Premier chapitre: Estimation non paramétriques :

Introduction :

La statistique non paramétrique regroupe l'ensemble des méthodes statistiques qui permettent de tirer de l'information pertinente de données sans faire l'hypothèse que la loi de probabilité de ces observations appartient à une famille paramétrée connue.

1. Le modèle statistique :

Un modèle statistique est un objet mathématique associé à l'observation de données issues d'un phénomène aléatoire.

Une expérience statistique consiste à recueillir une observation x d'un élément aléatoire X, à valeurs dans un espace $\mathcal{X}$ et dont on ne connait pas exactement la loi de probabilité $\mathbb{P}$. Des considérations de modélisation du phénomène observé amènent à admettre que $\mathbb{P}$ appartient à une famille $\mathcal{P}$ de lois de probabilité possibles.

Définition 1 : Le modèle statistique (ou la structure statistique) associé à cette expérience est le triplet $(\mathcal{X}; \mathcal{A}; \mathbb{P})$, où :

_ $\mathcal{X}$ est l'espace des observations, ensemble de toutes les observations possibles.

_ $\mathcal{A}$ est la tribu des évènements observables associée.

_ $\mathcal{P}$ est une famille de lois de probabilité possibles définie sur $\mathcal{A}$.

Remarque :

i) L'intérêt de cette notion de modèle statistique est qu'elle permet de traiter avec le même formalisme tous les types d'observations possibles.

ii) On dit que le modèle est discret quand $\mathcal{X}$ est fini ou dénombrable. Dans ce cas, la tribu $\mathcal{A}$ est l'ensemble des parties de $\mathcal{X}$: $\mathcal{A} = \mathcal{P}(\mathcal{X})$. C'est le cas quand l'élément aléatoire observé X a une loi de probabilité discrète.

iii) On dit que le modèle est continu quand $\mathcal{X} \in IR^p$ et $\forall \mathbb{P} \in \mathcal{P}$, $\mathbb{P}$ admet une densité (par rapport à la mesure de Lebesgue) dans IR^p . Dans ce cas, $\mathcal{A}$ est la tribu des boréliens de $\mathcal{X}$ (tribu engendrée par les ouverts de $\mathcal{X}$) : $\mathcal{A} = \mathcal{B}(\mathcal{X})$.

vi) On peut aussi envisager des modèles ni continus ni discrets, par exemple si l'observation a certains éléments continus et d'autres discrets. $\mathcal{X}$ et $\mathcal{A}$ sont alors plus complexes.

Exemple 1 : On a recueilli les durées de vie, supposées indépendantes et de même loi exponentielle, de n ampoules électriques. L'observation est de la forme $x = (x_1; \dots ; x_n)$, où les x_i sont des réalisations de variables aléatoires X_i indépendantes et de même loi exponentielle de paramètre λ inconnu.

$\forall i, x_i \in IR^+$, donc l'espace des observations est $\mathcal{X} = IR^{+n}$. Alors la tribu associée est $\mathcal{A} = \mathcal{B}(IR^{+n})$ => Le modèle est continu.

Comme on admet que la loi est exponentielle mais que son paramètre est inconnu, l'ensemble des lois de probabilités possibles pour chaque X_i est $\{exp(\lambda);\ \lambda \in\ IR^{+}\}$.

Comme les X_i sont indépendantes, la loi de probabilité du vecteur $(X_1; \dots ; X_n)$ est la loi produit $\mathcal{P} =$ ensemble des lois de probabilité des vecteurs aléatoires de taille n dont les composantes sont indépendantes et de même loi exponentielle de paramètre inconnu.

Finalement, le modèle statistique associé est : $(IR^{+n}, \mathcal{B}(IR^{+n}), \{exp(\lambda)^{\otimes n};\ \lambda \in\ IR^{+}\})$, qu'on peut aussi écrire:

$$(IR^{+}, \mathcal{B}(IR^{+}), \{exp(\lambda);\ \lambda \in\ IR^{+}\})^{n},$$

Exemple 2 : Contrôle de qualité. Une chaîne de production produit un très grand nombre de pièces et on s'intéresse à la proportion inconnue de pièces défectueuses. Pour l'estimer, on prélève indépendamment n pièces dans la production et on les contrôle. L'observation est $x\ =\ (x_1; \dots ;\ x_n)$, où :

$x_i\ = \begin{cases} 1 \text{ si la } i^{\text{ème}} \text{ pièce est défectueuse} \\ 0 \qquad\qquad \text{sinon} \end{cases}$.

- Trouver la structure statistique.

- L'espace des observations est $\mathcal{X} = \{0,1\}^n$ (Il est fini)=> Le modèle est discret.
- $\mathcal{A} = \mathcal{P}(\{0,1\}^n)$
- Les X_i sont i.i.d de loi de Bernoulli $\mathcal{B}(p)$ où $p = \mathbb{P}(X_i = 1)$ qui est la probabilité qu'une pièce soit défectueuse.
- Alors le modèle statistique est

$(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathcal{B}(p)^{\otimes n};\ p \in\ [0,1]\})$ ou $(\{0,1\}, \mathcal{P}(\{0,1\}), \{\mathcal{B}(p);\ p \in\ [0,1]\})^n$.

Remarque : Quand l'élément aléatoire X est numérique, il admet une fonction de répartition F. La fonction de répartition caractérisant une loi de probabilité, l'ensemble $\mathcal{P}$ des lois de probabilité possibles pour X est en bijection avec l'ensemble $\mathcal{F}$ des fonctions de répartition possibles. Aussi le modèle statistique peut dans ce cas être noté $(\mathcal{X}; \mathcal{A}; \mathcal{F})$ au lieu de $(\mathcal{X}; \mathcal{A}; \mathcal{P})$.

1.1. Modèle paramétrique ou non paramétrique

Définition 2 : Un modèle ***paramétrique*** est un modèle où l'on suppose que le type de loi de X est connu, mais qu'il dépend d'un paramètre θ inconnu, de dimension d. Alors, la famille de lois de probabilité possibles pour X peut s'écrire $\mathcal{P}\ = \{P(\theta)\ ;\ \theta \in\ \Theta \subset\ IR^{d}\}$.

(le cas des deux exemples précédents)

Remarque : Le problème principal est alors de faire de l'inférence statistique sur θ : l'estimer, ponctuellement ou par régions de confiance (intervalles si $d\ =\ 1$), et effectuer des tests d'hypothèses portant sur θ.

On fait alors de *la statistique paramétrique*.

Exercice 1:

Soit $\theta \in]-1,1[$ un paramètre inconnu et $X_1,\dots,X_n$ un échantillon *iid* de densité

$$f_\theta(x) = \frac{1}{2}(1+\theta x)1_{]-1,1[}(x).$$

1. Donner un estimateur $\hat{\theta}_n$ de θ par la méthode des moments.

2. Donner le biais de $\hat{\theta}_n$ ainsi que son risque quadratique.

3. Déterminer la loi limite de $\hat{\theta}_n$ (correctement centré et normalisé).

4. En déduire un intervalle de confiance bilatère asymptotique de niveau $(1-\alpha)$.

5. On veut tester $H_0 : \theta \geq 0$ contre $H_1 : \theta < 0$. Proposer un test de niveau asymptotique α.

Définition 3: Un modèle **non paramétrique** est un modèle où P ne peut pas se mettre sous la forme ci-dessus. Par exemple, P peut être :

_ L'ensemble des lois de probabilité continues sur IR,

_ L'ensemble des lois de probabilité dont le support est [0; 1],

_ L'ensemble des lois de probabilité sur IR symétriques par rapport à l'origine, etc...

Dans ce cadre, il est possible de déterminer des estimations, des intervalles de confiance, d'effectuer des tests d'hypothèses. Mais les objets sur lesquels portent ces procédures statistiques ne sont plus des paramètres de lois de probabilité. On peut vouloir estimer des quantités réelles comme l'espérance et la variance des observations. On a vu en Statistique inférentielle du Licence qu'on pouvait utiliser la moyenne et la variance empirique des données $(\overline{x} \; et \; S^2)$. On peut aussi vouloir estimer des fonctions, comme la fonction de répartition et la densité des observations. On a vu en statistique descriptive du 1ère année qu'un histogramme est une estimation de densité.

En termes de tests d'hypothèses, on peut effectuer des tests sur la valeur d'une espérance, tester si les observations sont indépendantes, si elles présentent une croissance, si elles proviennent d'une loi normale, tester si plusieurs échantillons proviennent de la même loi, etc...

On fait alors de *la statistique non paramétrique*.

2. Les outils de la statistique non paramétrique

On se place dans le cadre d'un modèle d'échantillon : l'observation x est un vecteur $(x_1; \dots ; x_n)$, constitué de réalisations de variables aléatoires réelles $X_1; \dots ; X_n$ indépendantes et de même loi, de fonction de répartition F. On notera f leur densité, si elle existe.

2.1 Statistiques d'ordre et de rang

Rappelons que si $x_1; \dots ; x_n$ sont n réels, on note $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ ces n réels rangés dans l'ordre croissant.

Définition 4 : La statistique d'ordre associée à l'échantillon $X_1; \dots ; X_n$ est le vecteur $X^* = (X_{(1)}; \dots ; X_{(n)})$. $X_{(i)}$ est appelée la **$i^{ème}$ statistique d'ordre**.

Remarques :

i) On note parfois X_i^* ou $X_{(i:n)}$ au lieu de $X_{(i)}$.

ii) X^* est à valeurs dans $\widetilde{IR}^n = \{(y_1; \dots ; y_n) \in IR^n ; y_1 \leq y_2 \leq \dots \leq y_n\}$.

iii) $X_{(1)} = Min(X_1; \dots ; X_n)$; $X_{(n)} = Max(X_1; \dots ; X_n)$.

vi) La statistique d'ordre contient toute l'information de l'échantillon de départ, sauf l'ordre dans lequel les observations ont été obtenues. Cet ordre est indiqué par les rangs r_i des observations.

