République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

MESSAOUDI Houssameddine

Titre :

Mode : estimations et application

Membres du Comité d'Examen :

Pr.	**BRAHIMI Brahim**	U. Biskra	Président
Pr.	**YAHIA Djabrane**	U. Biskra	Encadreur
Dr.	**ROUBI Afaf**	U. Biskra	Examinatrice

Septembre 2020

Dédicace

Je dédie ce modeste travail :

A mes parents, mes sentiments pour eux sont immenses,

je vous remercie pour tout ce que vous avez fait pour moi.

Que dieu vous préserve une longue vie heureuse.

Avec toute ma tendresse.

Pour vous, mes chères frères et mes chères sœurs.

Aux personnes que je n'oublierais jamais mes amies :

Brahim, Morad, Boubeker ,Khaled et Houssam,

pour leurs soutien moral et présence dans les bons et les mauvais moments.

A tous les membres de ma promotion.

A tous mes professeurs.

Houssameddine

REMERCIEMENTS

Je remercie DIEU, de m'avoir accordé santé, courage et patience afin d'accomplir ce travail.

Je tiens à exprime ma profonde reconnaissance à mon directeur de thèse Monsieur YAHIA Djabrane, Professeur à l'Université de Biskra, pour m'avoir donné l'opportunité de mener ce travail de recherche, l'avoir dirigé, sa disponibilité, son écoute, ses conseils et ses encouragements.

Je veux exprime aussi tout mon respect aux membres du jury, Professeur Brahimi B. en qualité de Président et Mme Roubi A. examinatrice, qui ont acceptés d'évaluer et de juger mon travail.

Enfin, je remercie ma famille, mes proches, mes amis et mes collègues se spécialisant en statistique, analyse et probabilités.

Table des matières

Table des figures

Liste des tableaux

Introduction

Dans ce mémoire, nous intéressons à la l'estimation et les différentes propriétés du mode comme l'une des mesures de tendance centrales. Le mode d'une distribution de probabilité continue est la valeur à laquelle sa fonction de densité de probabilité a sa valeur maximale. C'est la valeur qui apparaît le plus souvent dans un ensemble de données.

Parzen (1962) a été l'un des premiers à s'intéresser au problème de l'estimation du mode. L'estimation du mode par noyau proposé par Parzen dépend d'un seul paramètre de lissage. Sous certains conditions, il à montrer la cohérence, la normalité asymptotique et l'erreur quadratique moyenne du mode simple du noyau.

Pour les estimations non paramétriques du mode d'une fonction de densité, deux estimations de mode sont définies à partir d'une estimation de densité de noyau globale (resp. d'une estimation locale), tandis que les deux autres sont définies à partir d'une estimation de noyau global (ou à partir d'une estimation de noyau locale) de la dérivée de la fonction de densité, que chaque estimations de mode atteint le même taux de convergence que le mode simple habituel et que l'estimation de mode la plus efficace est celle basée sur l'estimation de densité de dérivation locale.

La méthode du noyau a été largement étudiée au cours des quatre dernières décennies. Il existe de nombreux livres et articles qui traitent de cette méthode et de ses applications dans différents domaines. La méthode du noyau est également utilisée pour étudier certains aspects de la fonction de densité comme la moyenne, les quantiles et le mode.

Ce mémoire se compose de deux chapitres. Dans le premier chapitre nous proposons d'étudier certains définitions, concepts de base, faits et concept d'estimation paramétrique de mode (mode d'une série statistique et mode de certaines variables aléatoires de loi continue et discrète).

Dans le deuxième chapitre, nous proposons d'étudier certaines propriétés asymptotiques de l'estimateur non paramétrique du mode. Nous construisons l'estimation du noyau de densité de probabilité. L'estimateur du mode de Parzen est étudié, la convergence et la normalité de ces estimateurs est donnée aussi. Finalement, ce chapitre se termine par une application en utilisant des études simulées sous le logiciel R.

Chapitre 1

Généralités

La statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques. L'analyse des données est utilisée pour d'écrire les phénomènes étudiés, faire des prévisions et prendre des décisions à leur sujet. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes. Les statistiques descriptives visent à étudier les caractéristiques d'un ensemble d'observations comme les mesures obtenues lors d'une expérience.

1.1 Population

En statistique, on travaille sur des populations. Ce terme vient du fait que la démographie, étude des populations humaines, a occupé une place centrale aux débuts de la statistique, notamment au travers des recensements de population. Mais, en statistique, le terme de population s'applique à tout objet statistique étudié, qu'il s'agisse d'étudiants (d'une université ou d'un pays), de ménages ou de n'importe quel autre ensemble sur lequel on fait des observations statistiques. Nous définissons la notion de population.

Définition 1.1.1 *On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté .*

1.1.1 Caractère : variable statistique

La statistique «descriptive», comme son nom l'indique cherche à décrire une population donnée. Nous nous intéressons au caractéristique des unités qui peuvent prendre différentes valeurs.

Définition 1.1.2 *On appelle caractère (ou variable statistique, notée V.S) toute application :*

$$X : \Omega \to C$$

L'ensemble C est dit : ensemble des valeurs du caractère X (c'est ce qui est mesuré ou observé sur les individus).

Exemple 1.1.1 *Taille, température, nationalité, couleur des yeux, catégorie socioprofessionnelle ... etc.*

1.2 Variable statistique discrète

On rappelle que une variable statistique est dite discrète lorsqu'elle ne peut prendre que des valeurs isolées dans son intervalle de variation.

Définition 1.2.1 ***(Effectif)***. *L'effectif d'une modalité x_i d'un caractère x est le nombre d'individus présentant cette modalité. L'effectif correspondant à la i^{eme} modalité du caractère x est noté n_i. L'effectif total est le nombre d'individus appartenant à la population statistique étudiée. L'effectif total sera noté N :*

$$\sum_{i=1}^{n} n_i = N$$

Définition 1.2.2 ***(Effectif cumulé croissant).*** *Les modalités d'un caractère variant de $_1$ à $_k$, l'effectif cumulé croissant d'une modalité i est le nombre d'individus de la population présentant une modalité d'indice inférieur ou égal à $_i$:*

$$N_i = n_1 + n_2 + ... + n_i = \sum_{j=1}^{i} n_j$$

Définition 1.2.3 ***(Fréquence relative).*** *La fréquence relative d'une modalité est la proportion d'individus de la population totale qui présentent cette modalité : elle est obtenue en divisant l'effectif de cette modalité du caractère par l'effectif total et notée f_i soit*

$$f_i = \frac{n_i}{N}$$

Definition 4 (Fréquence relative cumulée croissante). La fréquence relative cumulée croissante de la valeur x_i de la distribution statistique X comme suit :

$$F_i = f_1 + f_2 + ... + f_i = \sum_{j=1}^{i} f_j$$

Cette somme représente la proportion d'individus dans la population pour lesquels X prend une valeur inférieure ou égale à x_i.

1.2.1 Tableau statistique

Un tableau statistique est juste une liste de chifres relative au caractère de la population que l'on souhaite étudier, présentée de façon la plus compréhensible possible. Les données peuvent être présentées individuellement, sous la forme suivante :

Valeur observiées x_i	Effectif n_i	Effectif cumulé N_i	Fréquence relative f_i	Fréquence relative cumulée F_i
x_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$
x_2	n_2	$N_2 = n_1 + n_2$	f_2	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	N	f_k	$F_k = 1$
	N		1	

TAB. 1.1 – Tableau statistique d'un caractère quantitatif discret

Exemple 1.2.1 *Considérons l'exemple d'un groupe de 40 étudiants. On calcule les éffectifs cumulés et les fréquences relatives cumulées correspondant à l'âge des étdiants dans le tableau suivant :*

Age	Effectif n_i	Effectif cumulé N_i	Fréquence relative f_i	Fréquence relative cumulée F_i
21	7	7	0.175	0.175
22	10	17	0.25	0.425
23	20	37	0.5	0.925
24	3	40	0.075	1
Total	40		1	

TAB. 1.2 – Effectifs et fréquences relatives cumulées correspondant à l'âge des étdiants

1.2.2 Paramètres de position d'une variable quantitative

Ils visent à résumer la zone des réels où se trouvent les observations faites sur l'échantillon.

– **Moyenne :** Soit $\overline{x}$ la moyenne des n données observées $\{x_i =, i = 1, .., n\}$:

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

– **Médiane :** Soit X une variable quantitative observée sur n individus d'un échantillon :

★ Ranger les valeurs mesurées par ordre croissant : on obtient $\{x_i, i = 1,..., n\}$.

★ La médiane correspond à l'observation placée au milieu

$$Médiane = \begin{cases} x_{(n+1)/2} & \text{si } n \text{ impair} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2} & \text{si } n \text{ pair} \end{cases}.$$

Définition 1.2.4 (Mode) *On appelle mode ou valeur dominante d'une série statistique la valeur observée de la variable ayant le plus grand effectif (ou la fréquence la plus élevée). On note généralement le mode par Mo .*

Exemple 1.2.2 *On considère les notes obtenues en examen par un groupe de 20 étudiants : 11, 15, 19, 15, 14, 3, 19, 11, 15 ,16, 15, 3, 15, 11, 15, 12, 7, 7, 11, 15.*

La représentation en diagramme des notes ainsi que la détermination graphique du mode est présentée dans la figure suivante :

Le mode de cette série correspond à la note la plus fréquente, soit $Mo = 15$, valeur qui apparaît 6 fois. L'interprétation en est que la note la plus fréquente est 15.

