# Chapitre 3

# Estimation non-paramétrique d'une fonction de répartition et d'une densité

## 3.1 La fonction de répartition empirique

Soit $X \sim F$, avec $F(x) = P\{X \leq x\}$ la fonction de répartition de $X$.

Soit $X_1, X_2, \ldots, X_n$ un échantillon i.i.d. de $F$ (i.i.d.= indépendantes et identiquement distribuées) et

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

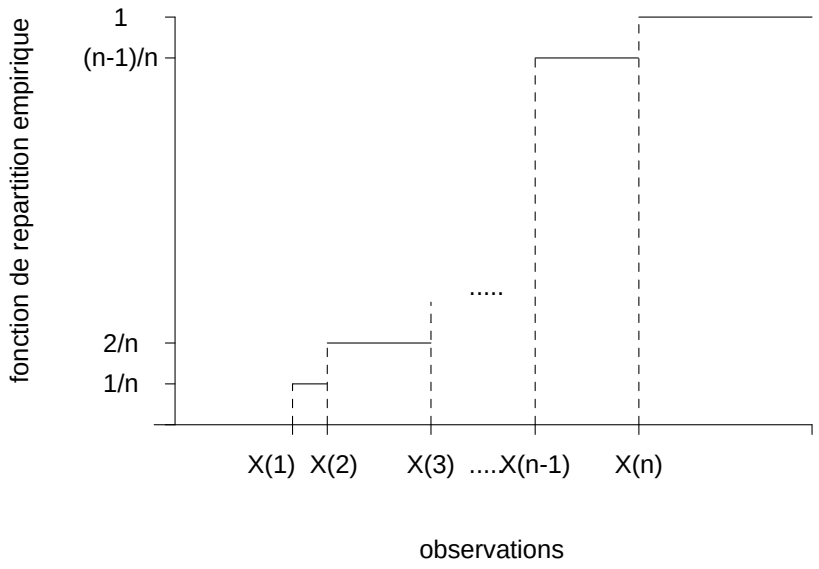
les observations ordonnées.
Supposons que $F$ soit complètement inconnue.

Comment estimer $F$, en se basant sur les observations $X_1, \cdots, X_n$?

Un bon estimateur pour $F$ est la *fonction de répartition empirique*, notée $F_n$, et définie

par

$$
\begin{aligned}
F_n(x) &= \frac{\text{nombre d'observations} \leq x}{n} \\
&= \frac{\#\{i : X_i \leq x\}}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\} \\
&= \frac{1}{n}\sum_{i=1}^{n} I\{X_{(i)} \leq x\} \\
&= \begin{cases} 0 & \text{si } x < X_{(1)} \\ \dfrac{k}{n} & \text{si } X_{(k)} \leq x < X_{(k+1)} \qquad k = 1, \ldots, n-1 \\ 1 & \text{si } x \geq X_{(n)}. \end{cases}
\end{aligned}
$$




**Exemple**: 'Old Faithful geyser data'

durée en minutes de 107 éruptions presque consécutives du geyser Old Faithful au Parc National du Yellowstone, USA (Weisberg (1985), Silverman (1986)).

*Figure 1.1*

### 3.1.1 Propriétés élémentaires de la fonction de répartition empirique

- Biais de l'estimateur $F_n(x)$

  $F_n(x)$ est-elle un estimateur sans biais de $F(x)$?

$$E\{F_n(x)\} = \frac{1}{n}\sum_{i=1}^{n} E\{I\{X_i \leq x\}\} = P\{X \leq x\} = F(x).$$

  Donc, pour tout point $x$, $F_n(x)$ est un estimateur sans biais de $F(x)$.

- Variance de l'estimateur $F_n(x)$.
  Il est facile de montrer que, pour tout $x$, la variance de l'estimateur $F_n(x)$ est donnée par:

$$\mathrm{Var}\{F_n(x)\} = F(x)(1 - F(x)).$$

- La loi des grands nombres nous donne

$$\forall x \in I\!R : \qquad F_n(x) \xrightarrow{\mathrm{P}} F(x), \qquad \text{si } n \to \infty.$$

- Le théorème central-limite donne

$$\frac{nF_n(x) - nF(x)}{\sqrt{nF(x)(1 - F(x))}} \xrightarrow{L} \mathrm{N}(0;1)$$

$$\Longrightarrow \quad \sqrt{n}(F_n(x) - F(x)) \xrightarrow{L} \mathrm{N}\left(0; F(x)(1 - F(x))\right).$$

- La distance de Kolmogorov-Smirnov est définie par

$$\sup_x |F_n(x) - F(x)|.$$

## 3.2 La fonction quantile empirique

Le $p\underline{\text{ème}}$ quantile (ou quantile d'ordre $p$) de la population

$$F^{-1}(p) = \inf\{x : F(x) \geq p\} \qquad 0 < p < 1$$

peut être estimé par

$$F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\},$$

le $p\underline{\text{ème}}$ quantile de la fonction de répartition empirique.