Exemple 1 (sans ex-aequos) : n = 5

x_i	2.3	-3.5	0.5	1.7	-1.4
$x_{(i)}$	-3.5	-1.4	0.5	1.7	2.3
r_i	5	1	4	3	2

Exemple 2 (avec ex-aequos) : n = 5

x_i	0.5	-3.5	1.7	0.5	-1.4
$x_{(i)}$	-3.5	-1.4	0.5	2.3	1.7
r_i	3	1	5	3	2

Définition 5. La statistique de rang associée à l'échantillon $(X_1; \dots ; X_n)$ est le vecteur $R = (R_1; \dots ; R_n)$ où $\forall i \in \{1; \dots ; n\}$,

$$R_i = 1 + \sum_{i=1}^{n} 1_{\{X_j < X_i\}}$$

= 1 + nombre d'observations strictement inférieures à X_i

= rang de X_i dans l'échantillon ordonné.

Le rang R_i de la $i^{ème}$ observation X_i est aussi appelé la $i^{ème}$ statistique de rang.

Remarque : On ne définit pas R_i comme le nombre d'observations inférieures ou égales à X_i, pour pouvoir traiter le cas des ex-aequos.

Propriété 1. Si on connait les statistiques d'ordre et de rang, on peut reconstruire l'échantillon initial car $X_i = X_{(R_i)}$.

Remarques :

i) On constate que s'il n'y a pas d'ex-aequos dans l'échantillon, les rangs seront les entiers de 1 à n dans un ordre quelconque.

ii) On est sûrs de ne pas avoir d'ex-aequos si et seulement si

$\forall (i; j) \in \{1; \dots ; n\}^2 ; i \neq j \Rightarrow P(X_i = X_j) = 0$. En théorie, c'est bien ce qui se passe si la loi des X_i est continue. Mais en pratique, même si cette loi est continue, il est possible qu'il y ait des ex-aequos, du fait de la limitation de la précision des mesures et des erreurs d'arrondis. Il faudra donc être très attentif à la présence d'ex-aequos dans les données.

iii) Sur le plan théorique, nous éviterons cette difficulté en nous limitant aux lois continues.

Théorème 1. Soit $X_1; \dots ; X_n$ un échantillon d'une loi continue. Alors :

1. La loi de R est la loi uniforme sur l'ensemble Σ_n des permutations des entiers de 1 à n.

2. Les statistiques d'ordre et de rang sont indépendantes.

Démonstration :

1) $\forall r = (r_1, \dots, r_n) \in \Sigma_n, \mathbb{P}(R = r) = \mathbb{P}(R_1 = r_1, \dots, R_n = r_n) = \frac{1}{card\,\Sigma_n}$?

La loi est continue => Il n'y a pas d'ex-aquos

$R_i \in \{1, \dots, n\} => R_n\ est\ bien$ à $valeurs\ dans\ \Sigma_n$.

Les X_i sont iid => Elles sont interchangeables et les permutations sont équiprobables d'où

$$\forall r = (r_1, \dots, r_n) \in \Sigma_n, \mathbb{P}(R = r) = \mathbb{P}(R_1 = r_1, \dots, R_n = r_n) = \frac{1}{card\Sigma_n} = \frac{1}{n!}$$

2) Il faut démontrer que : $\forall B \in \widetilde{\mathbb{R}}_n\ et\ \forall r \in \Sigma_n : \mathbb{P}(\{X^* \in B\} \cap \{R = r\}) = \mathbb{P}(\{X^* \in B\})\mathbb{P}(R = r)$

On a $\forall B \in \widetilde{\mathbb{R}}_n$, $\forall r \in \Sigma_n$; $\mathbb{P}(\{X^* \in B\} \cap \{R = r\}) = \mathbb{P}(\{X \in B\})$

D'autre part d'après le théorème de probabilités totales

$\mathbb{P}(\{X^* \in B\}) = \sum_{r \in \Sigma_n} \mathbb{P}(\{X^* \in B\} \cap \{R = r\}) = \sum_{r \in \Sigma_n} \mathbb{P}(\{X \in B\}) = n!\,\mathbb{P}(\{X \in B\})$

=> $\mathbb{P}(\{X \in B\}) = \frac{1}{n!}\,\mathbb{P}(\{X^* \in B\}) = \mathbb{P}(R = r)\mathbb{P}(\{X^* \in B\}) = \mathbb{P}(\{X^* \in B\} \cap \{R = r\})$

Conséquence importante : La principale conséquence de ce théorème est que la loi de R ne dépend pas de la loi des X_i. On en déduit que toute variable aléatoire qui ne s'exprime qu'à l'aide des rangs des observations a une loi de probabilité indépendante de la loi de ces observations. C'est bien ce qu'on cherche à obtenir en statistique non paramétrique, où la loi des observations n'appartient pas à une famille paramétrée connue. On pourra donc faire de l'estimation et des tests non paramétriques à partir des rangs des observations.

Remarques :

1) Il n'y a pas d'équivalent de ce théorème pour les lois non continues, ce qui limite beaucoup l'intérêt de la statistique non paramétrique basée sur les rangs dans ce cas.

2) Toute fonction symétrique des observations initiales est une fonction des statistiques d'ordre. Par exemple, $\sum_{i=1}^n X_i = \sum_{i=1}^n X_{(i)}$.

Propriété 2 : Si la loi des X_i est continue, X^* admet pour densité :

$$f_{(X_{(1)};\dots;X_{(n)})}(x_{(1)}; \dots ; x_{(n)}) = n! \prod_{i=1}^{n} f(x_{(i)}) 1_{\widetilde{IR}^n}(x_{(1)}; \dots ; x_{(n)})$$

Démonstration :

Etant donnée que pour tout borélien $B \in \widetilde{\mathbb{R}}_n$, on a $\mathbb{P}(\{X^* \in B\}) = n!\,\mathbb{P}(\{X \in B\})$, on obtient pour tout B :

$$\int_B f_{(X^*_{(1)};\dots;X^*_{(n)})}(x_{(1)};\dots;x_{(n)})dx_{(1)}\dots dx_{(n)} = n!\int_B f_{(X_{(1)};\dots;X_{(n)})}(x_{(1)};\dots;x_{(n)})dx_{(1)}\dots dx_{(n)}$$

$$= \int_B n!\prod_{i=1}^{n} f_{X_i}(x_i)\,dx_1\dots dx_n$$

$$= \int_B n!\prod_{i=1}^{n} f_{X_{(i)}}(x_{(i)})\,dx_{(1)}\dots dx_{(n)}$$

D'où le résultat.

Propriété 3 : $\forall i \in \{1;\dots;n\}$, la fonction de répartition de la i^{eme} statistique d'ordre $X_{(i)}$ est :

$$\forall x \in \text{IR};\ \mathrm{F}_{X_{(i)}}(\mathrm{x}) = \sum_{k=i}^{n} C_n^k\,[F(x)]^k[1-F(x)]^{n-k}$$

Démonstration :

$$F_{X_{(i)}}(x) = \mathbb{P}(X_i \le x) = \mathbb{P}\big(i\ au\ moins\ des\ X_j\ sont\ inférieurs\ à\ x\big)$$

$$= \sum_{k=i}^{n} \mathbb{P}\big(k\ exactement\ des\ X_j\ sont\ inférieurs\ à\ x\big)$$

$$= \sum_{k=i}^{n} C_n^k\,\mathbb{P}(X_1 \le x,\dots,X_k \le x, X_{k+1} > x,\dots,X_n > x)$$

$$= \sum_{k=i}^{n} C_n^k\,[\mathbb{P}(X_i \le x)]^k\,[\mathbb{P}(X_i > x)]^{n-k}$$

$$= \sum_{k=i}^{n} C_n^k\,[F(x)]^k[1-F(x)]^{n-k}$$

Corollaire 1 : Si la loi des X_i est continue, alors $\forall i \in \{1,\dots,n\}, X_{(i)}$ admet pour densité :

$$\forall x \in \text{IR};\ f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1}[1-F(x)]^{n-i}f(x).$$

Démonstration :(Exercice)

Exercice 2:

1/ Trouver les fonctions de répartition de $X_{(1)}\ et\ X_{(n)}$.

2/ Soit $X_1;\dots;X_n$ un $n-$échantillon *iid.* de loi $\mathcal{U}_{[a,b]}$

- Montrer que $X_{(1)}\ et\ X_{(n)}$ sont les estimateurs du maximum vraisemblance de $a\ et\ b$.

Propriété 4 : Pour tous $r_1; \dots ; r_k$ entiers tels que $1 \leq r_1 \leq r_2 \leq \cdots \leq r_k \leq n$, on a :

$$f_{X_{(r_1)}\dots X_{(r_k)}}(x_1, \dots, x_k) = \frac{n!}{(r_1-1)!\prod_{i=2}^{k}(r_i - r_{i-1} - 1)!\,(n-r_k)!}[F(x_1)]^{r_1-1}\prod_{i=1}^{k} f(x_i)$$

$$\left[\prod_{i=2}^{k}[F(x_i) - F(x_{i-1})]^{r_i - r_{i-1}-1}\right][1-F(x_k)]^{n-r_k} 1_{\widetilde{\mathrm{IR}^k}}(x_1, \dots, x_n).$$

2.2. Loi de probabilité empirique

La loi de probabilité empirique est une loi de probabilité créée directement à partir de l'échantillon observé $x_1, \dots, x_n$.

