad (2.2)$$

- L'*EMV* peut n'exister pas.
- Si un *EMV* existe, il n'est pas toujours unique, par exemple : modèle de Cauchy et modèle de Laplace.

1. Dans le cas discrète :

Si les X_i sont de loi *Bernoulli*(p), on a :

$$P(X_i = x, p) = \begin{cases} p & \text{si } x_i = 1 \\ 1 - p & \text{si } x_i = 0 \end{cases} = p^{x_i}(1 - p)^{1-x_i}.$$

Donc la fonction de vraisemblance est :

$$\mathcal{L}(p, x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(X_i = x_i, p) = \prod_{i=1}^n p^{x_i}(1 - p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1 - p)^{\sum_{i=1}^n (1-x_i)}.$$

D'où

$$\ln \mathcal{L}(p, x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i \ln p + \left(n - \sum_{i=1}^n x_i \right) \ln (1 - p). \quad (2.3)$$

Alors on va dériver [2.3](#)

$$\begin{aligned} \frac{\partial}{\partial p} \ln \mathcal{L}(p, x_1, x_2, \dots, x_n) &= \frac{\sum_{i=1}^n x_i}{p} - \frac{\left(n - \sum_{i=1}^n x_i \right)}{(1-p)} \\ &= \frac{\sum_{i=1}^n x_i - np}{p(1-p)} = 0. \end{aligned}$$

Donc on obtient :

$$p = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n.$$

Par conséquent, L'EMV de p est $\hat{p} = \bar{X}_n$.

2. Dans le cas continue :

Soit $X \rightsquigarrow N(\mu, 1)$ et

$$\begin{cases} H_0 & : \mu \leq 0. \\ H_1 & : \mu > 0. \end{cases}$$

Avec $\alpha = 0.1, n = 65$. pour $\mu_1 > \mu_2$ on écrit :

$$\frac{L_1}{L_2} = \exp \left\{ n(\mu_1 - \mu_2) \left[\bar{x} - \frac{1}{2}(\mu_1 + \mu_2) \right] \right\},$$

où $t = \bar{x}$, $a := n(\mu_1 - \mu_2) > 0$ et $b = a - \frac{1}{2}n(\mu_1 + \mu_2)$. Il est clair que la fonction $t \rightarrow \exp\{at + b\}$ est une fonction croissante en t . Donc la distribution de X possède un rapport de vraisemblance croissante. Comme X est continue, le test upp le plus puissant est défini par :

$$\delta = \begin{cases} 1 & \text{si } \bar{x} \geq c \\ 0 & \text{si } \bar{x} < c, \end{cases}$$

avec $P_{\mu=0}(\bar{X} \geq c) = 0.1$, ce qui implique $c = 1.28/\sqrt{65} = 0.16$.

On a donc

$$\delta = \begin{cases} 1 & \text{si } \bar{x} \geq 0.16 \\ 0 & \text{si } \bar{x} < 0.16, \end{cases}$$

la fonction puissance du test δ est définie par :

$$\pi(\mu) = \begin{cases} \alpha(\mu) & \text{si } \mu \geq 0, \\ 1 - \beta(\mu) & \text{si } \mu < 0. \end{cases}$$

En d'autres termes

$$\begin{aligned} \pi(\mu) &= P(\bar{X} \geq 0.16 | \mu \in \mathbb{R}) \\ &= 1 - P(\sqrt{65}(\bar{X} - \mu) < 1.28 - \sqrt{65}\mu | \mu \in \mathbb{R}) \\ &= 1 - P(Z < 1.28 - \sqrt{65}\mu | \mu \in \mathbb{R}) \\ &= 1 - \Phi(1.28 - \sqrt{65}\mu), \mu \in \mathbb{R} \\ &= \Phi(\sqrt{65}\mu - 1.28). \end{aligned}$$

La fonction puissance π étant croissante sur \mathbb{R} , on trouve bien le fait que :

$$\begin{aligned} \alpha &= \sup_{\mu \geq 0} \pi(\mu) = \pi(0) \\ &= \Phi(-1.28) = 1 - \Phi(1.28) \\ &= 1 - 0.899 = 1, \end{aligned}$$

qui correspond en effet au seuil (niveau) de signification de test. Le graphe de la fonction puissance

$\mu \rightarrow \pi(\mu) = \Phi(\sqrt{65}\mu - 1.28)$, $\mu \in \mathbb{R}$ est donné par la figure suivant :

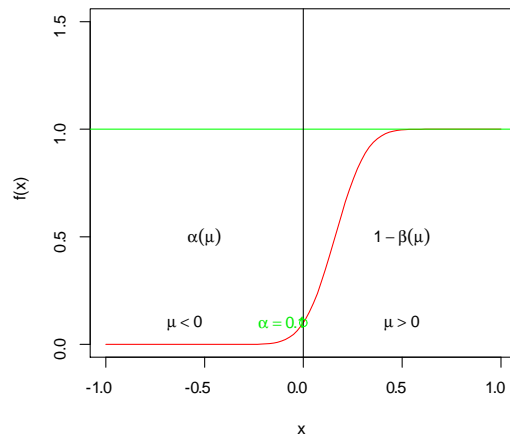


FIG. 2.1: Fonction puissance du test δ

Voici code R utilisée

```
f<-function(x){pnorm(sqrt(65)*x-1.28)}
x<-seq(-1,1,length=100)
plot(x,f(x),type="l",col="red",ylim=c(0,1.5))
xlab=expression(mu)
ylab=expression(pi~(mu))
abline(h=1,col="yellow")
abline(v=0)
points(0,0.1,col="green")
text(-0.1,0.1,expression(alpha==0.1),col="green2")
text(-0.5,0.5,expression(alpha(mu)))
text(0.5,0.5,expression(1-beta(mu)))
text(-0.6,0.1,expression(mu<=0))
text(0.5,0.1,expression(mu>0)).
```

2.2 Estimation non paramétrique

Comment choisir un estimateur de distribution non paramétrique : cela peut inclure des méthodes telles que la méthode de la noyau la méthode du densité, la méthode de l'histogramme la méthode de densité probabilité de l'estimateur empirique, ect...

Le choix de l'estimateur de distribution non paramétrique dépend de plusieurs facteurs tell que la taille de l'échantillon, la forme le niveau de lissage souhaite d'estimation de distribution non paramétrique le plus courants :

- **Méthode de l'histogramme** : cette méthode divise l'ensemble de données en plusieurs intervalles et estime la densité de probabilité en comptant le nombre d'observation dans chaque intervalle. La largeur de l'intervalle peut être choisie en utilisant différente méthode.
- **Méthode de noyau** : cette méthode utilise des fonctions de noyau pour estimer la densité de probabilité sous-jacente. Les choix courants de fonction de noyau incluent la fonction gaussienne la fonction uniforme et la fonction triangulaire, etc...

2.2.1 Un estimateur simple de la densité : l'histogramme

L'estimateur le plus ancien pour estimer une densité est l'histogramme des fréquences. D'après [Lejeune \(2004\)](#), L'origine des histogrammes est attribuée à *John Graunt* au *XVII^{eme}* siècle répondant à l'objectif d'une représentation de la distribution de données. À ce titre, il peut être considéré comme un estimateur de la densité de probabilité sous-jacente à un ensemble fini d'observations. Supposons que l'on ait n observations x_1, \dots, x_n issues d'une même loi de probabilité inconnue de densité f , où f est à support borné $]a_0, a_k]$. Estimer cette densité f par la méthode de l'histogramme Il s'agit d'approximation de f par une fonction étagée, qui permet d'étudier la répartition des données et parfois de justifier certains choix de modèles

Comte (2017).

Supposons pour simplifier qu'on soit en dimension 1 et que les variables de l'échantillon soient à valeurs dans $[0, 1]$ donc $f : [0, 1] \rightarrow \mathbb{R}_+$.