La note obtenues	Effectif n_i	Fréquence relative f_i
3	1	0.05
7	3	0.15
11	4	0.2
12	1	0.05
14	2	0.1
15	6	0.3
16	1	0.05
19	2	0.1
Total	20	1

TAB. 1.3 – Notes obtenues en examen par un groupe d'étudiants

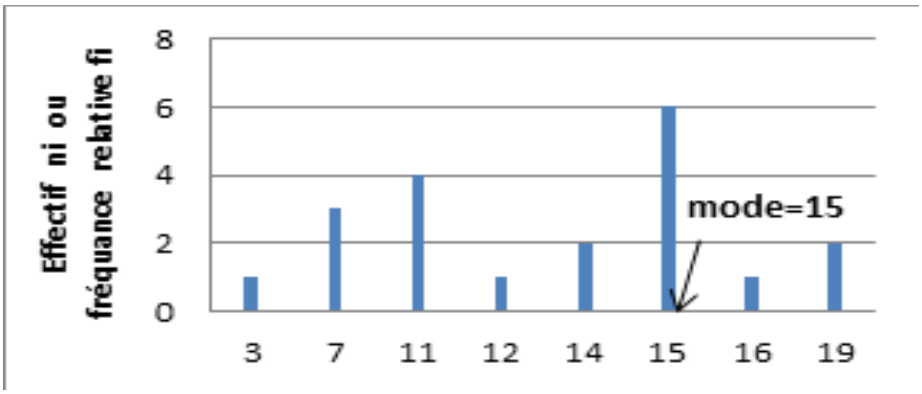

FIG. 1.1 – Détermination graphique du mode pour une variable statistique discréte

– **Variance :** La variance, notée Var, où :

$$Var = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

La racine carrée de Var est l'écart-type .

1.3 Variable statistique continue

Lorsque le caractère quantitatif discret comprend un grand nombre de valeurs, nous préférons regrouper les valeurs en intervalles appelées classes pour rendre la statistique plus lisible. Nous partageons alors l'ensemble des valeurs du caractère en classes $[e_{i-1}, e_i[$ avec $e_{i-1} < e_i$.

Remarque 1.3.1 *Dans le cas d'une variable continue en classes, ce critère est peu objectif. On*

parlera plutôt de classe modale : classe ayant la fréquence la plus élevée. Le mode n'est pas unique.

On choisit les classes pas trop nombreuses, mais suffisamment pour qu'il n'y ait pas de perte d'information. On peut fixer le nombre de classes selon l'un des deux formules suivantes :

Règle de Sturge : nb.de classe$= 1 + (3.3logN)$

Règle de Yule : nb.de classe$= 2.5\sqrt[4]{N}$.

L'amplitude de classe est alors donnée par :

$$\frac{\text{valeur } max \ - \ \text{valeur } min}{\text{nb.de classes}}.$$

Une classe modéle est définie par ses extrémités e_{i-1}, e_i et son effectif n_i . Chaque classe est caractérisé par son centre et son amplitude :

Le centre de la classe $[e_{i-1}, e_i[$ noté C_i se définit de manière évidente par la valeur :

$$C_i \ = \frac{e_{i-1} + e_i}{2}.$$

La différence entre les deux extrémités est appelé amplitude de la classe. L'amplitude d'une classe i est :

$$a_i = e_i - e_{i-1}.$$

1.3.1 Tableau statistique

Classes $[e_{i-1}, e_i[$	Centres c_i	Amplitude a_i	Effectifs n_i	Effectifs cumulés N_i	Fréquences f_i	Fréquence cumulée F_i
$[e_0, e_1[$	c_1	a_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$
$[e_1, e_2[$	c_2	a_2	n_2	$N_2 = n_1 + n_2$	f_2	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[e_{K-1}, e_K[$	c_K	a_k	n_k	N_k	f_k	$F_k = 1$
Total				N		

TAB. 1.4 – Tableau statistique d'un caractère quantitatif continue

Exemple 1.3.1 *Les 60 observations relatives à la variable « poids (en kg) » obtenues sur la population des étudiants sont comme suit :*

68 ,89 ,67 ,75 ,72 ,71 ,67 ,65 ,101 ,60 ,65 ,65 ,87 ,95 ,85 ,70 ,70 ,72 ,66 ,75 ,90 ,65 ,62 ,70 ,49 ,60 ,59 ,65 ,68 ,71 ,97 ,65 ,57 ,75 ,87 ,75 ,85 ,56 ,77 ,47 ,62 ,52 ,67 ,72 ,89 ,60 ,72 ,69 ,58 ,55 ,75 ,85 ,78 ,65 ,75 ,65 ,90 ,72 ,72 ,60. Ainsi, le tableau statistique corespondant est le suivant :

poids $[e_{i-1}, e_i[$	Centres c_i	Amplitude a_i	Effectifs n_i	Effectifs cumulés N_i	Fréquences f_i	Fréquence cumulée F_i
$[40, 50[$	45	10	2	2	0.033	0.033
$[50, 60[$	55	10	6	8	0.1	0.133
$[60, 70[$	65	10	21	29	0.35	0.483
$[70, 80[$	75	10	19	48	0.317	0.8
$[80, 90[$	85	10	7	55	0.117	0.917
$[90, 100[$	95	10	4	59	0.07	0.987
$[100, 110[$	105	10	1	60	0.017	1
Total			60		1	

TAB. 1.5 – Table statistique relatives à la variable ńpoidsż

1.4 Mode et classe modale

La classe modale est celle ayant le plus grand éffectif par unité d'amplitude. Dans le cas d'une classe modale unique, on parle de distribution continue unimodale. Graphiquement la classe modale est la base du rectangle ayant la hauteur la plus élevée.

Remarque 1.4.1 *Cependant, on distingue deux cas selon que les amplitudes des classes sont :*

1) Effectifs groupés par classes d'amplitudes égales

2) Effectifs groupés par classes d'amplitudes inégales

1.4.1 Effectifs groupés par classes d'amplitudes égales

Dans ce cas, la classe modale est la classe d'effectif ni le plus élevé, soit $[e_{i-1}, e_i[$. L'effectif de la classe qui précède la classe modale est n_{i-1} et celui de la classe qui suit la classe modale est n_{i+1} alors :

$$Mo = e_{i-1} + a_i \left(\frac{m_1}{m_1 + m_2} \right)$$

avec

$$m_1 = n_i - n_{i-1}$$

$$m_2 = n_i - n_{i+1}$$

e_{i-1} : Limite inférieure de la classe modale.

m_1 : La différence entre la classe modale est la classe avant.

m_2 : La différence entre la classe modale est la classe suivante.

On va appliquer dans l'exemple qui suive cette formule en utilisant les effectifs et les densités de classes.

Exemple 1.4.1 *Soit le tableau suivant où les données sont présentées par classes d'amplitudes égales :*

Classes	Centresc_i	Effectifsn_i
$[6, 12[$	9	2
$[12, 18[$	15	7
$[18, 24[$	21	12
$[24, 30[$	27	16
$[30, 36[$	33	14
$[36, 42[$	39	9
$[42, 48[$	45	5
$[48, 54[$	51	1
Total		

TAB. 1.6 – Données présentées par classes d'amplitudes égales

Ces données sint présentées graphiquement par histogramme, comme suit :

D'après les données du taleau et l'histogramme on a :

$$\begin{aligned} Mo &= e_{i-1} + a_i \left(\frac{m_1}{m_1 + m_2} \right) \\ &= 24 + 6 \left(\frac{16 - 12}{(16 - 12) + (16 - 14)} \right) = 28 \in [24, 30[. \end{aligned}$$

Donc, le mode $Mo = 28$ et la classe modale est $[24, 30[$.

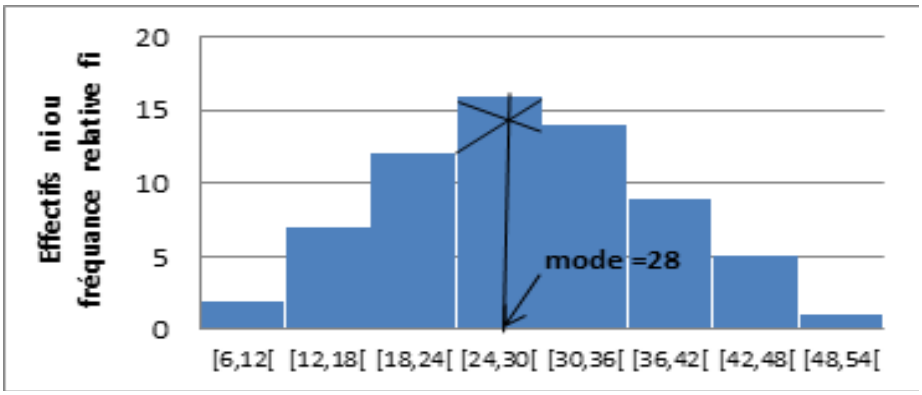

FIG. 1.2 – Détermination graphique du mode pour les classes sont d'égales amplitudes.