Définition 6. La loi de probabilité empirique $\mathbb{P}_n$ associée à l'échantillon $x_1, \dots, x_n$ est la loi uniforme (discrète) sur $\{x_1, \dots, x_n\}$. Si X_e est une variable aléatoire de loi $\mathbb{P}_n$, alors :

_ X_e est à valeurs dans $\{x_1, \dots, x_n\}$.

_ $\forall i \in \{1, \dots, n\}; P(X_e = x_i) = \mathbb{P}_n(x_i) = \frac{1}{n}$.

On peut aussi écrire $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$.

Les caractéristiques essentielles de la loi de probabilité empirique sont, en fait, des quantités bien connues:

i) La fonction de répartition de la loi de probabilité empirique est la fonction de répartition empirique F_n :

$P(X_e \leq x) = \sum_{x_i \leq x} P(X_e = x_i) = \frac{1}{n} \times nombre\ de\ x_i \leq x$

$$= \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}} = F_n(x).$$

ii) L'espérance de la loi de probabilité empirique est la moyenne empirique $\overline{x_n}$:

$$E(X_e) = \sum_{i=1}^{n} x_i P(X_e = x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x_n}.$$

iii) La variance de la loi de probabilité empirique est la variance empirique s_n^2:

$$Var(X_e) = E[(X_e - \mathrm{E}[\mathrm{X_e}])^2] = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 P(X_e = x_i) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = s_n^2.$$

iv) Le moment empirique d'ordre k est :

$$m_e^k = E[X_e^k] = \frac{1}{n}\sum_{i=1}^{n} {x_i}^k.$$

v) Le moment empirique centré d'ordre k est :

$$\mu_e^k = E[(X_e - \mathrm{E}[\mathrm{X_e}])^k] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^k$$

vi) Les quantiles de la loi de probabilité empirique sont les quantiles empiriques :

$$\forall p \in]0;\ 1[\ ;\ \tilde{q}_{n;p} = \begin{cases} \frac{1}{2}\left(x^*_{np} + x^*_{np+1}\right) & si\ np\ est\ entier \\ x^*_{[np]+1} & sinon \end{cases}$$

Remarques :

i) Puisqu'on considère les observations $x_1; \ldots ; x_n$ comme des réalisations de variables aléatoires $X_1; \ldots ; X_n$, toutes les quantités définies dans cette section sont elles mêmes des réalisations de variables aléatoires :

$\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}$; $\overline{X_n} =$; $\frac{1}{n}\sum_{i=1}^{n} X_i$; $S_{\mathrm{n}}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$

$$\tilde{Q}_{n;p} = \begin{cases} \frac{1}{2}\left(X^*_{np} + X^*_{np+1}\right) & si\ np\ est\ entier \\ X^*_{[np]+1} & sinon \end{cases}$$

ii) Ces quantités sont considérées comme des statistiques exhaustives ou même des estimateurs non paramétriques pour les quantités souhaitées (exp ; la moyenne , la variance,…)

Exercice 3:

Pour α désignant un réel positif, considérons la fonction de répartition suivante

$$F_\alpha(x) = \begin{cases} 0 & si\ x < 1 \\ 1 - (2-x)^\alpha & si\ 1 \leq x < 2 \\ 1 & si\ x \geq 2 \end{cases}$$

Pour un échantillon $X_1, X_2, \ldots, X_n$ issu de la loi F_α, On note $X_{(1)} = \min(X_1, X_2, \ldots, X_n)$ et $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$.

1- Trouver les fonctions de répartitions de $X_{(1)}$ et de $X_{(n)}$.

2- Déterminer la valeur vers laquelle $X_{(n)}$ converge en probabilité.

3 - L'estimation fonctionnelle

Les hypothèses de cette section sont les mêmes que celles du celle précédente : on suppose que les observations $x_1; \ldots ; x_n$ sont des réalisations de variables aléatoires réelles $X_1; \ldots ; X_n$ indépendantes et de même loi, de fonction de répartition F, et de densité f, si elle existe.

Nous allons maintenant nous intéresser à l'estimation de la fonction de répartition et, si elle existe, de la densité de l'échantillon. Il s'agit d'estimer des fonctions, d'où le nom d'**estimation fonctionnelle**. De plus l'une comme l'autre de ces fonctions caractérisent entièrement la loi de probabilité de l'échantillon.

Remarque :

1- La fonction de répartition empirique est un estimateur simple et performant de la fonction de répartition de l'échantillon. Il est beaucoup plus difficile d'estimer une densité.

2- l'estimation des quantiles peut être considérée comme de l'estimation fonctionnelle dans la mesure où estimer $q_p = F^{-1}(p)$ quel que soit p revient à estimer la fonction F^{-1}.

Estimer une fonction g, c'est d'abord estimer $g(x)$ pour tout x donnée. Il faut ensuite juger de la qualité de l'estimation de $g(x)$ pour chaque x, puis de l'estimation de g dans son ensemble.

Définition 7 : l'Erreur Quadratique Moyenne (ou risque quadratique)

Si $\hat{g}(x)$ est un estimateur de $g(x)$, la qualité de l'estimation pour un x donné est usuellement mesurée par le biais, la variance et l'Erreur Quadratique Moyenne (ou risque quadratique), qu'on notera $EQM_x(\hat{g})$:

$$EQM_x(\hat{g}) = \mathrm{E}[(\hat{g}(x) - g(x))^2] = [E(\hat{g}(x) - g(x))]^2 + Var(\hat{g}(x))$$

Pour juger de la qualité de l'estimation de g dans son ensemble, il faut utiliser des mesures de l'écart entre g et $\hat{g}$. Suivant les cas, on utilisera :

- L'Erreur Quadratique Moyenne Intégrée $(EQMI)$:

$$EQMI(\hat{g}) = \int_{-\infty}^{+\infty} EQM_x(\hat{g})dx$$

- l'écart maximum entre les deux fonctions :

$$sup\{|\hat{g}(x) - g(x)|; x \in \mathbb{R}\}$$

3.1. Estimation de la fonction de répartition

3.1.1 Estimation ponctuelle

Rappelons que la fonction de répartition empirique IF_n de l'échantillon est définie par :

$$IF_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}} = pourcentage\ d'observations\ inférieures\ \text{à}\ x.$$

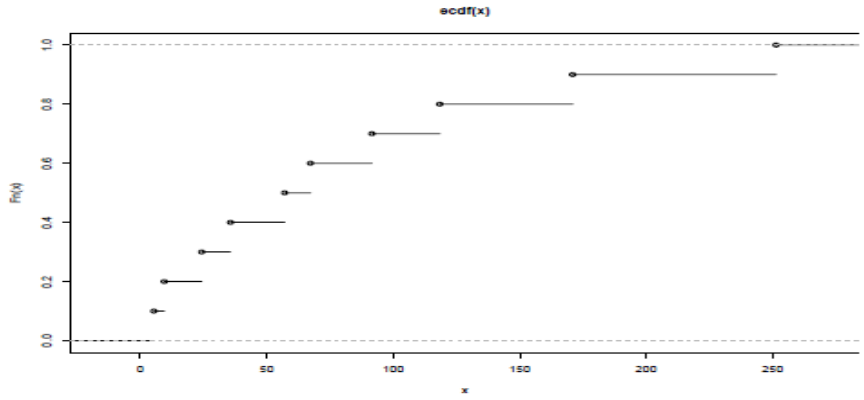
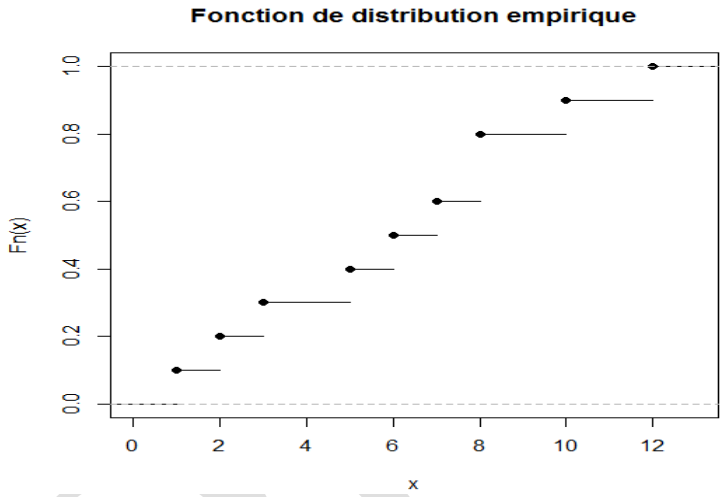


Figure 3.1 Fonction de répartition empirique.

Exemple : Soit les 10 réalisations suivantes {8,2,6,5,3,8,10,7,1,12} . Trouver la fonction de répartition empirique.

$$IF_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \le x\}} = \begin{cases} 0 & si\ x < 1 \\ 0.1 & si\ 1 \le x < 2 \\ 0.2 & si\ 2 \le x < 3 \\ 0.3 & si\ 3 \le x < 5 \\ 0.4 & si\ 5 \le x < 6 \\ 0.5 & si\ 6 \le x < 7 \\ 0.6 & si\ 7 \le x < 8 \\ 0.8 & si\ 8 \le x < 10 \\ 0.9 & si\ 10 \le x < 12 \\ 1 & si\ \ 12 \le x \end{cases}$$



Il s'avère que $\mathbb{F}_n$ est un excellent estimateur de F, ce que l'on peut montrer en plusieurs étapes.

Propriété 5 . $\forall x \in IR$; $n\mathbb{F}_n(x)$ est de loi binomiale $B(n;\ F(x))$.

Démonstration.

