Statistique non paramétrique

Meriem Henkouche

Département de mathématiques
U.S.T.O.M.B

Table des matières

Introduction

L'objectif de ce polycopié de cours est de présenter des notions en statistique non paramétrique. En opposition à la statistique paramétrique, la statistique non paramétrique n'est pas décrite par un nombre fini de paramètres : Les hypothèses sont qu'il n'y a pas d'hypothèses sur la nature de la distribution des variables aléatoires étudiées. Des exemples d'utilisation sont dans le cas où on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique et dans le cas où on n'a aucune idée du modèle. La statistique non paramétrique traite le cas où le nombre de variables est trop grand et qu'un modèle paramétrique est non utilisable (trop de paramètres à estimer).

Ce document est destiné aux étudiants de niveau master 2 mathématiques de l'Université des Sciences et de la Technologie d'Oran Mohammed Boudiaf (U.S.T.O.M.B), ayant déjà suivi un cours de statistique inférentielle (master 1) où les tests statistiques ont été abordés (p-valeur etc ...). Des notions en langage **R** introduites en travaux pratiques peuvent également aider à approfondir les résultats présentés de ce polycopié qui s'organise en quatre chapitres. Un premier traite de l'estimation de la fonction de répartition et de la densité et de ses propriétés. Le second traite des tests non paramétriques fondamentaux. Le troisième concerne les éléments de base de la régression non paramétrique. Le quatrième chapitre est consacré aux méthodes de rééchantillonnage.

Chapitre 1

Estimations relatives à une distribution

1.1 Introduction

En statistique non paramétrique, on veut estimer la fonction de répartition F, ou la fonction de densité f dans le cas continue, alors que dans le cas paramétrique, on estime un paramètre θ inconnu. Dans ce chapitre selon ([10], [11], [12]) et autres ([23], [33], [35]), les estimations relatives à une distribution, fournissent des propriétés intéressantes comme le biais, la variance et les erreurs quadratiques moyenne et intégrée, ainsi que les développements asymptotiques associés.

1.2 Estimation d'une fonction répartition

On observe $X_1, \ldots, X_n$ n variables aléatoires réelles définies sur un espace probabilisable $(\Omega, \mathscr{A}, P)$ de loi F. On cherche à estimer la loi F. Or P est entièrement décrite par sa fonction de répartition F :

$$F : \mathbb{R} \longrightarrow [0,1]$$

$$x \longrightarrow F(x) = P\left(X^{-1}(]-\infty, x])\right)$$

On note $X \sim F$. On construit un estimateur $\widehat{F}_n$ de F à l'aide des n observations $X_1, \ldots, X_n$. On définit les variables des observations ordonnées par ordre croissant :

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

Supposons que F soit inconnue. La question est : Comment estimer F en se basant sur les observations $X_1, \ldots, X_n$.
La réponse. Un bon estimateur pour F est la fonction de répartition empirique, notée $\widehat{F}_n$, et définie par

$$\widehat{F}_n(x) = \frac{\text{Nombre d'observations } \leq x}{n} = \frac{\#\{i | X_i \leq x\}}{n}$$

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

où

$$\mathbb{1}_{\{X_i \leq x\}} = \begin{cases} 1 & X_i \leq x \\ 0 & \text{sinon} \end{cases}$$

$$\widehat{F}_n(x) = \begin{cases} 0 & x \le X_{(1)} \\ \frac{k}{n} & X_{(k)} \le x < X_{(k+1)} \quad k = 1, \dots, n-1 \\ 1 & x \ge X_{(n)} \end{cases}$$

Où $\#\ A$ désigne le nombre d'éléments de l'ensemble A, ou le cardinal de A.

1.2.1 Propriétés de la fonction empirique

1) **Biais de l'estimateur $\widehat{F}_n(x)$**

On a

$$E(\widehat{F}_n(x)) = \frac{1}{n}\sum_{i=1}^{n} E\left[\mathbb{1}_{\{X_i \le x\}}\right] = n\frac{P(X_i \le x)}{n} = P(X \le x) = F(x).$$

On sait que

$$\mathbb{1}_{\{X_i \le x\}} = \begin{cases} 1 & \text{si } X_i \le x \\ 0 & \text{sinon} \end{cases}$$

donc

$$\mathbb{1}_{\{X_i \le x\}} \sim \mathscr{B}(p)$$

$p = P(X_i \le x)$, $\forall i = 1, \dots, n$. $\mathscr{B}(p)$ désigne la loi de Bernoulli de paramètre p.

Notations 1.2.1 Une variable aléatoire X suit la loi de Bernoulli de paramètre p est notée $X \sim \mathscr{B}(p)$ ou $X \sim \mathscr{B}(1, p)$.

$\mathbb{1}_{\{X_i \le x\}}$	1	0
	p	$1-p$

On a également :

$$E(\mathbb{1}_{X_i \le x}) = p, \quad Var(\mathbb{1}_{X_i \le x}) = p(1-p)$$

avec $P(X_i \le x) = p$

D'où la proposition suivante.

Proposition 1.2.1 *Pour tout $x \in \mathbb{R}$, l'estimateur $\widehat{F}_n(x)$ est un estimateur sans biais de $F(x)$.*

2) **Variance de $\widehat{F}_n(x)$**

Proposition 1.2.2

$$\forall x \in \mathbb{R},\ Var\left[\widehat{F}_n(x)\right] = \frac{F(x)\,(1-F(x))}{n} \tag{1.1}$$

En effet : $\forall n \in \mathbb{N}$, $\widehat{F}_n(x)$ est une somme de variables aléatoires indicatrices indépendantes, chacune suivant la loi de Bernoulli de même paramètre, donc elle suit une loi binomiale $\mathscr{B}(n, p)$ de paramètres n, p. On sait que si $Y \sim \mathscr{B}(n, p)$ alors $E(Y) = np$, $Var(Y) = np(1-p)$.

$$Var\left[\widehat{F}_n(x)\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(\mathbb{1}_{\{X_i \le x\}}) = \frac{nVar(\mathbb{1}_{\{X_i \le x\}})}{n^2}$$

$$= \frac{1}{n}F(x)\,(1-F(x))\,.$$

□

3)

Convergences asymptotiques

Proposition 1.2.3 (*Convergence en probabilité*)

$$\forall\, x \in \mathbb{R},\ \widehat{F}_n(x) \xrightarrow{P} F(x), \quad n \longrightarrow +\infty$$

$$\forall\, x \in \mathbb{R}, \forall\, \epsilon > 0,\ P\big(\big[\big|\widehat{F}_n(x) - F(x)\big| > \epsilon\big]\big) \underset{n\longrightarrow+\infty}{\longrightarrow} 0$$

En effet : Par l'inégalité de Tchebychev on a

$$\forall\, x \in \mathbb{R}, \forall\, \epsilon > 0,\ P\big(\big[\big|\widehat{F}_n(x) - F(x)\big| > \epsilon\big]\big) \leq \frac{\sigma^2}{\epsilon^2}$$

où $\sigma^2 = Var\left[\widehat{F}_n(x)\right]$

$$\frac{\sigma^2}{\epsilon^2} = \frac{F(x)(1-F(x))}{n\epsilon^2} \qquad \underset{n\longrightarrow+\infty}{\longrightarrow} 0.$$

□

Proposition 1.2.4 (*Convergence en loi*)

$$\forall\, x \in \mathbb{R},\ \frac{n\widehat{F}_n(x) - nF(x)}{\sqrt{n(F(x))(1-F(x))}} \xrightarrow[n\longrightarrow+\infty]{\mathscr{L}} N(0,1) \tag{1.2}$$

En effet : Par le Théorème Central Limite on a :

$$\frac{\widehat{F}_n(x) - F(x)}{\sqrt{\frac{F(x)(1-F(x))}{n}}} = \frac{\sqrt{n}(\widehat{F}_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \xrightarrow[n\longrightarrow+\infty]{\mathscr{L}} N(0,1)$$

On déduit

$$\sqrt{n}\Big(\widehat{F}_n(x) - F(x)\Big) \xrightarrow[n\longrightarrow+\infty]{\mathscr{L}} N\Big(0, F(x)(1-F(x))\Big) \tag{1.3}$$

□

4) **La fonction quantile empirique**

Le $p^{\text{ième}}$ quantile d'ordre p de la population défini par

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

peut être estimé par $F_n^{-1}(p) = \inf\{x | F_n(x) \geq p\}$ appelé $p^{\text{ième}}$ quantile empirique.

1.3 Estimation d'une densité de probabilité

On observe toujours $X_1, \ldots, X_n$, n variables réelles de loi F, mais on suppose en plus que F est absolument continue par rapport à la mesure de Lebesgue et on voudrait estimer sa densité f. En général la dérivée de $\widehat{F}_n$ n'est pas une bonne estimation.

1.3.1 Histogramme de densité

L'histogramme tient son origine à John Graunt au XVII siècle. Il fut avec son ami William Petty un des premiers démographes londoniens (1620-1674). La méthode consiste à choisir un point d'origine t_0 et une longueur de classe h $(h > 0)$ les classes sont définies par $B_k = [t_k, t_{k+1}[$, $t \in \mathbb{Z}$. On pose $t_{k+1} = t_k + h$, $t \in \mathbb{Z}$. Un estimateur de f est donné par :

$$\widehat{f}_n(x) = \frac{1}{nh} \# \{i | X_i \text{ est dans la classe qui contient } x\}$$

On note $\nu_k = \mathbb{1}_{\{[t_k, t_{k+1}[\}}(X_i)$ et on aura

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \nu_k, \qquad x \in B_k$$

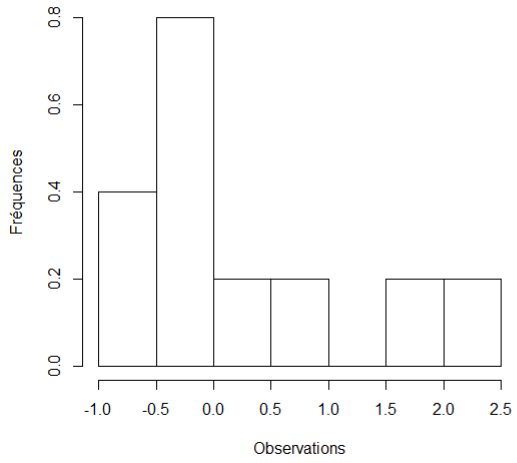

FIGURE 1.1 – Histogramme $(n = 10, h = 0.5)$

1.3.2 Propriété

L'histogramme de densité estimateur très élémentaire, est une fonction étagée donc discontinue. $\widehat{f}_h$ *dépend de deux paramètres, le point d'origine* t_0 *et la largeur* h *de la classe.*

1.3.3 Estimateur simple

$$f(x) = \frac{dF(x)}{dx}$$

$$f(x) = \lim_{h \longrightarrow 0} \frac{F(x+h) - F(x-h)}{2h}. \tag{1.4}$$

Un estimateur de $f(x)$ est alors

$$\widehat{f}(x) = \frac{1}{2h}\frac{\#\{i|x-h<X_i\leq x+h\}}{n} \tag{1.5}$$

$$= \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{1}_{\{x-h<X_i\leq x+h\}} \tag{1.6}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w(\frac{x-X_i}{h}) \tag{1.7}$$

avec

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

Remarque 1.3.1 *Le paramètre de lissage h a une influence sur la valeur de $\widehat{f}(x)$.*

1.3.4 Propriétés de l'estimateur simple

1. ***Le biais***
   On définit

$$\widehat{f}(x) = \lim_{h\longrightarrow 0}\frac{F_n(x+h)-F_n(x-h)}{2h}. \tag{1.8}$$

On estime $f(x)$ par $\widehat{f}(x)$ définie plus haut. On sait que F_n est la fonction de répartition empirique de paramètres $h=h_n$. On a

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

$$nF_n(x) = \sum_{i=1}^{n}\mathbb{1}_{\{X_i\leq x\}}$$

$$nF_n(x) \sim \mathscr{B}\big(n,F(x)\big). \tag{1.9}$$

On rappelle que la notation $X\sim\mathscr{B}\big(n,p\big)$ désigne le fait que X suit la loi binomiale de paramètres n et p. En reprenant (1.8) on a

$$2nh_n\widehat{f}(x) = nF_n(x+h_n)-nF_n(x-h_n).$$

Donc $2nh_n\widehat{f}(x)\sim\mathscr{B}(n,F(x+h_n)-F(x-h_n))$ *car* $nF_n(x)=\sum_{i=1}^{n}\mathbb{1}_{\{X_i\leq x\}}$.

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

$$E\big[\widehat{f}(x)\big] = \frac{1}{2h_n}\left[F(x+h_n)-F(x-h_n)\right] \tag{1.10}$$

Donc :

$$E\big[\widehat{f}(x)\big]\longrightarrow f(x) = \frac{F(x+h)-F(x-h)}{2h}. \tag{1.11}$$

Comme $Biais(\widehat{f}(x)) = E\big[\widehat{f}(x)\big]-f(x)$, *on déduit que le biais tend vers zéro.*

2. ***Variance***
En utilisant (1.9)

$$Var\left[2nh_n\widehat{f}(x)\right] = n\big(F_n(x+h) - F_n(x-h)\big)\big(1 - (F_n(x+h) - F_n(x-h))\big)$$

$$Var\left[\widehat{f}(x)\right] = \frac{n}{4n^2h_n^2}\Big(F(x+h_n) - F(x-h_n)\Big)\Big(1 - (F(x+h_n) - F(x-h_n))\Big)$$

$$Var\left[\widehat{f}(x)\right] \underset{n\longrightarrow+\infty}{\longrightarrow} \frac{1}{2}\left[\frac{F_n(x+h) - F_n(x-h)}{2h_n}\right] \times \left[1 - \Big(F(x+h_n) - F(x-h_n)\Big)\right]$$

Quand h_n *tend vers zéro* $(1 - (F(x+h_n) - F(x-h_n))) \longrightarrow 1$ *et*

$$\frac{F_n(x+h) - F_n(x-h)}{2h_n} \longrightarrow f(x).$$

Donc

$$Var\left[\widehat{f}(x)\right] \underset{h_n\longrightarrow 0}{\longrightarrow} \frac{1}{2}f(x). \tag{1.12}$$

3. ***L'erreur quadratique moyenne (MSE)***
Pour évaluer l'erreur quadratique moyenne ((MSE) Mean Square Error) on calcule la quantité $E\left[\left(\widehat{f}(x) - f(x)\right)^2\right]$

Proposition 1.3.1

$$E\left[(\widehat{f}(x) - f(x))^2\right] = Var(\widehat{f}(x)) + \left[Biais(\widehat{f}(x))\right]^2 \tag{1.13}$$

où $Biais(\widehat{f}(x)) = E\left[\widehat{f}(x)\right] - f(x)$. □

En effet :

$$E\left[(\widehat{f}(x) - f(x))^2\right] = E\left[\left((\widehat{f}(x) - E(\widehat{f}(x)) + E(\widehat{f}(x)) - f(x)\right)^2\right]$$

$$E\left[(\widehat{f}(x) - f(x))^2\right] = E\left[\left(\widehat{f}(x) - E(\widehat{f}(x))\right)^2\right]$$

$$+E\left[2(\widehat{f}(x) - E(\widehat{f}(x))(E(\widehat{f}(x)) - f(x))\right] + E\left[E^2(\widehat{f}(x) - f(x))\right].$$

$$E\left[(\widehat{f}(x) - f(x))^2\right] = E\left[\left(\widehat{f}(x) - E(\widehat{f}(x))\right)^2\right] + E^2\left(\widehat{f}(x) - f(x)\right)$$

Donc

$$E\left[(\widehat{f}(x) - f(x))^2\right] = Var(\widehat{f}(x)) + \Big(Biais(\widehat{f}(x))\Big)^2$$

car le second terme est nul.