On se donne un découpage de $[0, 1]$, on partitionne l'intervalle de référence $]a_0, a_k]$ en $k \in \mathbb{N}$] Pour simplifier encore, on suppose que les classes sont de même longueur $a_{k+1} - a_k = a_k - a_{k-1}$. Cette longueur est notée h . Estimer f par la méthode de l'histogramme consiste simplement à estimer f par une fonction constante sur chaque classe, cette constante étant liée à la proportion de X_k tombant dans cette classe. Plus exactement on pose, pour $t \in]a_k, a_{k+1}]$

$$f_h(x) = \frac{1}{nh} \# \{k : X_k \in]a_k, a_{k+1}]\}.$$

Pour voir très exactement d'où vient cette formule : on a, si f est égale à une constante c_k constante sur $]a_k, a_{k+1}]$

$$F(a_{k+1}) - F(a_k) = \int_{a_k}^{a_{k+1}} f(t) dt = c_k h.$$

Ensuite on approche la probabilité $F(a_{k+1}) - F(a_k)$, qui correspond à la probabilité que $X \in]a_k, a_{k+1}]$, par la proportion de X_k se trouvant dans $]a_k, a_{k+1}]$. On a alors :

$$c_k = \frac{F(a_{k+1}) - F(a_k)}{h} \simeq \frac{1}{nh} \# \{k, X_k \in]a_k, a_{k+1}]\}.$$

La performance de cet estimateur dépend fortement du nombre de classes.

Exemple 2.2.1 *Pour illustrer l'importance du nombre de classes utilisé dans l'histogramme, nous proposons dans la Figure [2.2](#) un histogramme de basé sur 10000 observations tirées aléatoirement d'une loi khi-deux*

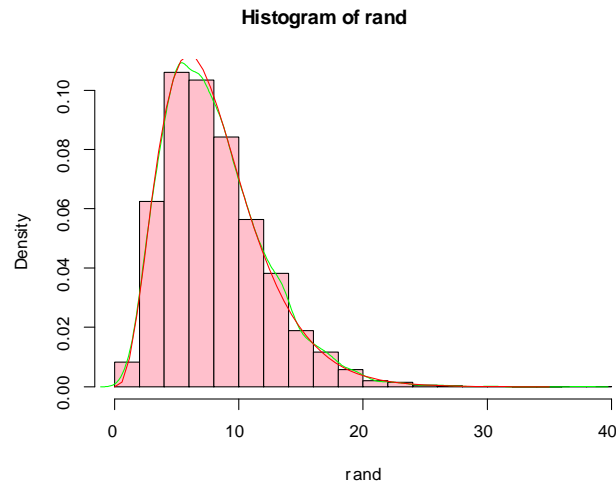


FIG. 2.2: Estimation par histogramme de densité de khi-deux

Voici code R utilisée

```
rand=rchisq(10000,8)

rand

hist(rand,freq=FALSE,col="pink")

lines(density(rand),col="green")

seq=seq(0,35,length.out = 60)

ych2=dchisq(seq,8)

lines(seq,ych2,type = "l",col="red")
```

2.2.2 Propriétés statistiques de l'histogramme

Nous présentons, quelques propriétés statistiques de l'estimateur par histogramme f_h .

En statistiques, il est nécessaire de mesurer la qualité d'un estimateur. Pour cela, on évalue, d'une part, l'écart entre la moyenne de l'estimateur et la densité à estimer,

ce critère d'évaluation est appelé biais, et d'autre part, la variance de l'estimateur (due au caractère aléatoire d'observations) qui caractérise la dispersion des valeurs de l'estimateur dans l'ensemble d'observations. On essaye généralement de réduire au mieux ces deux quantités.

- Le biais de l'estimateur est donné par [Simonoff et Smoothing \(2012\)](#), pour tout $x \in [a_{k-1}, a_k]$ par :

$$\begin{aligned} \text{Biais}(f_h(x)) &= \mathbb{E}(f_h(x)) - f(x) \\ &= \frac{1}{2}f'(x)(h - 2(x - a_k)) + O(h^2), \end{aligned}$$

ou O est un terme résiduel et f' est la dérivée de f . f doit être une fonction de $L^2([a, b])$ absolument continue et carrée intégrable.

- La variance de l'estimateur est donnée par [Simonoff et Smoothing \(2012\)](#), pour tout $x \in [a_{k-1}, a_k]$ par :

$$\text{Var}(f_h(x)) = \mathbb{E}(f_h(x)^2) - \mathbb{E}(f_h(x))^2 = \frac{f(x)}{nh} + O\left(\frac{1}{n}\right).$$

Discussion du comportement du biais et de la variance

- Le biais décroît si h diminue mais la variance augmente.
- Pour que la variance tende vers 0, il faut que $nh \rightarrow \infty$.
- La variance diminue si h augmente mais le biais augmente.

Afin d'apprécier la qualité de l'estimateur, il est usuel d'évaluer la distance entre l'estimateur et la densité à estimer. La distance la plus couramment utilisée est l'erreur quadratique moyenne MSE .

- Le MSE de l'estimateur f_h est donné [Simonoff et Smoothing \(2012\)](#), pour tout

$x \in]a_{k-1}, a_k]$ par :

$$\begin{aligned} MSE(f_h(x)) &= Var(f_h(x)) + Biases(f_h(x))^2 \\ &= \frac{f(x)}{nh} + \frac{f'(x)^2}{4} [h - 2(x - a_k)]^2 + O\left(\frac{1}{n}\right) + O(h^3). \end{aligned}$$

– Le *MISE* de l'estimateur f_h est donné [Simonoff et Smoothing (2012)], pour tout

$x \in]a_{k-1}, a_k]$ par :

$$MISE(f_h(x)) = \frac{1}{nh} + \frac{h^2 \int f'(x)^2 dx}{12} + O\left(\frac{1}{n}\right) + O(h^3). \quad (2.4)$$

Remarque 2.2.1 Dans [2.4], on voit que : un petit h donne un histogramme peu biaisé, tandis qu'un grand h et un grand n déterminent un histogramme moins variable.

Choix du paramètre h

Définition 2.2.1 (Règle de Scott) La valeur qui minimise l'erreur quadratique moyenne intégrée, *MISE* est :

$$h_{opt} = \left[\frac{6}{\int f'(t) dt} \right]^{\frac{1}{3}} n^{-\frac{1}{3}} \simeq C n^{-\frac{1}{3}}.$$

Définition 2.2.2 (Règle de Sturges) Prendre le nombre k de classes égal à $1 + \log_2 n$. En pratique, cela revient à prendre $h = \frac{x_k - x_1}{h}$, x_k sont les valeurs d'observations d'un échantillon ordonné par ordre croissant. La règle de Sturges a tendance à produire des histogrammes trop lisses.