On a pour $x \in IR; n\mathbb{F}_n(x) = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \le x\}}$ est une somme de n v. a. i.i.d. de loi de Bernoulli de paramètre $\mathbb{P}(X_i \le x) = F(x)$, donc $n\mathbb{F}_n(x) \rightsquigarrow \mathcal{B}(n, F(x))$.

On en déduit facilement les qualités de l'estimation de $F(x)$ par $\mathbb{F}_n(x)$.

Propriété 6 .

1. $\forall x \in IR$, $\mathbb{F}_n(x)$ est un estimateur sans biais et convergent en moyenne quadratique de $F(x)$.

2. $\forall x \in IR$; $\mathbb{F}_n(x) \xrightarrow{p.s} \mathrm{F(x)}$.

Démonstration :

1) $\mathbb{E}(\mathbb{F}_n(x)) = \frac{1}{n}\mathbb{E}(n\mathbb{F}_n(x)) = \frac{1}{n}nF(x) = F(x)$.

$$var\big(\mathbb{F}_n(x)\big) = \frac{1}{n^2} var\big(n\mathbb{F}_n(x)\big) = \frac{1}{n^2} nF(x)\big(1-F(x)\big)$$
$$= \frac{F(x)\big(1-F(x)\big)}{n} \xrightarrow[n\to+\infty]{} 0$$

2) Il suffit d'appliquer la loi des grands nombres au v.a. de loi de Bernoulli $\mathbb{I}_{\{X_i \le x\}}$:

$\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \le x\}} \overset{p.s}{\to} \mathbb{E}\big(\mathbb{I}_{\{X_i \le x\}}\big) = F(x).$

Remarque : La loi des grands nombres dit que la probabilité d'un événement est la limite de la fréquence d'occurrence de cet événement dans une suite d'expériences identiques et indépendantes. On en déduit que l'on peut estimer la probabilité que X soit inférieure à x, $F(x)$, par le pourcentage d'observations inférieures à x, $\mathbb{F}_n(x)$. Cette estimation est d'excellente qualité.

Pour juger de la qualité globale de l'estimation de F par $\mathbb{F}_n(x)$., on utilise le théorème de Glivenko-Cantelli, qui dit que $\mathbb{F}_n(x)$ est un estimateur convergent uniformément et presque sûrement de F :

Théorème 2 . Théorème de Glivenko-Cantelli.

$$D_n = sup\{|\mathbb{F}_n(x) - F(x)|;\ x \in IR\} \overset{PS}{\to} 0$$

Par ailleurs, l'erreur quadratique moyenne intégrée est :

$$EQMI(\mathbb{F}_n) = \int_{-\infty}^{+\infty} Var(\mathbb{F}_n(x))dx = \frac{1}{n}\int_{-\infty}^{+\infty} F(x)[1-F(x)]dx.$$

On ne peut pas calculer explicitement cette erreur, mais on sait qu'elle tend vers 0 quand n tend vers l'infini à la vitesse $1/n$.

3.1.2. Intervalle de Confiance

Soit x fixé. Un intervalle de confiance de seuil α pour $F(x)$ est un intervalle aléatoire I tel que $P(F(x) \in I) = 1-\alpha$. On va chercher un intervalle de confiance de la forme

$I = [\mathbb{F}_n(x) - a_\alpha;\ \mathbb{F}_n(x) + a_\alpha]$, où a_α est déterminé en écrivant :

$$P(F(x) \in I) = P(\mathbb{F}_n(x) - a_\alpha \le F(x) \le \mathbb{F}_n(x) + a_\alpha)$$
$$= P(F(x) - a_\alpha \le \mathbb{F}_n(x) \le F(x) + a_\alpha)$$
$$= P\big(n(F(x) - a_\alpha) \le n\mathbb{F}_n(x) \le n(F(x) + a_\alpha)\big)$$
$$= \sum_{k=[n(F(x)-a_\alpha)]+1}^{[n(F(x)+a_\alpha)]} C_n^k\,[F(x)]^k[1-F(x)]^{n-k} = 1-\alpha$$

On ne peut pas déduire la valeur de a_α de cette expression car elle implique $F(x)$, qui est inconnue. En revanche, on peut obtenir un résultat asymptotique.

En effet, l'application du théorème central-limite sur les $1_{\{X_i \le x\}}$, variables aléatoires indépendantes de loi de Bernoulli, d'espérance $F(x)$ et de variance $F(x)(1-F(x))$ permet d'écrire :

$$\frac{\sum_{i=1}^{n} 1_{\{X_i \le x\}} - nE(1_{\{X_i \le x\}})}{\sqrt{nVar(1_{\{X_i \le x\}})}} = \frac{n\mathbb{F}_n(x) - nF(x)}{\sqrt{nF(x)(1-F(x))}} = \sqrt{n}\frac{\mathbb{F}_n(x) - F(x)}{\sqrt{F(x)\big(1-F(x)\big)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Grâce au théorème de Slutsky et à la convergence presque sûre de$\mathbb{F}_n(x)$ vers $F(x)$, on a également :

$$\sqrt{n}\frac{\mathbb{F}_n(x) - F(x)}{\sqrt{\mathbb{F}_n(x)\big(1-\mathbb{F}_n(x)\big)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Alors on obtient que, pour n suffisamment grand :

$$P(F(x) \in I) = P(-a_\alpha \le \mathbb{F}_n(x) - F(x) \le a_\alpha) = P(|\mathbb{F}_n(x) - F(x)| \le a_\alpha)$$

$$= P\left(\sqrt{n}\frac{|\mathbb{F}_n(x) - F(x)|}{\sqrt{\mathbb{F}_n(x)\big(1-\mathbb{F}_n(x)\big)}} \le \sqrt{n}\frac{a_\alpha}{\sqrt{\mathbb{F}_n(x)\big(1-\mathbb{F}_n(x)\big)}}\right)$$

$$= P\left(|U| \le \sqrt{n}\frac{a_\alpha}{\sqrt{\mathbb{F}_n(x)\big(1-\mathbb{F}_n(x)\big)}}\right) = 1-\alpha$$

Où Uest de loi $\mathcal{N}(0,1)$. D'où $\sqrt{n}\frac{a_\alpha}{\sqrt{\mathbb{F}_n(x)\big(1-\mathbb{F}_n(x)\big)}} = u_\alpha$ et $a_\alpha = \frac{u_\alpha}{\sqrt{n}}\sqrt{\mathbb{F}_n(x)\big(1-\mathbb{F}_n(x)\big)}$

Et on obtient finalement :

Propriété 7 : $\forall x \in IR$, un intervalle de confiance asymptotique de seuil α pour $F(x)$ est :

$$\left[\mathbb{F}_n(x) - \frac{u_\alpha}{\sqrt{n}}\sqrt{\mathbb{F}_n(x)\big(1-\mathbb{F}_n(x)\big)}\,,\qquad \mathbb{F}_n(x) + \frac{u_\alpha}{\sqrt{n}}\sqrt{\mathbb{F}_n(x)\big(1-\mathbb{F}_n(x)\big)}\right]$$

Exercice 4:

Soit (X_n) une suite de variables aléatoires *i.i.d.* et F la fonction de répartition de X_1. On considère la fonction de répartition empirique $\mathbb{F}_n$ définie par :

$$\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \le t\}}, \qquad t \in \mathbb{R}$$

1. Quelle est la loi de $n\mathbb{F}_n(t)$? La loi limite de $\sqrt{n}(\mathbb{F}_n(t) - F(t))$?

2. Calculer $E([\mathbb{F}_n(t) - F(t)]^2)$ et en déduire que $\mathbb{F}_n(t)$ converge en moyenne quadratique vers $F(t)$ lorsque $n \to +\infty$.

3. Montrer que cette convergence est aussi presque sûre.

3.2. Estimation de la densité

Dans cette section, on suppose que la loi de l'échantillon est continue et on cherche à estimer sa densité f. f est la dérivée de F, mais la fonction de répartition empirique $\mathbb{F}_n$ n'est pas dérivable, puisque c'est une fonction en escalier. On ne peut donc pas utiliser directement les résultats sur la fonction de répartition empirique pour estimer la densité.

On peut se demander quelle est l'utilité d'estimer la densité alors que l'on a déjà un très bon estimateur de la fonction de répartition. La principale raison est que la forme d'une densité est beaucoup plus facile à interpréter que celle d'une fonction de répartition. Par exemple, on pourra facilement avoir, grâce à une estimation de densité, des informations sur la symétrie ou la multimodalité de la loi de l'échantillon, alors que ce n'est pas du tout facile au seul vu de la fonction de répartition empirique. De même, une estimation de densité est une aide importante au choix d'un modèle approprié pour la loi de l'échantillon. Par exemple, une densité estimée en forme de cloche symétrique peut conduire à l'adoption d'un modèle de loi normale.

Nous allons commencer par donner des rappels sur la méthode d'estimation de densité la plus élémentaire, celle de l'histogramme. Puis nous présenterons la méthode plus sophistiquée du noyau.

3.2.1 Rappels sur les histogrammes

On se fixe une borne inférieure de l'échantillon $a_0 < x_{(1)}$ et une borne supérieure $a_k > x_{(n)}$.

On partitionne l'intervalle $]a_0;\ a_k]$, contenant toutes les observations, en k classes $]a_{j-1};\ a_j\ ]$.

La largeur de la classe j est $h_j = a_j - a_{j-1}$.

L'effectif de la classe j est le nombre d'observations appartenant à cette classe : $n_j = \sum_{i=1}^{n} 1_{]a_{j-1};\, a_j\ ]}(x_i)$.

La fréquence de la classe j est $\hat{p}_j = \frac{n_j}{n}$.