Remarque 1.3.2 *Si* $h_n \longrightarrow 0, nh_n \longrightarrow \infty$ *quand* $n \longrightarrow +\infty$ *on a*

$$E\left[\left(\widehat{f}(x) - f(x)\right)^2\right] \longrightarrow 0 \quad \textit{pour tout point} x.$$

On dit que l'estimateur simple $\widehat{f}(x)$ *est un estimateur consistant de* $f(x)$.

1.4 L'estimateur à noyau

L'estimateur à noyau introduit par ([28], 1956) est une classe générale très utilisée d'estimateurs non paramétriques de la densité.

Définition 1.4.1

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

où

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

$w(.)$ *peut être une fonction générale appelée noyau.*

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\Big(\frac{x-X_i}{h}\Big) \tag{1.14}$$

La fonction poids ("weight function") K, h le paramètre de lissage, h la fenêtre.

Remarque 1.4.1

1. *Le paramètre h a une grande influence sur la performance de l'estimateur.*
2. *Avec le logiciel $\boldsymbol{R}$ on peut représenter pour un vecteur x, la fonction de répartition empirique $\widehat{F}_n$ et la densité $\widehat{f}_n$ respectivement par*
   $>ecdf(x)$
   $>plot(ecdf(x))$
   $>density(x)$

Exemples 1.4.1 (i) Le Noyau de Rosenblatt

$$K(u) = \frac{1}{2}\mathbb{1}_{\{|u|\leq 1\}}.$$

Ce noyau est appelé noyau uniforme.

(ii) Le noyau triangulaire

$$K(u) = (1-|u|)\mathbb{1}_{\{|u|\leq 1\}}$$

(iii) Le noyau d'Epanechnikov

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{\{|u|\leq 1\}}$$

(iv) Le noyau quadratique

$$K(u) = \frac{15}{16}(1-u^2)\mathbb{1}_{\{|u|\leq 1\}}$$

(v) Le noyau cubique

$$K(u) = \frac{35}{32}(1-u^2)^3\mathbb{1}_{\{|u|\leq 1\}}$$

(vi) Le noyau gaussien

$$K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$$

(vii) Le noyau circulaire

$$K(u) = \frac{\pi}{4}\cos(\frac{\pi}{2}u)\mathbb{1}_{\{|u|\leq 1\}}$$

Propriétés 1.4.1 *Soit*

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

1. *Si K est une densité alors $\widehat{f}$ est une densité ($\int \widehat{f}(x)dx = 1$)*
2. *$\widehat{f}$ a les mêmes propriétés de continuité et de différentiabilité que K à cause du fait qu'elle comporte un nombre fini d'éléments de la forme K.*
   - *Si K est continue alors $\widehat{f}$ sera une fonction continue.*
   - *Si K est différentiable alors $\widehat{f}$ sera une fonction différentiable.*
   - *Si K peut prendre des valeurs négatives, alors $\widehat{f}$ pourra aussi prendre des valeurs négatives.*

1.4.1 Biais et variance

Soit l'estimateur à noyau précédent

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

On notera $K_h(.) = \frac{1}{h}K(\frac{.}{h})$.

Notations 1.4.1

1. $$E\left[\widehat{f}(x)\right] = E\left[K_h(x-X)\right] = \int K_h(x-y)f(y)dy$$
2. *La convolution entre deux fonctions intégrables f et g est définie par*

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

Nous avons alors

Proposition 1.4.1

1.

$$E\left[\widehat{f}(x)\right] - f(x) = (K_h * f)(x) - f(x) \tag{1.15}$$

2.

$$Var\left[\widehat{f}(x)\right] = \left[\frac{1}{n}(K_h^2 * f)(x) - (K_h * f)^2(x)\right] \tag{1.16}$$

Démonstration 1.

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

$$E\left[\widehat{f}(x)\right] = \frac{n}{n}E(K_h(x-X)) = (K_h * f)(x)$$

2.

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

□

La proposition suivante va nous donner l'expression de l'erreur quadratique moyenne.

Proposition 1.4.2 *L'erreur quadratique moyenne (" MSE Mean Squared Error" (page 8)) de l'estimateur à noyau est*

$$MSE\left[\widehat{f}(x)\right] = \frac{1}{n}(K_h^2 * f)(x) + (1-\frac{1}{n})(K_h * f)^2(x) - 2(K_h * f)(x)f(x) + f^2(x). \tag{1.17}$$

Démonstration

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

□

Définition 1.4.2 *On définit MISE ("Mean Integrated Squared Error") l'erreur quadratique moyenne intégrée par l'expression*

$$MISE\left[\widehat{f}(x)\right] = \int MSE\left[\widehat{f}(x)\right] dx \tag{1.18}$$

Son expression vaut

$$MISE\left[\widehat{f}(x)\right] = \frac{1}{n}\int (K_h^2 * f)(x)dx + (1-\frac{1}{n})\int (K_h * f)^2(x)dx$$
$$-2\int (K_h * f)(x)f(x)dx + \int f^2(x)dx.$$

On peut trouver l'expression finale du MISE par la proposition suivante.

Proposition 1.4.3

$$MISE\left[\widehat{f}(.)\right] = \frac{1}{nh}\int K^2(u)du + (1-\frac{1}{n})\int (K_h * f)^2(x)dx \tag{1.19}$$
$$-2\int (K_h * f)(x)f(x)dx + \int f^2(x)dx. \tag{1.20}$$

En effet : Comme

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

Nous obtenons la proposition 1.4.3 en posant $u = \frac{x-y}{h}$, donc $y = x-uh$ et $\int f(x-uh)dx = 1$ car f est une densité de probabilité ($\int f(y)dy = 1$). □

Remarque 1.4.2 *Les expressions du $MSE\left[\widehat{f}(x)\right]$ et du $MISE\left[\widehat{f}(x)\right]$ sont complexes on cherchera des expressions asymptotiques qui pourraient dépendre de h de manière plus simple.*

1.5 Biais et variance asymptotique

Hypothèses 1.5.1 On suppose que le noyau K satisfait les conditions suivantes

$$K \geq 0, \qquad \int K(u)du = 1, \qquad \int K(u)udu = 0, \qquad 0 < \int K(u)u^2du < \infty. \tag{1.21}$$

Un tel noyau existe car on pourra prendre le noyau gaussien par exemple.

En faisant un développement de Taylor de f au voisinage de x on obtient

$$f(x-uh) = f(x) - f'(x)uh + \frac{1}{2}f''(x)u^2h^2 + o(h^2),$$

où $o(h^2)$ tend vers zéro quand h tend vers zéro.

Proposition 1.5.1 *Sous les conditions 1.21 l'espérance de l'estimateur* $E\left[\widehat{f}(x)\right]$ *est donnée par*

$$E\left[\widehat{f}(x)\right] = f(x)\int K(u)du - f'(x)h\int K(u)udu + \frac{1}{2}f''(x)h^2\int K(u)u^2du + o(h^2). \tag{1.22}$$

En effet :

$$\begin{aligned} E\left[\widehat{f}(x)\right] &= \int K_h(x-y)f(y)dy \\ &= \int K(u)f(x-uh)du \\ &= \int K(u)\left[f(x) - f'(x)uh + \frac{1}{2}u^2h^2 + o(h^2)\right]du \\ &= f(x)\int K(u)du - f'(x)h\int K(u)udu\frac{1}{2}f''(x)h^2\int K(u)u^2du + o(h^2) \end{aligned}$$

D'après (1.21), on obtient

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

Pour le calcul de la variance on utilise (1.16) et le fait que

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

Proposition 1.5.2

$$Var\left[\widehat{f}(x)\right] = \frac{1}{nh}f(x)\int K^2(u)du + o\left(\frac{1}{nh}\right) \tag{1.23}$$

Démonstration

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

On a

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

Nous trouvons que

$$Var\left[\widehat{f}(x)\right] = \frac{1}{nh}f(x)\int K^2(u)du + o\left(\frac{1}{nh}\right) - \frac{1}{n}\left[E(K_h(x-X))\right]^2$$

La dernière quantité précédée du signe moins est un $o(\frac{1}{n})$

Si on note $\mu_2 = \int K(u)u^2du, \quad R(K) = \int K^2(u)du,$ alors on a

$$\text{Biais}(\widehat{f}(x)) = \frac{1}{2}f"(x)\mu_2 h^2 + o(h^2) \tag{1.24}$$

$$Var\left[\widehat{f}(x)\right] = \frac{1}{nh}f(x)R(K) + o(\frac{1}{nh}) \tag{1.25}$$

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

$$\text{Biais}(\widehat{f}(x)) \longrightarrow 0 \qquad \text{quand } n \longrightarrow \infty$$

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

$$Var\left[\widehat{f}(x)\right] \longrightarrow 0 \qquad \text{quand } n \longrightarrow +\infty.$$

□

Remarque 1.5.1 *Si h décroît alors le carré du biais décroît et la variance croît.*
Si h augmente alors le carré du biais croît et la variance décroît. Il faut trouver un compromis entre le biais et la variance car ils évoluent dans le sens contraire.

On déduit les expressions asymptotiques de MSE et de MISE.

Proposition 1.5.3

$$MSE\left[\widehat{f}_n(x)\right] = \frac{1}{4}h^4\mu_2^2(f"(x))^2 + \frac{1}{nh}f(x)R(K) + o\left(h^4 + \frac{1}{nh}\right) \tag{1.26}$$

qui va désormais être notée $AMSE(\widehat{f}(x))$.

$$MISE\left[(\widehat{f}_n(x)\right] = \frac{1}{4}h^4\mu_2^2\int(f"(x))^2dx + \frac{1}{nh}R(K) + o\left(h^4 + \frac{1}{nh}\right) \tag{1.27}$$

Sous des conditions d'intégrabilité de f et de ses dérivées, cette dernière expression va désormais être notée $AMISE\left[\widehat{f}(x)\right]$.

Démonstration En utilisant (1.13), ainsi que les expressions de (1.24) au carré avec (1.25), on obtient $MSE\left[\widehat{f}_n(x)\right]$. Le $MISE\left[\widehat{f}_n(x)\right]$ est déduit par intégration (peut être traité à titre d'exercice). □

1.5.1 Choix optimal du paramètre de lissage

On introduit deux paramètres de lissage dont on fait la distinction
h paramètre de lissage constant (ou global)
$h(x)$ paramètre de lissage variable (local). On notera

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

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

On considère le problème du choix du paramètre de lissage constant h. Le critère de sélection est la MISE asymptotique (1.27). Le paramètre de lissage optimal est la valeur de h qui minimise la MISE que l'on notera h_{MISE}. Depuis l'expression de $AMISE(\widehat{f}(x))$ on peut montrer que : (En annulant la dérivée de la MISE)

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2^2 R(f")}\right]^{1/5} n^{-1/5} \tag{1.28}$$

Une approximation asymptotique de h_{MISE} est donnée par h_{AMISE}.

$$h_{MISE} \simeq \left[\frac{R(K)}{\mu_2^2 R(f")}\right]^{1/5} n^{-1/5}$$

d'où $\lim_{n\longrightarrow\infty} \frac{h_{\text{MISE}}}{h_{\text{AMISE}}} = 1$.
Un critère pour sélectionner un paramètre de lissage variable (local) $h(x)$ est la mesure de performance locale $MSE\left[\widehat{f}_{n,L}(x)\right]$. Nous adaptons les notations suivantes

$$h_{MSE}(x) = Arg \min_h MSE(\widehat{f}_{n,L}(x))$$

et

$$h_{AMSE}(x) = Arg \min_h AMSE(\widehat{f}_{n,L}(x))$$

Nous avons en utilisant (1.28)

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

avec $f''(x) \neq 0$.

Exemple 1.5.1 **Cas normal** *:*
On veut calculer h_{AMISE} dans le cas où f appartient à une famille de distributions normales $N(\mu,\sigma^2)$, de moyenne μ et de variance σ^2 inconnues. On notera alors

$$f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \text{ avec } \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$\varphi(x)$ la densité de probabilité normale centrée réduite.

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

$$\begin{aligned} R(f'') = \int (f''(x))^2 dx &= \frac{1}{\sigma^6}\int \left[\varphi\left(\frac{x-\mu}{\sigma}\right)\right]^2 dx \\ &= \frac{1}{\sigma^5}\int (\varphi''(v))^2 dv, \ v = \frac{x-\mu}{\sigma} \end{aligned}$$

Comme

$$\varphi(v) = \frac{1}{\sqrt{2\pi}} e^{-v^2/2} \quad \text{alors} \ \ \varphi'(v) = -\frac{v}{\sqrt{2\pi}} e^{-v^2/2}$$

$$\text{et } \varphi''(v) = \frac{1}{\sqrt{2\pi}}(v^2-1)e^{-v^2/2}$$

Quand on remplace dans $R(f'')$ on obtient

$$\begin{aligned} R(f'') &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left(\int_{-\infty}^{+\infty} v^4 e^{-v^2}dv - 2\int_{-\infty}^{+\infty} v^2 e^{-v^2}dv + \int_{-\infty}^{+\infty} e^{-v^2}dv\right) \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\left(-\frac{1}{2}\int_{-\infty}^{+\infty} v^2 e^{-v^2}dv + \int_{-\infty}^{+\infty} e^{-v^2}dv\right) \end{aligned}$$

car $I = \int_{-\infty}^{+\infty} v^4 e^{-v^2}dv = \frac{3}{2}\int_{-\infty}^{+\infty} v^2 e^{-v^2}dv$. En effet et en passant par une intégration par parties on a

$$dw = -2ve^{-v^2}dv \Longleftrightarrow w = e^{-v^2}.$$
$$u = -\frac{1}{2}v^3 \Longleftrightarrow u' = -\frac{3}{2}v^2.$$

et

$$I = \left(-\frac{1}{2}v^3 e^{-v^2}\right)_{-\infty}^{+\infty} + \frac{3}{2}\int_{-\infty}^{+\infty} v^2 e^{-v^2}dv$$

$\lim\limits_{v\longrightarrow+\infty}\frac{1}{2}v^3e^{-v^2} = 0$, de même $\lim\limits_{v\longrightarrow-\infty}\frac{1}{2}v^3e^{-v^2} = 0$.
En poursuivant le calcul de $R(f'')$, on pose $u = \sqrt{2}v$, $du = \sqrt{2}dv$

$$\begin{aligned} R(f'') &= \frac{1}{\sigma^5}\frac{1}{2\pi}\Big(-\frac{1}{2}\int_{-\infty}^{+\infty}\frac{u^2}{2}e^{-u^2/2}\frac{du}{\sqrt{2}} + \frac{1}{\sqrt{2}}\int_{-\infty}^{+\infty} e^{-u^2/2}du\Big) \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\Big(-\frac{1}{4}\cdot\sqrt{\pi} + \sqrt{\pi}\cdot 1\Big) \\ &= \frac{1}{\sigma^5}\frac{1}{2\pi}\frac{3}{4}\sqrt{\pi} = \frac{3}{8\sigma^5\sqrt{\pi}}. \end{aligned}$$

Le paramètre de lissage optimal asymptotique devient

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

Le paramètre de lissage du type "Normal reference" est défini par

$$\widehat{h}_{NR} = \Big(\frac{8\pi R(K)}{3\mu_2^2}\Big)^{1/5}\widehat{\sigma}\, n^{-1/5}$$

où $\widehat{\sigma}$ est un estimateur de σ l'écart type de la population. Des choix possibles pour $\widehat{\sigma}$

1. L'écart type empirique
$$S = \sqrt{\frac{1}{n-1}\sum_{i=2}^{n}(X_i - \overline{X})^2}$$
2. L'écart interquartile empirique standardisé
$$\frac{\text{l'écart interquartile standardisé}}{\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})} = \frac{R}{\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})}$$

   Cette expression est équivalente à $\frac{R}{1.349}$, où $\Phi(.)$ est la fonction de répartition d'une normale centrée réduite. $\Phi^{-1}(\frac{3}{4}) - \Phi^{-1}(\frac{1}{4})$ est l'écart interquartile d'une variable aléatoire normale réduite.