**Exemple**:

*Figure 1.2*

## 3.3 Estimation non-paramétrique d'une densité de probabilité

Comment estimer non-paramétriquement la densité de probabilité $f$, en se basant sur les observations $X_1, \cdots, X_n$? Il existe plusieurs méthodes d'estimation non-paramétrique d'une densité. La méthode la plus simple est celle de l'histogramme. L'objectif de cette section est de décrire quelques autres méthodes importantes d'estimation non-paramétrique d'une densité.

### 3.3.1 Histogramme de densité

On choisi un *point d'origine* $t_0$ et une *longueur de classe* $h$ ($h > 0$).
Les *classes* sont définies par:

$$B_k = [t_k, t_{k+1}[, \qquad k \in \mathbb{Z} \qquad (\text{ la } k\underline{\text{ème}} \text{ classe})$$

avec

$$t_{k+1} = t_k + h, \qquad k \in \mathbb{Z}.$$

Un estimateur de $f$ est donné par

$$\widehat{f}_{\text{H}}(x) = \frac{1}{nh}\#\{i : X_i \text{ est dans la classe qui contient } x\}.$$

Si nous notons le *nombre d'observations dans une classe* $B_k$ par $\nu_k$, l'estimateur du type *histogramme de densité* s'écrit

$$\widehat{f}_{\mathrm{H}}(x) = \frac{\nu_k}{nh} = \frac{1}{nh}\sum_{i=1}^{n} I_{[t_k,t_{k+1}[}(X_i) \qquad \text{pour } x \in B_k$$

- L'histogramme de densité est un estimateur très élémentaire, mais peut quand même déjà donner une première idée assez bonne de la forme de la densité $f$. Par contre, si on voulait utiliser cet estimateur dans d'autres analyses statistiques (comme par exemple l'analyse discriminante, l'estimation d'un taux de hasard, etc) il vaudrait mieux démarrer avec un estimateur plus précis.

- L'histogramme de densité est une fonction étagée, et donc discontinue.

L'estimateur $\widehat{f}_{\mathrm{H}}$ dépend de deux paramètres: le point d'origine $t_0$ et la largeur de classe $h$. Ces deux paramètres peuvent avoir une influence importante sur l'histogramme. Ceci est illustré dans les exemples suivants.

**Exemple**: Old Faithful geyser

*Figure 2.2*

**Exemple**: 'suicide data'

longueurs de 86 périodes d'un traitement psychiatrique subi par des patients utilisés comme référence dans une étude sur les risques de suicide (Copas and Fryer (1980))

*Figure 2.3*

**Exemple**: Buffalo snowfall data

chute de neige annuelle à Buffalo, New York, 1910 – 1972, en pouces (Carmichael (1976) and Parzen (1979))

*Figure 2.4*

*Figure 2.5*

### 3.3.2 Estimateur simple

Rappelons que la densité de probabilité $f$ est égale à la dérivée de la fonction de répartition $F$ (si cette dérivée existe). On peut donc écrire

$$\begin{aligned} f(x) &= \lim_{h\to 0} \frac{F(x+h)-F(x-h)}{2h} \\ &= \lim_{h\to 0} \frac{P\{x-h<X\leq x+h\}}{2h} \end{aligned}$$

Un estimateur de $f(x)$ est alors

$$\begin{aligned} \widehat{f}(x) &= \frac{1}{2h}\frac{\#\{i : x-h<X_i\leq x+h\}}{n} \\ &= \frac{1}{2hn}\sum_{i=1}^{n} I\{x-h<X_i\leq x+h\} \\ &= \frac{1}{2hn}\sum_{i=1}^{n} I\{-1\leq \frac{x-X_i}{h}<1\}. \end{aligned}$$

Notons que cet estimateur peut encore s'écrire comme

$$\widehat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left(\frac{x-X_i}{h}\right)$$

où

$$w(y)=\begin{cases} 1/2 & \text{si } y\in[-1,1[ \\ 0 & \text{sinon.} \end{cases}$$

La construction de l'estimateur $\widehat{f}(\cdot)$ est illustrée dans l'exemple ci-dessous.

*Figure 2.8*

L'influence du paramètre $h$, le *paramètre de lissage* est montrée dans l'exemple ci-dessous.

*Figure 2.9*

**Exemple**: Old Faithful geyser data

*Figure 2.10*

Quelles sont les propriétés de l'estimateur simple $\widehat{f}(x)$?
Remarquons que

$$\widehat{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}$$

avec $F_n$ la fonction de répartition empirique. Le paramètre de lissage $h$ dépend de la taille de l'échantillon $n$, c'est-à-dire $h = h_n$.
Nous savons que

$$nF_n(x) = \sum_{i=1}^{n} I\{X_i \leq x\} \sim \text{ Bin } (n, F(x))$$

et

$$2nh_n\widehat{f}(x) = nF_n(x+h_n) - nF_n(x-h_n) \sim \text{ Bin } (n, F(x+h_n) - F(x-h_n))$$

$$\Rightarrow E\{2nh_n\widehat{f}(x)\} = n[F(x+h_n) - F(x-h_n)]$$

$$\Rightarrow E\{\widehat{f}(x)\} = \frac{1}{2h_n}[F(x+h_n) - F(x-h_n)].$$

Pour la variance nous trouvons

$$\text{Var}\{2nh_n\widehat{f}(x)\} = n[F(x+h_n) - F(x-h_n)][1 - F(x+h_n) + F(x-h_n)]$$

$$\Rightarrow \text{Var}\{\widehat{f}(x)\} = \frac{1}{4nh_n^2}[F(x+h_n) - F(x-h_n)][1 - F(x+h_n) + F(x-h_n)].$$