L'histogramme est constitué de rectangles dont les bases sont les classes et dont les aires sont égales aux fréquences de ces classes. Donc l'histogramme est la fonction en escalier constante sur les classes et qui vaut $\frac{n_j}{n\mathrm{h_j}}$ sur la classe $]a_{j-1};\ a_j\ ]$. Cette fonction peut s'écrire :

$$\hat{f}(x) = \sum_{j=1}^{k} \frac{n_j}{n\mathrm{h_j}} 1_{]a_{j-1};\, a_j\ ]}(x) = \sum_{j=1}^{k} \frac{\hat{p}_j}{\mathrm{h_j}} 1_{]a_{j-1};\, a_j\ ]}(x)$$

$$= \frac{1}{n}\sum_{j=1}^{k} \frac{1}{\mathrm{h_j}} 1_{]a_{j-1};\, a_j\ ]}(x) \sum_{i=1}^{n} 1_{]a_{j-1};\, a_j\ ]}(x_i).$$

Remarque : *i)* Dans l'histogramme à pas fixe, les classes sont de même largeur $h = \frac{a_k - a_0}{k}$ (qu'on appelle une fenêtre). Dans ce cas, la hauteur d'un rectangle est proportionnelle à l'effectif de sa classe. Par substitution, nous définissons l'estimateur de f par histogramme à k classes comme suit :

$$\widehat{f_h}(x) = \sum_{j=1}^{k} \frac{n_j}{nh} 1_{]a_{j-1};\, a_j\ ]}(x) = \sum_{j=1}^{k} \frac{\hat{p}_j}{h} 1_{]a_{j-1};\, a_j\ ]}(x).$$

Exercice. Vérifier que l'estimateur par histogramme $\widehat{f_h}$ est une densité de probabilité.

1) $\forall x \in]a_0, a_k], \widehat{f_h}(x) \geq 0.$
2) $\int \widehat{f_h}(x)dx = \int \sum_{j=1}^{k} \frac{\hat{p}_j}{\mathrm{h_j}} 1_{]a_{j-1};\, a_j\,]}(x)dx = \sum_{j=1}^{k} \frac{\hat{p}_j}{\mathrm{h_j}} \int_{a_{j-1}}^{a_j} dx = \sum_{j=1}^{k} \frac{\hat{p}_j}{\mathrm{h_j}} \mathrm{h_j} = \sum_{j=1}^{k} \hat{p}_j = 1.$

ii) Il est pertinent de choisir un histogramme à classes de même effectif. Admettons pour simplifier que n soit divisible par k. Alors chaque classe doit contenir n/k observations. Les limites des classes seront alors les j/k quantiles empiriques :

$$a_j = \tilde{q}_{n,j/k} = \frac{1}{2}\left(x_{\left(\frac{n_j}{k}\right)} + x_{\left(\frac{n_j}{k}+1\right)}\right), \qquad j = 1, \ldots, k-1\,;$$

Les bornes des classes sont donc cette fois aléatoires, puisqu'elles sont fonction des observations.

Enfin, le polygone des fréquences est la ligne brisée reliant les milieux des sommets des rectangles, et prolongée de part et d'autre de l'histogramme de façon à ce que l'aire totale délimitée par le polygone soit égale à 1, comme pour une densité.

Exemple :

Dans une rue passante de Montréal, on a mesuré le niveau de bruit en décibels émis par 20 véhicules pris au hasard. Les données ordonnées sont les suivantes :

54.8 55.4 57.7 59.6 60.1 61.2 62.0 63.1 63.5 64.2

65.2 65.4 65.9 66.0 67.6 68.1 69.5 70.6 71.5 73.4

Les histogrammes à classes de même largeur et de même effectif avec leurs polygones des fréquences, sont donnés par :

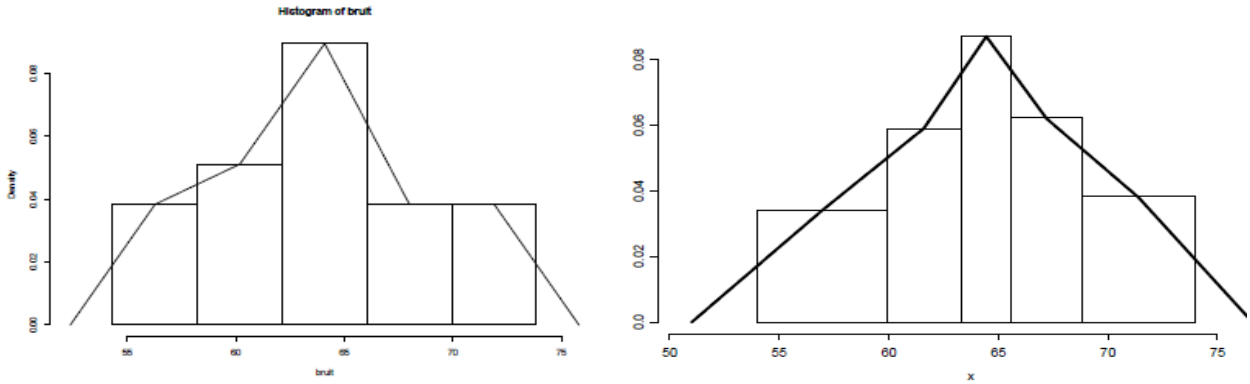


Figure1 : Histogramme à classes de même largeur et à classes de même effectif pour les niveaux de bruit à Montréal.

La forme de ces histogrammes est assez proche d'une cloche symétrique, ce qui nous amène à envisager l'hypothèse que les données proviennent d'une loi normale.

3.2.2 Risque de l'estimateur par histogramme

la qualité de l'estimateur par histogramme dépend fortement de la fenêtre h. Afin de quantifier cette dépendance, nous introduisons le risque quadratique de $\hat{f}_h$ au point $x \in [a_0, a_k]$ comme étant la moyenne de l'erreur quadratique :

$$EQM_x(\hat{f}_h) = \mathrm{E}\left[(\hat{f}_h(x) - f(x))^2\right] = \underbrace{\left[E\left(\hat{f}_h(x)\right) - f(x)\right]^2}_{carré\ du\ biais} + \underbrace{Var(\hat{f}_h(x))}_{variance},$$

Plaçons-nous maintenant dans le cadre d'un n−échantillon aléatoire de loi mère de densité f continue sur tout IR et étudions les propriétés d'échantillonnage de la v.a. (notée simplement comme précédemment) $\hat{f}_h(x) = N_j/nh$ où N_k est le nombre aléatoire de valeurs tombant dans $]a_{j-1}, a_j]$, x étant fixé dans cet intervalle. On a $N_j \sim B(n, p_j)$, d'où :

On a pour $x \in]a_{j-1}, a_j]$, $E\left(\hat{f}_h(x)\right) = E\left(\frac{N_j}{nh}\right) = \frac{n\,p_j}{nh} = \frac{p_j}{h}$

$$=> E\left(\hat{f}_h(x)\right) = \sum_{j=1}^{k} \frac{p_j}{h} 1_{]a_{j-1};\, a_j\,]}(x)\ sur\ ]a_0, a_k]$$

On a pour $x \in]a_{j-1}, a_j]$, $Var\left(\hat{f}_h(x)\right) = Var\left(\frac{N_j}{nh}\right) = \frac{n\,p_j(1-p_j)}{n^2h^2} = \frac{p_j(1-p_j)}{nh^2}$

$$=> Var\left(\hat{f}_h(x)\right) = \sum_{j=1}^{k} \frac{p_j(1-p_j)}{nh^2} 1_{]a_{j-1};\, a_j\,]}(x)\ sur\ ]a_0, a_k]$$

Conséquences :

i) Comme $p_j = \int_{a_{j-1}}^{a_j} f(x)dx$, $\frac{p_j}{h}$ est la valeur moyenne de f sur $]a_{j-1}, a_j]$, et $\hat{f}_h(x)$ n'est donc sans biais que pour la ou les valeurs de x dans $]a_{j-1}, a_j]$, où f prend cette valeur moyenne.

ii) Le risque EQM_x est supérieur au carré du biais $\left[\frac{p_j}{h} - f(x)\right]^2$. Par conséquent, si la fenêtre h est choisie indépendamment de la taille de l'échantillon n, l'estimateur par histogramme ne convergera pas vers la vraie densité lorsque $n \to \infty$, excepté la situation peu fréquente où f est constante sur l'intervalle $]a_{j-1}, a_j]$. Afin d'élargir la classe des densités pour lesquelles $\hat{f}_h$ est convergent, nous devons choisir h comme une fonction de n ; $h = h_n$ doit tendre vers 0 lorsque n tend vers +∞. A partir de maintenant, on suppose que cette condition est satisfaite.

Rappelons que le but de ce paragraphe est d'évaluer le risque de l'estimateur $\hat{f}_h$. Afin d'avoir une évaluation globale valable pour tout point $x \in [a_0, a_k]$, on considère le risque quadratique intégré :

$$EQMI(\hat{f}_h) = \int_{-\infty}^{+\infty} EQM_x(\hat{f}_h)dx = \mathrm{E}\left[\int_{-\infty}^{+\infty} (\hat{f}_h(x) - f(x))^2 dx\right]$$

(Pour obtenir la dernière égalité nous avons utilisé le théorème de Fubini).