1.4.2 Effectifs groupés par classes d'amplitudes inégales

Dans ce cas, pour calculer le mode, m_1 et m_2 change, car il faut remplacer les effectifs n_i par les amplitudes corrigées :

$$m_1 = h_i - h_{i-1}$$

$$m_2 = h_i + h_{i+1}$$

$$s_i = n_i = a_i \times h_i$$

$$h_i = \frac{n_i}{a_i}$$

s_i : L'aire du rectangle correspond à la classe i .

n_i : L'effectif de la classe i .

a_i : Amplitude de la classe i .

h_i : La hauteur de la classe i .

Exemple 1.4.2 *Soit le tableau suivant où les données sont présentées par classes d'amplitudes inégales.*

L'histogramme des données est :

Classes	Effectifs n_i	Amplitude a_i	La hauteur h_i
$[0, 10[$	7	10	0.7
$[10, 18[$	12	8	1.5
$[18, 23[$	20	5	4
$[23, 35[$	24	12	2
$[35, 44[$	6	9	0.667
$[44, 50[$	10	6	1.667

TAB. 1.7 – Données présentées par classes d'amplitudes inégales

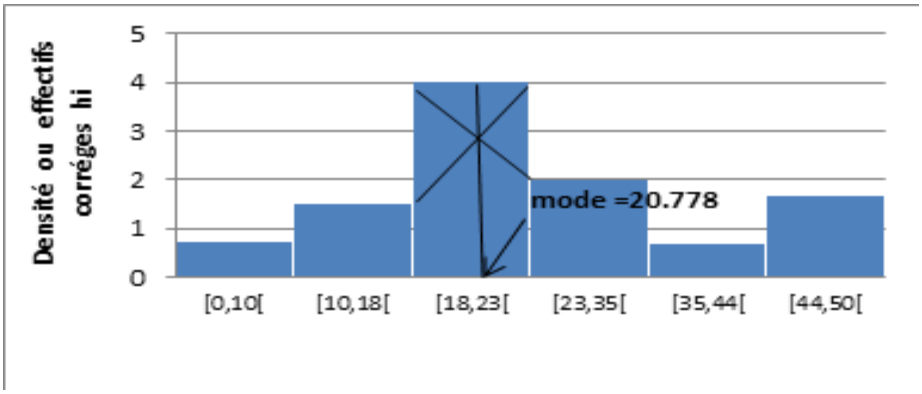

FIG. 1.3 – Détermination graphique du mode pour les classes sont d'inégales amplitudes.

D'après les données du taleau et l'histogramme on a :

$$m_1 = h_i - h_{i-1} = 4 - 1.5 = 2.5$$

$$m_2 = h_i - h_{i+1} = 4 - 2 = 2$$

Donc,

$$Mo = e_{i-1} + a_i \left(\frac{m_1}{m_1 + m_2} \right) = 18 + 5 \left(\frac{2.5}{2.5 + 2} \right) = 20.7778$$

avec

$$20.7778 \in [18, 23[.$$

1.5 Le mode d'une variable aléatoire de loi de probabilité

Définition 1.5.1 *Le mode d'une variable aléatoire X est la valeur la plus vraisemblable. C'est le maximum de la densité $f(x)$ pour les variable de loi de probabilité absolument continue.*

1.5.1 La fonction de mode

Soit $X_1, ..., X_n$ une suite de v.a. de même loi qu'une v.a. X ayant de fonction de répartition réelle F et densité continue f. On appelle mode simple de f tout point noté θ défini par :

$$f(\theta) = \sup_{t \in R} f(t) \tag{1.1}$$

θ est la solution de l'équation :

$$\begin{cases} f^{'}(t) = 0 \\ f^{''}(t) > 0 \end{cases} \tag{1.2}$$

Il s'agit d'une caractéristique de f qui joue un rôle important tant en probabilité qu'en statistique.

Exemple 1.5.1 (loi gamma) *Si X une va qui suit une loi gamma, sa densité de probabilité f est :*

$$f(x) = \frac{1}{\Gamma(p)} x^{p-1} e^{-x} 1_{\mathbb{R}_+^*}(x), \text{ où } p \geq 1, \Gamma(p) = \int_0^{+\infty} e^{-y} y^{p-1} dy,$$

pour $p = 2$, on à

$$f(x) = \frac{1}{\Gamma(2)} x e^{-x}.$$

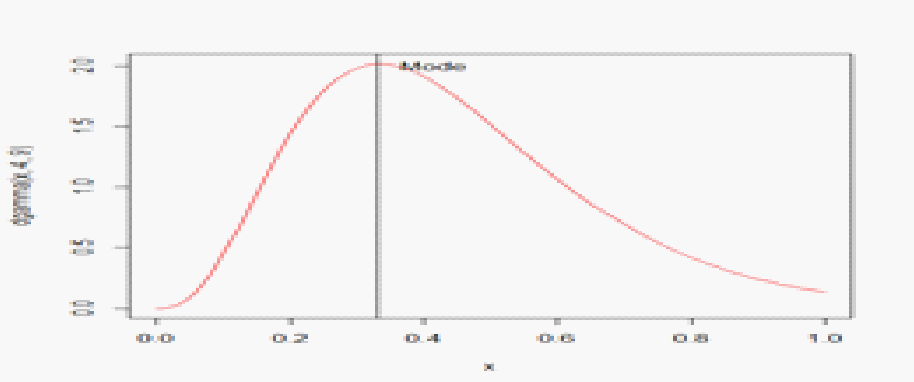

FIG. 1.4 – Détermination graphique du mode pour la loi gamma

D'apres l'équation (1.1) et θ est la solution de l'équation (1.2) on à :

$$\begin{aligned} f^{'}(x) &= \frac{1}{\Gamma(2)}\left[e^{-x} - te^{-x}\right] \\ &= \frac{1}{\Gamma(2)}\left[1-x\right]e^{-x} \\ f^{'}(x) = 0 &\Longrightarrow 1 - x = 0 \Longrightarrow x = 1 \\ &\Longrightarrow x = (p-1) = 1 \\ f^{''}(x) &= \frac{1}{\Gamma(2)}\left[-e^{-x} - (1-x)e^{-x}\right] \\ &= \frac{1}{\Gamma(2)}e^{-x}\left[x-2\right] \\ f^{''}(x) > 0 &\Longrightarrow x > 2 \end{aligned}$$

Le mode de $f(x)$ est défini par : $\theta = (p-1) = 1$.

Exemple 1.5.2 (loi exponentielle) *Soit $\lambda > 0$. On dit qu'une variable aléatoire X suit une loi exponentielle de paramètre λ si sa fonction de densité f est donnée par :*

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

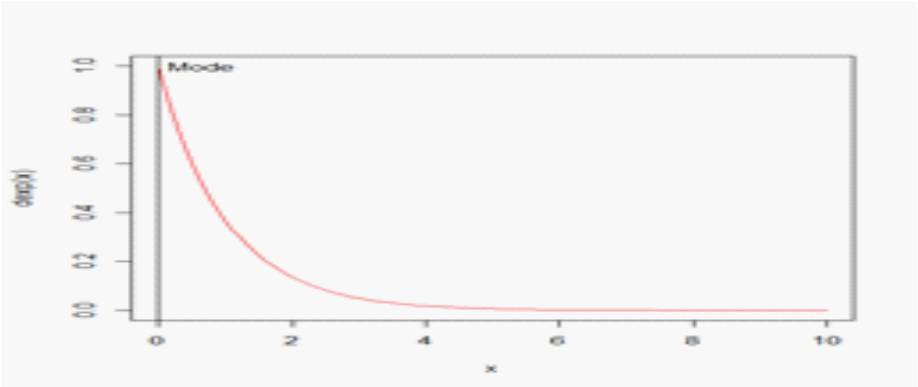

FIG. 1.5 – Détermination graphique du mode pour la loi exponentielle

Des équations (1.1) et (1.2), on obtient :

$$f^{'}(x) = -\lambda^2 e^{-\lambda x} \neq 0$$
$$f^{''}(x) = \lambda^3 e^{-\lambda x} > 0$$

Le mode de $f(x)$ est défini donc par : $\theta = 0$.

Exemple 1.5.3 (loi normale) *Soient $\mu \in \mathbb{R}$, $\sigma > 0$. On dit qu'une variable aléatoire X suit une loi de normale ou gaussienne si sa fonction de densité f est donnée par :*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} 1_{\mathbb{R}}(x)$$

Les équations (1.1) et (1.2) permettes d'avoir :

$$f^{'}(x) = C^{te}(x-\mu)e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$
$$f^{'}(x) = 0 \Longrightarrow x = \mu$$
$$f^{''}(x) = C^{te}((\frac{x-\mu}{\sigma})^2 - 1)e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$
$$f^{''}(x) > 0 \Longrightarrow \mu - \sigma < x < \mu + \sigma$$

Alors, le mode de $f(x)$ est défini par : $\theta = \mu$.