Si $X \sim N(\mu, \sigma^2)$, alors l'écart interquatile de X est

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

ce qui justifie l'estimateur proposé.
On propose d'utiliser le minimum entre S et $R/1.349$. Le paramètre de lissage

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

1.6 La validation croisée

Rudemo(1982)([29]) et Bowman(1984)([2]) ont introduit la méthode de validation croisée (¨cross-validation¨) du type moindres carrés. Elle permet d'obtenir un paramètre de lissage simple.

Proposition 1.6.1

$$MISE\Big[(\widehat{f_n}(.)\Big] = E\left[\int (\widehat{f_n}(x) - f(x))^2 dx\right] \tag{1.29}$$

$$= E\int \left[\widehat{f_n^2}(x) dx\right] - 2E\left[\int \widehat{f_n}(x) f(x) dx\right] + \int f^2(x) dx \tag{1.30}$$

$\int f^2(x)dx$ ne dépend pas de h. Minimiser $MISE\left[\widehat{f_n}(.)\right]$ par rapport à h revient à minimiser

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

De plus $\int f^2(x)dx = E\left[\int f^2(x)dx\right]$ du fait que le terme dépend seulement de la densité inconnue f.
On estime $E\left[\int \widehat{f_n}(x) f(x) dx\right]$ par

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i) \tag{1.31}$$

où

$$\widehat{f}_{-i}(x) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_h(x - X_j) \tag{1.32}$$

est l'estimateur à un noyau basé sur l'échantillon réduit $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, où l'observation X_i a été supprimée. On l'appelle estimateur ¨leave-one-estimator¨.

Proposition 1.6.2 *L'estimateur (1.31) est un estimateur sans biais de $E\left[\int \widehat{f_n}(x) f(x) dx\right]$*

Démonstration

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\widehat{f}_{-i}(X_i)\right]$$

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

□

Corollaire 1.6.1 *Un estimateur sans biais pour*

$$MISE\left[\widehat{f_n}(.)\right] - \int f^2(x)dx = E\left[\int \widehat{f_n^2}(x)dx - 2\int \widehat{f_n}(x)f(x)dx\right] \tag{1.33}$$

est donné par

$$LSCV(\widehat{f}_n(.)) = \int \widehat{f_n^2}(x)dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(X_i)$$

Cette quantité est désignée par $LSCV(h)$*, elle dépend de* h *et de la méthode de validation croisée (" Least square Cross-validation" (CV) en anglais).*

Définition 1.6.1 *Cette dernière quantité s'appelle validation croisée. Le paramètre de lissage du type " validation croisée" est la valeur de* h *qui minimise cette quantité, c'est à dire*

$$\widehat{h}_{LSCV} = Arg\min_{h} LSCV(h)$$

Chapitre 2

Tests non paramétriques

Les tests non paramétriques sont utilisés quand on n'a pas d'hypothèses sur la distribution de l'échantillon.

2.1 Quantiles d'une loi et quantiles empiriques

Définition 2.1.1 *Soit F une fonction de répartition continue et strictement croissante. Pour tout $p \in]0,1[$, on appelle quantile d'ordre p et on note q_p la racine unique de l'équation $F(x) = p$, c'est-à-dire $q_p = F^{-1}(p)$. En particulier*
Si $p = \frac{1}{2}$, $q_{1/2}$ est appelé la médiane de la loi F.
Si $p = \frac{1}{4}$, $q_{1/4}$ est appelé le premier quartile de la loi F.

Remarque 2.1.1

1. *Quand on considère les autres quantiles, le second $q_{1/2}$ et le troisième $q_{3/4}$, ces trois quartiles partagent la distribution en parties de même poids,*

$$P(X \leq q_{1/4}) = P(q_{1/4} \leq X < q_{1/2}) = P(q_{1/2} \leq X < q_{3/4}) = P(X \geq q_{3/4}) = \frac{1}{4}$$

Le quantile $q_{1/10}$ s'appelle le premier décile, les autres sont les quantiles $q_{k/10}$ où $k = 2,\ldots,9$; ils divisent la distribution en 10 régions de même probabilité $1/10$.

2. *On peut définir les quantiles pour une loi discrète ou une loi continue telle que F ne soit pas strictement croissante. Dans le cas continue, on définit le quantile d'ordre p par la formule*

$$q_p = \inf\{x;\ F(x) \geq p\} \tag{2.1}$$

On supposera dans la suite que la fonction de répartition est continue et strictement croissante.

Définition 2.1.2 *Soit $(X_1,\ldots,X_n)$ un n-échantillon issu d'une distribution F et $(X_{(1)},\ldots,X_{(n)})$ l'échantillon ordonné. Soit $p \in]0,1[$, la statistique d'ordre $X_{([np]+1)}$ (où $[np]$ désigne la partie entière de np), s'appelle le quantile empirique d'ordre p de l'échantillon. En particulier $X_{([np])}$ est la médiane de l'échantillon.*

Exemple 2.1.1 Si $(X_1, X_2, X_3, X_1, X_5)$ est un échantillon rangé de taille 5, alors la médiane empirique est $X_{(3)}$, c'est la valeur qui partage l'effectif de l'échantillon en deux parties de même effectif. Le théorème suivant montre que le quantile empirique est un quantile théorique.

Théorème 2.1.1 *Si F est continue et strictement croissante alors on a*

$$X_{([np]+1)} \longrightarrow q_p \qquad \mathbb{P}-\textit{presque sûrement si } n \longrightarrow +\infty \tag{2.2}$$

Autrement dit $X_{([np]+1)}$ est une estimation de q_p d'autant meilleure que la taille n de l'échantillon est grande.

Démonstration Soit F_n la fonction de répartition empirique de l'échantillon $(X_1,\ldots,X_n)$ ou de l'échantillon ordonné $(X_{(1)},\ldots,X_{(n)})$. Par définition de la fonction de répartition empirique on a

$$F_n^{\omega}(X_{([np]+1)}) = \frac{1}{n}([np]+1) \longrightarrow p \qquad \text{si } n \longrightarrow +\infty \tag{2.3}$$

Par ailleurs par le théorème fondamental de la statistique [27].

$$\text{Pour } \mathbb{P}-\text{presque tout } \omega,\ \sup_{x\in\mathbb{R}} \left| F(X_{([np]+1)}(\omega)) - F_n^{\omega}(X_{([np]+1)}(\omega)) \right| \underset{n\longrightarrow+\infty}{\longrightarrow} 0$$

D'où compte tenu de (2.3)

$$F(X_{([np]+1)}(\omega)) \longrightarrow p \qquad (n \longrightarrow +\infty)$$

et puisque la fonction F^{-1} est continue, on a donc

$$X_{([np]+1)}(\omega) \longrightarrow F^{-1}(p) = q_p \qquad (n \longrightarrow +\infty)$$

□

Théorème 2.1.2 *Si la loi F a une densité de probabilité f strictement positive sur $\mathbb{R}$, alors en posant $D = \frac{\sqrt{p(1-p)}}{f(q_p)}$. La variable aléatoire $\sqrt{n}\left(\frac{X_{([np]+1)}-q_p}{D}\right)$ converge en loi vers la loi normale $N(0,1)$ quand $n \longrightarrow +\infty$.*

Démonstration Admise. La démonstration nécessite une preuve détaillée (cf. ([27])).

Exemple 2.1.2 On considère l'échantillon $(X_1,\ldots,X_n)$ i-d-d selon la loi de Cauchy de densité

$$f(x) = \frac{1}{\pi(1+(x-\theta)^2)}$$

i-d-d désigne le fait que les variables aléatoires X_i, $i = 1\ldots n$ sont indépendantes identiquement distribuées. Sa médiane est clairement le paramètre de translation θ, que l'on estime donc par la médiane empirique $x_{1/2}(n)$. En utilisant le résultat précédent on a

$$x_{1/2}(n) \underset{n\longrightarrow\infty}{\overset{p.s.}{\longrightarrow}} x_{1/2}(n) = \theta$$

qui s'écrit

$$\sqrt{n}(x_{1/2}(n)-\theta) \underset{n\longrightarrow\infty}{\overset{\mathscr{L}}{\longrightarrow}} N(0,\frac{\pi^2}{4}).$$

Comme $\Phi^{-1}(0.975) \approx 2$, on déduit un intervalle de confiance de niveau asymptotique 95% pour θ donné par

$$\left[x_{1/2}(n) - \frac{\pi}{\sqrt{n}}; x_{1/2}(n) + \frac{\pi}{\sqrt{n}} \right],$$

Φ^{-1} désignant la fonction cumulative inverse de la loi normale centrée réduite.

2.2 Test des signes

2.2.1 Définition du test

On considère un échantillon de n observations. Chaque observation est constituée d'une paire de valeurs. On forme une nouvelle variable $D = X_1 - X_2$ dont les valeurs

$$d_i = x_{i1} - x_{i2}$$

Le test des signes consiste à s'intéresser uniquement au sens des écarts positifs.
Soit $\pi = P(D > 0)$ le test bilatéral

$$\begin{array}{ll} H_0: & \pi = \frac{1}{2} \\ H_1: & \pi \neq \frac{1}{2} \end{array} \tag{2.4}$$

où H_0 désigne l'hypothèse nulle et H_1 l'hypothèse alternative. Pour un test unilatéral : $H_1 : \pi > \frac{1}{2}\left(resp. H_1 : \pi < \frac{1}{2}\right)$ lorsqu'on veut savoir si X_1 est stochastiquement plus grand (resp. plus petit) que X_2.
Le test des signes consiste à tester si le nombre des écarts positifs est différent du nombre des écarts négatifs (dans le cas des tests unilatéraux, on teste si le nombre des écarts positifs est plus élevé). Pour un échantillon de taille n, une statistique naturelle pour le test des signes est

$$S = \sum_{i=1}^{n} \mathbb{1}_{d_i}$$

Sous H_0 : S suit une loi binomiale $\mathscr{B}\left(n, \frac{1}{2}\right)$. *La p-valeur (" p-value") vaut*

$$p = 2 \times Pr\left(\mathscr{B}\left(n, \frac{1}{2}\right) \geq \max\left(S, n - S\right)\right) \tag{2.5}$$

Pour un test unilatéral $H_1 : \pi > \frac{1}{2}$, la probabilité critique sera $p = Pr\left(\mathscr{B}\left(n, \frac{1}{2}\right) \geq \max\left(S, n - S\right)\right)$.
La probabilité cumulée d'une loi binomiale s'écrit

$$P\left(\mathscr{B}\left(n, \pi\right) \geq S\right) = \sum_{j=S}^{n} C_n^j \pi^j (1-\pi)^{n-j} \tag{2.6}$$

$$= \sum_{j=S}^{n} \frac{n!}{j!(n-j)!} \pi^j (1-\pi)^{n-j}. \tag{2.7}$$

Remarque 2.2.1 *Avec le **R** on a les commandes suivantes :*
*>**binom.test**($sum(x > 0), n = length(x); p = 0.5,$ **alternative**= " two.sided")*
Pour l'alternative $H_1 : m < 0$ on remplace "two.sided" par "less". Si $H_1 : m > 0$ alternative = "greater".

2.2.2 Le test pour échantillons appariés

On rencontre les échantillons appariés ("paired data") par exemple pour comparer les effets de deux traitements sur deux populations que l'on veut apparier.
On dit que les données sont appariées quand "l'individu" i du premier échantillon est lié à "l'individu" i du second échantillon. Pour chaque i, U_i et V_i sont liés, on n'a pas l'indépendance entre U_i et V_i. Cependant, on a toujours l'indépendance entre les (U_i, V_i) pour différents i.

C'est le cas quand les individus sont observés avant et après un traitement. Soient $(U_i, V_i)_{i=1}^n$ les n-couples d'observations correspondantes et $p = P(V_i > U_i)$. L'hypothèse nulle est que la distribution est identique avant et après un traitement, à savoir qu'il n'y a pas d'effet du traitement implique que $p = \frac{1}{2}$. La statistique nulle est alors le nombre de variables aléatoires réelles $d_i = V_i - U_i$ négatives ou positives, les valeurs nulles ne sont pas comptabilisées. Le test peut être étendu à un test de quantile en remplaçant la valeur $p = \frac{1}{2}$ dans H_0. La fonction de décision est

$$\mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq V_i\}} > q_{1-\alpha}^{\mathscr{B}\left(n,\frac{1}{2}\right)}\right\}$$

Exemple 2.2.1 On teste l'effet du glucose sur la mémoire de 16 patients âgés. Au réveil on leur donne une boisson sucrée, on leur narre une histoire. Quelques temps après, on leur demande de la raconter. On note la qualité de la restitution par un juge. Après un délai raisonable on recommence la même expérience avec une boisson contenant de la saccharine. L'objectif de l'étude est de vérifier que la boisson au glucose a un effet sur la qualité de mémorisation. On a le tableau suivant : S est la statistique qui correspond au nombre de signes positifs des d_i.

FIGURE 2.1 – Effet du glucose sur la mémorisation-Test des signes

X_1(glucose)	X_2(saccharine)	Écart	Signe(Écart)
0	1	–1	—
10	9	1	+
9	6	3	+
4	2	2	+
8	5	3	+
6	5	1	+
9	7	2	+
3	2	1	+
12	8	4	+
10	8	2	+
15	11	4	+
9	3	6	+
5	6	–1	—
6	8	–2	—
10	8	2	+
6	4	2	+

Quand on calcule $S = 13$, à partir de la Fig (2.1) de l'Exemple (2.2.1). On calcule

$$P\left(\mathscr{B}\left(16, \frac{1}{2}\right) = k\right) \text{ pour } k = 13, 14, 15, 16.$$

$P\left(\mathscr{B}\left(16, \frac{1}{2}\right) = 13\right) = 0.00854$. etc....

$$P\left(\mathscr{B}\left(16, \tfrac{1}{2}\right) \geq 13\right) = \sum_{k=13}^{16} P\left(\mathscr{B}\left(16, \frac{1}{2}\right) = k\right) = 0.01064$$

j	Pr
13	0.00854
14	0.00183
15	0.00024
16	0.00002
Somme	0.01064

p-value	0.02127

Dans le cas bilatéral, la p-value est $2 \times 0.01064 = 0.02127$. Au risque $\alpha = 0.05$ on a RH_0 (on rejette H_0). Donc il y a un effet du glucose sur la mémorisation. Dans le cas d'un test unilatéral $H_1 : \pi > \frac{1}{2}$, la probabilité critique est

$$p = P\left(\mathscr{B}(16, \frac{1}{2}) \geq 13\right) = 0.01064.$$

RH_0 est plus important que dans le cas bilatéral précédent.