Nous avons donc le résultat suivant :

Lemme 1. Si $X_1, \ldots, X_n$ sont indépendantes de même loi de densité f supportée par $[a_0, a_k]$, et $\hat{f}_h$ est l'estimateur par histogramme avec $k = 1/h$ classes, alors :

$$EQMI(\hat{f}_h) = \mathrm{E}\left[\int_{-\infty}^{+\infty} \left(\hat{f}_h(x) - f(x)\right)^2 dx\right] = \int_{-\infty}^{+\infty} (f(x))^2 dx + \frac{1}{nh} - \frac{n+1}{nh} \sum_{j=1}^{k} p_j^2.$$

Démonstration :

D'une part, en vertu de la propriété : $\sum_{j=1}^{k} p_j = \int f(x)dx = 1$, on a

$$\int_{-\infty}^{+\infty} Var\left(\hat{f}_h(x)\right) dx = \sum_{j=1}^{k} \int_{a_{j-1}}^{a_j} Var\left(\hat{f}_h(x)\right) dx$$

$$= \sum_{j=1}^{k} \frac{p_j(1-p_j)}{nh} = \frac{1}{nh} - \frac{1}{nh} \sum_{j=1}^{k} p_j^2$$

D'autre part,

$$\int_{-\infty}^{+\infty} \left(\mathrm{E}\left(\hat{f}_h(x)\right) - f(x)\right)^2 dx = \sum_{j=1}^{k} \int_{a_{j-1}}^{a_j} \left(\frac{p_j}{h} - f(x)\right)^2 dx$$

$$= \sum_{j=1}^{k} \left[\frac{p_j^2}{h} - 2\frac{p_j}{h} \int_{a_{j-1}}^{a_j} f(x)dx + \int_{a_{j-1}}^{a_j} \left(f(x)\right)^2 dx\right]$$

$$= \int_{a_0}^{a_k} \left(f(x)\right)^2 dx - \frac{1}{h} \sum_{j=1}^{k} p_j^2$$

Nous avons donc :

$$EQMI(\hat{f}_h) = \int_{a_0}^{a_k} \left(f(x)\right)^2 dx - \frac{1}{h} \sum_{j=1}^{k} p_j^2 + \frac{1}{nh} - \frac{1}{nh} \sum_{j=1}^{k} p_j^2$$

$$= \int_{-\infty}^{+\infty} (f(x))^2 dx + \frac{1}{nh} - \frac{n+1}{nh} \sum_{j=1}^{k} p_j^2$$

Le résultat du Lemme 1 est non-asymptotique : il est valable pour tout h et pour tout n. Ce qui nous intéresse maintenant c'est le comportement du risque $EQMI$ lorsque $h = h_n$ décroit vers zéro quand $n \to +\infty$, qui est donné dans le théorème suivant :

Théorème 2.1. Supposons que la densité de l'échantillon $X_1, \dots, X_n$ est deux fois continûment différentiable et s'annule en dehors de l'intervalle $[a_0, a_k]$. Si la fenêtre h de l'estimateur par histogramme $\widehat{f_h}$ est telle que $h_n \to 0$ lorsque $n \to \infty$, alors

$$EQMI(\widehat{f_h}) = \underbrace{\frac{h^2}{12}\int_0^1 f'^{(x)^2} dx + \frac{1}{nh}}_{terme\ \ principal\ \ du\ risque} + \underbrace{o(h^3) + O\left(\frac{1}{n}\right)}_{terme\ \ résiduel}$$

Lorsque $n \to \infty$.

Démonstration :

On vérifie aisément que :

$$\int_{a_{j-1}}^{a_j} \big(f(x)\big)^2 dx - \frac{1}{h}p_j^2 = \int_{a_{j-1}}^{a_j} \left(f(x) - \frac{1}{h}\int_{a_{j-1}}^{a_j} f(u)du\right)^2 dx \ \ (\boldsymbol{exercice})$$

$$= \frac{1}{h^2}\int_{a_{j-1}}^{a_j} \left(\int_{a_{j-1}}^{a_j} \big(f(x) - f(u)\big)du\right)^2 dx$$

Comme f est supposée deux fois continûment différentiable, on a

$$f(u) - f(x) = (u - x)\, f'(a_j) + o(h^2) \text{ pour tout } u, x \in]a_{j-1}, a_j],$$

Par conséquent,

$$\int_{a_{j-1}}^{a_j} \big(f(x)\big)^2 dx - \frac{1}{h}p_j^2 = \frac{f'(a_j)^2}{h^2}\int_{a_{j-1}}^{a_j} \left(\int_{a_{j-1}}^{a_j} (x - u)du\right)^2 dx + o(h^4).$$

En utilisant le changement de variable $(x, u) = (a_j + yh, a_j + zh)$, on obtient :

$$\int_{a_{j-1}}^{a_j} \left(\int_{a_{j-1}}^{a_j} (x - u)du\right)^2 dx = h^5 \int_0^1 \left(\int_0^1 (y - z)dy\right)^2 dz = \frac{h^5}{12}.$$

Nous avons donc démontré que :

$$\int_{a_{j-1}}^{a_j} \big(f(x)\big)^2 dx - \frac{1}{h}p_j^2 = \frac{h^3}{12}f'(a_j)^2 + o(h^4) = \frac{h^2}{12}\int_{a_{j-1}}^{a_j} f'(x)^2 dx + o(h^4).$$

En conséquence,

$$EQMI(\widehat{f_h}) = \left(\sum_{j=1}^{k}\int_{a_{j-1}}^{a_j} \big(f(x)\big)^2 dx - \frac{1}{h}p_j^2\right) + \frac{1}{nh} - \frac{n+1}{nh}\sum_{j=1}^{k} p_j^2$$

$$= \frac{h^2}{12}\int_0^1 f'^2(x)dx + o(h^3) + \frac{1}{nh} + O\left(\frac{1}{n}\right),$$

où nous avons utilisé la relation $ko(h^4) = o(h^3)$.

Supposons un instant qu'on connaît la quantité $\int_0^1 f'^{(x)^2} dx$. Dans ce cas, on peut calculer le terme principal du risque $EQMI(\widehat{f_h})$. Cela nous permet de trouver la valeur idéale de la fenêtre qui minimise le terme principal du risque. En effet, on voit aisément que le minimum de la fonction

$$h \to \frac{h^2}{12}\int_0^1 f'^2(x)dx + \frac{1}{nh}$$

est atteint au point

$$h_{opt} = \left(\frac{n}{6}\int_0^1 f'(x)^2 dx\right)^{-1/3}$$

Exercice 5: Vérifier la valeur de h_{opt}.

Cette fenêtre optimale est en général inaccessible au statisticien, car la densité f (ainsi que sa dérivée) est inconnue. Cependant, elle a le mérite de nous indiquer que la fenêtre optimale doit être de l'ordre de $n^{-1/3}$ lorsque n est grand. De plus, en injectant cette valeur de h dans l'expression de EQMI obtenue dans le théorème précédent, on obtient :

$$EQMI(\widehat{f_h}) = (3/4)^{2/3}\left(\int_0^1 f'(x)^2 dx\right)^{1/3} n^{-2/3} + O(n^{-1})$$

Ce résultat nous indique les limites de l'estimateur par histogramme : pour les densités deux fois différentiables, la meilleure vitesse de convergence qu'on puisse espérer atteindre avec un estimateur par histogramme est de $n^{-2/3}$. C'est une vitesse honorable, mais elle est nettement moins bonne que la vitesse de convergence $1/n$ qui apparaît typiquement dans des problèmes paramétriques. Ceci n'est pas très surprenante, car l'estimation de densité est un problème non-paramétrique et, par conséquent, est plus difficile à résoudre qu'un problème paramétrique.

En revanche, on verra par la suite que, sous les mêmes hypothèses que celles considérées dans ce paragraphe, on peut construire un autre estimateur de la densité f qui converge à une meilleure vitesse $n^{-4/5}$. L'estimateur qui atteint cette vitesse s'appelle estimateur à noyau et on peut démontrer que cette vitesse ne peut pas être améliorée sans imposer de nouvelles conditions sur f.

3.3 La méthode du noyau

Les histogrammes et les polygones des fréquences ne sont pas des estimations très satisfaisantes de la densité de l'échantillon car ce sont des fonctions en escalier et des lignes brisées alors que la densité à estimer est en général plus lisse, avec au moins sa dérivée continue.

D'autre part, l'aléa dû au choix du nombre de classes et des bornes des classes est un élément très perturbant de l'analyse, puisque des choix différents peuvent aboutir à des histogrammes d'allures assez nettement différentes.

L'estimation par noyau a pour but de répondre à ces deux écueils et de proposer des estimations de densité ayant de bonnes propriétés. L'origine de la méthode des noyaux est due à Rosenblatt (1956). Celui-ci a proposé une sorte d'histogramme mobile où la fenêtre de comptage des observations se déplace avec la valeur de x. La densité en x est estimée par la fréquence relative des observations dans l'intervalle $[x - h, x + h]$, donc centré sur x, divisée naturellement par la largeur de l'intervalle $2h$. On appelle h la largeur de fenêtre (bien que cette largeur soit en fait égale à $2h$).

Pour cela, on commence par remarquer que la densité est la dérivée de la fonction de répartition, ce qui permet d'écrire pour tout x :

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

Donc pour un $h > 0$ fixé « petit », on peut penser à estimer $f(x)$ par :

$$\hat{f}(x) = \frac{1}{2h}\left(\mathbb{F}_n(x+h) - \mathbb{F}_n(x-h)\right) = \frac{1}{2nh}\sum_{i=1}^{n} 1_{]x-h;\, x+h\,]}(X_i)$$

On a alors :

$$\mathbb{E}[\hat{f}(x)] = \frac{1}{2h}\left(\mathbb{E}[\mathbb{F}_n(x+h)] - \mathbb{E}[\mathbb{F}_n(x+h)]\right) = \frac{1}{2h}(F(x+h)] - F(x-h)]) \underset{h\to 0}{\longrightarrow} f(x)$$

Il faut donc faire dépendre h de la taille de l'échantillon, et le faire tendre vers 0 quand $n \to +\infty$, de sorte que $\hat{f}(x)$ soit un estimateur asymptotiquement sans biais de $f(x)$. h sera donc dorénavant noté h_n.