Remarquons que, si $n \to \infty$ et $h_n \to 0$, alors

$$E\{\widehat{f}(x)\} \to f(x)$$

et

$$nh_n \cdot \text{Var}\{\widehat{f}(x)\} \to \frac{1}{2}f(x).$$

Le risque quadratique moyen de l'estimateur $\widehat{f}(x)$ de $f(x)$ est donné par

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

Donc, si $h_n \to 0$ et $nh_n \to \infty$ quand $n \to \infty$, on a que

$$E\{\widehat{f}(x) - f(x)\}^2 \to 0$$

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

**Remarques**:

- On n'a plus le problème du choix d'un point d'origine (un point $t_0$) comme dans le cas d'un histogramme de densité.
- L'estimateur

  $$\widehat{f}(x) = \frac{1}{2hn}\sum_{i=1}^{n} I\{x - h < X_i \leq x + h\} = \frac{1}{2hn}\sum_{i=1}^{n} I\{X_i - h \leq x < X_i + h\}$$

  est une fonction discontinue, avec des discontinuités aux points $X_i \pm h$, et constante entre ces points.

### 3.3.3 L'estimateur à noyau

#### Définition et construction

Rappelons l'estimateur simple:

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right)$$

avec

$$w(y) = \begin{cases} 1/2 & \text{si } \ y \in [-1, 1[ \\ 0 & \text{sinon,} \end{cases}$$

la densité de probabilité uniforme sur l'intervalle $[-1, 1[$. Cet estimateur peut être généralisé en remplaçant la fonction de poids $w(\cdot)$ (la densité de probabilité uniforme) par une fonction de poids plus générale $K$ (par exemple une densité de probabilité quelconque). Ceci

résulte en l'estimateur

$$\boxed{\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)}$$

$K$ { la fonction de poids ("weight function")
le noyau ("the kernel function")

$h$ { le paramètre de lissage ("smoothing parameter")
la fenêtre ("the window width")

Souvent on prend pour $K$ une densité de probabilité symétrique.

Construction de l'estimateur:

> En chaque observation $X_i$ on place une 'bosse' (la densité de probabilité $K$). L'estimateur qui en résulte est simplement la somme de ces 'bosses'.
>
> Le noyau $K$ détermine la forme des 'bosses', et la fenêtre $h$ détermine la largeur des 'bosses'.

Le paramètre de lissage $h$ a une grande influence sur la performance de l'estimateur.
Un $h$ trop petit résulte en un estimateur avec une 'bosse' en chaque observation. Un $h$ trop grand résulte en un estimateur qui montre peu de détails.

*Figure 2.11*

*Figure 2.12*

**Exemple**: exemple d'estimateur à noyau pour une densité bimodale.

*Figure 2.13*

**Exemple**: estimateur à noyau pour les données 'Old Faithful' et pour les données de suicide.

*Figure 2.14*

*Figure 2.15*

**Quelques propriétés de l'estimateur à noyau**:
Il est facile de voir que l'estimateur à noyau

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)$$

possède les propriétés suivantes:

- Si $K$ est une densité de probabilité, alors $\widehat{f}$ est aussi une densité de probabilité.
- $\widehat{f}$ a les mêmes propriétés de continuité et de différentiabilité que $K$:
  - •• Si $K$ est continue, $\widehat{f}$ sera une fonction continue.
  - •• Si $K$ est différentiable, $\widehat{f}$ sera une fonction différentiable.
  - •• Si $K$ peut prendre des valeurs négatives, alors $\widehat{f}$ pourra aussi prendre des valeurs négatives.

**Expressions du biais et de la variance**

Considérons l'estimateur à noyau

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}K_h(x-X_i),$$

où nous avons introduit la notation

$$K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right),$$

pour une version transformée de $K$.

Pour calculer le biais de l'estimateur à noyau, remarquons d'abord que

$$\begin{aligned} E\{\widehat{f}(x)\} &= E\{K_h(x-X)\} \qquad \text{car les } X_i \text{ sont identiquement distribuées} \\ &= \int K_h(x-y)f(y)dy. \end{aligned}$$

La convolution entre deux fonctions $f$ et $g$ est définie par

$$(f*g)(x) = \int f(x-y)g(y)dy.$$

Dès lors, nous avons

$$E\{\widehat{f}(x)\} - f(x) = \underbrace{(K_h * f)(x)}_{\substack{\text{'version lissée'} \\ \text{de } f}} - f(x).$$