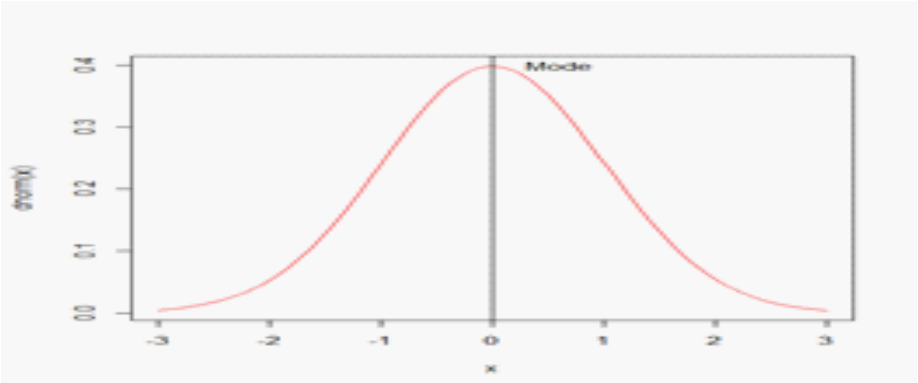

FIG. 1.6 – Détermination graphique du mode pour la loi normale

1.5.2 Mode d'une variable aléatoire de loi discrète

Dans le tableau ci-dessous, nous présentons le mode de certaines variables aléatoires de loi discrètes. On suppose $n \in \mathbb{N}^*, p \in]0,1[$ et $\lambda \in \mathbb{R}_+^*$:

La loi	Probabilités $p(X=k)$	Le mode θ
Bernoulli $\beta(p)$	$\begin{cases} p(X=0)=1-p \\ p(X=1)=p \end{cases}$	$\begin{cases} 0 & \text{si } p<q \\ 0.1 & \text{si } p=q \\ 1 & \text{si } p>q \end{cases}$
Binomiale $\beta(n,p)$	$p(X=k)=C_k^n\ p^k(1-p)^{n-k}1_{\{0,..,n\}}(k)$	$(n+1)p$
Poisson$\rho(\lambda)$	$p(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}1_{\mathbb{N}}(k)$	$\begin{cases} \lambda & \text{si } \lambda \in \mathbb{R} \\ \lambda \text{ et } \lambda-1 & \text{si } \lambda \in \mathbb{N} \end{cases}$
Géométrique $g(p)$	$p(X=k)=p(1-p)^{k-1}1_{\mathbb{N}^*}(k)$	1

1.5.3 Mode d'une variable aléatoire de loi continue

Supposons que $[a,b] \in \mathbb{R}$, $m \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$, $\lambda \in \mathbb{R}_+^*$, $\alpha \in \mathbb{R}_+^*$, $b \in \mathbb{R}_+^*$, $n \in \mathbb{R}^*$. Dans le tableau suivant, nous présentons le mode de certaines variables aléatoires de loi continues :

La loi	La densité	Le mode θ
Uniforme $U[a,b]$	$f_X(x)=\frac{1}{b-a}1_{[a,b]}(x)$	toute valeur dans $[a,b]$
Normale $N(m,\sigma^2)$	$f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}1_{\mathbb{R}}(x)$	m
Exononentielle $\zeta(\lambda)$	$f_X(x)=\lambda e^{-\lambda x}1_{\mathbb{R}_+(x)}$	0
Gamma $G(\alpha)$	$f_X(x)=\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x}1_{\mathbb{R}_+^*}(x)$	$\alpha-1$
Cauchy	$f_X(x)=\frac{1}{\pi(1+x^2)}1_{\mathbb{R}}(x)$	0
Khi-deux $\varkappa_n^2$	$f_X(x)=\frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}1_{\mathbb{R}_+}(x)$	$n-2$ si $n \geq 2$
Béta$\beta(\alpha,b)$	$f_X(x)=\frac{x^{\alpha-1}(1-x)^{b-1}}{B(\alpha,b)}1_{[0,1]}(x)$	$\frac{\alpha-1}{\alpha+b-2}$ pour $\alpha>1, b>1$
Laplace	$f_X(x)=\frac{1}{2}e^{-\lvert x\rvert}1_{\mathbb{R}}(x)$	0
Gumbel	$e^{-e^{-x}}1_{\mathbb{R}}(x)$	0

TAB. 1.8 – Mode d'une variable aléatoire de loi continue

Chapitre 2

Estimation non paramétrique du mode

Estimer le mode est souvent une conséquence directe de l'estimation de la densité. Son importance est dûe au fait que c'est une mesure naturelle de tendance centrale, qui n'est pas influencée par les queues des distributions. Le mode est la valeur la plus probable : pour une densité de probabilité f, c'est la valeur pour laquelle f admet un maximum (global ou local). Dans la première partie de ce chapitre en va étudier la densité à noyau, puis en passera dans la deuxième partie à l'estimation du mode.

2.1 Estimation de la densité par noyau

Soit X une variable aléatoire de densité de probabilité inconnue f. Supposons que nous avons n observations $x_1, x_2, ..., x_n$ provenant de X. Le problème consiste à trouver un estimateur pour la fonction f à partir de cet échantillon issu de X. Pour cela, l'approche non paramétrique est la plus adéquate lorsqu'on ne possède aucune information précise sur la forme et la classe de la vraie densité. Dans cette approche, ce sont les observations qui vont nous permettre de déterminer un estimateur pour la densité f. Dans cette section, on s'intéresse à la méthode du noyau pour l'estimation de la densité de probabilité. L'estimateur à noyau sera présenté ainsi que ses différentes propriétés statistiques.

En 1956, Rosenblatt [10] a proposé le premier estimateur à noyau pour la densité de probabilité

$f(x)$. Six ans après, cet estimateur a été généralisé par Parzen(1962) ; à partir de cette date, cet estimateur a pris le nom de l'estimateur de Parzen-Rosenblatt. L'idée de l'estimateur par la méthode du noyau consiste à évaluer la densité $f(x)$ au point x en comptant le nombre d'observations tombées dans un certain voisinage de x sur $\mathbb{R}$.

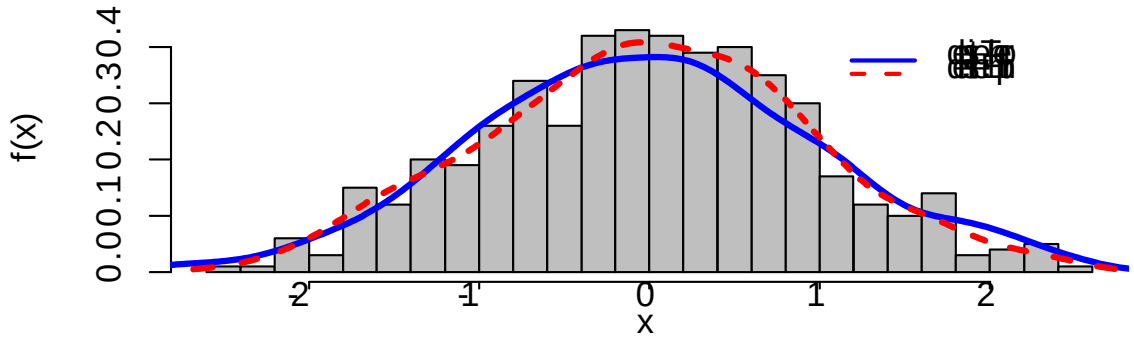

FIG. 2.1 – Estimation de la densité normale par noyau et histogramme

Définition 2.1.1 *Soient $x_1, ..., x_n$ n observations d'une variable aléatoire X de densité de probabilité $f(x)$ et de fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)dt$. On appelle fonction de répartition empirique associé à $x_1, ..., x_n$, la fonction aléatoire $F_n : \mathbb{R} \to [0, 1]$ définie, pour tout $x \in \mathbb{R}$, par :*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i < x\}} \tag{2.1}$$

A partir de la définition d'une densité de probabilité et en utilisant l'équation (2.1), on aura :

$$\hat{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} \text{ avec } h \to 0$$

Cette derniére peut étre récrite, en ses points de continuité, sous la forme suivante :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) \tag{2.2}$$

avec $h = h(n)$ est le paramètre de lissage ("bandwidth") choisi en fonction de n telle que

$$\lim_{n\to\infty} h(n) = 0, \tag{2.3}$$

et K est la fonction des poids ("noyau, kernel") où :

$$k(t) = \frac{1}{2} 1_{(|t|<1)}$$

Ce dernier est l'estimateur à noyau uniforme dit de Rosenblatt (1956).