Remarque 2.2.2 *Dans le cas où les d_i sont nuls, la modélisation des signes par la loi binomiale pose un problème, il faut supprimer ces observations. On réduit l'effectif.*

2.2.3 L'approximation normale pour les grands effectifs

Pour n assez grand $(n > 30)$, on approche la distribution de S par une loi normale qui, sous $H_0 : \pi = \frac{1}{2}$, on a

$$E(S) = n \times \pi = \frac{n}{2} \tag{2.8}$$

$$Var(S) = n \times \pi \times (1 - \pi) = \frac{n}{4} \tag{2.9}$$

La statistique centrée et réduite

$$Z = \frac{S - \frac{n}{2}}{\sqrt{n/4}} = \frac{2S - n}{\sqrt{n}} \tag{2.10}$$

Pour un test bilatéral au risque $\alpha = 0.05$, la région critique (RC) du test est définie

$$R.C. : |Z| \geq u_{1-\alpha/2} \tag{2.11}$$

où $u_{1-\alpha/2}$ désigne la quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

Remarque 2.2.3 *Pour les effectifs modérés on améliore la précision avec la correction de continuité.*

1. *Pour le test bilatéral*
$$|Z| = \frac{|2S - n| - 1.0}{\sqrt{n}}$$
2. *Pour les tests unilatéraux*
$$Z = \frac{2S - n \pm 1.0}{\sqrt{n}}$$

On rajoute $+1$ pour un test unilatéral à gauche ($H_1 : \pi < \frac{1}{2}$) ; on retranche 1 pour un test unilatéral à droite.

Exemple 2.2.2 *On a demandé à* 39 *footballeurs de shooter dans des ballons gonflés à l'hélium et à l'air. On se demande si les gonfler à l'hélium est plus intéressent que les gonfler à l'air pour les envoyer plus loin. Donc on considère le test :* $H_0 : \pi = \frac{1}{2}$ *contre* $H_1 : \pi > \frac{1}{2}$.

- $n = 39$. On supprime les observations avec les écarts nuls, on a donc $n = 37$.
- On calcule S elle vaut 20.
- La statistique est centrée et réduite, avec la correction de continuité -1 (cas unilatéral),

$$Z = \frac{(2S - n) - 1}{\sqrt{n}} = \frac{(2 \times 20 - 37) - 1}{\sqrt{37}} = 0.32880.$$

La probabilité critique associée est $p = 0.37115$. Pour $\alpha = 0.05$, on a $\overline{R}\, H_0$. Gonfler les ballons à l'hélium ne permet pas de les envoyer plus loin. Le tableau suivant de la Fig (2.2) donne les valeurs du gonflage à l'hélium et à l'air.

FIGURE 2.2 – Gonflage à l'hélium de ballon-Test des signes

Essai	Helium	Air	Écart	Signe
1	25	25	0	0
2	16	23	–7	—
3	25	18	7	+
4	14	16	–2	—
5	23	35	–12	—
6	29	15	14	+
7	25	26	–1	—
8	26	24	2	+
9	22	24	–2	—
10	26	28	–2	—
11	12	25	–13	—
12	28	19	9	+
13	28	27	1	+
14	31	25	6	+
15	22	34	–12	—
16	29	26	3	+
17	23	20	3	+
18	26	22	4	+
19	35	33	2	+
20	24	29	–5	—
21	31	31	0	0
22	34	27	7	+
23	39	22	17	+
24	32	29	3	+
25	14	28	–14	—
26	28	29	–1	—
27	30	22	8	+
28	27	31	–4	—
29	33	25	8	+
30	11	20	–9	—
31	26	27	–1	—
32	32	26	6	+
33	30	28	2	+
34	29	32	–3	—
35	30	28	2	+
36	29	25	4	+
37	29	31	–2	—
38	30	28	2	+
39	26	28	–2	—

2.3 Test des rangs signés de Wilcoxon

Il traite la comparaison d'échantillons appariés. Par rapport au test des signes il est plus puissant et plus riche.

Construction de la statistique de test

On considère les écarts $d_i = x_{i1} - x_{i2}$ et les $|d_i|$, les étapes de l'étude sont

1. On calcule les rangs r_i des valeurs absolues des écarts.
2. On calcule $T^+ = \sum_{i:d_i>0} r_i$.
3. On calcule les écarts négatifs T^-. Comme la somme totale est égale à $\frac{n(n+1)}{2}$, on a :

$$T^- = \frac{n(n+1)}{2} - T^+ \tag{2.12}$$

Sous H_0, les variables X_i et X_j ont la même fonction de répartition et donc ont les mêmes caractéristiques. Pour un test bilatéral, la région critique du test au risque α sera donc

$$R.C.: (T^+ \leq \frac{n(n+1)}{2} - T_\alpha) \text{ ou } T^+ \geq T_\alpha) \tag{2.13}$$

où le seuil critique T_α est lu dans la table (Fig 4.1) spécifique due à Wilcoxon (cf. Annexe (Fig 4.1)).

2.3.1 Statistiques d'ordre et de rang

Définition 2.3.1 *Soient $(X_1, X_2, \ldots, X_n)$ un échantillon de n- variables aléatoires de taille n la statistique d'ordre $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ est définie par la condition $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. On pose $X^* = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$, il existe une permutation $\widehat{\sigma} \in \mathfrak{S}_n$ telle que $(X_{(1)}, X_{(2)}, \ldots, X_{(n)}) = (X_{\widehat{\sigma}(1)}, \ldots, X_{\widehat{\sigma}(n)})$.*
$\mathfrak{S}_n$ étant l'ensemble de toutes les permutations de l'ensemble $\{1, \ldots, n\}$. Comme on peut avoir $X_i = X_j$ pour $i \neq j$, il n'y a pas toujours unicité de la permutation. Le vecteur des rangs R_X est la permutation inverse de $\widehat{\sigma}$. Il peut exister i et j tels que $i \neq j$ et $X_i = X_j$.

Exemple 2.3.1 $x = (2, 5, 5, 0, 1, 8)$

x	2	5	5	0	1	8
R_X	3	4	5	1	2	6

2.3.2 Cas d'un échantillon

On considère des variables $(X_1, X_2, \ldots, X_n)$ diffuses (cf. définition (2.3.2) plus loin) et indépendantes, le vecteur de rangs de X est alors unique presque sûrement.

Proposition 2.3.1 *Si les variables $X_1, \ldots, X_n$ sont indépendantes et diffuses alors*

$$P\left(\exists\, i \neq j : |X_i| = \left|X_j\right|\right) = 0$$

Démonstration Pour tout $i \neq j$ on a

$$\begin{aligned} P\big(|X_i| = |X_j|\big) &\leq P\big(X_i = X_j\big) + P\big(X_i = -X_j\big) \\ &= \int P\big(X_i = x\big) dP_{X_j}(x) + \int P\big(X_i = -x\big) dP_{X_j}(x) = 0 \end{aligned}$$

car les variables sont indépendantes et diffuses. Donc

$$P\big(\exists\, i \neq j : |X_i| = |X_j|\big) \leq \sum_{i \neq j, (i,j) \in [1,n]^2} P\big(|X_i| = |X_j|\big) = 0.$$

□

Définition 2.3.2 *On dit qu'une variable aléatoire U est diffuse si*

$$\forall x \in \mathbb{R},\ P\big(U = x\big) = 0$$

Cette notion est équivalente au fait que la distribution est continue.

Hypothèses 2.3.1 On suppose que les observations $(X_1, X_2, \ldots, X_n)$ vérifient les conditions suivantes

1. Les X_i sont indépendantes entre elles.
2. Les X_i sont diffuses.
3. Les X_i ont une médiane commune m.
4. Les lois des X_i sont symétriques par rapport à m.

On considère la statistique associée aux $(|X_i|)_{i=1}^n$, on a donc

$$|X|_{(1)} < |X|_{(2)} < \ldots < |X|_{(n)},\ \text{presque sûrement}$$

On note $R_{|X|}$ le vecteur des rangs associés. On pose

$$T_n^+ = \sum_{i=1}^{n} R_{|X|}(i)\, \mathbb{1}_{\{X_i > 0\}} \tag{2.14}$$

Exemple 2.3.2 L'exemple suivant fait un calcul des rangs

X_i	–1	–0.5	0.3	0.7	–0.1
$\lvert X_i\rvert$	1	0.5	0.3	0.7	0.1
$R_{\lvert X\rvert}(i)$	5	3	2	4	1

Remarque 2.3.1 *1.*

$$0 \leq T_n^+ \leq \frac{n(n+1)}{2} \tag{2.15}$$

Le cas $T_n^+ = 0$ correspond au cas où tous les X_i sont strictement négatifs, le cas $T_n^+ = \frac{n(n+1)}{2}$ correspond au cas où tous les $X_i > 0$.

2. Si on pose en plus

$$T_n^- = \sum_{i=1}^{n} R_{|X|}(i)\, \mathbb{1}_{\{X_i < 0\}} \tag{2.16}$$

et si $P(X_i = 0) = 0$ (v.a diffuses), alors

$$T_n^+ + T_n^- = \frac{n(n+1)}{2} \tag{2.17}$$

Théorème 2.3.1 *Les hypothèses (2.3.1) précédentes étant satisfaites. Sous* $H_0 : m = 0$, *on a*

1. T_n^+ *et* T_n^- *ont la même distribution.*
2. $E(T_n^+) = \frac{n(n+1)}{4}$.
3. T_n^+ *et* T_n^- *sont libres en loi de* X.
4. $Var(T_n^+) = \frac{n(n+1)(2n+1)}{24}$.
5. *Asymptotiquement on a*

$$\frac{T_n^+ - E(T_n^+)}{\sqrt{Var(T_n^+)}} \xrightarrow[n \longrightarrow +\infty]{\mathscr{L}} N(0,1) \tag{2.18}$$

La convergence est en loi.

Démonstration 1.

$$\begin{aligned} T_n^+ &= \sum_{i=1}^{n} R_{|X|}(i) \mathbb{1}_{X_i > 0} \\ T_n^+ &= \sum_{j=1}^{n} j \mathbb{1}_{\left\{X_{\sigma_{|X|}(j) > 0}\right\}} \end{aligned}$$

On a noté : $\sigma_{|X|} = R_{|X|}^{-1}$. De même

$$T_n^- = \sum_{j=1}^{n} j \mathbb{1}_{\left\{X_{\sigma_{|X|}(j) < 0}\right\}}$$

Comme la loi des X_i est symétrique par rapport à 0, donc

$$X_1^n \sim -X_1^n$$

$\sim$ désigne suivre la même distribution (loi). Le vecteur T_n^+ est une fonction du vecteur X_1^n, donc

$$T_n^+ = \sum_{j=1}^{n} j \mathbb{1}_{\left\{X_{\sigma_{|X|}(j) > 0}\right\}} \sim \sum_{j=1}^{n} j \mathbb{1}_{\left\{-X_{\sigma_{|-X|}(j) > 0}\right\}}.$$

Or $\sigma_{|X|} = \sigma_{|-X|}$ donc

$$T_n^+ \sim \sum_{j=1}^{n} j \mathbb{1}_{\left\{-X_{\sigma_{|X|}(j) > 0}\right\}} = \sum_{j=1}^{n} j \mathbb{1}_{\left\{X_{\sigma_{|-X|}(j) < 0}\right\}} = T_n^-$$

On déduit que T_n^+ et T_n^- sont de même loi.

2. La symétrie par rapport à zéro de la distribution de X_1^n (2.3.1) implique que pour tout $1 \leq j \leq n$, $X_{\sigma_{|X|}(j)} \sim -X_{\sigma_{|X|}(j)}$ et donc

$$P(X_{\sigma_{|X|}(j)} > 0) = P(X_{\sigma_{|X|}(j)} < 0)$$

Ainsi si $P(X_{\sigma_{|X|}(j)} = 0) = 0$, alors

$$P(X_{\sigma_{|X|}(j)} > 0) = \frac{1}{2}$$

et donc

$$E(T_n^+) = \sum_{i=1}^{n} j P(X_{\sigma_{|X|}(j)} > 0) = \frac{n(n+1)}{4}$$

Comme les X_i sont diffuses, les variables $X_{\sigma_{|X|}(j)}$ sont également diffuses car nous avons $\forall\ x \in \mathbb{R}$

$$P(X_{\sigma_{|X|}(j)} = x) = \sum_{i=1}^{n} P(X_i = x, \sigma_{|X|}(j) = i) = 0$$

3. La symétrie par rapport à 0 implique que les vecteurs $(|X_1|,\dots,|X_n|)$ et $(\mathbb{1}_{\{X_1>0\}},\dots,\mathbb{1}_{\{X_n>0\}})$ sont indépendants. Les $X_{\sigma_{|X|}(i)}$ étant diffuses ceci implique que

$$(\mathbb{1}_{X_{\sigma_{|X|}(1)>0}},\dots,\mathbb{1}_{X_{\sigma_{|X|}(n)>0}}) \sim (Y_1,\dots,Y_n)$$

Y_i suit une loi de Bernoulli $\mathscr{B}\big(1/2\big), \forall\, i = 1,\dots,n.$ $\forall i = 1,\dots,n$, les Y_i, sont donc indépendantes identiquement distribuées.

En effet : Soit $(\epsilon_1,\dots,\epsilon_n) \in \{0,1\}^n$, on a

$$P\big((\mathbb{1}_{X_{\sigma_{|X|}(1)>0}},\dots,\mathbb{1}_{X_{\sigma_{|X|}(n)>0}}) = (\epsilon_1,\dots,\epsilon_n)\big)$$

$$\begin{aligned}
&= \sum_{s\in\mathfrak{S}_n} P\big((\mathbb{1}_{X_1>0},\dots,\mathbb{1}_{X_n>0}) = (\epsilon_1,\dots,\epsilon_n)\big)\\
&= \sum_{s\in\mathfrak{S}_n} P\big((\mathbb{1}_{X_{s(1)}>0},\dots,\mathbb{1}_{X_{s(n)}>0}) = (\epsilon_1,\dots,\epsilon_n), \sigma_{|X|} = s\big)\\
&= \sum_{s\in\mathfrak{S}_n} P\big((\mathbb{1}_{X_{s(1)}>0},\dots,\mathbb{1}_{X_{s(n)}>0}) = (\epsilon_1,\dots,\epsilon_n) P(\sigma_{|X|} = s\big)\\
&= \frac{1}{2^n} \sum_{s\in S_n} P(\sigma_{|X|} = s)\\
&= \frac{1}{2^n}\\
&= P\big((Y_1,\dots,Y_n) = (\epsilon_1,\dots,\epsilon_n)\big).
\end{aligned}$$

1. La première égalité provient du Théorème des probabilités totales.
2. La seconde vient de l'indépendance de $(|X_1|,\dots,|X_n|)$ et de $(\mathbb{1}_{X_{s(1)}>0},\dots,\mathbb{1}_{X_{s(n)}>0})$ à cause de l'indépendance avec le vecteur $(\mathbb{1}_{\{X_1>0\}},\dots,\mathbb{1}_{\{X_n>0\}})$ car s est fixe. ($\sigma_{|X|}$ est une fonction de $|X|$).
3. s est fixe et les variables $X_1,\dots,X_n$ sont indépendantes, donc

$$P(\mathbb{1}_{X_{s(1)}>0},\dots,\mathbb{1}_{X_{s(n)}>0}) = (\epsilon_1,\dots,\epsilon_n))$$
$$= P\big(\mathbb{1}_{X_{s(1)}>0} = \epsilon_1\big) \times \dots \times P\big(\mathbb{1}_{X_{s(n)}>0} = \epsilon_n\big).$$

   On a aussi $P(X_{\sigma_{|X|}(i)} > 0) = 1/2$. Pour montrer que le 3. du Théorème (2.3.1) est vrai on a : $T_n^+ \sim \sum_{j=1}^{n} jY_j$. La loi de T_n^+ et de T_n^- ne dépend que des rangs des observations, non des lois des X_i, d'où le fait qu'elles soient libres.
4. Pour la variance on a

$$\begin{aligned}
Var\left(\sum_{j=1}^{n} j\mathbb{1}_{\left\{X_{\sigma_{|X|}(j)}>0\right\}}\right) &= \sum_{j=1}^{n} j^2 Var(Y_j)\\
&= \sum_{j=1}^{n} \frac{j^2}{4} = \frac{n(n+1)(2n+1)}{24}
\end{aligned}$$

5. La convergence se déduit par le Théorème Central Limite. □

Remarque 2.3.2 *1. Sous H_0, T_n^+ a une loi symétrique par rapport à sa moyenne $\frac{n(n+1)}{4}$.*

En effet : *Comme $T_n^+ \sim T_n^-$ et $T_n^+ + T_n^- = \frac{n(n+1)}{2}$, on a :*

$$T_n^+ \sim \frac{n(n+1)}{2} - T_n^-$$

et avec $b = \frac{n(n+1)}{4}$,

$$T_n^+ \sim 2b - T_n^- \tag{2.19}$$

2. *On utilise le test exact de T_n^+ sous H_0 quand $n \leq 20$ (en raison des tables statistiques associées). Quand $n > 20$, on utilise un test asymptotique.*
3. Avec le logiciel **R** si l'échantillon est saisi dans un vecteur x, pour tester $H_0 : m = 0$ contre $H_1 : m \neq 0$, on utilise le test

wilcox.test(x,alternative="two.sided")

Exemple 2.3.3 On veut savoir si un entraînement modifie la tension artérielle de personnes. On a recueilli la tension systolique de 8 personnes, on leur a fait suivre un programme d'entraînement spécifique pendant 6 mois, puis on leur a mesuré à nouveau la tension.