Pour la variance on calcule

$$\begin{aligned}
\mathrm{Var}\{\widehat{f}(x)\} &= E\{\widehat{f}^2(x))\} - [E\{\widehat{f}(x)\}]^2 \\
&= E\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(x-X_i)K_h(x-X_j)\right\} - \{EK_h(x-X)\}^2 \\
&= \frac{1}{n}E\{K_h^2(x-X)\} + \frac{1}{n^2}n(n-1)\{EK_h(x-X)\}^2 - \{EK_h(x-X)\}^2 \\
&= \frac{1}{n}E\{K_h^2(x-X)\} - \frac{1}{n}[EK_h(x-X)]^2 \\
&= \frac{1}{n}\{EK_h^2(x-X) - [EK_h(x-X)]^2\} \\
&= \frac{1}{n}\{(K_h^2 * f)(x) - (K_h * f)^2(x)\}.
\end{aligned}$$

L'erreur quadratique moyenne (en anglais: "Mean squared error", MSE) de l'estimateur à noyau est donnée par:

$$\begin{aligned}
\mathrm{MSE}\{\widehat{f}(x)\} &= E\{\widehat{f}(x) - f(x)\}^2 \\
&= \mathrm{Var}\{\widehat{f}(x)\} + [\mathrm{Biais}(\widehat{f}(x))]^2 \\
&= \frac{1}{n}\{(K_h^2 * f)(x) - (K_h * f)^2(x)\} \\
&\quad + \{(K_h * f)(x) - f(x)\}^2 \\
&= \frac{1}{n}(K_h^2 * f)(x) + \left(1 - \frac{1}{n}\right)(K_h * f)^2(x) - 2(K_h * f)(x)f(x) + f^2(x).
\end{aligned}$$

L'expression exacte de l'erreur quadratique moyenne integrée (en anglais: "Mean Integrated Squared Error", MISE) peut être obtenue à partir de

$$\mathrm{MISE}\{\widehat{f}\} = \int \mathrm{MSE}\{\widehat{f}(x)\}dx$$

et est égale à

$$\begin{aligned}
\mathrm{MISE}\{\widehat{f}(\cdot)\} &= \frac{1}{n}\int (K_h^2 * f)(x)dx + (1 - \frac{1}{n})\int (K_h * f)^2(x)dx \\
&\quad -2\int (K_h * f)(x)f(x)dx + \int f^2(x)dx.
\end{aligned}$$

Comme

$$\begin{aligned}\int (K_h^2 * f)(x)dx &= \int \frac{1}{h^2}\left\{\int K^2\left(\frac{x-y}{h}\right) f(y)dy\right\}dx \\ &= \frac{1}{h}\int\int K^2(u)f(x-uh)dudx\,, \qquad \text{avec} \quad u=\frac{x-y}{h} \\ &= \frac{1}{h}\int K^2(u)\left\{\int f(x-uh)dx\right\}du \\ &= \frac{1}{h}\int K^2(u)du\,,\end{aligned}$$

nous trouvons

$$\begin{aligned}\text{MISE}\{\widehat{f}(\cdot)\} &= \frac{1}{nh}\int K^2(u)du + \left(1-\frac{1}{n}\right)\int (K_h * f)^2(x)dx \\ &\quad -2\int (K_h * f)(x)f(x)dx + \int f^2(x)dx.\end{aligned}$$

Malgré le fait qu'on ait des expressions exactes pour $\text{MSE}\{\widehat{f}(x)\}$ et $\text{MISE}\{\widehat{f}(\cdot)\}$, ces expressions ne sont pas très attrayantes, car elles dépendent de manière très complexe du paramètre de lissage $h$. Pour cette raison on cherche des expressions asymptotiques qui pourraient dépendre de $h$ de manière plus simple.

### Expressions asymptotiques du biais et de la variance

Une approximation asymptotique de l'espérance de l'estimateur $\widehat{f}(x)$ est donnée (sous certaines conditions sur $f$ et $K$) par

$$\begin{aligned}E\{\widehat{f}(x)\} &= \int K_h(x-y)f(y)dy \\ &= \int K(u)f(x-uh)du\,, \qquad \text{avec} \quad u=\frac{x-y}{h} \quad du=-\frac{1}{h}dy \\ &= \int K(u)[f(x)-f'(x)uh+\frac{1}{2}f''(x)u^2h^2+\ldots]du \qquad \text{par Taylor} \\ &= f(x)\int K(u)du - f'(x)h\int K(u)udu \\ &\quad +\frac{1}{2}f''(x)h^2\int K(u)u^2du + o(h^2).\end{aligned}$$

Supposons maintenant que le noyau $K$ satisfait

$$K \geq 0 \qquad \int K(u)du = 1 \qquad \int K(u)udu = 0 \qquad 0 < \int K(u)u^2du < \infty.$$