2.2.3 Estimateurs à noyaux

En raison de leur simplicité, les praticiens ont utilisé les noyaux classiques (symétriques), cette méthode, elle à été proposé initialement par [Parzen (1979)] et [Rosenblatt (1956)] pour l'estimation de la fonction densité de probabilité f à support non borné. Citons le noyau gaussien, uniforme, biweight ainsi que le noyau optimal d'[Epanechnikov (1969)]

La fonction de répartition empirique

On considère un ensemble de n réalisations $\{x_1, x_2, \dots, x_n\}$ de variables aléatoires réelles iid $\{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n\}$, ayant la fonction de répartition F telle que $\forall x \in \mathbb{R}, F(x) = P(X_i \leq x), i = 1, 2, \dots, n$. L'estimateur naturel de F est la fonction en escalier définie pour une valeur donnée d'un réel x par :

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(X_i \leq x)} = \begin{cases} 0 & \text{si } X_{(1)} > x, \\ \frac{i}{n} & \text{si } X_{(i)} \leq x < X_{(i+1)}, \\ 0 & \text{si } X_{(n)} < x, \end{cases} \quad (2.5)$$

où $\mathbb{I}_{(X_i \leq x)}$ est la fonction indicatrice définie par :

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

Exemple 2.2.2 *Illustration :soit les 10 réalisations aléatoire. Pour construire \hat{F} à application \mathbf{R} ,*

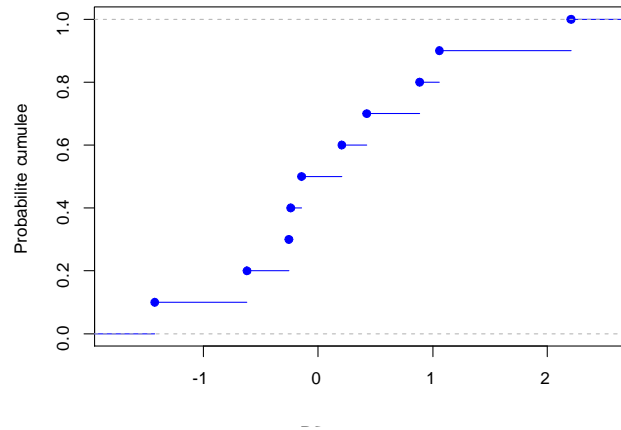


FIG. 2.3: Fonction de répartition empirique $n = 10$

Voici code R utilisé

```
# Créer un vecteur de données

data <- rnorm(10)

# Estimation de la fonction de répartition empirique

ecdf_data <- ecdf(data)

# Tracer la fonction de répartition empirique

plot( ecdf_data,main=" ",

      xlab="Données",

      ylab="Probabilité cumulée",

      col="blue")
```

Propriétés de \hat{F}

Pour une valeur x donnée, $\mathbb{I}_{(X_i \leq x)}$ et pour tout i , il y'a une variable de Bernoulli telle que :

$$P(\mathbb{I}_{(X_i \leq x)} = 1) = P(X_i \leq x) = F(x).$$

Et donc, $E(\mathbb{I}_{(X_i \leq x)}) = F(x)$: cette *Bernoulli* est un estimateur sans biais de $F(x)$.

En conséquence \hat{F} est également un estimateur sans biais de F :

$$\hat{F}_n(x) = \mathbb{E} \left(\frac{1}{n} \mathbb{I}_{(X_i \leq x)} \right) = F(x).$$

Biais Estimateur de F est un estimateur sans biais :

$$Biais(\hat{F}_n(x)) = \mathbb{E}(F(x)) - F(x) = 0.$$

Variance Remarquons que $n\hat{f}_n(x)$ est une somme d'une suite de variables aléatoires iid U_i avec $U_i = \mathbb{I}_{(X_i \leq x)}, i = 1, 2, \dots, n$. Les variables aléatoires U_i suivent une loi de *Bernoulli* de paramètre $p = F(x)$, ($U_i \rightsquigarrow B(p), i = 1, 2, \dots, n$).

Ainsi $n\hat{F}_n(x) = \sum_{i=1}^n \mathbb{I}_{(X_i \leq x)} = \sum_{i=1}^n U_i$ suit une loi *Binomiale* de paramètre n et $p = F(x)$.

Ce qui donne :

$$Var(n\hat{F}_n(x)) = Var \left(\sum_{i=1}^n U_i \right) = n(F(x)(1 - F(x))).$$

Ainsi

$$Var(\hat{F}_n(x)) = \frac{1}{n} F(x) [1 - F(x)].$$

La dernière égalité utilisant l'hypothèse d'indépendance des n aléatoires de *Bernoulli*.

On peut remarquer qu'asymptotiquement,

$$Var(\hat{F}_n(x)) \xrightarrow[n \rightarrow \infty]{} 0.$$

C'est à dire que F_n est un estimateur convergent de F [Parzen (1962), page 5].

On dispose d'observations $X_i, i = 1, 2, \dots, n$ iid, de loi inconnue P_f de densité f telle que : $f \in \mathcal{F}$, où \mathcal{F} espace fonctionnel à estimation non-paramétrique de f .

Construction d'un estimateur : l'estimateur à noyau [Rosenblatt \(1956\)](#).

D'après *Glivenko-Cantelli*, on a (uniformément en x) :

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(X_i \leq x)} \xrightarrow[p.s.]{} F(x) = P(X \leq x).$$

Un inconvénient de l'estimateur par histogramme précédent est que la fonction de densité résultante f_n est irrégulière : c'est une fonction constante par morceaux et présente donc des sauts à la fin de chaque classe. En général, la densité à estimer est plus lisse, au moins continue. L'estimation par noyau vise à remédier à cet écueil [Lejeune \(2004\)](#).

Principe : Si f est continue en x (basée sur la dérivée de la fonction de répartition) et en utilisant l'équation [2.5](#) ce qui permet d'écrire pour tout x :

$$f(x) = F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x-h)}{2h}.$$

L'idée est donc d'utiliser l'approximation suivante, pour h petit,

$$f(x) \approx \lim_{h \rightarrow 0} \frac{F(x+h) - F(x-h)}{2h}.$$

Pour estimer la densité f on peut donc passer par un estimateur \hat{F}_n de la F . Voyons ce qui se passe si on choisit comme estimateur la fonction de répartition empirique F_n . (On rappelle que $F_n(x) = \frac{1}{n} \sum \mathbb{I}_{(X_i \leq x)}$). On choisit un $h > 0$ petit pour que

l'approximation ci-dessus soit valable, et on pose :

$$\begin{aligned}
 \hat{f}_n(x) &= \frac{F(x+h) - F(x-h)}{2h} \\
 &= \frac{1}{2h} \left[\frac{1}{n} \left(\sum_{i=1}^n \mathbb{I}_{(X_i \leq x+h)} - \sum_{i=1}^n \mathbb{I}_{(X_i \leq x-h)} \right) \right] \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{1}{2h} \mathbb{I}_{(x-h \leq X_i \leq x+h)} \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{1}{2h} \mathbb{I}_{(-1 \leq \frac{X_i - x}{h} \leq 1)} \\
 &= \frac{1}{nh} \sum_{i=1}^n \mathbb{K}_0 \left(\frac{X_i - x}{h} \right), \tag{2.6}
 \end{aligned}$$

où

$$\mathbb{K}_0(u) = \frac{1}{2} \mathbb{I}_{(-1 \leq u \leq 1)},$$

\mathbb{K}_0 est appelé le noyau de *Rosenblatt*. Cet estimateur a le même inconvénient de régularité que l'estimateur par histogramme.

On a donc l'idée d'utiliser des noyaux plus réguliers.

Ce cas particulier nous amène à la généralisation suivante, qui revient à [Parzen \(1962\)](#)

et qui consiste à prendre comme estimateur de la densité :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n \mathbb{K} \left(\frac{X_i - x}{h} \right). \tag{2.7}$$

Où $\mathbb{K} : \mathbb{R} \rightarrow \mathbb{R}$ est une fonction intégrable, et telle que $\int_{-\infty}^{+\infty} \mathbb{K}(u) du = 1$. C'est l'estimateur à noyau de la densité ou estimateur de *Parzen-Rosenblatt*. La fonction \mathbb{K} est dite noyau (*Kernel en anglais*) et h est appelée paramètre de lissage ou fenêtre (*bandwidth en anglais*).

Pour $n \in \mathbb{N}^*$; On supposera que la fenêtre h dépend de n et on la notera h_n ; la suite $(h_n)_{n \geq 1}$ tendant vers 0 lorsque $n \rightarrow \infty$.

Définition 2.2.3 Soit $\mathbb{K} : \mathbb{R} \rightarrow \mathbb{R}$ intégrable et tel que :

$$\int_{\mathbb{R}} \mathbb{K}(u) = 1.$$

Alors \mathbb{K} est appelé noyau (*kernel*).