TABLE 2.1 – Tension systolique

N°	Avant	Après	Écart	$\lvert$Écart$\rvert$	Rang$\left(\lvert$Écart$\rvert\right)$	Rang signé
1	130	120	10	10	5	5
2	170	163	7	7	4	4
3	125	120	5	5	2	2
4	170	135	35	35	7	7
5	130	143	–13	13	6	–6
6	130	136	–6	6	3	–3
7	145	144	1	1	1	1
8	160	120	40	40	8	8

– On calcule pour $n = 8$ les écarts entre les valeurs :

$$T_n^+ = 5 + 4 + 2 + 7 + 1 + 8 = 27.$$

– Dans la table des seuils critiques (cf. Annexe Fig 4.1), pour un test bilatéral à 5 %, nous lisons $T_{0.05} = 4$, pour un échantillon $n = 8$. On déduit que nous sommes dans la zone de rejet $(27 > 4)$ donc on a un RH_0. La tension artérielle n'est pas la même après un entraînement de 6 mois.

Remarque 2.3.3 *1. Dans le cas des écarts nuls, la solution est de supprimer les observations correspondantes.*

2. Lorsque nous avons des ex aequo, nous devons attribuer un rang identique aux observations qui présentent une valeur identique $|d_i|$. La statistique T_n^+ n'est pas modifiée. Sa variance le sera. Dans le cas où n devient grand, il faut utiliser l'approximation normale pour définir la région critique de Z variable normale centrée réduite.

En utilisant le Théorème (2.3.1) 5. pour n assez grand $(n > 15)$ on peut approximer la distribution de T_n^+ sous H_0 par une loi normale de paramètres

$$E(T_n^+) = \frac{1}{4}n(n+1).$$

$$Var(T_n^+) = \frac{1}{24}n(n+1)(2n+1).$$

La statistique du test centrée réduite suit une loi normale

$$Z = \frac{T_n^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}.$$

La région critique pour un test bilatéral au risque α s'écrit

$$R.C. : |Z| \geq u_{1-\alpha/2}.$$

Pour la correction de continuité, pour améliorer la précision, on utilise pour le test bilatéral :

$$|Z| = \frac{|T_n^+ - \frac{1}{4}n(n+1)| - 0.5}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}. \tag{2.20}$$

2.3.3 La correction de la variance pour les ex aequo

Soit G l'ensemble des valeurs distinctes $|d_i|$ et le cardinal de G (noté $|G|$) leur nombre, pour la valeur numéro g appartenant à G, on observe t_g observations, la variance corrigée s'écrit

$$\widetilde{Var}(T_n^+) = \frac{n(n+1)}{4} - \frac{1}{48}\sum_{g=1}^{|G|} t_g(t_g-1)(t_g+1). \tag{2.21}$$

La statistique centrée réduite servant à définir la région critique modifiée est

$$\widetilde{Z} = \frac{T_n^+ - \frac{1}{4}n(n+1)}{\sqrt{\widetilde{Var}(T_n^+)}}. \tag{2.22}$$

Exemple 2.3.4 Dans l'exemple des ballons gonflés à l'hélium et à l'air (Exemple 2.2.2 page 26), on applique le test de Wilcoxon des rangs signés. On a le tableau suivant :

n	37
T^+	398.5
$E(T^+)$	351.5

1. Avec correction pour ex aequo :

   (a) Les rangs moyens sont obtenus en faisant la moyenne des rangs des ex aequo.

   (b) La statistique du test T_n^+, somme des rangs positifs

$$T_n^+ = 2.5 + 9.5 + \ldots + 35.5 + 37 = 398.5$$

FIGURE 2.3 – Tableau des $|d_i|$ distinctes et comptage

g	valeur $\|d_i\|$	t_g	$t_g(t_g-1)(t_g+1)$
1	1	4	60
2	2	10	990
3	3	4	60
4	4	3	24
5	5	1	0
6	6	2	6
7	7	3	24
8	8	2	6
9	9	2	6
10	12	2	6
11	13	1	0
12	14	2	6
13	17	1	0
		Somme	1188

(c) Sous H_0 : $E(T_n^+) = \frac{1}{4}n(n+1) = \frac{1}{4}37(37+1) = 351.5$

(d) La variance non corrigée $Var(T_n^+) = \frac{1}{24}n(n+1)(2n+1) = \frac{1}{4}37(37+1)(2\times37+1) = 4393.75$
La statistique centrée réduite Z est égale à

$$Z = \frac{|T^+ - \frac{1}{4}n(n+1)| - 0.5}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} = \frac{398.5 - 351.3}{\sqrt{4393.75}} = 0.70906$$

Pour $\alpha = 0.05$, on doit comparer cette valeur à $u_{0.95} = 1.645$. $Z = 0.70906 < u_{0.95} = 1.645$, $\overline{RH_0}$. La p-valeur vaut 0.23914.

(e) **Calcul de $\widetilde{Var}(T_n^+)$**

$$\sum_{g=1}^{|G|} t_g(t_g-1)(t_g+1) = 1188$$

$$\begin{aligned}\widetilde{Var}(T_n^+) &= Var(T_n^+) - \frac{1}{48}\sum_{g=1}^{|G|} t_g(t_g-1)(t_g+1) \\ &= 4393.75 - \frac{1}{48}\times 1188 \\ &= 4369\end{aligned}$$

(f) La statistique centrée et réduite du test

$$\widetilde{Z} = \frac{398.5 - 351.5}{\sqrt{4369}} = 0.711.6$$

La probabilité critique du test unilatéral est $p = 0.23852$. $\overline{RH_0}$. Gonfler à l'hélium et à l'air sont identiques.

Les tableaux suivants résument les valeurs des variances dans les deux cas.

2. **Sans correction pour ex aequo**

$Var(T_n^+)$	4393.75
Z	0.70906
p-value	0.23914

3. **Les valeurs finales sont :**

$\widetilde{Var}(T_n^+)$	4369
$\widetilde{Z}$	0.71106
p-valeur	0.23852

FIGURE 2.4 – Gonflage à l'hélium de ballon-Test des signes

Essai	Hélium	Air	Écart	$\lvert d_i \rvert$	r_i	r'_i	Rang signé
1	25	25	0	0	–	–	–
21	31	31	0	0	–	–	–
7	25	26	–1	1	1	2.5	–2.5
13	28	27	1	1	2	2.5	2.5
26	28	29	–1	1	3	2.5	–2.5
31	26	27	–1	1	4	2.5	–2.5
4	14	16	–2	2	5	9.5	–9.5
8	26	24	2	2	6	9.5	9.5
9	22	24	–2	2	7	9.5	9.5
10	26	28	–2	2	8	9.5	–9.5
19	35	33	2	2	9	9.5	9.5
33	30	28	2	2	10	9.5	9.5
35	30	28	2	2	11	9.5	9.5
37	29	31	–2	2	12	9.5	–9.5
38	30	28	2	2	13	9.5	9.5
39	26	28	–2	2	14	9.5	9.5
16	29	26	3	3	15	16.5	16.5
17	23	20	3	3	16	16.5	16.5
24	32	29	3	3	17	16.5	16.5
34	29	32	–3	3	18	16.5	–16.5
18	26	22	4	4	19	20	20
28	27	31	–4	4	20	20	–20
36	29	25	4	4	21	20	20
20	24	29	–5	5	22	22	–22
14	31	25	6	6	23	23.5	23.5
32	32	26	6	6	24	23.5	23.5
2	16	23	–7	7	25	26	–26
3	25	18	7	7	26	26	26
22	34	27	7	7	27	26	26
27	30	22	8	8	28	28.5	28.5
29	33	25	8	8	29	28.5	28.5
12	28	19	9	9	30	30.5	30.5
30	11	20	–9	9	31	30.5	–30.5
5	23	35	–12	12	32	32.5	–32.5
15	22	34	–12	12	33	32.5	–32.5
11	12	25	–13	13	34	34	–34
6	29	15	14	14	35	35.5	35.5
11	29	15	14	14	35	35.5	35.5
25	14	28	–14	14	36	35.5	35.5
23	39	22	17	17	37	37	37

2.3.4 Échantillons appariés

On dispose de deux échantillons appariés de taille n : $(U_1,\ldots,U_n)$ et $(V_1,\ldots,V_n)$. On veut savoir si l'un des échantillons a une tendance à prendre des valeurs plus grandes que l'autre. On pose : $X_i = V_i - U_i$ et on utilise le test de Wilcoxon des rangs signés sur l'échantillon X_1^n. Les hypothèses sont

1. Les X_i sont indépendants entre eux (mais pas forcément de même loi).
2. Les X_i sont diffuses.
3. Les X_i ont une médiane.
4. Les X_i sont de loi symétrique par rapport à m.

$$H_0 : m = 0 \quad H_1 : m \neq 0$$

Avec le logiciel **R** on a : **>wilcox.test(x,y, paired=T, alternative="two.sided")** ou **>wilcox.test(x-y, alternative="two.sided")**.
Dans le cas où les données sont dans un data.frame on peut utiliser :

wilcox.test(data=nom du fichier, $y \sim x$, paired=T, alternative="greater").

2.4 Test de Mann et de Whitney

2.4.1 Préliminaires

Théorème 2.4.1 *Soient $X_1, X_2, \ldots, X_n$, n v.a identiquement distribuées de loi continue, de vecteur de statistiques d'ordre X^* et de vecteur des rangs R_X (cf. définition (2.3.1)), alors X^* et R_X sont indépendants, de plus R_X est distribué uniformément sur $\mathfrak{S}_n$. σ suit une loi uniforme sur $\mathfrak{S}_n$, par ailleurs σ et X^* sont indépendants.*

$$\forall\, s \in \mathfrak{S}_n,\ P(\sigma = s) = \frac{1}{card\,\mathfrak{S}_n} = \frac{1}{n!}$$

$\mathfrak{S}_n$ étant l'ensemble de toutes les permutations de l'ensemble $\{1,\ldots,n\}$

$$X^* = \left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right)$$

Exemple 2.4.1 Pour $n = 3$ on a

$$P(X_1 < X_2 < X_3) = P(X_1 < X_3 < X_2)$$

$$P(X_2 < X_3 < X_1) = P(X_3 < X_2 < X_1)$$

$$P(X_3 < X_2 < X_1) = P(X_3 < X_1 < X_2)$$

$$P(X_3 < X_1 < X_2) = 1/6$$

σ et X^* sont indépendants.

Remarque 2.4.1 *La principale conséquence de ce théorème est que la loi de R_X ne dépend pas de la loi des X_i. On pourra donc faire de l'estimation et des tests non paramétriques à l'aide des rangs des observations.*

Proposition 2.4.1 *Soient $X_1, X_2, \ldots, X_n$ n variables identiquement distribuées de loi continue de vecteur des rangs $R_X = (R_1, \ldots, R_n)$. Pour tout s tel que $1 \leq s \leq n$, et pour toute suite d'entiers distincts $(r_1, r_2, \ldots, r_s)$ dans $\{1, \ldots, n\}$ on a*

$$P\Big((R_1, R_2, \ldots, R_s) = (r_1, \ldots, r_s)\Big) = \frac{1}{n(n-1)\ldots(n-s+1)} \tag{2.23}$$

$\forall\, i \in \{1, \ldots, n\}$, R_i *suit une loi uniforme sur* $\{1, \ldots, n\}$.