Alors

$$E\{\widehat{f}(x)\} - f(x) = \frac{1}{2}f''(x)h^2 \int K(u)u^2du + o(h^2)$$

Comme

$$\mathrm{Var}\{\widehat{f}(x)\} = \frac{1}{n}\left\{EK_h^2(x-X) - [EK_h(x-X)]^2\right\}$$

et

$$\begin{aligned} EK_h^2(x-X) &= \frac{1}{h^2}\int K^2\left(\frac{x-y}{h}\right)f(y)dy \\ &= \frac{1}{h}\int K^2(u)f(x-uh)du\,, \qquad \text{avec} \quad u = \frac{x-y}{h} \\ &= \frac{1}{h}\int K^2(u)[f(x) - f'(x)hu + \ldots]du\,, \qquad \text{par Taylor} \\ &= \frac{1}{h}f(x)\int K^2(u)du - f'(x)\int K^2(u)udu + o(1) \end{aligned}$$

nous trouvons que

$$\mathrm{Var}\{\widehat{f}(x)\} = \frac{1}{nh}f(x)\int K^2(u)du + o\left(\frac{1}{nh}\right).$$

Nous avons donc établi que

$$\mathrm{Biais}\{\widehat{f}(x)\} = \frac{1}{2}f''(x)\mu_2h^2 + o(h^2) \qquad \mu_2 = \int K(u)u^2du$$

$$\mathrm{Var}\{\widehat{f}(x)\} = \frac{1}{nh}f(x)R(K) + o\left(\frac{1}{nh}\right) \qquad R(K) = \int K^2(u)du$$

Si $h = h_n \to 0$ quand $n \to \infty$, alors

$$\mathrm{Biais}\{\widehat{f}(x)\} \to 0 \ \text{ si } \ n \to \infty.$$

Si $h = h_n \to 0$ et $nh_n \to \infty$ quand $n \to \infty$, alors

$$\mathrm{Var}\{\widehat{f}(x)\} \to 0 \ \text{ as } \ n \to \infty.$$

Remarquons que

Si $h$ décroît alors le (bias)$^2$ $\searrow$ et la variance $\nearrow$

Si $h$ augmente alors le (bias)$^2$ $\nearrow$ et la variance $\searrow$

Il faut donc essayer de choisir un $h$ qui fasse un compromis entre le (bias)$^2$ et la variance.

Les expressions asymptotiques du biais et de la variance de $\widehat{f} = \widehat{f}_n$ nous premettent de trouver des expressions asymptotiques pour la MSE et la MISE. Rappelons ces expressions asymptotiques du biais et de la variance:

$$\begin{aligned} \text{Biais}\{\widehat{f}_n(x)\} &= \frac{1}{2} f''(x) h^2 \mu_2 + o(h^2) \\ \text{Var}\{\widehat{f}_n(x)\} &= \frac{1}{nh} f(x) R(K) + o\left(\frac{1}{nh}\right), \end{aligned} \tag{3.1}$$

où $\mu_2 = \int K(u) u^2 du$ et $R(K) = \int K^2(u) du$, où $R(g) = \int g^2(u) du$, pour une fonction $g$ de carré intégrable.

Ces expressions ont été obtenues sous certaines conditions sur $K$:

$$K(t) \geq 0 \qquad \int K(u) du = 1 \qquad \int K(u) u du = 0, \qquad 0 < \int u^2 K(u) du < \infty$$

et en supposant que la densité de probabilité $f$ avait toutes les dérivées (continues) nécessaires.

A partir de (3.1) on peut obtenir facilement les approximations asymptotiques suivantes pour la MSE et la MISE

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

sous des conditions appropriées d'intégrabilité de $f$ et ses dérivées.

On note l'approximation asymptotique de la MSE par

$$\text{AMSE}\{\widehat{f}_n(x)\} = \frac{1}{4} h^4 \mu_2^2 \{f''(x)\}^2 + \frac{1}{nh} f(x) R(K), \tag{3.2}$$

et l'approximation asymptotique de la MISE par

$$\text{AMISE}\{\widehat{f}_n(\cdot)\} = \frac{1}{4} h^4 \mu_2^2 R(f'') + \frac{1}{nh} R(K). \tag{3.3}$$

### Choix théoriques optimaux du paramètre de lissage

Pour le paramètre de lissage on fait la distinction entre

| | |
|---|---|
| $h$ | paramètre de lissage constant (ou global) |
| $h(x)$ | paramètre de lissage variable (local). |

Ces choix différents du paramètre de lissage résultent en les estimateurs à noyau suivants:

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)$$
$$\widehat{f}_{n,L}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(x)}K\left(\frac{x-X_i}{h(x)}\right).$$

Le choix $h(x)$ implique qu'un noyau différent est utilisé en chaque point. Ceci est illustré dans l'exemple ci-dessous.

*Figure 3.1*

Nous allons ensuite décrire des choix théoriques optimaux des paramètres de lissage $h$ et $h(x)$.

Un critère approprié pour sélectionner un paramètre de lissage constant $h$ est la MISE. Le paramètre de lissage optimal est la valeur de $h$ qui minimise la MISE. Notons cette valeur par

$$h_{\text{MISE}}.$$

Une approximation asymptotique de $h_{\text{MISE}}$ est donnée par

$$h_{\text{AMISE}},$$

la valeur de $h$ qui minimise $\text{AMISE}\{\widehat{f}_n(\cdot)\}$.