- Un noyau est positif ou nul ; c'est à-dire $\mathbb{K}(x) \geq 0$ pour tout x ; en d'autres termes \mathbb{K} est une fonction de densité de probabilité.
- On définit les moments d'un noyau \mathbb{K} par $\mu_j(\mathbb{K}) = \int_{-\infty}^{+\infty} x^j \mathbb{K}(x) dx$, $j = 1, 2, \dots, n$.
- Une fonction noyau est symétrique : $\mathbb{K}(u) = \mathbb{K}(-u)$ pour tout x : Dans ce cas, tous les moments impairs sont nuls :

$$\int_{-\infty}^{+\infty} x^m \mathbb{K}(x) dx = 0 \quad \text{si } m \text{ impaire.}$$

Les noyaux usuels sont répertoriés dans le tableau suivant :

<i>Noyaux</i>	<i>support</i>	<i>équations</i>
Uniform	$[-1, 1]$	$\mathbb{K}_0(x) = \frac{1}{2}$
Epanechnikov	$[-1, 1]$	$\mathbb{K}_1(x) = \frac{3}{4}(1 - x^2)$
Biweight	$[-1, 1]$	$\mathbb{K}_2(x) = \frac{15}{16}(1 - x^2)^2$
Triweight	$[-1, 1]$	$\mathbb{K}_3(x) = \frac{35}{32}(1 - x^2)^3$
Gaussien	\mathbb{R}	$\mathbb{K}_\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$
Cosinus	$[-1, 1]$	$\mathbb{K}(t) = \frac{\pi}{4} \cos(\frac{\pi}{2}x)$

TAB. 2.2: Noyaux usuelles

Nous représentons 4 noyaux par le programme **R** comme suite :

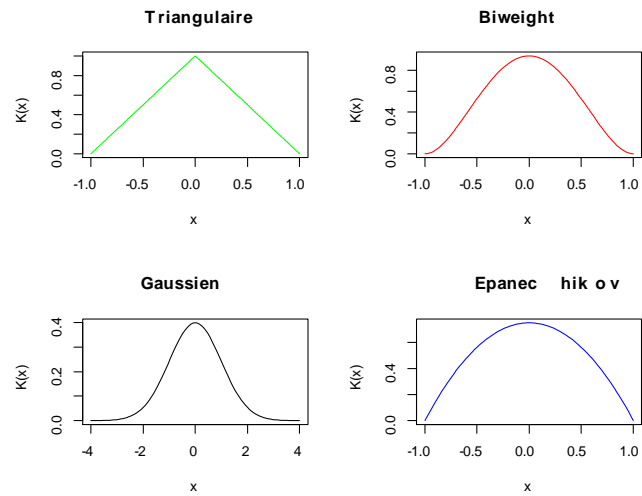


FIG. 2.4: Courbes des noyaux usuels

Voici Code R utilisée :

```

K1=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}
K2=function(t){(15/16)*((1-t^2)^2)*ifelse(abs(t)<=1,1,0)}
K3=function(t){dnorm(t)}
K4=function(t){ifelse(abs(t)<=1,(3/4)*(1-t^2),0)}

op=par(mfrow=c(2,2))

curve(K1(x),-1,1,ylab="K(x)",main="Triangulaire",col="green")
curve(K2(x),-1,1,ylab="K(x)",main="Biweight",col="red")
curve(K3(x),-4,4,ylab="K(x)",main="Gaussien",col="black")
curve(K4(x),-1,1,ylab="K(x)",main="Epanechnikov",col="blue")

par(op)
    
```

Définition 2.2.4 Etant donné \mathbb{K} un noyau et $h > 0$, on pose :

$$\forall x \in \mathbb{R}, \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} \mathbb{K} \left(\frac{X_i - x}{h} \right).$$

Proposition 2.2.1 Si \mathbb{K} est un noyau positif et $\int_{\mathbb{R}} \mathbb{K}(t) dt = 1$, alors, \hat{f}_n est une densité de probabilité. De plus, \hat{f}_n a les mêmes propriétés de continuité et de différentiabilité que \mathbb{K} . De plus, \hat{f}_n est continue si \mathbb{K} est continue.

Preuve. on a $\hat{f}_n > 0$ et

$$\begin{aligned} \int_{\mathbb{R}} \hat{f}_n(x) dx &= \int_{\mathbb{R}} \frac{1}{nh} \sum_{i=1}^n \mathbb{K} \left(\frac{x - X_i}{h} \right) dx = \frac{1}{nh} \int_{\mathbb{R}} \sum_{i=1}^n h \mathbb{K}(t) dt \\ &= \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} \mathbb{K}(t) dt = \frac{1}{n} \sum_{i=1}^n 1 \\ &= 1, \end{aligned}$$

avec le changement variable ($t = \frac{x - X_i}{h}$).

Donc \hat{f}_n est bien une densité de probabilité. ■

2.2.4 Propriétés asymptotiques de l'estimateur

Nous étudions deux modes de convergence dans cette section, la convergence quadratique sur la moyenne et presque complètement convergente. On suppose que \mathbb{K} est un noyau vérifiant les conditions suivantes :

1. \mathbb{K} est borné, c-à-d $\sup_{x \in \mathbb{R}} |\mathbb{K}(x)| < \infty$.
2. $\lim_{|x| \rightarrow +\infty} |x| \mathbb{K}(x) \rightarrow 0$ quand $|x| \rightarrow +\infty$.
3. $\mathbb{K} \in L_1(\mathbb{R})$ c-à-d $\int_{\mathbb{R}} |x| \mathbb{K}(x) dx < \infty$.
4. $\int_{\mathbb{R}} |\mathbb{K}(x)| dx = 1$.
5. $\int_{\mathbb{R}} x |\mathbb{K}(x)| dx = 0$.

6. $\int_{\mathbb{R}} x^2 |\mathbb{K}(x)| dx < \infty$.
7. \mathbb{K} est bornée et intégrable et à support compact.

Etude de biais

Soit x fixé dans \mathbb{R} : Le biais de l'estimateur à noyau est :

$$\text{Biais}(\hat{f}_n(x)) = \mathbb{E}(\hat{f}_n(x)) - f(x).$$

En calcule $\mathbb{E}(\hat{f}_n(x))$ on obtient :

$$\begin{aligned} \mathbb{E}(\hat{f}_n(x)) &= \mathbb{E}\left(\frac{1}{nh} \sum_{i=1}^n \mathbb{K}\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{nh} \sum_{i=1}^n \mathbb{E}\left(\mathbb{K}\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{nh} \sum_{i=1}^n \mathbb{E}\left(\mathbb{K}\left(\frac{X - x}{h}\right)\right), \end{aligned}$$

car les X_i sont iid

$$\mathbb{E}(\hat{f}_n(x)) = \frac{1}{h} \int_{\mathbb{R}} \mathbb{K}\left(\frac{Y - x}{h}\right) f_y(y) dy,$$

on pose $\frac{Y - x}{h} = t$.

$$\begin{aligned} \mathbb{E}(\hat{f}_n(x)) &= \frac{1}{h} \int_{\mathbb{R}} \mathbb{K}(t) f_y(th + x) h dt \\ &= \int_{\mathbb{R}} \mathbb{K}(t) f_y(th + x) dt. \end{aligned}$$

En utilisant le développement de Taylor de f au voisinage de x ; on obtient :

$$f_y(th + x) = f(x) + ht f'(x) + \frac{(ht)^2}{2} f''(x) + O(h^2).$$

Alors

$$\begin{aligned}
 \mathbb{E}(\hat{f}_n(x)) &= \int_{\mathbb{R}} \mathbb{K}(t) \left[f(x) + ht f'(x) + \frac{(ht)^2}{2} f''(x) + O(h^2) \right] dt \\
 &= \int_{\mathbb{R}} \mathbb{K}(t) f(x) dt + \int_{\mathbb{R}} \mathbb{K}(t) ht f'(x) dt + \int_{\mathbb{R}} \mathbb{K}(t) \frac{(ht)^2}{2} f''(x) dt + O(h^2) \\
 &= f(x) + \frac{h^2}{2} f''(x) \int_{\mathbb{R}} t^2 \mathbb{K}(t) dt + O(h^2) \\
 &= f(x) + \frac{h^2}{2} f''(x) \mu_2(\mathbb{K}) + O(h^2),
 \end{aligned}$$

où $\mu_2(\mathbb{K}) = \int_{\mathbb{R}} t^2 \mathbb{K}(t) dt$.