Démonstration $r = (r_1, r_2, \ldots, r_n) \in \mathfrak{S}_n$

$$P\Big((R_1, R_2, \ldots, R_n) = r\Big) = \frac{1}{n!}$$

$$s = 1, \quad P\Big(R_1 = s\Big) = \frac{1}{n!}$$

$$s = 2, \quad P\Big((R_1, R_2) = (s_1, s_2)\Big) = P\Big(R_1 = s_1\Big) P\Big(R_2 = s_2 | R_1 = s_1\Big) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$$

Par récurrence on déduit le résultat. □

Proposition 2.4.2

$$\forall\, i \in \{1, \ldots, n\}, \; E\Big(R_i\Big) = \frac{n+1}{2} \tag{2.24}$$

$$\forall\, i \in \{1, \ldots, n\}, \quad Var\Big(R_i\Big) = \frac{n^2 - 1}{12} \tag{2.25}$$

$$Cov\Big(R_i, R_j\Big) = -\frac{n+1}{12} \tag{2.26}$$

Démonstration 1.

$$E\Big(R_i\Big) = \frac{1}{n} \sum_{i=1}^{n} i = \frac{n+1}{2}$$

2.

$$\mathrm{Var}\Big(R_i\Big) = E\Big(R_i^2\Big) - E^2\Big(R_i\Big)$$

$$E\Big(R_i^2\Big) = \sum_{i}^{n} i^2 = \frac{(n+1)(2n+1)}{6}$$

$$\mathrm{Var}\Big(R_i\Big) = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$\mathrm{Var}\Big(R_i\Big) = \frac{(n+1)}{12}\left(2(2n+1) - 3(n+1)\right) = \frac{(n+1)(n-1)}{12}$$

ou

$$\mathrm{Var}\Big(R_i\Big) = \frac{n^2 - 1}{12}$$

3.

$$\mathrm{Cov}\Big(R_i, R_j\Big) = \sum_{l} \sum_{k=1; k \neq l}^{n} \left(l - \frac{(n+1}{2}\right)\left(k - \frac{(n+1}{2}\right) P\left((R_i, R_j) = (k, l)\right)$$

$$\mathrm{Cov}\Big(R_i, R_j\Big) = \sum \sum_{l,k=1}^{n} \left(l - \frac{(n+1)}{2}\right)\left(k - \frac{(n+1}{2}\right) P\left((R_i, R_j) = (k, l)\right) -$$

$$\sum_{l=1}^{n}\left(l-\frac{(n+1)}{2}\right)^2 P\left((R_i,R_j)=(k,l)\right)$$

$$\mathrm{Cov}\Big(R_i,R_j\Big)=\left[\sum_{l=1}^{n}\left(l-\frac{(n+1}{2}\right)\cdot\sum_{k=1}^{n}\left(k-\frac{(n+1)}{2}\right)-\sum_{l=1}^{n}\left(l-\frac{(n+1)}{2}\right)^2\right]P\Big(\big(R_i,R_j\big)=\big(k,l\big)\Big)$$

Comme $P\Big(\big(R_i,R_j\big)=(k,l)\Big)=P\Big((R_1,R_2)=(k,l)\Big)$ et $\sum_{k=1}^{n}\left(k-\frac{(n+1)}{2}\right)=0$, on a :

$$\mathrm{Cov}\Big(R_i,R_j\Big)=-\sum_{l=1}^{n}\left(l-\frac{(n+1)}{2}\right)^2 P\Big(R_1=k,R_2=l\Big)$$

De plus $\sum_{l=1}^{n}\left(l-\frac{(n+1)}{2}\right)^2=n\mathrm{Var}(R_i)$ et $P\Big(R_1=k,R_2=l\Big)=\frac{1}{n(n-1)}$, en utilisant la proposition (2.23) avec $s=2$, on a

$$\mathrm{Cov}\Big(R_i,R_j\Big)=-nVar\Big(R_i\Big)\times\frac{1}{n(n-1)}=-n\frac{n^2-1}{12}\times\frac{1}{n(n-1)}$$

d'où :

$$\mathrm{Cov}\Big(R_i,R_j\Big)=-\frac{n^2-1}{12(n-1)}$$

donc

$$\mathrm{Cov}\Big(R_i,R_j\Big)=-\frac{n+1}{12}.$$

□

2.4.2 Le test de Mann et de Whitney

Étant donnés deux échantillons indépendants U_1^n et V_1^p de tailles respectives n et p. Ils sont chacun de loi notée U et V, on suppose U et V diffuses, n et p peuvent être différents. Le test non paramétrique peut être formulé par :

$$H_0:\ F=G\ \text{ contre } H_1:\exists\,\theta\neq 0 \text{ tel que } \mathrm{F}(.)=\mathrm{G}(.-\theta) \tag{2.27}$$

Exemple 2.4.2 On veut tester un nouveau médicament par rapport à un ancien. On donne le premier médicament à un groupe de n personnes et le deuxième à un groupe de p personnes, ces deux groupes sont indépendants. On veut savoir si le deuxième médicament est plus efficace que l'ancien.

La procédure de traitement consiste à fusionner les deux échantillons pour former un seul global de taille $n+p$, soit $(U_1,U_2,\ldots,U_n,V_1,\ldots,V_p)$.
Les vecteurs (U_i,V_i) sont indépendants, on les classe par leur rang global. On forme le vecteur de rangs global noté $R_{(U,V)}$. Il est constitué de

- $R_1^1,R_2^1,\ldots,R_n^1$ pour les rangs associés aux variables U_i.
- $R_1^2,\ldots,R_p^2$ pour les rangs associés aux variables V_j.

Exemple 2.4.3 Soient $U_1=3.5,\ U_2=4.7,\ U_3=1.2$ et $V_1=0.7,\ V_2=3.9$, on a donc

$$V_1<U_3<U_1<V_2<U_2$$

$R_1^1=3,\ R_2^1=5,\ R_3^1=2,\ R_1^2=1,R_2^2=4.$
Les vecteurs $(R_1^1,R_2^1,R_3^1),(R_1^2,R_2^2)$ sont les vecteurs des rangs associés aux (U_i,V_j).

Soient les statistiques qui vont nous permettre de définir les statistiques du test

$$S_1 = R_1^1 + R_2^1 + \ldots + R_n^1 \tag{2.28}$$
$$S_2 = R_1^2 + R_2^2 + \ldots + R_p^2 \tag{2.29}$$

Proposition 2.4.3

$$\frac{n(n+1)}{2} \leq S_1 \leq np + \frac{n(n+1)}{2} \tag{2.30}$$
$$\frac{p(p+1)}{2} \leq S_2 \leq np + \frac{p(p+1)}{2} \tag{2.31}$$

Démonstration Sous $H_0 : F = G$ sous les conditions données plus haut on a

$$\forall\, i = 1,\ldots,n, \forall\, j = 1,\ldots,p,\ E\left(R_i^1\right) = E\left(R_j^2\right) = \frac{n+p+1}{2}$$

$$\forall\, i = 1,\ldots,n, \forall\, j = 1,\ldots,p,\ Var\left(R_i^1\right) = Var\left(R_j^2\right) = \frac{(n+p)^2-1}{12}$$

$$E(S_1) = \frac{n(n+p+1)}{2};\ E(S_2) = \frac{p(n+p+1)}{2}$$

$$Var\left(S_1\right) = Var\left(S_2\right) = \frac{np(n+p+1)}{12}.$$

De plus on a

$$S_1 = R_1^1 + R_2^1 + \ldots + R_n^1 \geq \frac{n(n+1)}{2}$$

$$S_2 = R_1^2 + R_2^2 + \ldots + R_p^2 \geq \frac{p(p+1)}{2}$$

$$S_1 + S_2 = \frac{(n+p)(n+p+1)}{2}$$

$$S_1 \leq \frac{(n+p)(n+p+1)}{2} - \frac{p(p+1)}{2}$$

$$S_1 \leq np + \frac{n(n+1)}{2}$$

De même pour S_2,

$$S_2 \leq \frac{(n+p)(n+p+1)}{2} - \frac{n(n+1)}{2} = np + \frac{p(p+1)}{2}$$

Pour chaque composante R_i^1 ou R_i^2 du vecteur $R_{(U,V)}$ on a, $\forall\, i = 1,\ldots,N$

$$E(R_i^k) = \frac{(N+1)}{12}$$

$$Var(R_i^k) = \frac{N^2-1}{12}$$

k pouvant être égal à 1 ou 2, $N = n+p$ étant l'effectif total. Les sommes partielles dans 2.30 et dans 2.31 sont majorées par les sommes totales. □

Proposition 2.4.4 *1.*

$$E(S_1) = \frac{n(n+p+1)}{2} \tag{2.32}$$

2.

$$E(S_2) = \frac{p(n+p+1)}{2} \tag{2.33}$$

3.

$$Var(S_1) = \frac{(n+p+1)n(N-n)}{12} \tag{2.34}$$

Comme $N-n=p$ *et* $Var(S_1) = Var(S_2)$ *on a*

$$Var(S_1) = \frac{np(n+p+1)}{12}$$

De même

$$Var(S_2) = \frac{np(n+p+1)}{12}$$

Démonstration 1.

$$E(S_1) = E(R_1^1 + R_2^1 + \ldots + R_n^1) = nE(R_i^1) = n\frac{(N+1)}{2} = \frac{n(n+p+1)}{2}$$

2. De même

$$E(S_2) = E(R_1^2 + R_2^2 \ldots + R_p^2) = pE(R_i^2) = \frac{p(N+1)}{2} = \frac{p(n+p+1)}{2}.$$

Remarque 2.4.2 *Les* R_i^1 *et* R_j^2 *sont de même loi mais non indépendantes.*

3.

$$Var(S_1) = \sum_{i=1}^{n} Var(R_i^1) + \sum_{i=1}^{n}\sum_{j\neq i} \mathrm{Cov}((R_i, R_j)$$

$$Var(S_1) = \sum_{i=1}^{n} Var(R_i^1) - \frac{n(n-1)}{N-1} Var(R_1^1)$$

$Var(R_i^1) = \frac{N^2-1}{12}$, $\forall\, i = 1, \ldots, N$. *Les variances* ($Var(R_i^1)$, $\forall\, i$) *sont identiques.*

$$Var(S_1) = \frac{n(N-1)(N+1)}{12} - \frac{n(n-1)}{N-1} \cdot \frac{(N-1)(N+1)}{12}$$

$$Var(S_1) = \frac{(N+1)}{12} n(N-n)$$

$p = N-n$, *donc*

$$Var(S_1) = \frac{np(n+p+1)}{12}$$

Le calcul pour $Var(S_2)$ *se déduit du précédent de manière analogue, on a donc*

$$Var(S_2) = \frac{np(n+p+1)}{12}$$

Sous H_0 *: On constitue pour la suite les statistiques*

$$M_U = \sum_{i=1}^{n} R_i^1 - \frac{n(n+1)}{2} \quad \in \{0, 1, \ldots, np\} \tag{2.35}$$

$$M_V = \sum_{i=1}^{p} R_j^2 - \frac{p(p+1)}{2} \quad \in \{0, 1, \ldots, np\} \tag{2.36}$$

Proposition 2.4.5 *i)*

$$M_U + M_V = np \tag{2.37}$$

ii) Sous $H_0 : F = G$ *la loi de* M_U *est symétrique par rapport à* $\frac{np}{2}$.

iii) Sous $H_0 : F = G$, M_U *et* M_V *suivent la même distribution* ($M_U \sim M_V$).

iv) M_V *est égal au nombre de paires* (U_i, V_j) *parmi toutes les paires possibles telles que* $U_i < V_j$.

Démonstration *La démonstration est une conséquence de la démonstration de la proposition précédente.*

Exemple 2.4.4

1. On peut étudier le poids du cerveau d'un homme (en grammes). On relève les poids du cerveau de 10 hommes et de 10 femmes. La question est de savoir si la variable étudiée diffère entre les sexes.
2. La variable mesurée est le temps de survie (en jours) de patients atteints d'un cancer et traités avec un médicament donné. Cette variable dépend-elle du type de cancer ?

Le logiciel **R** nous fournit la fonction wilcox.test avec la p-value avec les instructions suivantes :

- >males <−$c(1381, 1349, 1258, 1355, 1335, 1416, 1475, 1421, 1383)$
  >females <−$c(1055, 1305, 1155, 1120, 1252, 1208, 1154, 1197, 1229, 1212)$
  wilcox.test(males,females,exact=F,correct=F).
- >estomac <−$c(124, 42, 25, 45, 412, 51, 1112, 46, 103, 876, 146, 340, 396)$
  >poumon <−$c(1235, 24, 1581, 1166, 40, 727, 3808, 791, 1804, 3460, 719)$
  wilcox.test(estomac,poumon,exact=F,correct=F)

Remarque 2.4.3 *On posera dans la suite* $M_U = U_1$, $M_V = U_2$ *et on étudiera la statistique* $U = \min(U_1, U_2)$ *qui vaut soit* U_1, *soit* U_2.

Exemple 2.4.5 On dispose de deux groupes d'individus, un premier exerçant une activié sportive journalière et le second, n'ayant jamais fait de sport. On étudie pour chaque groupe l'indice de masse corporelle (IMC). C'est une variable qui mesure la corpulence d'une personne. On peut le calculer par la formule : $IMC = \frac{poids(Kg)}{(taille(cm))^2}$. Une IMC normale est telle que :

$$18.5 < IMC < 25$$

On a le tableau des valeurs :

TABLE 2.2 – Indice de masse corporelle (IMC)

Numéro global	Numéro dans chaque groupe	IMC	Sport	Rang
1	1	22.8	DAILY	1
2	2	23.4	DAILY	3
3	3	23.6	DAILY	4
4	4	23.7	DAILY	5
5	5	24.8	DAILY	6
6	6	26.1	DAILY	8
7	7	30.2	DAILY	12
8	1	23	NEVER	2
9	2	26	NEVER	7
10	3	26.3	NEVER	9
11	4	27.3	NEVER	10
12	5	28.7	NEVER	11
13	6	33.5	NEVER	12
14	7	35.3	NEVER	14

n_1	7
$\overline{r}_1$	5.571

n_2	7
$\overline{r}_2$	9.429

Pour le premier groupe

$$S_1 = \sum_{i=1}^{7} r_{i_1}^1 = 1+3+4+5+6+8+12 = 39$$

Le rang moyen $= \frac{S_1}{n_1} = \frac{39}{7} = 5.571$ Pour le groupe 2, $n_2 = 7$, $S_2 = 66$, le rang moyen $= \frac{66}{7} = 9.429$
Test bilatéral : $H_0 : \theta = 0 \qquad H_0 : F_1(X) = F_2(X+0)$
Test bilatéral : $H_1 : \theta \neq 0 \qquad H_1 : F_1(X) = F_2(X+\theta)$

Nous allons calculer les quantités suivantes

$$\left\{ \begin{array}{rcl} U_1 & = & S_1 - \frac{n_1(n_1+1)}{2} \\ U_2 & = & S_2 - \frac{n_2(n_2+1)}{2} \end{array} \right. \tag{2.38}$$

La statistique de Mann et de Whitney correspond à $U = \min(U_1, U_2)$

Sous H_0

$$E(U) = \frac{1}{2} n_1 n_2 \tag{2.39}$$

$$Var(U) = \frac{1}{12}(n_1 + n_2 + 1) n_1 n_2. \tag{2.40}$$

La région critique du test correspond aux valeurs exagérément élevées ou exagérément faibles de U par rapport à son espérance.

$$\begin{cases} U_1 = 39 - \frac{7(7+1)}{2} = 11 \\ U_2 = 66 - \frac{7(7+1)}{2} = 38 \end{cases}$$
$$U = \min(U_1, U_2) = \min(11, 38) = 11.$$

La probabilité critique vaut $2P(MW \leq U) = 2P(MW \leq 11) = 2 \times 0.049 = 0.098$, au seuil $\alpha = 0.05$.

En utilisant la table de la Figure 2 (cf. Annexe) de Mann et Whitney pour les échantillons indépendants, on a un non rejet de H_0. On ne peut pas rejeter l'hypothèse d'égalité de l'amplitude des réactions des individus selon leur groupe d'appartenance (jeunes ou personnes âgées).

2.4.3 Approximation normale

Lorsque les tailles des échantillons deviennent assez élevées ($n \geq 8$, $p \geq 8$ cf. [31] page 343), la distribution de M_U et de M_V convergent vers la loi normale de moyenne $E(M_U)$ et de variance $Var(M_U)$. p étant la taille de l'échantillon 2, n celle l'échantillon 1.

Proposition 2.4.6 *Pour un test bilatéral, nous définissons la statistique centrée et réduite*

$$Z = \frac{U - \frac{1}{2}np}{\sqrt{\frac{1}{12}(n+p+1)np}} \tag{2.41}$$

La région critique du test au niveau de signification α est

$$R.C : |Z| \geq u_{1-\alpha/2}, \tag{2.42}$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

2.4.4 La correction de continuité

Quand les effectifs sont modérés, on peut améliorer l'approximation normale en introduisant la correction de continuité

$$|Z| = \frac{|U - E(U)| - 0.5}{\sqrt{Var(U)}}$$

La région critique R.C est $|Z| \geq u_{1-\frac{\alpha}{2}}$.
Dans l'exemple précédent

$$|Z| = \frac{|11 - 24.5| - 0.5}{\sqrt{61.25}} = 1.6611$$

p-value$= 0.0967$, car $2P(N > 1.66) = 0.0967$. N désignant la loi normale centrée réduite.