Les noyaux les plus utilisés dans l'estimation de la densité de probabilité sont donnés dans le tableau suivant :

Noyau	Fonction $k(t)$
Rectongulaire	$\frac{1}{2} 1_{(\|t\|<1)}$
Trangulaire	$(1-\|t\|) 1_{(\|t\|<1)}$
Gaussien	$\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} 1_{\mathbb{R}}(t)$
Quartique	$\frac{15}{16}(1-t^2)^2 1_{(\|t\|<1)}$
Epanechinkov	$\frac{3}{4}(1-t^2) 1_{(\|t\|<1)}$

TAB. 2.1 – Noyaux usuels

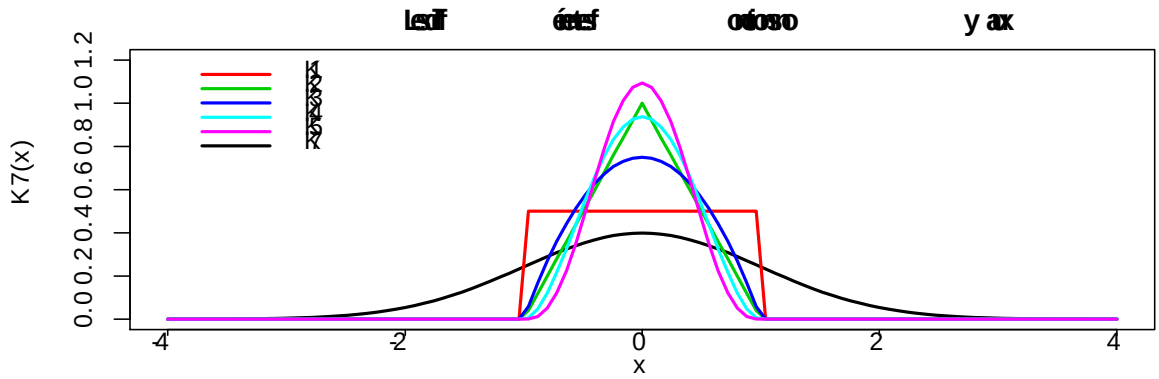

FIG. 2.2 – Courbes des noyaux usuels

2.1.1 Propriétés de l'estimateur à noyau de la densité

Proposition 2.1.1 *L'estimateur à noyau est une fonction de densité. De plus, $\hat{f}$ a les même propriétés de continuité et de différentiabilité que K :*

Si K est continue, $\hat{f}$ sera une fonction continue.

Si K est différentiable, $\hat{f}$ sera une fonction différentiable.

Sous les conditions suivantes sur f, h et K :

1. La dérivé seconde $f^{"}(x)$ est continue, de carré intégrable et monotone sur $[-\infty, M]$ et $[M; +\infty]$ pour $M > 0$.

2. $\lim\limits_{h\to 0} h = 0$ et $\lim\limits_{n\to 0} nh = 0$.

3. Pour que $\hat{f}(x)$ soit une densité, on suppose que $K(t) > 0$ et $\int_{\mathbb{R}} k(t)dt = 1$. La fonction noyau est supposée être symétrique autour de zéro, c.à.d $\int_{\mathbb{R}} tK(t)dt = 0$ et posséde un moment d'ordre 2 fini, c.à.d $\int_{\mathbb{R}} t^2K(t)dt < \infty$.

Par développement de Taylor de la densité f, supposée deux fois dérivable, on trouve les expressions de l'espérance, biais et de la variance de l'estimateur à noyau sont :

$$\begin{aligned} E(\hat{f}_n(x)) &= f(x) + \frac{h^2}{2} f^{"}(x)\mu_2(k) + o(h^2) \\ Biais(\hat{f}_n(x)) &= E(\hat{f}_n(x)) - f(x) = \frac{h^2}{2} f^{"}(x)\mu_2(k) + o(h^2) \\ Var(\hat{f}_n(x)) &= \frac{f(x)}{nh}\int_{\mathbb{R}} k^2(t)dt - \frac{f^{'}(x)}{n}\int_{\mathbb{R}} tk^2(t)dt - \frac{1}{n}(f(x) - Biais(\hat{f}_n(x))^2 \\ &= \frac{f(x)}{n} R(k) + o(\frac{1}{nh}) \end{aligned}$$

où $0 < \mu_2 = \int_{\mathbb{R}} t^2k(t)dt$ et $R(g(t)) = \int_{\mathbb{R}} g^2(t)dt$ pour une fonction g de carré intégrable.

Remarque 2.1.1 *L'estimateur $\hat{f}_n(x)$ est asymptotiquement sans biais, i.e,*

$$\begin{aligned} \lim_{n\to\infty} E(\hat{f}_n(x)) &= f(x) \\ \lim_{n\to\infty} nhVar(\hat{f}_n(x)) &= f(x)\int_{\mathbb{R}} k^2(t)dt \end{aligned}$$

telle que, f est une densité continue $\forall x \in \mathbb{R}$.

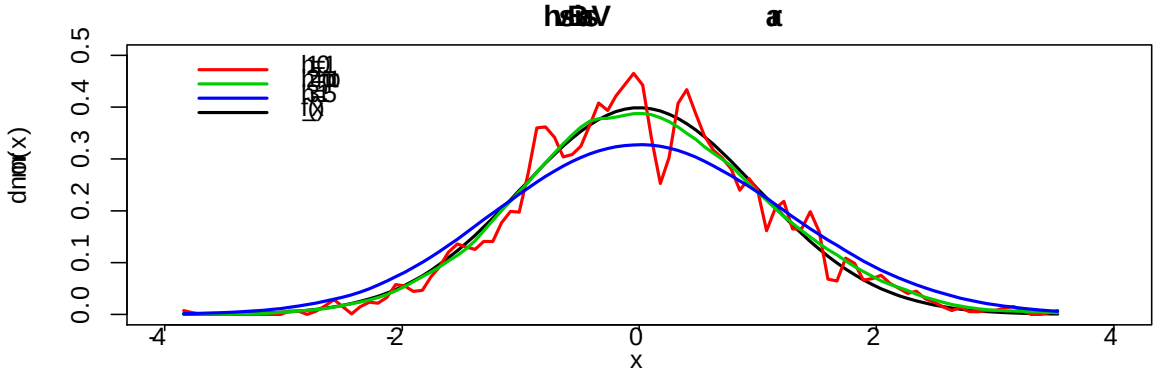

FIG. 2.3 – Biais, Var et choix de h

2.1.2 Erreure quadratique moyenne (MSE) et integré (MISE)

Ces expressions ont été obtenues sous la condition (3) sur K et en supposant que la densité de probabilité f avait toutes les dérivéés (continues) nécessaires. On peut obtenir facilement les approximations asymptotiques suivantes pour la MSE et la $MISE$:

$$\begin{aligned} MSE(\hat{f}_n(x)) &= E[(f(x) - \hat{f}_n(x))^2] \\ &= [Biais(\hat{f}_n(x))]^2 + Var(\hat{f}_n(x)) \\ &= \frac{h^4}{4}(f"(x))^2\mu_2^2(k) + \frac{1}{nh}f(x)\int_{\mathbb{R}} k^2(t)dt + o(\frac{1}{nh}) + o(h^2) \end{aligned}$$

$$\begin{aligned} MISE(\hat{f}_n(x)) &= \int_{\mathbb{R}} MSE(\hat{f}_n(x))dx \\ &= \frac{1}{nh}\int_{\mathbb{R}} k^2(t)dt + \frac{h^4}{4}\mu_2^2(k)\int_{\mathbb{R}}(f"(x))^2dx \end{aligned}$$

Sous les conditions $h \to 0$ et $nh \to \infty$ quand n tend vers l'infini, on a le développement asymptotique de $MISE$, que l'on note $AMISE$:

$$AMISE(h) = n^{-1}h^{-1}R(k) + h^4R(f")\left(\int_{\mathbb{R}} t^2k(t)dt\right)^2$$

Une autre caractéristique utile de l'$AMISE(h)$ est que sa minimisation est donnée par :

$$AMISE(\hat{f}_n(x)) = n^{-1}h^{-1}R(k) + \frac{h^4}{4}R(f")\mu_2^2(k) \tag{2.4}$$

L'estimateur de h qui minimise cette critère est donné par :

$$h_{AMISE} = \left(\frac{R(k)}{nR(f")\mu_2^2(k)}\right)^{\frac{1}{5}}. \tag{2.5}$$

Remarque 2.1.2 ***Cas particuliers :*** *Soit $X_1, X_2, ..., X_n$ une suite de variables aléatoires de densité de probabilité f, supposons que f appartient à une famille de distributions normales, $N(\mu;\sigma^2)$:*

– Si $X \sim N(\mu;\sigma^2)$, $k \sim N(0,1)$, alors $h_{opt} = 1.06\hat{\sigma}n^{-\frac{1}{5}}$.