Il est facile de vérifier à partir de (3.3) que

$$\boxed{h_{\text{AMISE}} = \left\{\frac{R(K)}{\mu_2^2 R(f'')}\right\}^{1/5} n^{-1/5}}$$

et

$$h_{\text{MISE}} \sim \left\{ \frac{R(K)}{\mu_2^2 R(f'')} \right\}^{1/5} n^{-1/5},$$

c'est-à-dire $\lim_{n\to\infty} \frac{h_{\text{MISE}}}{h_{\text{AMISE}}} = 1$.

Remarquons que si $f$ montre des changements rapides, alors $R(f'')$ sera grand, et $h_{\text{AMISE}}$ sera petit.

Un critère approprié pour sélectionner un paramètre de lissage variable (local) $h(x)$ est la mesure de performance locale $\text{MSE}\{\widehat{f}_{n,L}(x)\}$. Nous introduisons les notations suivantes:

$$h_{\text{MSE}}(x) = \text{argmin}_h \text{MSE}\{\widehat{f}_{n,L}(x)\}$$

et

$$h_{\text{AMSE}}(x) = \text{argmin}_h \text{AMSE}\{\widehat{f}_{n,L}(x)\}.$$

A partir de (3.2) nous trouvons que

$$\boxed{h_{\text{AMSE}}(x) = \left\{ \frac{f(x)R(K)}{\mu_2^2\{f''(x)\}^2} \right\}^{1/5} n^{-1/5},}$$

sous condition que $f''(x) \neq 0$.

Les choix $h_{\text{AMISE}}$ et $h_{\text{AMSE}}(x)$ sont des choix théoriques, qui ne sont pas utilisables en pratique car ils dépendent des quantités inconnues $f$ et $f''$. Nous allons maintenant décrire quelques choix optimaux pratiques pour un paramètre de lissage constant et un paramètre de lissage variable (local).

**Choix pratiques du paramètre de lissage**

**La règle simple de référence à une distribution normale**

Rappellons l'expression pour le paramètre de lissage optimal constant:

$$h_{\text{AMISE}} = \left\{ \frac{R(K)}{\mu_2^2 R(f'')} \right\}^{1/5} n^{-1/5}. \tag{3.4}$$

Supposons que $f$ appartient à une famille de distributions normales $\mathrm{N}(\mu;\sigma^2)$, de moyenne $\mu$ et variance $\sigma^2$ inconnues. Alors

$$f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \qquad \text{avec } \varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^{2/2}},$$
la densité de probabilité normale réduite

et

$$f''(x) = \frac{1}{\sigma^3}\varphi''\left(\frac{x-\mu}{\sigma}\right).$$

La quantité inconnue $R(f'')$ s'écrit alors

$$\begin{aligned}
R(f'') = \int (f''(x))^2 dx &= \frac{1}{\sigma^6}\int \left\{\varphi''\left(\frac{x-\mu}{\sigma}\right)\right\}^2 dx \\
&= \frac{1}{\sigma^5}\int \{\varphi''(v)\}^2 dv \\
&\qquad \varphi(v) = \frac{1}{\sqrt{2\pi}}e^{-v^2/2} \\
&\qquad \Rightarrow \varphi'(v) = -\frac{v}{\sqrt{2\pi}}e^{-v^2/2} \\
&\qquad \Rightarrow \varphi''(v) = \frac{1}{\sqrt{2\pi}}(v^2-1)e^{-v^2/2} \\
&= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{\int_{-\infty}^{+\infty} v^4 e^{-v^2}dv - 2\int_{-\infty}^{+\infty} v^2 e^{-v^2}dv + \int_{-\infty}^{+\infty} e^{-v^2}dv\right\} \\
&= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{2}\int_{-\infty}^{+\infty} v^2 e^{-v^2}dv + \int_{-\infty}^{+\infty} e^{-v^2}dv\right\} \\
&\qquad \text{posons } u = \sqrt{2}v \Rightarrow du = \sqrt{2}dv \\
&= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{2}\int_{-\infty}^{+\infty} \frac{u^2}{2}e^{-u^2/2}\frac{du}{\sqrt{2}} + \frac{1}{\sqrt{2}}\int_{-\infty}^{+\infty} e^{-u^2/2}du\right\} \\
&= \frac{1}{\sigma^5}\frac{1}{2\pi}\left\{-\frac{1}{4}\cdot\sqrt{\pi} + \sqrt{\pi}\cdot 1\right\} \\
&= \frac{1}{\sigma^5}\frac{1}{2\pi}\frac{3}{4}\sqrt{\pi} = \frac{1}{\sigma^5}\frac{3}{8\sqrt{\pi}}.
\end{aligned}$$

Donc, en faisant référence à une densité de probabilité normale, l'expression du paramètre de lissage optimal asymptotique devient

$$h_{\mathrm{AMISE}} = \left\{\frac{8\sqrt{\pi}\,R(K)}{3\mu_2^2}\right\}^{1/5}\sigma n^{-1/5}.$$

Le paramètre de lissage du type *"normal reference"* est défini par

$$\boxed{\widehat{h}_{\mathrm{NR}} = \left\{ \frac{8\sqrt{\pi}\, R(K)}{3\mu_2^2} \right\}^{1/5} \widehat{\sigma} n^{-1/5},} \tag{3.5}$$

où $\widehat{\sigma}$ est un estimateur de $\sigma$, l'écart-type de la population $X$. Ce paramètre de lissage est très simple ("Rule-of-Thumb" bandwidth selector).
Quelques choix possibles pour $\widehat{\sigma}$ sont donnés ci-dessous.