Donc

$$\begin{aligned}
 \text{Biais}(\hat{f}_n(x)) &= \mathbb{E}(\hat{f}_n(x)) - f(x) \\
 &= \frac{h^2}{2} f''(x) \mu_2(\mathbb{K}) + O(h^2).
 \end{aligned} \tag{2.8}$$

Etude de variance

Soit x fixée dans \mathbb{R} . La variance de l'estimateur à noyau est :

$$\begin{aligned}
 \text{Var}(\hat{f}_n(x)) &= \text{Var} \left(\frac{1}{nh} \sum_{i=1}^n \mathbb{K} \left(\frac{X_i - x}{h} \right) \right) \\
 &= \frac{1}{n^2 h^2} \sum_{i=1}^n \text{Var} \left(\mathbb{K} \left(\frac{X_i - x}{h} \right) \right) \\
 &= \frac{1}{nh^2} \text{Var} \left(\mathbb{K} \left(\frac{X - x}{h} \right) \right),
 \end{aligned}$$

puisque les X_i sont iid.

$$\begin{aligned}
 \text{Var}(\hat{f}_n(x)) &= \frac{1}{nh^2} \left[\mathbb{E} \left(\mathbb{K} \left(\frac{X_i - x}{h} \right)^2 \right) - \left(\mathbb{E} \left(\mathbb{K} \left(\frac{X_i - x}{h} \right) \right) \right)^2 \right] \\
 &= \frac{1}{nh^2} \left[\int_{\mathbb{R}} \mathbb{K}^2 \left(\frac{Y - x}{h} \right) f_y(y) dy - \left(\int_{\mathbb{R}} \mathbb{K} \left(\frac{Y - x}{h} \right) f_y(y) dy \right)^2 \right],
 \end{aligned}$$

En utilisant la substitution $\frac{Y-x}{h} = t$, on obtient :

$$Var(\hat{f}_n(x)) = \frac{1}{nh} \int_{\mathbb{R}} \mathbb{K}^2(t) f_y(th+x) dt - \frac{1}{n} \left(\int_{\mathbb{R}} \mathbb{K}(t) f_y(th+x) \right)^2 dt.$$

On effectuant un développement de Taylor de l'ordre 0, il vient :

$$f_y(th+x) = f(x) + O(1) \text{ et le terme } \frac{1}{n} \left(\int_{\mathbb{R}} \mathbb{K}(t) f_y(th+x) \right)^2 dt \rightarrow 0, \text{ quand } n \rightarrow \infty,$$

alors

$$\begin{aligned} Var(\hat{f}_n(x)) &= \frac{1}{nh} \int_{\mathbb{R}} \mathbb{K}^2(t) f(x) dt + O(1) \\ &= \frac{1}{nh} f(x) \int_{\mathbb{R}} \mathbb{K}^2(t) dt + O\left(\frac{1}{nh}\right) \\ &= \frac{f(x)R(\mathbb{K})}{nh} + O\left(\frac{1}{nh}\right), \end{aligned} \tag{2.9}$$

où $R(\mathbb{K}) = \int_{\mathbb{R}} \mathbb{K}^2(t) dt$.

2.2.5 Erreur quadratique moyenne

Une des mesures courantes et pratiques de la précision de l'estimation est l'erreur quadratique moyenne (en anglais "Mean Squared Error") donnée par :

$$\begin{aligned} MSE(\hat{f}_n(x)) &= \text{Biais}^2(\hat{f}_n(x)) + Var(\hat{f}_n(x)) \\ &= \frac{h^4}{4} f''(x)^2 \mu_2^2(\mathbb{K}) + \frac{1}{nh} f(x) R(\mathbb{K}) + O\left(\frac{1}{nh}\right) + O(h^4). \end{aligned} \tag{2.10}$$

Asymptotiquement, on a :

$$AMSE(\hat{f}_n(x)) = \frac{h^4}{4} (f''(x))^2 \mu_2^2(\mathbb{K}) + \frac{1}{nh} f(x) R(\mathbb{K}). \tag{2.11}$$

2.2.6 Erreur quadratique moyenne intégrée(MISE)

L'erreur quadratique moyenne intégrée (en anglais "*Mean Integrated Squard Error*").on intégrons l'équation (2.10)

$$\begin{aligned} MISE(\hat{f}_n(x)) &= \int_{\mathbb{R}} MSE(\hat{f}_n(x)) \\ &= \frac{h^4}{4} \mu_2^2(\mathbb{K}) R(f''(x)) + \frac{1}{nh} R(\mathbb{K}) + O\left(\frac{1}{nh}\right) + O(h^4). \end{aligned} \quad (2.12)$$

Telle que $R(f'') = \int_{\mathbb{R}} (f''(x))^2 dx$.

Asymptotiquement, on a :

$$AMISE(\hat{f}_n(x)) = \frac{h^4}{2} \mu_2^2(\mathbb{K}) R(f'') + \frac{1}{nh} R(\mathbb{K}). \quad (2.13)$$

2.2.7 Choix du paramètre de lissage

Choix théorique optimal du paramètre de lissage

Il y a essentiellement deux approches pour trouver un choix optimal, la première approche consiste à trouver le paramètre de lissage (**local**) qui minimise l'*AMSE* ; c-à-d qu'on cherche la solution en prenant la dérivée de l'*AMSE* par rapport à h et en la fixant égale à zéro :

$$h_{opt} = \arg \min_h (\hat{f}_n(x)) = \frac{\partial AMSE}{\partial h} = 0,$$

ce qui est équivalent à :

$$h_{opt} = \left(\frac{f(x) R(\mathbb{K})}{\mu_2^2(\mathbb{K}) f''(x)^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}. \quad (2.14)$$

La seconde approche consiste à trouver le paramètre de lissage (**global**) qui est la

valeur de h qui minimise l' $AMISE$:

$$h_{opt}^* = \arg \min_h AMISE = \frac{\partial AMISE}{\partial h} = 0, \quad (2.15)$$

supposons d'abord que le noyau \mathbb{K} est fixé en dérivant (2.13) par rapport à h ; on obtient :

$$h_{opt}^* = \left(\frac{R(\mathbb{K})}{\mu_2^2(\mathbb{K})R(f'')} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}. \quad (2.16)$$

Remarque 2.2.2 Les choix h_{opt} et h_{opt}^* ont des choix théoriques, qui ne sont pas utilisables en pratique car ils dépendent des quantités f et f'' .

Estimation de la densité à l'aide d'une méthode de lissage

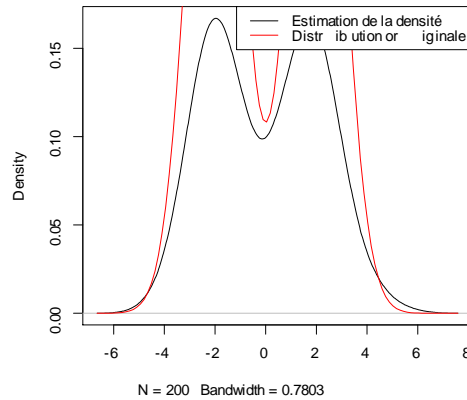


FIG. 2.5: La densité théorique et la densité estimée pour $n = 200, Bandwidth = 0.7803$.