– Si $f \sim N(\mu;\sigma^2)$, alors $f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$, avec $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$,

et $f''(x) = \frac{1}{\sigma^3}\varphi''\left(\frac{x-\mu}{\sigma}\right)$, $\varphi''(x) = \frac{1}{\sqrt{2\pi}}(x^2-1)e^{-\frac{x^2}{2}}$

La quantité inconnue $R(f")$ sécrit alors

$$\begin{aligned} R(f") &= \int_{-\infty}^{+\infty}\left[f''(x)\right]^2 dx \\ &= \frac{1}{\sigma^6}\int_{-\infty}^{+\infty}\left\{\varphi''\left(\frac{x-\mu}{\sigma}\right)\right\}^2 dx \\ &= \frac{1}{\sigma^5}\int_{-\infty}^{+\infty}\left\{\varphi''(\upsilon)\right\}^2 d\upsilon \end{aligned}$$

Nous avons :

$$\begin{aligned} \varphi(\upsilon) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{\upsilon^2}{2}} \\ &\Rightarrow \varphi'(\upsilon) = -\frac{\upsilon}{\sqrt{2\pi}}e^{-\frac{\upsilon^2}{2}} \\ &\Rightarrow \varphi''(\upsilon) = \frac{1}{\sqrt{2\pi}}(\upsilon^2-1)e^{-\frac{\upsilon^2}{2}} \end{aligned}$$

$$
\begin{aligned}
R(f") &= \frac{1}{\sigma^5}\int_{-\infty}^{+\infty}\left\{\frac{1}{\sqrt{2\pi}}(\upsilon^2-1)e^{-\frac{\upsilon^2}{2}}\right\}^2 d\upsilon \\
&= \frac{1}{\sigma^5}\frac{1}{\sqrt{2\pi}}\left\{\int_{-\infty}^{+\infty}\upsilon^4 e^{-\upsilon^2}d\upsilon - 2\int_{-\infty}^{+\infty}\upsilon^2 e^{-\upsilon^2}d\upsilon + \int_{-\infty}^{+\infty}e^{-\upsilon^2}d\upsilon\right\} \\
&= \frac{1}{\sigma^5}\frac{1}{\sqrt{2\pi}}\left\{-\frac{1}{2}\int_{-\infty}^{+\infty}\upsilon^2 e^{-\upsilon^2}d\upsilon + \int_{-\infty}^{+\infty}e^{-\upsilon^2}d\upsilon\right\} \\
&= \frac{1}{\sigma^5}\frac{1}{\sqrt{2\pi}}\left\{-\frac{1}{2}\int_{-\infty}^{+\infty}\frac{\mu^2}{2}e^{-\frac{\mu^2}{2}}\frac{1}{\sqrt{2}}d\mu + \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2}}e^{-\frac{\mu^2}{2}}d\mu\right\} \qquad \text{avec } \mu=\sqrt{2\upsilon} \\
&= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{4}\sqrt{\pi}+\sqrt{\pi}\right\} \\
&= \frac{1}{\sigma^5}\frac{3}{8\sqrt{\pi}}.
\end{aligned}
$$

Donc, l'expression du paramètre de lissage optimal devient

$$h_{opt} = \left[\frac{8\sqrt{\pi}R(k)}{3\left(\mu_2(k)\right)^2}\right]^{\frac{1}{5}}\hat{\sigma}n^{-\frac{1}{5}}$$

où $\hat{\sigma}$ est un estimateur de σ, tel que

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - X)^2}, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i.$$

On a $k \sim N(0,1)$ alors,

$$
\begin{aligned}
R(k) &= \int_{-\infty}^{+\infty}\left[k(\mu)\right]^2 d\mu \\
&= \int_{-\infty}^{+\infty}\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{\mu^2}{2}}\right]^2 d\mu \\
&= \int_{-\infty}^{+\infty}\left[\frac{1}{2\pi}e^{-\mu^2}\right]^2 d\mu \\
&= \frac{1}{2\pi}\sqrt{\pi} = \frac{1}{2\sqrt{\pi}}
\end{aligned}
$$

avec $\mu_2(k) = \int_{-\infty}^{+\infty}\mu^2 k(\mu)d\mu = 1$.

Nous remplaçons dans l'équation (2.5) nous obtenons :

$$h_{opt} = \left(\frac{4}{3}\right)^{\frac{1}{5}} \hat{\sigma} n^{-\frac{1}{5}} = 1.06 \hat{\sigma} n^{-\frac{1}{5}}$$

Exemple 2.1.1 *Soit $X_1, X_2, ..., X_n$ une suite de variables aléatoires de densité de probabilité f, supposons que f appartient à une famille de distributions normales $N(\mu; \sigma^2)$, soit K un noyau d'Epanechnikov, si $X \sim N(\mu; \sigma^2)$, $k(\mu) = \frac{3}{4}(1-\mu^2)1_{(|\mu|<1)}$, alors $h_{opt} = 2.34\hat{\sigma}n^{-\frac{1}{5}}$.*

2.1.3 Choix du Noyaux

Pour mesurer l'efficacité d'un noyau symétrique on peut calculer le rapport de $AMISE$ des 2 noyaux :

$$eff(k_1, k_2) = \frac{AMISE(k_1, n, h)}{AMISE(k_2, n, h)} < 1$$

Le choix du noyaux n'influe pas trop dans le cas du noyaux symétriques.

Noyau	Fonction $k(t)$	$eff(K)$
Ectongulaire	$\frac{1}{2}1_{(\|t\|<1)}$	0.930
Trangulaire	$(1-\|t\|)1_{(\|t\|<1)}$	0.986
Gaussien	$\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}1_{\mathbb{R}}(t)$	0.951
Quartique	$\frac{15}{16}(1-t^2)^2 1_{(\|t\|<1)}$	0.994
Epanechinkov	$\frac{3}{4}(1-t^2)1_{(\|t\|<1)}$	1.000

TAB. 2.2 – Tableau des efficacités relatives de plusieurs noyaux

2.2 Esimation à noyau du mode

Pour une distribution symétrique, le mode coincide avec deux autres paramètres de position, la moyenne et la médiane. On peut distinguer plusieurs estimateurs du mode : on en citera quelques uns qui sont définis selon deux approches :

2.2.1 Méthode indirecte

Cet approche consiste à obtenir dans un premier temps une estimation de la densité f (pour le cas non paramétrique), et à prendre pour mode la valeur de x pour laquelle $f(x)$ est maximale. Parzen (1962) a été l'un des premiers à s'intéresser au problème de l'estimation du mode θ dans le cas d'une densité univariée. Il définit un estimateur comme la variable aléatoire qui maximise l'estimateur à noyau $f_n$de la densité :

$$\hat{\theta}_n = \arg\max_{x\in\mathbb{R}} f_n(x)$$

Il démontre que cet estimateur est uniformément convergent (en probabilité), asymptotiquement normal et donne une évaluation de l'erreur quadratique moyenne (MSE).

D'autres auteurs se sont penchés sur le sujet, parmi lesquels on peut citer : Vieu (1996) qui propose l'étude de quatre estimateurs à noyau (globaux et locaux) du mode, basés sur des estimateurs à noyau de la densité et de sa dérivée.

-Deux estimateurs globaux, $\theta_{n,1}$ défini par Parzen, et $\theta_{n,1}$ obtenu en annulant $f_n^{'}(x)$, l'estimateur de la dérivée $f^{'}(x)$.

-Deux estimateurs locaux $\theta_{n,2}$ et $\theta_{n,4}$ basés respectivement, sur l'estimateur

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

et sa dérivée, qui s'appuient sur une fenêtre locale, et non plus globale.

Bickel (2003) qui introduit une autre approche indirecte. Il étudie deux estimateurs paramétriques, en transformant les observations $x_1, x_2, ..., x_n$ en données approximativement normales $y_i = (x_i)^{\alpha}$, $\alpha \in \mathbb{R}$. Ensuite, on annule la dérivée de la densité normale ainsi obtenue (les paramètres de la loi normale sont estimés par la moyenne $\bar{y}$ et l'écart type σ des données transformées), et obtient pour estimateur

$$M = \left[\frac{1}{2}\left(\bar{y} + \sqrt{\bar{y}^2 + \frac{4\sigma^2(\alpha - 1)}{\alpha}}\right)\right]^{\frac{1}{\alpha}}.$$

Pour $\alpha = 1$, $M = \bar{y}$, ce qui correspond au mode dans le cas d'une distribution symétrique comme la loi normale.

Le deuxième estimateur, plus robuste, est obtenu en remplaçant les paramètres $\bar{y}$ et σ respectivement par la médiane et l'écart médian absolu standardisé ($\triangle(y_i) = C.med\,|(y_i) - med(y_i)|$) égal à σ dans le cas normal.

2.2.2 Méthode directe

Considérant que dans l'échantillon, on doit observer un groupement de valeurs dans le voisinage du mode :

• Chernoff (1964) présente l'estimateur noté $\hat{t}_{a_n}$ comme le milieu d'un intervalle de longueur $2a_n$ contenant le maximum d'observations $t_1, t_2, ..., t_n$, $(a_n)_n$ étant une suite de réels positifs décroissant lentement vers 0. Cet estimateur est inspiré de l'estimateur naïf de noyau $k_a(t) = \frac{1}{2a}$ si $|t| \leq a$, dont le mode est le milieu de l'intervalle$[-a, a]$.

• Wegman (1971) montre la consistence forte de cet estimateur.

• Grenander (1965) définit les estimateurs :

$$M_{p,k} = \left[\frac{1}{2}\sum_{i=1}^{n-k}\frac{(t_{i+k}+t_i)}{(t_{i+k}-t_i)^p}\right]\left[\sum_{i=1}^{n-k}\frac{1}{(t_{i+k}-t_i)^p}\right]^{-1}, 1 < p < k,$$

pour un échantillon d'observations $(t_1, t_2, ..., t_n)$ ordonnées par ordre croissant, $t_1 \leq t_2 \leq .... \leq t_n$.

• Venter (1967) élabore un estimateur à partir d'une suite $(k_n)_n$ d'entiers naturels : c'est le milieu du plus petit intervalle contenant k_n observations parmi $t_1, t_2, ..., t_n$.

• Hall (1982) établit la normalité asymptotique de $M_{p,k}$ pour $k > 2p$.

Remarque 2.2.1 *Ces estimateurs ont suscité peu d'intérêt, car hautement influencés par les valeurs extrêmes.*

• Bickel (2002) propose deux estimateurs plus robustes appelés :

-HSM (Half-Sample Mode) obtenu à partir d'algorithmes basés sur des demi-échantillons successifs.