- L'écart-type empirique

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

- L'écart interquartile empirique standardisé:

$$\begin{aligned} \frac{\text{l'ecart interquartile empirique}}{\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})} &\equiv \frac{R}{\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})} \\ &\simeq \frac{R}{1.349}. \end{aligned}$$

où $\Phi(\cdot)$ est la fonction de répartition d'une normale réduite.

Remarquons que $\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)$ est l'écart interquartile d'une variable aléatoire normale réduite. La motivation pour la standardisation utilisant cette quantité est simple:

Si $X \sim N(\mu; \sigma^2)$, alors $Z = \dfrac{X-\mu}{\sigma} \sim N(0;1)$ et

$$\begin{aligned} & P\{\Phi^{-1}\left(\frac{1}{4}\right) \leq Z \leq \Phi^{-1}\left(\frac{3}{4}\right)\} = 0.50 \\ \Longleftrightarrow \quad & P\{\Phi^{-1}\left(\frac{1}{4}\right) \leq \frac{X-\mu}{\sigma} \leq \Phi^{-1}\left(\frac{3}{4}\right)\} = 0.50 \\ \Longleftrightarrow \quad & P\{\sigma\Phi^{-1}\left(\frac{1}{4}\right) + \mu \leq X \leq \sigma\Phi^{-1}\left(\frac{3}{4}\right) + \mu\} = 0.50 \end{aligned}$$

Alors

l'écart interquartile de $X$ est

$$F^{-1}\left(\frac{3}{4}\right) - F^{-1}\left(\frac{1}{4}\right) = \sigma[\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)]$$

ce qui justifie l'estimateur proposé.

On propose d'utiliser le mimimum entre $S$ et $R/1.349$, c'est-à-dire d'utiliser le paramètre de lissage suivant:

$$\boxed{\widehat{h}_{\mathrm{NR}} = \left\{ \frac{8\sqrt{\pi}R(K)}{3\mu_2^2} \right\}^{1/5} \min(S, \frac{R}{1.349}) n^{-1/5}.} \tag{3.6}$$

Voici, pour quelques noyaux, l'expression de $\widehat{h}_{\mathrm{NR}}$:

| noyau $K$ | paramétre de lissage pratique $\widehat{h}_{\mathrm{NR}}$ |
|---|---|
| densité normale réduite $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ | $\widehat{h}_{\mathrm{NR}} = 1.06 \min(S, \frac{R}{1.349}) n^{-1/5}$ |
| noyau "Epanechnikov" $\frac{3}{4}(1-x^2) I\{\lvert x\rvert \leq 1\}$ | $\widehat{h}_{\mathrm{NR}} = 2.34 \min(S, \frac{R}{1.349}) n^{-1/5}$ |
| noyau "biweight" $\frac{15}{16}(1-x^2)^2 I\{\lvert x\rvert \leq 1\}$ | $\widehat{h}_{\mathrm{NR}} = 2.78 \min(S, \frac{R}{1.349}) n^{-1/5}$ |

**La méthode de validation croisée**

La méthode de validation croisée (en anglais: cross-validation) du type moindres carrés a été introduite par Rudemo (1982) et Bowman (1984). Cette méthode permet d'obtenir un paramètre de lissage simple et attrayant. La méthode est motivée par la décomposition suivante de l'erreur quadratique moyenne intégrée $\mathrm{MISE}\{\widehat{f}_n(\cdot)\}$ de l'estimateur à noyau:

$$\begin{aligned} \mathrm{MISE}\{\widehat{f}_n(\cdot)\} &= E[\mathrm{ISE}\{\widehat{f}_n(\cdot)\}] = E \int \{\widehat{f}_n(x) - f(x)\}^2 dx \\ &= E \int \widehat{f}_n^2(x) dx - 2E \int \widehat{f}_n(x) f(x) dx + \int f^2(x) dx. \end{aligned}$$

Remarquons que le terme $\int f^2(x)dx$ ne dépend pas de $h$, et donc minimiser $\text{MISE}\{\widehat{f}_n(\cdot)\}$ par rapport à $h$ est équivalent à minimiser

$$\text{MISE}\{\widehat{f}_n(\cdot)\} - \int f^2(x)dx = E\left[\int \widehat{f}_n^2(x)dx - 2\int \widehat{f}_n(x)f(x)dx\right].$$

L'expression à droite de cette équation est inconnue car elle dépend de la densité inconnue $f$. Un estimateur pour $\int \widehat{f}_n(x)f(x)dx$ est donné par

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i), \tag{3.7}$$

où

$$\widehat{f}_{-i}(x) = \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} K_h(x - X_j),$$

est l'estimateur à noyau basé sur l'échantillon 'réduit' $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, où l'observation $X_i$ à été supprimée. On appelle cet estimateur le *"leave-one-out estimator"*. Le terme "validation croisée" vient du fait qu'une partie de l'échantillon est utilisée pour obtenir l'information sur une autre partie: les observations $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ sont utilisées pour obtenir une idée de $f(X_i)$.