Remarque : La grande différence par rapport à l'histogramme est qu'il n'y a pas de classe fixée à l'avance : on crée une classe en chaque point où on veut estimer la densité.

L'estimateur $\hat{f}$ reste une fonction en escalier. Pour obtenir quelque chose de plus lisse, on peut remarquer que :

$$\widehat{f_h}(x) = \frac{1}{2nh_n}\sum_{i=1}^{n} 1_{]x-h_n;\, x+h_{\,n}\,]}(X_i) = \frac{1}{nh_n}\sum_{i=1}^{n}\frac{1}{2}1_{]x-h_n;\, x+h_{\,n}\,]}(X_i)$$

$$= \frac{1}{nh_n}\sum_{i=1}^{n}\frac{1}{2}1_{]-1;\,+1]}\left(\frac{x-X_i}{h_n}\right) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)$$

Où $K(u) = \frac{1}{2}1_{]-1;\,+1]}(u)$.

On remarque aisément que la discontinuité de l'estimateur défini ci-dessus est une conséquence de la discontinuité de la fonction indicatrice. Parzen (1962) a proposé une généralisation de l'idée de Rosenblatt permettant, entre autres, de lisser davantage l'estimation. A la fonction K ci-dessus on substitue une fonction que l'on pourra choisir continue ou dérivable partout, propriété qui se transfère à la fonction $\widehat{f_h}(x)$. En d'autres termes on fera entrer ou sortir les points x_i «en douceur» quand on déplace la fenêtre.

Définition 8 : Un estimateur à noyau de la densité f est une fonction $\hat{f}_h^K$ définie par :

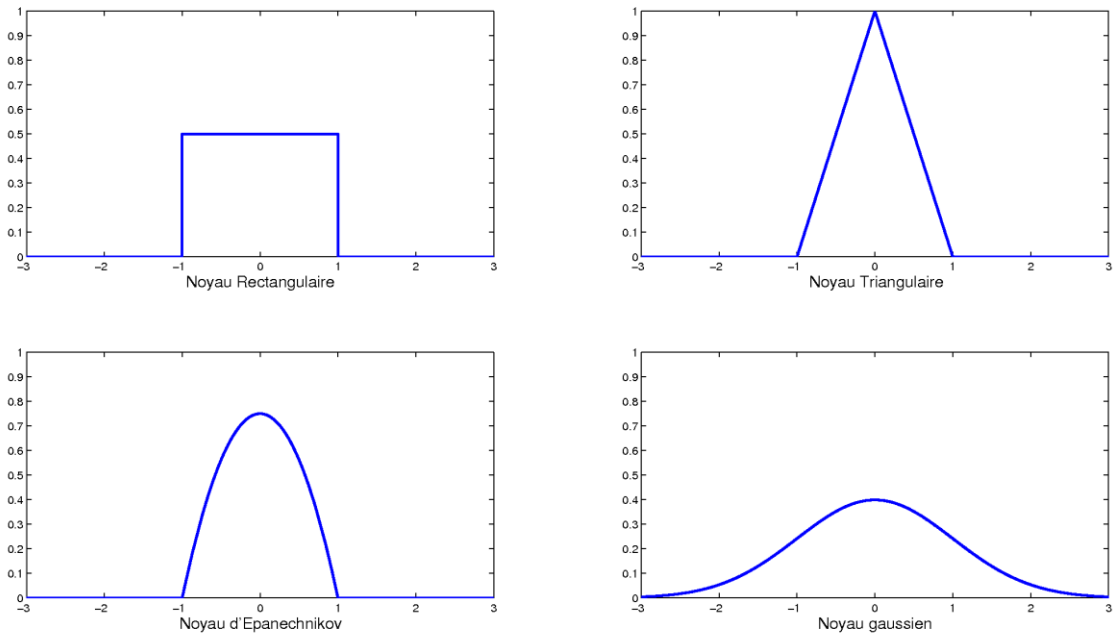
$$\hat{f}_h^K(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

où $\{h_n\}_{n\geq 1}$ est une suite de réels positifs appelés ***paramètres de lissage*** ou ***largeurs de la fenêtre***, qui tend vers 0 quand n tend vers l'infini, et K est une densité de probabilité appelée ***noyau***.

Les noyaux les plus communs sont :

- Le noyau rectangulaire (de Rosenblatt) : $K(u) = \frac{1}{2}1_{]-1;\,+1]}(u)$. C'est celui qui donne l'estimateur de type histogramme.
- Le noyau triangulaire : $K(u) = (1 - |u|)1_{]-1;\,+1]}(u)$.
- Le noyau gaussien : $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, u \in IR$.
- Le noyau d'Epanechnikov : $K(u) = \frac{3}{4}(1 - u^2)\, 1_{]-1;\,+1]}(u)$.
- Le noyau de Tukey ou biweight : $K(u) = \frac{15}{16}(1 - u^2)^2\, 1_{]-1;\,+1]}(u)$.

Les courbes de ces noyaux sont présentées ci-dessous :



Remarques : *i)* Dans l'estimation de $f(x)$ par le noyau rectangulaire, le même poids est accordé à toutes les observations comprises entre $x - h\ et\ x + h$. Dans les autres noyaux, le poids d'une observation est d'autant plus fort qu'elle est proche de x.

ii) $\hat{f}_h^K$ a les mêmes propriétés de continuité et de différentiabilité que K. Par exemple, si K est le noyau gaussien $\hat{f}_h^K$ admet des dérivées de tous ordres.

Lemme 2 : Un estimateur à noyau est une densité.

Démonstration :

1) L'estimateur à noyau est positif et continu car la somme des fonctions positives et continues et elle-même une fonction positive et continue

2)

$$\int \hat{f}_h^K(x)dx = \int \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)dx$$

$$= \frac{1}{nh_n}\sum_{i=1}^{n}\int K\left(\frac{x-X_i}{h_n}\right)dx$$

$$= \frac{1}{nh_n}\sum_{i=1}^{n}\int K(u)h_n\,du = \frac{1}{n}\sum_{i=1}^{n} 1 = 1$$

La troisième égalité est due au changement de variable $u = \frac{X_i - x}{h_n}$.

Pour choisir quel noyau prendre et surtout choisir le paramètre de lissage h_n, il faut étudier la qualité de l'estimation de f par $\hat{f}_h^K$.

Propriété 9 : Si K est la densité d'une loi de probabilité symétrique par rapport à l'origine et de variance μ_2, si f admet des dérivées continues de tous ordres, alors, quand n tend vers l'infini, on a :

- $\mathbb{E}\left[\hat{f}_h^K(x)\right] - f(x) \sim \frac{h_n^2\mu_2}{2} f''(x)$.
- $Var\left[\hat{f}_h^K(x)\right] \sim \frac{f(x)}{nh_n}\int_{-\infty}^{+\infty} K^2(u)\,du$.
- $EQMI\left(\hat{f}_h^K\right) \sim \frac{h_n^4\mu_2^2}{4}\int_{-\infty}^{+\infty} f''^2(x)\,dx + \frac{1}{nh_n}\int_{-\infty}^{+\infty} K^2(u)\,du$.

Démonstration :

Dans tout ce qui suit, on considère h au lieu de h_n

1) $\mathbb{E}\left[\hat{f}_h^K(x)\right] = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{x-X_i}{h_n}\right)\right] = \frac{1}{nh}\sum_{i=1}^{n}\int K\left(\frac{x-y}{h}\right)f(y)dy$

$$= \frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(y)dy = \frac{1}{h}\int K(u)\,f(x-hu)hdu$$

La dernière égalité est due en posant $u = \frac{x-y}{h} => y = x - hu\ et\ dy = -hdu$.

En effectuant un développement limité d'ordre 2, il vient :

$$\mathbb{E}\left[\hat{f}_h^K(x)\right] = \int K(u)\,f(x-hu)hdu$$

$$= \int K(u)\left[f(x) - (hu)f'(x) + \frac{(uh)^2}{2}f''(\zeta_u)\right]du$$

$$= f(x)\int K(u)du - hf'(x)\int uK(u)du + \frac{h^2}{2}\int u^2K(u)\,f''(\zeta_u)du$$

Il en résulte que :

$$\mathbb{E}\left[\hat{f}_h^K(x)\right] - f(x) = \frac{h^2}{2}f''(x)\int u^2K(u)\,du + o(h^2)$$

$$= \frac{h^2\mu_2}{2}f''(x) + o(h^2) \Rightarrow \mathbb{E}\left[\hat{f}_h^K(x)\right] - f(x) \sim \frac{h_n^2\mu_2}{2}f''(x)$$

2) $Var\left[\hat{f}_h^K(x)\right] = \frac{1}{(nh)^2}\left[Var\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\right] = \frac{1}{(nh)^2}\sum_{i=1}^n Var\left[K\left(\frac{x-X_i}{h}\right)\right]$

$$= \frac{n}{(nh)^2}Var\left[K\left(\frac{x-X_1}{h}\right)\right]$$

On a la dernière égalité puisque $y_i = \frac{x-X_i}{h}$ sont des v.a. *iid* . Posons $Z = \frac{1}{h}K\left(\frac{x-X_1}{h}\right)$ on aura

$$\frac{1}{n}Var(Z) = \frac{1}{n}\mathbb{E}(Z^2) - \frac{1}{n}\mathbb{E}(Z)^2$$

i) $\frac{1}{n}\mathbb{E}(Z^2) = \frac{1}{n}\int\frac{1}{h^2}\left[K\left(\frac{x-y}{h}\right)\right]^2 f(y)dy = \frac{1}{nh}\int[K(u)]^2 f(x+hu)du$

Et $\frac{1}{n}\mathbb{E}(Z)^2 = \frac{1}{n}[\int K(u)f(x+hu)du]^2$

On remarque que

$$\begin{cases}\frac{1}{n}\mathbb{E}(Z^2) & \xrightarrow[n\to+\infty]{} 0\\ \frac{1}{n}\mathbb{E}(Z)^2 & \xrightarrow[nh\to+\infty]{} 0\end{cases} \Rightarrow \hat{f}_h^K(x) \xrightarrow[\substack{h\to 0\\ nh\to+\infty}]{n\to+\infty}{}^{MQ} f(x)$$

Les mêmes conditions nécessaires pour l'histogramme.

ii) Le terme $\frac{1}{n}\mathbb{E}(Z)^2$ est d'ordre $\frac{1}{n}$ c.à.d. $\approx O\left(\frac{1}{n}\right)$.