Voici code R utilisé

```
# Générer un échantillon de données à partir d'une distribution bimodale
set.seed(123)

data <- c(rnorm(100, mean = -2), rnorm(100, mean = 2))

# Estimation de la densité à l'aide d'une méthode de lissage local
density <- density(data, bw = "nrd")
```

```
# Tracer l'estimation de la densité et la distribution originale
plot(density, main = "Estimation de la densité à l'aide d'une méthode
de lissage local")

curve(dnorm(x, mean = -2, sd = 1) + dnorm(x, mean = 2, sd = 1),
      add = TRUE, col = "red")

legend("topright", c("Estimation de la densité",
"Distribution originale"), col = c("black", "red"), lty = c(1, 1)).
```

Exemple 2.2.3 On suppose que l'on a observé un échantillon de taille $n = 100$; issu d'une v.a gaussienne et $\hat{f}_n(x)$ l'estimateur à noyau de la densité donné par :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n \mathbb{K} \left(\frac{X_i - x}{h} \right).$$

Nous prenons différentes valeurs de h : $h = 0.05, 0.2, 0.4$ et 0.8 . Pour le noyau \mathbb{K} , nous considérons :

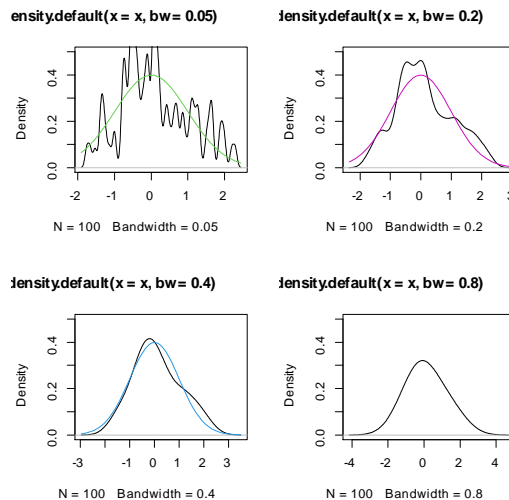


FIG. 2.6: La densité théorique $N(0,1)$ et la densité estimée avec noyau Gaussien pour $n = 100, h = 0.05, h = 0.2, h = 0.4$ et $h = 0.8$.

Voici code R utilisé

```
x=rnorm(100,0,1)

par(mfrow=c(2,2))

plot(density(x,bw=0.05),ylim=c(0,0.5))

curve(dnorm(x,0,1),add=T,col=3)

plot(density(x,bw=0.2),ylim=c(0,0.5))

curve(dnorm(x,0,1),add=T,col=6)

plot(density(x,bw=0.4),ylim=c(0,0.5))

curve(dnorm(x,0,1),add=T,col=4)

plot(density(x,bw=0.8),ylim=c(0,0.5))

curve(dnorm(x,0,1),add=T,col=2).
```

Remarque 2.2.3 Ces résultats (2.14) et (2.16) qui donnent le h_{opt} et h_{opt}^* sont basés sur un choix optimal théorique qui n'est pas utilisable en pratique, car il dépend de la quantité inconnue f'' . Plusieurs méthodes ont été développées pour résoudre se problème, nous citons :

1. Validation croisée des moindres carrés voir [Silverman (1986)] (p. 48-52)].
2. Validation croisée biaisée voir [Wand et Jones (1995)] (p. 65-66)].
3. Validation croisée des probabilités voir [Silverman (1986)] (p. 52-55)].
4. Validation croisée lissée voir [Wand et Jones (1995)], (p.75-79)].
5. La méthode Plug-in voir [Wand et Jones (1995)] (p. 71-75)].
6. Référence à une distribution standard voir [Silverman (1986)] (p. 45-48)].

2.2.8 Méthode de la validation croisée

Validation croisée par les moindres carrés

La méthode de validation croisée vise à trouver une fonction de score $CV(h)$ plus facile à calculer que la $MISE(h)$. Dans cette approche, le paramètre de lissage h est choisi de manière non déterministe, en fonction des observations. L'idée fondamentale des méthodes de validation combinées est de trouver une fonction de score $CV(h)$ qui soit plus simple à calculer que l' $MISE(h)$. Ainsi, le choix du paramètre de lissage h dépend des observations plutôt que d'être prédéterminé. Pour obtenir davantage d'informations sur ce sujet, plusieurs auteurs, tels que [Bowman en 1984], [Rudemo en 1982] et [Stone](1984), peuvent être consultés.

La méthode de validation croisée utilisant les moindres carrés est la plus couramment employée. Son objectif est de minimiser un estimateur raisonnablement défini de l' $ISE(h)$ raisonnable défini par :

$$\int_{\mathbb{R}} (\hat{f}_n(x) - f(x))^2 dx = \int_{\mathbb{R}} \hat{f}_n(x)^2 dx - 2 \int_{\mathbb{R}} \hat{f}_n(x) f(x) dx + \int_{\mathbb{R}} f(x)^2 dx.$$

Compte tenu du fait que la quantité $\int_{\mathbb{R}} f(x)^2 dx$ ne dépend pas de h , on choisit le paramètre de lissage de façon à ce qu'il minimise un estimateur de $CV(h)$ défini par :

$$CV(h) = \int_{\mathbb{R}} \hat{f}_n(x)^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{n,-i}(X_i),$$

où $\sum_{i=1}^n \hat{f}_{n,-i}(X_i) = \frac{1}{(n-1)h} \sum_{i \neq j} \mathbb{K}\left(\frac{x-X_j}{h}\right)$ est l'estimateur calculé à partir de l'échantillon privé de l'observation X_i . Et la valeur optimale de h notée h_{CV} est donnée par :

$$h_{CV} = \arg \min_{h>0} CV(h).$$

Validation croisée non biaisée (UCV)

La méthode de validation croisée non biaisée a été proposée par [\[Rudemo en 1982\]](#) et [\[Bowman en 1984\]](#).

Le critère consiste à choisir le paramètre de lissage qui minimise un estimateur convenable de :

$$\begin{aligned}
 UCV(h) &= \mathbb{E}(MISE[\hat{f}_n(\cdot)]) - \int_{\mathbb{R}} f^2(x) dx \\
 &= \mathbb{E} \left(\int_{\mathbb{R}} [\hat{f}_n(x) - f(x)]^2 dx \right) - \int_{\mathbb{R}} f^2(x) dx \\
 &= \mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n^2(x) dx \right) + \int_{\mathbb{R}} f^2(x) dx - 2\mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n(x) f(x) dx \right) - \int_{\mathbb{R}} f^2(x) dx \\
 &= \mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n^2(x) dx \right) - 2\mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n(x) f(x) dx \right).
 \end{aligned}$$

Remarquons que le terme $\int_{\mathbb{R}} f^2(x) dx$ ne dépend pas de h et donc minimiser $MISE[\hat{f}_n(\cdot)]$ par rapport à h est équivalent à minimiser

$$UCV(h) = \mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n^2(x) dx \right) - 2\mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n(x) f(x) dx \right).$$

Tout d'abord, nous voulons trouver une estimation pour $\int_{\mathbb{R}} \hat{f}_n(x) f(x) dx$.

Remarquons que :

$$\int_{\mathbb{R}} \hat{f}_n(x) f(x) dx = \mathbb{E} \left(\hat{f}_n(x) \right).$$

L'estimateur empirique de $\int_{\mathbb{R}} \hat{f}_n(x) f(x) dx$ est donnée par : $\frac{1}{n} \sum_{i=1}^n f_{n,i}(x_i)$.