-HRM (Half-Range Mode), obtenu en cherchant un "intervalle modal" contenu dans d'autres intervalles modaux (le mode sera alors l'étendue du petit intervalle modal).

Pour Hedges & Shah (2003), HRM est obtenu par une "méthode simple et rapide et produit moins de biais dans les simulations". Ces deux derniers auteurs utilisent cet estimateur dans le cadre d'une étude (à l'Institut d'Astrobiologie - NASA) sur l'horloge moléculaire indiquant une distribution asymétrique, la moyenne étant plus élevée que la médiane dans la plupart des cas, justifiant l'usage du mode, dont les vraies valeurs sont inconnues. (l'horloge moléculaire est un phénomène décrivant l'empreinte laissée par le temps dans les molécules du vivant. Cette liste d'estimateurs fait partie des plus cités dans la littérature, mais n'est pas exhaustive.

2.3 Propriétés asymptotiques

2.3.1 Consistance

Dans cette section, nous détrminons les conditions dans lesquelles la fonction de densité de probabilité estimée $f_n(x)$ tend uniformement (en probabilité) à la vraie fonction de densité de probabilité. En utilsant ce fait, nous sommes en mesure d'obtenir des estimations cohérentes du mode.

Supposons que le noyau $k(u)$ est absulement intégrable. Du fait que $f_n(x)$ est continue et tend vers 0 quand x tend vers $\pm\infty$. Par conséquente, il existe une variable alèatoire θ_n telle que

$$f_n(\theta_n) = \max_{-\infty<x<+\infty} f_n(x)$$

On appelle θ_n le mode d'èchantillon.

Supposons ensuite que la vraie fonction de densité de probabolité $f(x)$ est uniformèment continue en x (c'est le cas si elle a une fonction caractéristique absolument intégrable). Il s'ensuit que $f(x)$ possède un mode θ définie par

$$f(\theta) = \max_{-\infty<x<+\infty} f(x).$$

Supposons finalement que θ est unique.

Théorème 2.3.1 *Si h est fonction de n statisfaisant $\lim_{n\to\infty} nh^2 = \infty$, et si la densité de probabilité $f(x)$ est uniformément continue, alors pour chaque $\epsilon > 0$*

$$p\left[\sup_{-\infty<x<+\infty} |f_n(x) - f(x)| < \epsilon\right] \to 1, \text{ pour } n \to \infty. \tag{2.6}$$

Si $\{\theta_n\}$ sont les modes d'échantillonnage, et si le mode de population θ est unique, alors pour $\epsilon > 0$

$$P\left[|\theta_n - \theta| < \epsilon\right] \to 1, \text{ pour } n \to \infty. \tag{2.7}$$

Preuve. Voir [9] . ■

2.3.2 Normalité asymptotique

Dans ce qui suit, nous indiquons les conditions sur les constantes $h(n)$ et le noyau $k(u)$ telle que le mode θ_n estimé soit asymptotique normale. Considèrons une fonction de densité de probabilité $f(x)$ avec un mode unique à θ. Si $f(x)$ a dérivée seconde continue, alors

$$f^{'}(\theta) = 0, \qquad f^{''}(\theta) \leq 0.$$

De même, si la fonction de densité estimée $f_n(x)$ est choisie pour être deux foix diffèrenciable (c'est-à-dire que la fonction noyau $k(t)$ est choisie pour être deux foix diffèrenciable aussi), alors

$$f_n^{'}(\theta_n) = 0, \qquad f_n^{''}(\theta_n) \leq 0.$$

Si θ_n est le mode de $f_n(x)$, par le divelopement de Taylor,

$$0 = f_n^{'}(\theta_n) = f_n^{'}(\theta) + (\theta_n - \theta) f_n^{''}(\theta_n^*)$$

pour une variable aléatoire θ_n^* entre θ_n et θ. Donc

$$\theta_n - \theta = -f_n^{'}(\theta)/f_n^{''}(\theta_n^*). \tag{2.8}$$

Si le dénominateur ne disparait pas, en utilisant (2.8) comme base, nous indiquons les conditions dans lesquelles le mode θ_n est asymptotiquement normal.

Théorème 2.3.2 *Suppose qu'il existe δ, $0 < \delta < 1$, tel que la transformée $k(u)$ a exposant caractéristique $r \geq 2$ et satisfait*

$$\int_{-\infty}^{+\infty} u^{2+\delta} \left|k(u)\right| du < \infty.$$

et que h est une fonction de n, telle que

$$\lim_{n\to\infty} nh^5 = \infty, \qquad \lim_{n\to\infty} nh^{5+2\delta} = 0,$$

et que la fonction caractéristique $\varphi(u)$ satisfait

$$\int_{-\infty}^{+\infty} u^{2+\delta} \left|\varphi(u)\right| du < \infty.$$

Puis, quand $n \to \infty$,

$$E\left[\sup_{-\infty<x<+\infty} \left|f_n^{''}(x) - f^{''}(x)\right|^2\right] \to 0.$$

$$f_n^{''}(\theta_n^*) \to f^{''}(\theta) \quad \text{en probabilité}$$

$$\sqrt{nh^3} f_n^{'}(\theta) \to N(0, f(\theta)J) \quad \text{en distibution}$$

$$\sqrt{nh^3}(\theta_n - \theta) \to N\left(0, \left\{f(\theta)/\left[f^{''}(\theta)\right]^2\right\} J\right) \quad \text{en distibution} \tag{2.9}$$

avec J définie par

$$J = \int_{-\infty}^{+\infty} k'^2(t)dt = (2\pi)^{-1} \int_{-\infty}^{+\infty} u^2 k^2(u)du.$$

Preuve. Pour la preuve du théorème, voir [9] . ■

2.4 Applications

Nous terminons ce deuxièmes chapitre par trois exemples de simulation avec logiciel R, sur l'estimation non paramétrique du mode.

Exemple 2.4.1 (cas gaussien) *Soit X une v.a de loi normale $N(1,2)$. La valeur théorique du mode est donc $\theta = 1$. Simulons un échantions de taille $n = 300$ de la va X. En utilisant l'estimation a noyau de la densité avec un noyau d'Epanechinkov :*

$$K(t) = \frac{3}{4}(1-t^2)1_{(|t|<1)}$$

et pour h, nous prenons le h optimle :

$$h_{opt} = 2.34\hat{\sigma}n^{-\frac{1}{5}}.$$

Nous obtenons sous le logiciel R, la valeur

$$\theta_n = 1.116562$$

comme estimateur non paramétrique du mode qui est la solution de l'equation

$$f_n(\theta_n) = \max_{-\infty<x<+\infty} f_n(x).$$

La figure suivante est une ulistration graphique de cet exemple.

Exemple 2.4.2 (Loi Gamma) *De même que l'exemple précident, soit X une va de loi gamma*

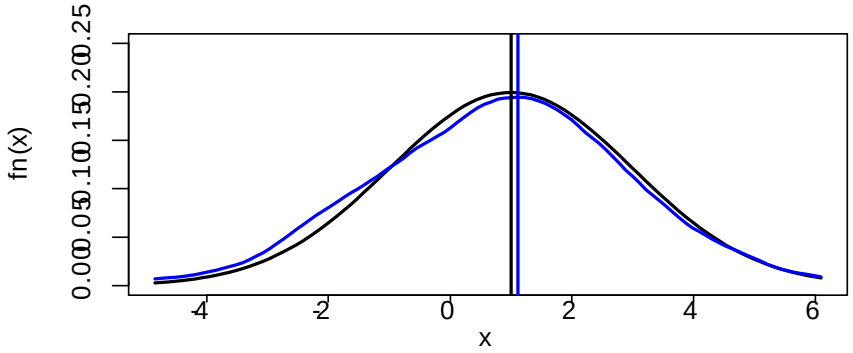

FIG. 2.4 – Estimation a noyau du mode : cas gaussien

de paramètres p *et* λ*, de densité :*

$$f(x) = \frac{1}{\Gamma(p)} x^{p-1} e^{-x/\lambda}, \qquad x > 0 \text{ et } p, \lambda \geq 1, \Gamma(p) = \int_0^{+\infty} e^{-y} y^{p-1} dy,$$

La valeur du mode théorique θ *est la solution de l'équation (1.2), c'est la valeur maximisant* $f(x)$ *:*

$$\theta = (p-1)\lambda$$

Simulons un échantions de taille $n = 500$ *de la v.a* $X \sim \Gamma(p = 2, \lambda = 1)$*. En utilisant l'estimation a noyau de la densité avec un noyau gaussien :*

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

et pour h*, nous prenons le* h *optimle selon le critère* $AMISE$ *:*

$$h_{opt} = 1.06 \hat{\sigma} n^{-\frac{1}{5}}.$$

La figure suivante est une ulistration graphique de cet exemple.