2.4.5 Cas des tests unilatéraux

Pour les test unilatéraux ($\theta < 0$, ou $\theta > 0$) (cf. (2.27))

$$Z = \frac{U - E(U) \pm 0.5}{\sqrt{Var(U)}} \tag{2.43}$$

Cas d'un test unilatéral à gauche $+0.5$
Cas d'un test unilatéral à droite -0.5
La région critique au risque α, R.C : $Z \leq u_\alpha$ pour le cas d'un test unilatéral à gauche et R.C : $Z \geq u_{1-\alpha}$ pour le cas d'un test unilatéral à droite.

2.4.6 Correction pour les ex aequo

Le problème des ex aequo peut surgir et les formules des statistiques données ne sont pas applicables. La variance de la statistique doit être corrigée par la variance suivante :

$$\widetilde{Var}\Big(U\Big) = Var\Big(U\Big)\left(1 - \frac{\sum_{g=1}^{|G|} t_g(t_g^2-1)}{n^3-n}\right)$$

$N = n+p$ est l'effectif total, $|G|$ est le nombre de valeurs distinctes dans l'échantillon Ω, t_g est le nombre d'observations associées à la valeur numéro g.

Exemple 2.4.6 On considère X la variable aléatoire qui correspond au niveau d'anxiété d'enfants face à la socialisation orale dans les sociétés primitives (table 2.3). On oppose le groupe $n_1 = 16$ enfants issus d'une société où la tradition orale expliquant les effets des différentes maladies est "absente", avec celui où elle est "présente", $n_2 = 23$. La somme et la moyenne des rangs pour chaque groupe :

1. Le groupe numéro 1 (tradition absente) : $n_1 = 16,\ S_1 = 200,\ \overline{r}_1 = 12.5$
2. Le groupe numéro 2 (tradition présente) : $n_1 = 23,\ S_2 = 580,\ \overline{r}_2 = 25.22$

Le calcul de U et de $|Z|$ sans la correction pour les ex aequo.

- On calcule $U_1 = 200 - \frac{16(16+1)}{2} = 64,\ U_2 = 580 - \frac{23(23+1)}{2} = 304$
  Nous déduisons $U = min(64, 304)$
- $E(U) = \frac{16\times 23}{2} = 184$
- $Var(U) = \frac{(16+23+1)\times 16\times 23}{12} = 1226.6667$
- La statistique $|Z|$ pour le test bilatéral s'obtient par :

$$|Z| = \frac{|64-184|}{\sqrt{1226.6667}} = 3.4262$$

- La p-value (la p-valeur) est $p = 0.00061$. Au risque 5%, nous concluons à une différence significative entre les niveaux d'anxiété des enfants.

Pour la correction dans le cas des ex aequo, nous calculons les t_g. Il y a 12 valeurs différentes. ($v_1 = 6, v_2 = 7, v_3 = 8, \ldots, v_{12} = 17$). On leur associe les effectifs associés qui sont : $t_1 = 2,\ t_2 = 5,\ \ldots,\ t_{12} = 1$

$$\sum_{g=1}^{|G|=12} t_g(t_g^2-1) = 6 + 120 + 60 + 0 + 336 + \ldots + 0 = 846$$

$N = 16 + 23 = 39$ La nouvelle variance devient :

$$\widetilde{Var}\Big(U\Big) = Var\Big(U\Big) \times \left(1 - \frac{\sum_{g=1}^{G} t_g(t_g^2-1)}{n^3-n}\right) \tag{2.44}$$

$$= 1226.6667 \times \left(1 - \frac{846}{39^3-39}\right) \tag{2.45}$$

$$= 1209.166 \tag{2.46}$$

Donc $|Z| = \frac{|64-184|}{\sqrt{1209.1606}} = 3.4510$ et la probabilité critique du test est $p = 0.00056$.
La correction est assez faible.
Le tableau de la table 2.3 (page 47) correspond aux données de l'Exemple 2.4.6.

TABLE 2.3 – Comparaison de l'anxiété

Tradition orale	Niveau anxiété	Rang brut	Rang moyen
absent	6	1	1.5
present	6	2	1.5
absent	7	3	5
absent	7	4	5
absent	7	4	5
absent	7	6	5
absent	7	7	5
absent	8	8	9.5
absent	8	9	9.5
present	8	10	9.5
present	8	11	9.5
absent	9	12	12
absent	10	13	16
absent	10	14	16
absent	10	15	16
absent	10	16	16
present	10	17	16
present	10	18	16
present	10	19	16
present	11	20	20.5
present	11	21	20.5
absent	12	22	24.5
absent	12	23	24.5
present	12	24	24.5
present	12	25	24.5
present	12	26	24.5
present	12	27	24.5
absent	13	28	29.5
present	13	29	29.5
present	13	30	29.5
present	13	31	29.5
present	14	32	33
present	14	33	33
present	14	34	33
present	15	35	36
present	15	36	36
present	15	37	36
present	16	38	38
present	17	39	39

2.5 Le test de Kruskal-Wallis

On généralise à présent à k échantillons avec $k>2$. Pour $i \in \{1,\dots,k\}$, le $i^{\text{ème}}$ échantillon est noté $(X_1^i,\dots,X_{n_i}^i)$, n_i étant la taille de l'échantillon numéro i. Le nombre total d'observations est $n=\sum_{i=1}^{k} n_i$. Le test considéré est :

H_0 : "Les k- échantillons sont de même loi"contre $H_1=\overline{H_0}$, où $\overline{H_0}$ désigne le fait qu'il existe un échantillon parmi les k qui ne soit pas de la même loi que les autres. On ordonne l'ensemble des n observations regroupées en un seul échantillon et on note par R_j^i le rang de X_j^i dans l'échantillon global, et par $R^i=\sum_{j=1}^{n_i} R_j^i$ la somme des rangs des observations du $i^{\text{ème}}$ échantillon dans l'échantilllon global.

Définition 2.5.1 *Le test de Kruskal-Wallis est basé sur la statistique de Kruskal-Wallis définie par*

$$K_n=\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{\left(R^i\right)^2}{n_i}-3(n+1) \tag{2.47}$$

Expression qui peut encore s'écrire :

$$\frac{12}{n(n+1)}\sum_{i=1}^{k} n_i(\overline{r_i}-\overline{r})^2 \tag{2.48}$$

$\overline{r}=\frac{(n+1)}{2}$, *représente la moyenne globale des rangs et* $\overline{r_i}$ *la moyenne des rangs de l'échantillon numéro* i.

Proposition 2.5.1 *Sous* H_0

$$K_n \xrightarrow[n\longrightarrow+\infty]{\mathscr{L}} \chi^2_{k-1} \tag{2.49}$$

Démonstration On pourra trouver la démonstration dans ([4] page 224). En pratique la valeur de n_k ([4] page 226) doit être supérieure à 5 pour tout k. Le but est dans l'application de ce test. Le test de Kruskal-Wallis consiste à rejeter l'hypothèse d'égalité des k lois si K_n est trop "grand". La région critique $W=\left\{K_n>z_{k-1,\alpha}\right\}$, $z_{k-1,\alpha}$ étant le quantile d'ordre $1-\alpha$ de la loi χ^2_{k-1}, où χ^2_{k-1} désigne la loi du Chi-deux à $k-1$ degrés de liberté.

Chapitre 3

Régression non paramétrique

3.1 Introduction

La régression non paramétrique est une méthode qui fournit une procédure automatique d'ajustement quel que soit le type de données. Elle fait partie des méthodes dites *adaptatives.* Elle a l'inconvénient qu'elle ne propose pas un modèle facilement réutilisable pour la prévision. Cependant, elle présente une description point par point de la fonction de régression. Le contexte est donc le suivant :
On cherche à expliquer les valeurs que peut prendre une variable Y à partir des valeurs que prend une variable X. On n'émet aucune hypothèse sur la forme de la corrélation entre X et Y.

Exemple 3.1.1 Y peut être le niveau du diplôme obtenu, qu'on explique par X =(âge, sexe, revenu des parents, métier des parents).

On suppose que la variable Y est intégrable $(E(|Y|) < \infty)$, on note r la fonction de régression de Y sur X.

$$r\Big(x\Big) = E\Big(Y|X = x\Big) \tag{3.1}$$

L'objectif est d'estimer la fonction r pour expliquer et prédire Y à partir de X. Pour cela, on dispose des réalisations de n couples de variables $(X_1, Y_1), \ldots, (X_n, Y_n)$.
On suppose que les (X_i, Y_i) sont indépendants. Les variables Y_i sont les variables à expliquer ou les variables réponses ou variables de sortie. Les variables X_i (design) sont les variables explicatives, les covariables ou variables d'entrée. Les X_i pouvant être aléatoires ou déterministes que l'on notera par x_i au lieu X_i. $r(x) = E(Y|X = x)$ peut s'écrire $Y = r(X) + \epsilon$ avec $E(\epsilon|X) = 0$.
On aura donc pour l'échantillon :

$$\forall\ i = 1, \ldots n, \qquad Y_i = r\Big(X_i\Big) + \epsilon_i, \quad E\Big(\epsilon_i|\, X_i\Big) = 0 \tag{3.2}$$

En particulier on a donc : $E(\epsilon) = 0$. Les ϵ_i sont appelés erreurs et jouent le rôle de bruit. On émet une hypothèse courante :

$$\forall\ i = 1, \ldots, n, \qquad Var(\epsilon_i) = \sigma^2 < \infty$$

3.2 Le modèle linéaire : rappels

3.2.1 Les EMC non paramétriques

La notation EMC désigne les estimateurs des moindres carrés (cf. [31]). On va supposer que r *s'écrit sous la forme* $r(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, *avec* $x = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^p$.

$$\forall\ i = 1, \ldots, n,\ r(X_i) = \beta_0 + \beta_1 X_{i_1} + \ldots + \beta_p X_{ip} = X_i^T \beta$$

où X_i^T *désigne la transposée de* X_i. *On note*

$$X = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{pmatrix} \tag{3.3}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

L'estimation de r *revient à l'estimation du vecteur* β. *Il s'agit d'un problème paramétrique. On utilise les moindres carrés ordinaires.*

$$\begin{aligned} \widehat{\beta} &= Arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 \\ &= Arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 \end{aligned}$$

Si X *est injective* $\left(rang(X) = p\right)$ *alors* $X^T X$ *est inversible et on a*

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \tag{3.4}$$

On a également

$$\widehat{Y} = X\widehat{\beta} = X(X^T X)^{-1} X^T Y = AY \tag{3.5}$$

où $A = X(X^T X)^{-1} X^T$. *On déduit que*

$$\widehat{r}(x) = (1, x^T)\widehat{\beta} \qquad \text{pour } x \in \mathbb{R}^p$$

$x = (x_1, x_2, \ldots, x_p)$.

Propriétés 3.2.1

1.

$$E\left[\widehat{\beta}\right] = \beta$$

2.

$$Var\left[\widehat{\beta}\right] = \sigma^2 (X^T X)^{-1}$$

Démonstration 1.

$$E\left[(X^T X)^{-1} X^T Y\right] = (X^T X)^{-1} X^T E\left[X\beta + \epsilon\right],$$

donc et comme $E[\epsilon] = 0$

$$E\left[\widehat{\beta}\right] = (X^T X)^{-1} X^T X\beta = \beta$$

2.

$$\begin{aligned} Var\left[\widehat{\beta}\right] &= Var\left[(X^TX)^{-1}X^TY\right] \\ &= (X^TX)^{-1}X^T\sigma^2 I_n X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

□

3.3 Estimateur de Nadaraya-Watson

On suppose que les (X_i, Y_i) admettent une densité $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, pour tout $x > 0$, $f_X(x) = \int f(x,y)dy > 0$ $\left(f_X \text{ est la densité de } X\right)$, on a donc

$$\forall\, x \in \mathbb{R},\ r(x) = E(Y|X = x) = \int \frac{yf(x,y)}{f_X(x)}dy \tag{3.6}$$

Donc pour estimer r, on peut passer par l'estimation de f et f_X et poser

$$\widehat{r}_n(x) = \begin{cases} \int \frac{y\widehat{f}_n(x,y)}{\widehat{f}_{n,X}(x)}dy & \text{si } \widehat{f}_{n,X}(x) \neq 0 \\ 0 & \text{si } \widehat{f}_{n,X}(x) = 0. \end{cases} \tag{3.7}$$

On utilise les estimateurs à noyau du chapitre précédent

$$\widehat{f}_{n,X}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{3.8}$$

$$\widehat{f}_n(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)K\left(\frac{Y_i - y}{h}\right)$$

Proposition 3.3.1 *Si K est un noyau d'ordre 1 alors $\forall\, x \in \mathbb{R}$ on a*

$$\widehat{r}_n(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K(\frac{X_i - x}{h})} & \text{si } \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0 \\ 0 & \text{sinon} \end{cases} \tag{3.9}$$

Démonstration

$$\widehat{f}_{n,X}(x) = 0 \Leftrightarrow \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) = 0$$

Supposons maintenant que $\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0$ donc

$$\begin{aligned}
\widehat{r}_n(x) &= \int \frac{y\widehat{f}_n(x,y)}{\widehat{f}_{n,X}(x)} dy && (3.10)\\
&= \frac{1}{\widehat{f}_{n,X}(x)} \int y \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right) dy && (3.11)\\
&= \frac{nh}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \int yK\left(\frac{Y_i - y}{h}\right) dy && (3.12)\\
&= \frac{1}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \frac{1}{h} \int yK\left(\frac{Y_i - y}{h}\right) dy && (3.13)\\
\widehat{r}_n(x) &= \frac{1}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i && (3.14)
\end{aligned}$$

On a utilisé le fait que

$$\frac{1}{h} \int yK\left(\frac{Y_i - x}{h}\right) dy = \frac{1}{h} \int (Y_i - uh) K(u) h du = Y_i \int K(u) du - h \int uK(u) du = Y_i.$$

Car $\int uK(u)du = 0$ et $\int K(u)du = 1$. □

Remarque 3.3.1 Avec le logiciel **R** on utilise,
>**kerf1 <-ksmooth(x,y,"normal",bandwidth=valeur de h)** ### valeur de h=2 ### par exemple ###
>**lines(kerf1, col="red")**

3.3.1 Biais et variance

– **Cas des x_i fixés $(x_i \in [a,b],\ a,\ b \in \mathbb{R})$**

On pose $w_i = K\left(\frac{x - x_i}{h}\right)$, $X_i = x_i$ on a donc

$$E\left[\widehat{r}_n(x)\right] - r(x) = \frac{\sum_{i=1}^{n} w_i \left(r(x_i) - r(x)\right)}{\sum_{i=1}^{n} w_i}. \quad (3.15)$$

$$Var\left[\widehat{r}_n(x)\right] = \sigma^2 \frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2}. \quad (3.16)$$

Comme les $x_i \in [a,b]$ $K(u)=0$, si $u \in [-1,+1]^c$ (cf. Exemples (1.4.1) du chapitre 1, $[-1,+1]^c$ étant le complémentaire de $[-1,+1]$) et $\forall u \in \mathbb{R}, K(u)=K(-u)$, on a

$$E\Big[\widehat{r}_n(x)\Big]-r\big(x\big) = \int_{-1}^{+1} K(u)\Big[r(x+uh)-r(x)\Big]du+o\Big(\frac{1}{nh}\Big) \tag{3.17}$$

$$Var\Big[\widehat{r}_n(x)\Big] = \frac{(b-a)}{nh}\sigma^2\int_{-1}^{+1}\big(K(u)\big)^2du+o\Big(\frac{1}{n^2}\Big) \tag{3.18}$$