L'estimateur (3.7) est un estimateur sans biais de $E\{\int \widehat{f}_n(x)f(x)dx\}$. En effet,

$$\begin{aligned}
E\{\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i)\} &= \frac{1}{n}\sum_{i=1}^{n} E\{\widehat{f}_{-i}(X_i)\} \\
\text{et } E\{\widehat{f}_{-i}(X_i)\} &= \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} E\{K_h(X_i - X_j)\} \\
&= E\{K_h(X_1 - X_2)\} \\
&= \int\int K_h(x-y)f(x)f(y)dxdy \\
&= \int\left\{\int K_h(x-y)f(y)dy\right\}f(x)dx \\
&= \int E\{\widehat{f}_n(x)\}f(x)dx \\
&= E\{\int \widehat{f}_n(x)f(x)dx\}.
\end{aligned}$$

Ainsi, un estimateur sans biais pour

$$\text{MISE}\{\widehat{f}_n(\cdot)\} - \int f^2(x)dx = E\left[\int \widehat{f}_n^2(x)dx - 2\int \widehat{f}_n(x)f(x)dx\right]$$

est donné par

$$\boxed{\text{LSCV}(h) = \int \widehat{f}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i)} \quad . \tag{3.8}$$

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

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

$$\boxed{\widehat{h}_{\text{LSCV}} = \text{argmin}_h \text{LSCV}(h)} \quad . \tag{3.9}$$

*Figure 4.5*

### 3.3.4 La méthode d'estimation des points les plus proches

Soit $x$ fixé .

Supposons que l'objectif est d'estimer $f(x)$

Considérons la distance $d(x, y) = |x - y|$

Notons par $d_1(x) \leq d_2(x) \leq \ldots \leq d_n(x)$ les distances ordonnées de $x$ aux points d'observation.

Considerons l'intervalle $]x - r, x + r[ \quad (r > 0)$

Le nombre attendu d'observations dans l'intervalle $]x-r, x+r[$ est

$$\begin{aligned} E\left\{\sum_{i=1}^{n} I\{x-r<X_i<x+r\}\right\} &= nP\{x-r<X<x+r\} \\ &= n\int_{x-r}^{x+r} f(t)dt \\ &\simeq 2nr\ f(x). \end{aligned}$$

Prenons $r = d_k(x)$, avec $k > 0$ un nombre entier, fixé. Dans ce cas, nous avons

$$2nd_k(x)f(x) \simeq k-1$$

ce qui peut motiver l'estimateur suivant de $f(x)$:

$$\widehat{f}_{\mathrm{NN}}(x) = \frac{k-1}{2nd_k(x)} = \frac{1}{nd_k(x)}\sum_{i=1}^{n} w\left(\frac{x-X_i}{d_k(x)}\right)$$

$$\downarrow$$

$$\boxed{\widehat{f}_{NN}(x) = \frac{1}{nd_k(x)}\sum_{i=1}^{n} K\left(\frac{x-X_i}{d_k(x)}\right)}$$

l'estimateur du type
des points les plus proches
("nearest neighbour estimator")

Quelques propriétés de cet estimateur:

- $\widehat{f}_{\mathrm{NN}}(x)$ resemble à un estimateur à noyau avec un paramètre de lissage variable $d_k(x)$:
  - $k$ détermine un paramètre de lissage discret
  - le paramètre de lissage $d_k(x)$ est déterminé par le nombre d'obervations dans le voisinage de $x$.
- $d_k(x)$ est une fonction continue et positive
  la fonction $d_k(x)$ est non-différentiable aux points $\frac{1}{2}(X_{(j)}+X_{(j+\ell)}), j = 1, \cdots, n-1$, $\ell = 2-j, \cdots, n-j$.

  $\Longrightarrow$ $\widehat{f}_{\mathrm{NN}}(x)$ est positive et continue, mais non-différentiable aux points $\frac{1}{2}(X_{(j)} + X_{(j+\ell)})$.
- pour $x < X_{(1)}$, $d_k(x) = X_{(k)} - x$
  pour $x > X_{(n)}$, $d_k(x) = x - X_{(n-k+1)}$

les queues de $\widehat{f}_{\mathrm{NN}}$ se comportent comme $\dfrac{1}{x}$, et dès lors $\displaystyle\int_{\mathbb{R}} \widehat{f}_{\mathrm{NN}}(x)dx = \infty$.

Par conséquent, l'estimateur du type "points les plus proches", n'est pas un très bon estimateur si l'objectif est une estimation **globale** de $f$.

*Figure 2.16*

**Exemple**: 'nearest neighbour estimator' pour les données 'Old Faithful geyser'

*Figure 2.17*