Nous obtenons sous le logiciel R, la valeur

$$\theta_n = 1.186564$$

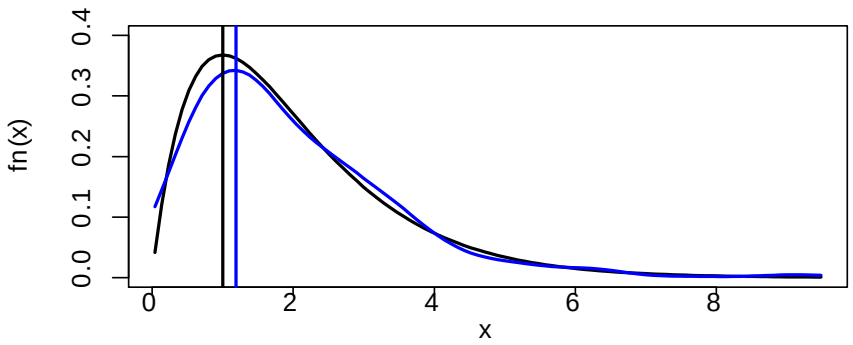

FIG. 2.5 – Estimation a noyau du mode : cas Gamma.

comme estimateur non paramétrique du mode $\theta = (p-1)\lambda = 1$, avec

$$f_n(\theta_n) = \max_{-\infty<x<+\infty} f_n(x).$$

Exemple 2.4.3 (Loi log-normale) *La loi log-normale de paramètres μ et σ admet pour densité de probabilité*

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0$$

Les paramètres μ et σ sont l'espérance et l'écart type du logarithme de la variable (puisque par définition, le logarithme de la variable est distribué selon une loi normale d'espérance μ et d'écart-type σ. Il est simple a montrer que le mode d'une variable de loi log-normale est

$$\theta = e^{\mu-\sigma^2} \quad où \quad f(\theta) = \max_{x>0} f(x).$$

Soit X une va de loi log-normale de paramètres $\mu = 0$ et $\sigma = 1$. Sous R, en simule un échantions de taille $n = 300$ de la v.a X. La figure suivante donne une aperçu sur les courbe de densité théorique et empirique, ainsi que les points maximal représantant le mode et lemode estimé. En utilisant le noyau gaussien :

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

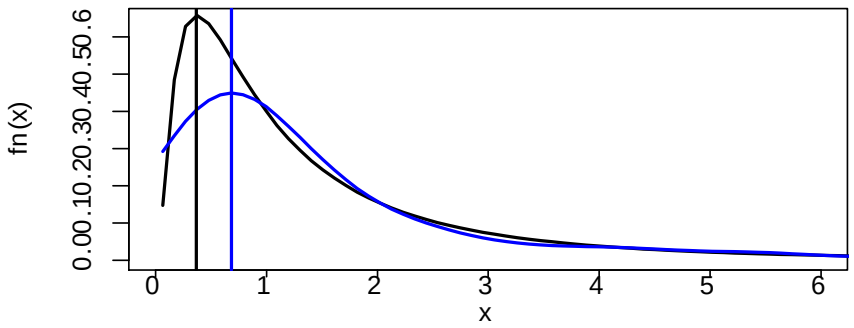

FIG. 2.6 – Estimation a noyau du mode : cas log-normale.

et le h optimle selon le critère $AMISE$:

$$h_{opt} = 1.06\hat{\sigma}n^{-\frac{1}{5}}.$$

Nous obtenons sous le logiciel R, la valeur

$$\theta_n = 0.660337$$

comme estimateur non paramétrique du mode théorique

$$\theta = e^{\mu-\sigma^2} = 0.3678794.$$

Conclusion

Dans ce mémoire plusieurs estimateurs du mode sont considérés, paramétrique et non paramétrique. L'estimateur de Parzen du mode présente un certains inconvénients car il dépend d'un paramétre de lissage en tout point, considérée à introduire un estimateur de la fonction de densité qui dépendent du paramètre de lissage différent de celui du mode considéré.

Tandis que cet estimateur est convergent, simplement, fortement, en moyenne quadratique et en loi. Cela permet de construire des intervalles de confiances pour l'estimation non paramétrique du mode.

La comparaison des différentes estimateurs du mode et l'application de l'estimation du mode pour des données réelles sont des sujets à étudier prochainement.

Bibliographie

[1] Bickel, D. R. (2002). Robust estimators of the mode and skewness of continuous data. Computational Statistics and Data Analysis 39 : 153-163.

[2] Bickel, D. R. (2003). Robust and efficient estimation of the mode of continuous data : The mode is a viable measure of central tendency. Journal of Statistical Computation and Simulation 73 : 899-912.

[3] Chernoff, H. (1964). Estimation of the mode. Ann. Instit. Statist. Math. 16 : 31-41.

[4] Hall, P. (1982). Limit theorems for estimators based on inverses of spacings of order statistics. Ann. Probab. 10 : 992–1003.

[5] Hedges, S.B., Shah, P. (2003). Comparison of mode estimation methods and application in molecular clock analysis. BMC Bioinformatics 4 : 1-11.

[6] Gaudoin, O. et Béguin, M. (2001). Principes et méthodes statistiques. Ensimag-2ème Année, INP.

[7] Grenander, U. (1965). Some direc estimates of the mode. Ann. Math. Statist. 36 : 131-138.

[8] Park, B. U., Marron, S. J. (1990) Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association 85 : 66-72.

[9] Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist, 33 :1065-1076.

[10] Rosenblatt, M. (1956). Remarks in some nonparametric estimates of a density function. Ann. Math. Statist, 27 :832-837.

[11] Silverman, B. W. (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

[12] Venter, J.H. (1967). On estimation of the mode. Ann. Math. Statist. 38 : 1446-1455.

[13] Wegman, E.J. (1971). A note on the estimation of the mode. Ann. Math. Statist. 42 : 1909-1915.

Annexe A : Logiciel *R*

• Le langage **R** est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• **R** a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team. L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

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

Ω	:	Population l'ensemble sur lequel notre étude statistique.
$Card(\Omega)$	:	Le cardinal : Nombre d'éléments de l'ensemble Ω.
$V.S$	:	La variable statistique.
X	:	Caractère.
C	:	Ensemble des valeurs du caractère X.
N	:	La taille de pupilation.
$\mathbb{N}$	:	Ensemble des nombres entiers naturels.
$\mathbb{Q}$	:	Ensemble des nombres entiers décimaux.
$\mathbb{Z}$	:	Ensemble des nombres entiers relatifs.
$\mathbb{R}$	:	Ensemble des nombres réels.
$:=$	:	Est défini comme étant (symbole d'affectation).
$\sum_{i=1}^{k}$	:	La somme pour i variant de 1 à k.
I	:	La fonction indicateur.
n_i	:	Effectif.
N_i	:	Effectif cumulé croissant.
f_i	:	la fréquence relative.
F_i	:	la fréquence relative cumulé croissant.
$\overline{x}$	:	La moynne arithmétique des n données.

Var	:	La variance.
σ	:	L'écrt-type.M
$Median$	:	La variable quantitative d'une échantion.
M_o	:	Le mode ou valeur dominate d'une série statistique.
$[]$	:	La partie entiére.
e_i	:	Limite de la classe modale.
a_i	:	L'amplitude d'une classe modale.
C_i	:	Centre de classe.
f	:	Fonction de densité.
$f^{(i)}$	:	La $i^{ème}$ dérivée de la fonction f.
F	:	Fonction de rèpartotion.
F_n	:	La fonction de rèpartotion empirique.
f_n	:	L'estimateur noyau de la fonction f.
f_L	:	L'estimateur noyau locale pour la fonction f.
$h(n)$	:	Paramétre de lissage.
$K(t)$	:	Noyau, kernel.
$Biais$	:	Biais.
MSE	:	Erreure quadratique moyenne.
$MISE$	:	Erreure quadratique moyenne intégrée
$AMISE$	:	Le développement asymptotiquede $MISE$
h_{AMISE}	:	L'estimateur de h qui minimise.
h_{opt}	:	L'estimateur de h qui minimise.
$\mu_2(k)$	:	$\int_{\mathbb{R}} t^2 k(t)dt$.
$R(g(t))$	:	$\int_{\mathbb{R}} g^2(t)dt$ pour une fonction g de carré intégrable.
$eff(k_1, k_2)$	:	Le rapport de $AMISE$ des deux noyaux.
φ	:	La fonction caractéristique
θ	:	Le mode.
θ_n	:	Estimation du mode variable.
$\hat{\theta}_n$	:	Estimation du mode de Parzen,le mode simple.

θ_n^* : Variable entre θ_n et θ.

M : Le mode d'une distribution symétrique.

$M_{p,k}$: Estimateur.

P : La fonction de probabilité.

$\xrightarrow{p}$: La convergence en probabilité.

$\xrightarrow{d}$: La convergence en distribution.

الملخص

إن الهدف الرئيسي من هذه المذكرة هو الحصول على نتائج حول تقدير المنوال المعلمي (منوال سلسلة إحصائية وبعض المتغيرات العشوائية للقانون المستمر أو المنفصل)، والتقدير لا معلمي للمنوال بالإضافة لمختلف الخصائص والتطبيقات.

Résumé

L'objet principale de ce mémoire est de trouver des résultats sur l'estimation paramétrique de mode (mode d'une série statistique et de certaines variables aléatoires de loi continue ou discrète), et l'estimation non paramétrique du mode, ainsi que ces différentes propriétés et applications.

Abstract

The main objective of this memory is to fined results about the parametric mode estimation (mode of a statistical series and of random variable of continuous and discrete law), and the nonparametric estimation with its different properties and applications.