En utilisant le développement de Taylor : $f(x+hu) = f(x) + uhf'(x) + o(h)$, on obtient :

$$\frac{1}{n}\mathbb{E}(Z^2) = \frac{1}{nh}f(x)\int[K(u)]^2du + O\left(\frac{1}{n}\right)$$

D'où

$$Var\left[\hat{f}_h^K(x)\right] = \frac{1}{nh}\,f(x)\int[K(u)]^2du + O\left(\frac{1}{n}\right)$$

$$\Rightarrow\ Var\left[\hat{f}_h^K(x)\right] \sim \frac{f(x)}{nh}\int_{-\infty}^{+\infty}K^2(u)\,du$$

3) Finalement, l'erreur quadratique moyenne intégrée est :

$$EQMI\left(\hat{f}_h^K\right) = \int_{-\infty}^{+\infty}EQM_x\left(\hat{f}_h^K\right)dx = \int_{-\infty}^{+\infty}\left[\left(\mathbb{E}\left[\hat{f}_h^K(x)\right] - f(x)\right)^2 dx + Var\left[\hat{f}_h^K(x)\right]\right]dx$$

$$= \int_{IR}\left(\frac{h^4}{4}[f''(x)]^2\left[\int u^2K(u)\,du\right]^2 + \frac{f(x)}{nh}\int_{-\infty}^{+\infty}K^2(u)\,du\right)dx + o(h^4) + O\left(\frac{1}{n}\right)$$

$$= \frac{h^4\mu_2^2}{4}\int_{-\infty}^{+\infty}f''^2(x)\,dx + \frac{1}{nh}\int_{-\infty}^{+\infty}K^2(u)\,du\int f(x)dx\ o(h^4) + O\left(\frac{1}{n}\right)$$

$$=> EQMI\left(\hat{f}_h^K\right) \sim \frac{h^4 \mu_2^2}{4} \int_{-\infty}^{+\infty} f''^2(x)\, dx + \frac{1}{nh} \int_{-\infty}^{+\infty} K^2(u)\, du$$

Remarque : On remarque que, dans l'erreur quadratique moyenne intégrée, le terme de biais est une fonction croissante de h_n, alors que le terme de variance est une fonction décroissante de h_n :

- Si h_n est grand, la variance sera faible, mais le biais sera fort.
- Si h_n est petit, c'est l'inverse.
- La valeur de h_n optimale, qui minimise l'EQMI, réalise donc un compromis entre biais et variance.

Cette valeur optimale est une fonction de f, qui est inconnue. On ne peut donc en donner qu'une valeur approchée. En pratique, on choisit souvent :

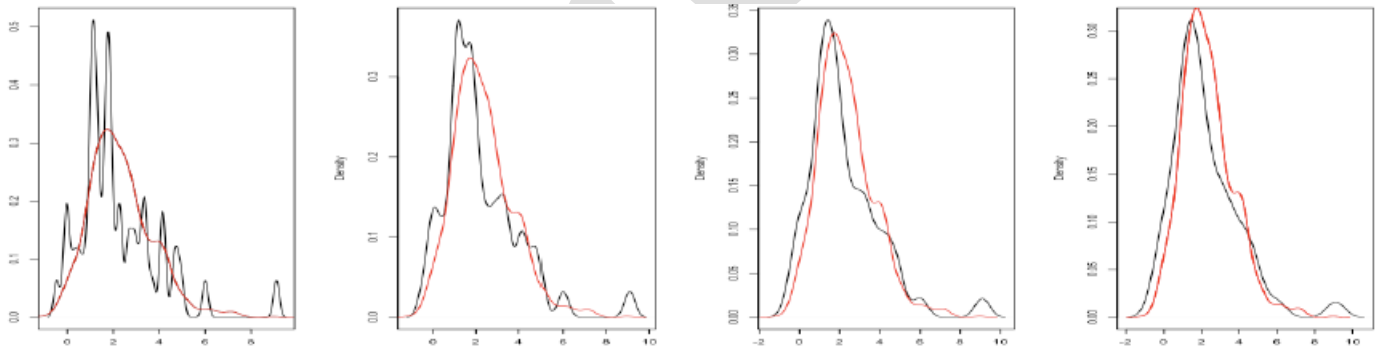
$$h_n = \left(\frac{4}{3}\right)^{1/5} n^{-1/5} \min\left(s'_n, \frac{1}{1.34}\left(\tilde{q}_{n,\frac{3}{4}} - \tilde{q}_{n,\frac{1}{4}}\right)\right)$$

Exemple1 :

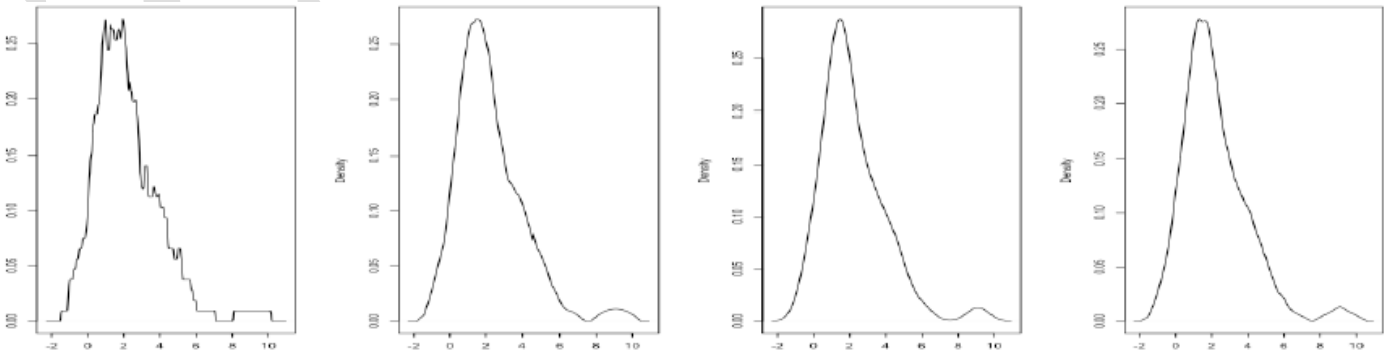
Pour un certain échantillon, on a déterminé une estimation de la densité pour plusieurs valeurs de h : 2, 3/2,

1, 1/2 et plusieurs noyaux gaussien, rectangulaire, triangulaire, epanechnikov.

1. Associer une valeur de h à chacun des graphes suivants :



2. Associer un noyau K à chacun des graphes suivants :



Exemple 2 : Dans l'exemple des niveaux de bruit, l'estimation de densité par la méthode du noyau gaussien avec le paramètre de lissage ci-dessus est donnée par la commande :

> lines(density(bruit,n=200))

On obtient la figure suivante, la densité estimée semble bien proche de celle d'une loi normale.

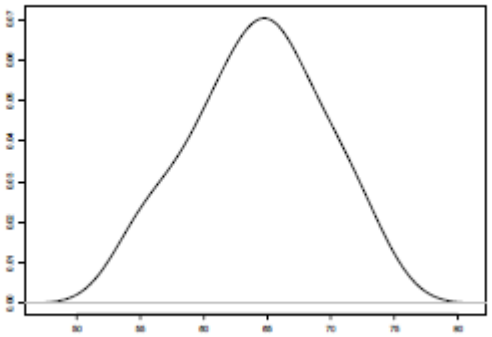


Fig : Estimation de densité par la méthode du noyau gaussien

Exercice 6:

1- Dans le cas de l'estimation de la fonction de densité par la méthode d'histogramme réécrire $\widehat{f_h}(x)$ en considérons la grille d'intervalles de largeur h suivante : $\dots, a_0 - 2h, a_0 - h, a_0, a_0 + h, a_0 + 2h, \dots$.

2- Dans le cas d'une grille régulière $\{a_k\}$ de largeur d'intervalle h, déterminer $\widehat{f_h}(x)$ pour un déplacement de la grille $\{a_k + t\}$ où $t > 0$ et $t < h$.

3- Calculer la valeur moyenne de $\widehat{f_h}(x)$ quand t varie de 0 à h.
4- Montrer qu'on obtient ainsi une estimation par noyau triangulaire.

Exercice 7 : Lissage de F_n.

Considérons le noyau intégré $H(u) = \int_{-\infty}^{u} K(v)dv$.

1) Trouver l'estimateur de la fonction de répartition $\hat{F}_n^K$ en utilisant l'estimateur par noyau de f (c.à.d. $\hat{f}_h^K$).
2) Montrer qu'il y a une analogie entre la fonction de répartition empirique $\mathbb{F}_n$ et l'estimateur par noyau intégré.
3) Trouver le noyau intégré de Rosemblatt et celui de Tuckey.
4) Calculer le biais et la variance de $\hat{F}_n^K$.
5) En déduire l' $EQMI\left(\hat{F}_n^K\right)$.