Le critère à optimiser est alors :

$$UCV(h) = \int_{\mathbb{R}} \hat{f}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^n f_{n,i}(x_i). \quad (2.17)$$

Preuve. Montrons maintenant que $UCV(h)$ est un estimateur sans biais de

$MISE - \int_{\mathbb{R}} f^2(x)dx$. On a :

$$\mathbb{E}(MISE[\hat{f}_n(x)]) - \int_{\mathbb{R}} f^2(x)dx = \mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n^2(x)dx \right) - 2\mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n(x)f(x)dx \right)$$

Il suffit de montrer que $\int_{\mathbb{R}} \hat{f}_n^2(x)dx$ et $\frac{1}{n} \sum_{i=1}^n f_{n,i}(x_i)$ sont des estimateur sans biais de $\mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n^2(x)dx \right)$ et $\mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n(x)f(x)dx \right)$ respectivement. Or $\mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n^2(x)dx \right)$ admet l'estimateur sans biais trivial .Il reste donc à montrer que l'estimateur $\frac{1}{n} \sum_{i=1}^n f_{n,i}(x_i)$ est un estimateur sans biais de $\mathbb{E} \left(\int_{\mathbb{R}} \hat{f}_n(x)f(x)dx \right)$.

En effet,

$$\mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n f_{n,i}(x_i) \right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(f_{n,i}(x_i)),$$

et

$$\begin{aligned} \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n f_{n,i}(x_i) \right) &= \mathbb{E} \left(\frac{1}{n(n-1)h} \sum_{n=1}^n \sum_{i=1, j \neq 1}^n \mathbb{K} \left(\frac{X_i - X_j}{h} \right) \right) \\ &= \frac{1}{(n-1)h} \sum_{i=1, j \neq 1}^n \mathbb{E} \left(\mathbb{K} \left(\frac{X_i - X_j}{h} \right) \right) \\ &= \frac{1}{h} \mathbb{E} \left(\mathbb{K} \left(\frac{X_i - X_j}{h} \right) \right) \\ &= \frac{1}{h} \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{K} \left(\frac{x-y}{h} \right) f(x)f(y)dx dy \\ &= \int_{\mathbb{R}} \left(\frac{1}{h} \int_{\mathbb{R}} \mathbb{K} \left(\frac{x-y}{h} \right) f(y)dy \right) f(x)dx \\ &= \int_{\mathbb{R}} \left(\frac{1}{nh} \sum_{j=1}^n \mathbb{E} \left(\mathbb{K} \left(\frac{x - X_j}{h} \right) \right) \right) f(x)dx \\ &= \int_{\mathbb{R}} \mathbb{E} \left(\hat{f}_n(x)f(x)dx \right) \\ &= \mathbb{E} \left(\int_{\mathbb{R}} f_n(x)f(x)dx \right). \end{aligned}$$

Finalement, un estimateur sans biais de $MISE - \int_{\mathbb{R}} f^2(x)dx$ est donnée donc par

$UCV(h)$. ■

La formule est donnée par :

$$UCV(h) = \int_{\mathbb{R}} \hat{f}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^n f_{n,i}(x_i). \quad (2.18)$$

Cette quantité est appelée la quantité de "validation croisée".

En utilisant l'équation [2.17](#), le critère $UCV(h)$ devient :

$$UCV(h) = \frac{R(\mathbb{K})}{nh} + \sum_{i=1}^n \sum_{i=1, j \neq 1}^n \int_{\mathbb{R}} \frac{1}{n^2 h^2} \mathbb{K}\left(\frac{x - X_i}{h}\right) \mathbb{K}\left(\frac{x - X_j}{h}\right) dx - \frac{2}{n(n-1)h} \mathbb{K}\left(\frac{X_i - X_j}{h}\right).$$

Le paramètre de lissage du type "validation croisée" est la valeur de h qui minimise cette quantité de validation croisée, c'est-à-dire :

$$h_{UCV}^* = \arg \min_h (UCV(h)). \quad (2.19)$$

La popularité de cette méthode est due à la motivation intuitive et au fait que cet estimateur est asymptotiquement optimal sous de faible condition. L'optimalité asymptotique de la Validation Croisée non biaisée obtenue par [Stone](#).

Validation croisée biaisée (BCV)

[Scot et Terrell 1987](#) ont introduit le critère de validation croisée biaisée en 1987 pour résoudre le problème de la validation croisée impartiale, qui consiste à introduire un biais dans UCV pour réduire sa variance. L'erreur intégrale quadratique moyenne asymptotique $AMISE$ s'écrit sous la forme suivante :

$$AMISE = \frac{h^4}{4} \sigma_{\mathbb{K}}^4(\mathbb{K}) R(\hat{f}_n'') + \frac{R(\mathbb{K})}{nh}.$$

Le paramètre de lissage basé sur la méthode de validation croisée biaisée est la valeur de h qui minimise l'estimateur $AMISE$. $AMISE$ peut être estimé si $R(f'')$ est estimé. L'estimateur naturel pour ce terme est donné par $R(f_n'')$ ou \hat{f}_n l'estimateur de la densité qui utilise la méthode du noyau.

Finalement, Scott et Terrell ont proposé la forme de l'estimateur d' $AMISE$ à minimiser (critère de $BCV(h)$), qui se résume comme suit :

Proposition 2.2.2 (*Scott et Trell*)

Soit (X_1, X_2, \dots, X_n) un n -échantillon iid issu d'une variable aléatoire X de fonction de densité f . Pour un noyau \mathbb{K} , on obtient :

$$BCV(h) = \frac{R(\mathbb{K})}{nh} + h^4 \frac{\mu_2^2(\mathbb{K})}{4n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \mathbb{K}^{(2)} \mathbb{K}^{(2)}(X_i - X_j).$$

Le paramètre de lissage h_{BCV} qui minimise $BCV(h)$:

$$h_{BCV} = \arg \min_h BCV(h).$$

Les résultats de simulation pour l'approche de validation croisée biaisée ont été obtenus dans les travaux de [Park et Marron](#). Les auteurs ont constaté que la méthode Biased Cross-Validation (BCV) présente les mêmes faiblesses que la méthode Unbiased Cross-Validation (UCV). Cette méthode fournit plusieurs minima locaux pour la fonction fonctionnelle objective à minimiser. Cependant, selon plusieurs simulations. Les auteurs recommandent de choisir une valeur inférieure parmi les minima locaux.

On va faire la simulation validation croisée non biaisée :

- Ce programme définit la quantité [2.17](#) donc on obtient la courbe suivante :

Code R utilisé :

```
set.seed(1)

n=100

X=rnorm(n)

J=function(h){

  fhat=Vectorize(function(x) density(X,from=x,to=x,n=1,bw=h)$y)

  fhati=Vectorize(function(i) density(X[-i],from=X[i],to=X[i],n=1,bw=h)$y)

  F=fhati(1 :length(X))

  return(integrate(function(x) fhat(x)^2,-Inf,Inf)$value-2*mean(F)) }

vx=seq(.1,1,by=.01)

vy=Vectorize(J)(vx)

df=data.frame(vx,vy)

library(ggplot2)

qplot(vx,vy,geom="line",data=df).
```

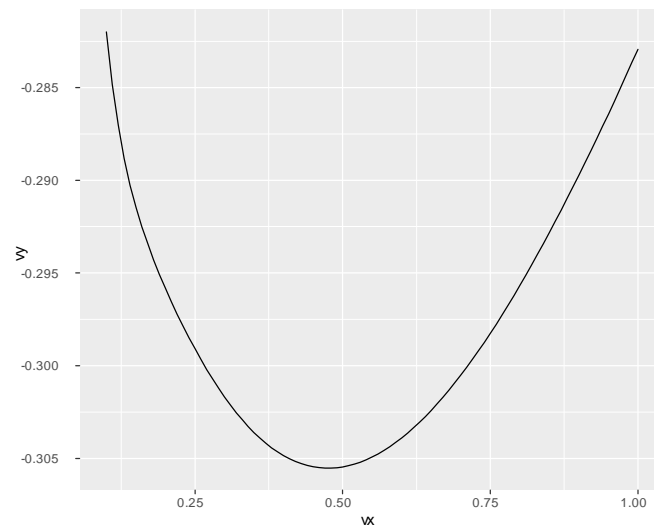


FIG. 2.7: La représentation graphique de UCV sur[0.1,1]

– on cherche à minimiser cette fonction par suite :

Code R utilisé :

```
m=optimize(J,interval=c(.1,1))

  hopt=m$minimum

k1=function(x) ifelse(abs(x)<1,(3/4)*(1-x^2),0)

x1=seq(min(X),max(X),length=n)

fn=rep(0,n)

for(i in 1 :n){fn[i]=(1/(n*hopt))*sum(k1((X[1 :n]-x1[i])/hopt))}

plot(density(X),col=2,lwd=2,xlab = "x",ylab = "f(x)")

plot(X,dnorm(X),col=2,lwd=2,xlab = "x",ylab = "f(x)")

lines(x1,fn,col=5,lwd=3)

legend("right", c("la densité de X f(x)", "l'estimateur de la density de
X fn(x)"), col = c(2,5), text.col = "green4", pch = c(1, 1), lty = c(1,1),
merge = TRUE, trace=TRUE)
```

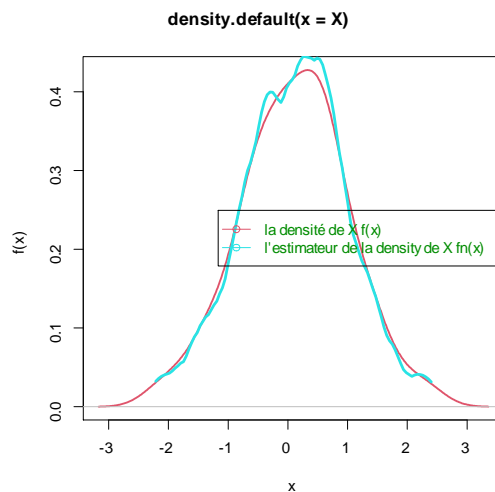


FIG. 2.8: l'estimateur de la densité par validation croisée non biaisée

2.2.9 Noyaux asymptotiquement optimales

Le noyau d'Epanechnikov est souvent appelé le "**noyau optimal**"(\mathbb{K}_{opt}). Pour comparer les noyaux (\mathbb{K}), on utilise le principe d'efficacité relative :

$$\begin{aligned} eff[\mathbb{K}] &= \left(\frac{AMISE[\mathbb{K}_{opt}]}{AMISE[\mathbb{K}]} \right) \\ &= \left(\frac{\mu_2^2(\mathbb{K}_{opt})R^4(\mathbb{K}_{opt})}{(\mu_2^2(\mathbb{K}))R^4(\mathbb{K})} \right)^{\frac{1}{5}} \leq 1. \end{aligned}$$

Les efficacités relatives des différents noyaux sont présentées dans le tableau suivants:

Noyau $\mathbb{K}(u)$	$eff(\mathbb{K})$
Noyau Epanechnikov	1
Noyau Quartique (biweight)	0.9939
Noyau Gaussien	0.9512
Noyau triangulaire	0.9859
Noyau Uniforme (Rosenblatt)	0.9295
Noyau Noyau unité	0.9424
Noyau Noyau de Picard	0.8051

TAB. 2.3: Efficacite des noyaux

Conclusion

En conclusion, la simulation des estimateurs de distribution non paramétrique offre une approche flexible et informative pour étudier les performances des méthodes d'estimation dans des situations où les distributions sont complexes, inconnues ou sujettes à des variations. Cela permet de guider le choix et l'utilisation des estimateurs dans une variété de domaines, tels que la finance, l'épidémiologie, la biostatistique et bien d'autres.

Bibliographie

A. B. Tsybakov, (2003). Introduction à l'estimation non paramétrique (Vol. 41). Springer Science & Business Media.

A.W.Bowman, (1984) .An alternative method of cross validation for the smoothing density estimates .Biometrika,(71,553-560).

B. W. Silverman, (1986) Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability, London. Chapman and Hall.

B.U.Park and S.J.Marron, (1990) .Comparaison of data-driven bandwidth selectors.,(85,66- 72).

C.Stone. (1984). An asymptotically optimal window selection rule for kernel density estimates.The Annals of Statistics,(12,P1285-1297).

D.W.Scott and G.R.Terrell, (1987). Biased and unbiased cross-validation in density estimation.Journal of the American Statistical Association,(82,P1131-1146).(7)

E. A. Epanechnikov, (1969). Nonparametric estimation of a multidimensional probability density. Theory of Probability and its Applications, 14 :153–158,

E.Parzen, (1979). Nonparemtric statistical data modeling. J. Amer. Statist. Assoc., 74, 105-131.

E. Parzen. On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33 :1065–1076, 1962.

- E. Fix, and J.L. Hodges, (1951). Discriminatory analysis, nonparametric estimation : consistency properties. Report No. 4, Project no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.
- F.Comte, (2017). Estimation non-paramétrique. Cours de master. Deuxième édition.
- G.Saporta, (2006). Probabilités, analyse des données et statistique. Editions Technip.
- G. Turini, (2019/20). Introduction à la statistique non paramétrique. M1 math university paris Dauphine-psl.
- J. S.Simonoff, (2012). Smoothing methods in statistics. Springer Science & Business Media,
- K. Newey, Whitney convergence rates and asymptotic normality for series estimators. *J.Econometrics* 79(1997), no.1, 147- -168.
- L. Devroye, and L. Györfi, (1985). Nonparametric Density Estimation : The L_1 View. New York : Wiley.
- M.Lejeune, (2004). Statistique. La théorie et ses applications. Springer, France.
- M. Rosenblatt, (1956) . Remarks in some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27 :832–837.
- M.P Wand, M .C. Jones, (1995). Kernel Smoothing. Monographs on Statistics and Applied Probability, London. Chapman and Hall.
- M.Rudemo, (1982). Empirical choice of histogram and kernel density estimators .*Scandinavian Journal of Statistics* ,(9,65-78).(1)
- P. Burman, (1985) .A date dependent approach to density estimation .*Zeitschrift Fr Wahr-scheinlichkeitstheorie and Verwandte Gebiete*,(69,600-628).(6)

P. Hall and J.S.Marron, (1991) .Local minima in cross-validation function .Journal of the royal statistical society,(90,149-173).(4)

P. Hall. (1982), Cross-validation in density estimation .Biometrika,(69,383-390).(5)

S. Loustau, (2010). Estimation non-paramétrique et Apprentissage statistique. Université de La Rochelle.

V.A. Epanechnikov, (1969). Nonparametric estimation of a multidimensional probability density. Theor. Probab. Appl., 14, 153-158.

ملخص

هذا البحث يقدم دراسة مفصلة لتقديرات التوزيع غير القابلة للتعريف بمعاملات، من خلال تحليل أدائها ومزاياها وعيوبها. هذه النتائج ذات أهمية بالغة للباحثين والممارسين الذين يرغبون في اختيار الطريقة الأكثر ملاءمة لتقدير البيانات الخاصة بهم وأهداف بحثهم

Abstract

This thesis provides an in-depth study of non-parametric distribution estimators, analyzing their performance, advantages, and disadvantages. These findings are important for researchers and practitioners who wish to choose the most appropriate estimation method based on the characteristics of their data and research objectives.

Résumé

Ce mémoire fournit une étude approfondie des estimateurs de distribution non paramétrique, en analysant leurs performances, leurs avantages et leurs inconvénients. Ces résultats sont importants pour les chercheurs et les praticiens qui souhaitent choisir la méthode d'estimation la plus appropriée en fonction des caractéristiques de leurs données et de leurs objectifs de recherche.