On considère le développement de Taylor de r :

$$r(x+uh)=r(x)+uhr'(x)+\frac{1}{2}(uh)^2r''(x)+o(h^2)$$

nous avons alors

$$E\Big[\widehat{r}_n(x)\Big]-r\big(x\big)=\frac{h^2}{2}r''(x)\int_{-1}^{+1}u^2K(u)du+o\Big(h^2\Big)+o\Big(\frac{1}{nh}\Big) \tag{3.19}$$

– **Cas des X_i aléatoires**

On suppose que la densité conjointe $f_{X,Y}$ du couple (X,Y) est continue dans $\mathbb{R}^2$, on peut montrer que $\widehat{r}_n(x)$ converge en probabilité vers $r(x)$ en tout point tel que $f_X(x)\neq 0$. Sous certaines hypothèses de $f_{(X,Y)}$ et de K on a

$$E\Big[\widehat{r}_n(x)\Big]-r\big(x\big) = h^2\int_{-1}^{+1}u^2K(u)du\left[r'(x)\frac{f'_X(x)}{f_X(x)}+\frac{1}{2}r''(x)\right]+o(h^2) \tag{3.20}$$

$$Var\Big[\widehat{r}_n(x)\Big] = \frac{\sigma^2}{nh}\frac{1}{f_X(x)}\int_{-1}^{+1}\Big[K(u)\Big]^2du+o\Big(\frac{1}{nh}\Big) \tag{3.21}$$

Remarque 3.3.2 1. *Si K est continue, positif et à support sur $\mathbb{R}$ (par exemple le noyau gaussien) alors $\widehat{r}_n(x)$ est aussi continue.*

2. *On peut aussi écrire*

$$\widehat{r}_n(x)=\sum_{i=1}^{n}\omega_{n,i}(x)Y_i$$

où

$$\omega_{n,i}(x)=\begin{cases}\dfrac{K\left(\frac{X_i-x}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)} & \text{si } \sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)\neq 0\\ 0 & \text{sinon}\end{cases}$$

3. *Si $\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)=0$ i.e. si x se trouve dans une zone où il n'y a pas de X_i, alors $\widehat{r}(x)=0$. Sinon $\sum_{i=1}^{n}\omega_{n,i}(x)=1$ alors Y_i est une moyenne pondérée des Y_i qui correspondent aux points X_i proches de x.*

4. *Il se peut que la densité f_X soit connue. Dans ce cas, on préfère utiliser*

$$\widetilde{r}_n(x)=\begin{cases}\int\frac{y\widehat{f}_n(x,y)}{f_X(x)}dy & \text{si } f_X(x)\neq 0\\ 0 & \text{si } f_X(x)=0\end{cases} \tag{3.22}$$

i.e. si K est un noyau d'ordre 1

$$\widetilde{r}_n(x)=\begin{cases}\frac{1}{nhf_X(x)}\sum_{i=1}^{n}Y_iK\left(\frac{X_i-x}{h}\right) & \text{si } f_X(x)\neq 0\\ 0 & \text{si } f_X(x)=0\end{cases} \tag{3.23}$$

3.3.2 Majoration du MSE

Pour le risque quadratique $E\Big[\big(\widetilde{r}_n(x)-r(x)\big)^2\Big]$ *nous allons établir une majoration.*

Proposition 3.3.2 *On suppose que* f_X *connue et on s'intéresse à l'estimation de* $r(x)$ *pour* x *fixé. Soit* K *un noyau d'ordre 1. On suppose de plus que*

Hypothèses 3.3.1 i) $f_X(x)>0,\ \forall x\in\mathbb{R}$.

ii) Il existe $\epsilon>0$ tel que les fonctions f_X et r sont continûment dérivables sur $[x-\epsilon,x+\epsilon]$.

iii) Pour tout y, si $|u|\le\epsilon$

$$\big|f(x+u,y)-f(x,y)\big|\le M(x,y)\epsilon, \tag{3.24}$$

où

$$\int y^2M(x,y)dy<\infty \quad et \quad \int y^2f(x,y)dy<+\infty.$$

iv) K est un noyau à support dans $[-1,+1]$ et de carré intégrable.

Alors, si $|h|\le\epsilon$ il existe une constante $C(x)$ (dépendant de x) telle que

$$E\left[\Big(\widetilde{r}_n(x)-r(x)\Big)^2\right]\le C(x)\left(h^2+\frac{1}{nh}\right) \tag{3.25}$$

De plus si on choisit une fenêtre h telle que $h\approx n^{-1/3}$ ($\approx$ signifie de l'ordre de), il existe une constante $C'(x)$ telle que :

$$E\left[\Big(\widetilde{r}_n(x)-r(x)\Big)^2\right]\le C'(x)n^{-2/3} \tag{3.26}$$

Démonstration On a la décomposition standard

$$E\left[\Big(\widetilde{r}_n(x)-r(x)\Big)^2\right]=(\text{Biais})^2+\text{Variance}$$

1. **Biais** : On va montrer que $Var\left[Y_1K\left(\frac{X_1-x}{h}\right)\right]<\infty$.
   On sait que $|X|\le 1+X^2$, donc si $E(X^2)<\infty$ alors $E(|X|)<\infty$ et $Var(X)<\infty$.

$$\begin{aligned}
E\Big[\widetilde{r}_n(x)\Big] &= E\left[\frac{1}{nhf_X(x)}\sum_{i=1}^n Y_iK\left(\frac{X_i-x}{h}\right)\right]\\
&= \frac{1}{hf_X(x)}E\left[Y_1K\left(\frac{X_1-x}{h}\right)\right]\\
&= \frac{1}{hf_X(x)}\int\int yK\left(\frac{t-x}{h}\right)f(t,y)dtdy\\
&= \frac{1}{f_X(x)}\int\int yK(v)f(x+vh,y)dvdy
\end{aligned}$$

   De plus

$$\begin{aligned}
r(x) &= E\Big[Y|X=x\Big]\\
&= \int yf_Y\big(y|X=x\big)dy\\
&= \int y\frac{f(x,y)}{f_X(x)}dy
\end{aligned}$$

On déduit

$$r(x)f_X(x) = \int yf(x,y)dy$$

De même

$$r(x+vh)f_X(x+vh) = \int yf(x+vh,y)dy$$

Donc

$$E\left[\widetilde{r}_n(x)\right] - r\left(x\right) =$$

$$\begin{aligned}
&= \frac{1}{f_X(x)}\left[\int\int yK(v)f(x+vh,y)dvdy - \int yf(x,y)dy\right] \\
&= \frac{1}{f_X(x)}\left[\int\int yK(v)f(x+vh,y)dvdy - \int\int yK(v)f(x,y)dvdy\right] \\
&= \frac{1}{f_X(x)}\left[\int K(v)f_X(x+vh)r(x+vh)dv - \int K(v)r(x)f_X(x)dv\right] \\
&= \frac{1}{f_X(x)}\left[\int K(v)\left[f_X(x+vh) - f_X(x) + f_X(x)\right]r(x+vh)dv\right] \\
&\qquad -\frac{1}{f_X(x)}\left[\int K(v)r(x)f_X(x)dv\right] \\
&= \frac{1}{f_X(x)}\left[\int_{-1}^{+1} K(v)\left[f(x+vh,y) - f_X(x)\right]r(x+vh)dv\right] \\
&\qquad +\frac{1}{f_X(x)}\left[\int_{-1}^{+1} K(v)f_X(x)\left[r(x+vh) - r(x)\right]dv\right]
\end{aligned}$$

car K est à support dans $[-1,+1]$. Par le théorème des accroissements finis appliqué à r et à f_X, et grâce au fait qu'elles sont continûment dérivables au voisinage de x, il existe une constante $C(x)$ telle que, pour tout $|u| \leq \epsilon$

$$\left|r(x+u) - r(x)\right| \leq C(x)u$$

et

$$\left|f_X(x+u) - f_X(x)\right| \leq C(x)u.$$

On applique ces inégalités avec $u = vh$ pour $|v| \leq 1$ et $|h| \leq \epsilon$

$$\left|E\left[\widetilde{r}_n(x)\right] - r(x)\right|$$

$$\begin{aligned}
&\leq \frac{1}{f_X(x)}\Big[\int_{-1}^{+1}\Big|K(v)\Big|\Big|f(x+vh,y) - f_X(x)\Big|\Big|r(x+vh)\Big|dv\Big] \\
&\qquad +\int_{-1}^{+1}\Big|K(v)\Big|\Big|r(x+vh) - r(x)\Big|dv \\
&\leq \frac{C(x)}{f_X(x)}\left[\int_{-1}^{+1}|K(v)||hv||r(x+vh)|\,dv\right] + C(x)\int_{-1}^{+1}\Big|K(v)\Big|\Big|hv\Big|dv.
\end{aligned}$$

Comme r est continue sur $[x-\epsilon, x+\epsilon]$, il existe une constante $c(x)$ telle que $|r(x+hv)| \leq c(x)$, pour tout $|h| \leq \epsilon$ et tout $|v| \leq 1$. On a donc :

$$\left|E\left[\widetilde{r}_n(x)\right] - r(x)\right| \leq C_1(x)h \tag{3.27}$$

si on pose $C_1(x) = C(x)\left(\frac{c(x)}{f_X(x)} + 1\right)\int |k(v)|\,dv$.

2. **Calcul de la variance**

$$\begin{aligned} Var\Big[\widetilde{r}_n(x)\Big] &= Var\Big[\frac{1}{nhf_X}\sum_{i=1}^{n} Y_i K(\frac{X_i - x}{h})\Big] \\ &= nVar\Big[\frac{1}{nhf_X(x)} Y_1 K(\frac{X_1 - x}{h})\Big] \\ &\leq n\frac{1}{n^2h^2f_X^2(x)} E\Big(Y_1^2 K^2(\frac{X_1 - x}{h})\Big) \\ &= \frac{1}{nh^2f_X^2(x)}\int\int y^2 K^2\Big(\frac{t-x}{h}\Big) f(t,y)dtdy \\ &= \frac{1}{nhf_X^2(x)}\int\int y^2 K^2(v) f(x+vh,y)dvdy \end{aligned}$$

$$\forall v \in [-1,+1], |h| \leq \epsilon, |hv| \leq \epsilon$$

En utilisant les hypothèses (iii) de la Proposition (3.3.2) on a

$$\big|f(x+hv,y) - f(x,y)\big| \leq M(x,y)\epsilon$$

on déduit

$$f(x+hv,y) \leq f(x,y) + M(x,y)\epsilon$$

Donc

$$\begin{aligned} Var\Big[\widetilde{r}_n(x)\Big] &\leq \frac{1}{nhf_X^2(x)}\Big(\int\int y^2K^2(v)M(x,y)\epsilon dvdy + \int\int y^2K^2(v)f(x,y)dvdy\Big) \\ &= \frac{\int K^2(v)dv}{nhf_X^2(x)}\Big(\epsilon\int y^2M(x,y)dy + \int y^2 f(x,y)dy\Big). \end{aligned}$$

Si $|h| \leq \epsilon$

$$Var\Big[\widetilde{r}_n(x)\Big] \leq \frac{C_2(x)}{nh} \tag{3.28}$$

où la constante $C_2(x) = \frac{\int K^2(v)dv}{f_X^2(x)}\Big(\epsilon\int y^2M(x,y)dy + \int y^2 f(x,y)dy\Big)$ est finie.
(Hypothèses iii). et iv) pour que les intégrales ci-dessus soient finies)

3. **Calcul du risque quadratique**

$$E\Big[\big(\widetilde{r}_n(x) - r(x)\big)^2\Big] \leq C_1^2(x)/h^2 + \frac{C_2(x)}{nh} \tag{3.29}$$

On a

$$h^2 \approx \frac{1}{nh} \Leftrightarrow h \approx n^{-\frac{2}{3}}$$

Si on choisit une fenêtre $h^* = cn^{-\frac{1}{3}}$ avec une constante positive, on a

$$E\Big[\big(\widetilde{r}_n(x) - r(x)\big)^2\Big] \leq C_3(x)n^{-2/3} \tag{3.30}$$

Les constantes $C(x)$ et $C'(x)$ des énoncés dépendent des constantes $C_1(x)$, $C_2(x)$, $C_3(x)$. □

Remarque 3.3.3 *L'estimateur de Nadaraya-Watson est un cas particulier des estimateurs par polynômes locaux.*

3.4 Estimateurs par polynômes locaux

Cette méthode régression pondérée locale (RPL) est une généralisation de la méthode Local Weighted Regression ou LOWESS de Cleveland ([5], 1979) proposée par Lejeune ([22], 1983) dans le cadre de l'estimation par noyau.

Définition 3.4.1 *Si K est un noyau positif, $h>0$ une fenêtre et p un entier, on définit*

$$\forall x \in \mathbb{R}, \widehat{\theta}(x) = Arg \min_{\theta=(\theta_0,\ldots,\theta_p)\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\left[Y_i - \sum_{k=0}^{p} \frac{\theta_k}{k!}\left(\frac{X_i - x}{h}\right)^k\right]^2 \tag{3.31}$$

On pose $\widehat{\theta} = (\widehat{\theta}_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_p)$. L'estimateur par polynôme local d'ordre p est alors défini par

$$\widehat{r}_n^p(x) = \widehat{\theta}_0 \tag{3.32}$$

Remarque 3.4.1 *Si $p=0$ alors $\widehat{r}_n^p(x)$ est égal à l'estimateur de Nadaraya-Watson.*

Proposition 3.4.1 *Si $\widehat{r}_n$ est l'estimateur de Nadaraya - Watson associé à un noyau $K \geq 0$ alors $\widehat{r}_n$ est solution de*

$$\widehat{r}_n(x) = Arg \min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\left(Y_i - \theta\right)^2,$$

$\widehat{r}_n(x)$ est donc un estimateur des moindres carrés pondérés si $\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0$.

Démonstration

$$\widehat{r}_n(x) = Arg \min_{\theta\in\mathbb{R}} \tau(\theta)$$

$$\tau(\theta) = \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\left(Y_i - \theta\right)^2$$

τ est un polynôme de second degré en θ. Le point critique :

$$\tau'(\theta) = 0 \Leftrightarrow \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i = \theta \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

$$\Leftrightarrow \theta = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)},$$

c'est un minimum car $\tau''(\theta) = 2\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \geq 0$. □

3.4.1 Construction des estimateurs localement polynomiaux

Pour p un entier naturel fixé, par la méthode des moindres carrés on considère l'ajustement du polynôme

$$\beta_0 + \beta_1(. - x) + \beta_2(. - x)^2 + \ldots + \beta_p(. - x)^p$$

aux données (X_i, Y_i), par la méthode des moindres carrés pondérés.

1. On suppose l'existence de la $(p+1)$-ième dérivée de la fonction de régression $r(.)$ au point x. (Hypothèse essentielle pour valider théoriquement la construction de l'estimateur localement polynomial). On peut alors approximer localement la fonction de régression $r(x)$ par un polynôme d'ordre p. On a donc autour du point x,

$$r(u) \approx r(x) + r'(x)(u-x) + \frac{r"(x)}{2}(u-x)^2 + \ldots + \frac{r^{(p)}(x)}{p!}(u-x)^p \tag{3.33}$$

$$= \sum_{j=0}^{p} \frac{r^{(j)}(x)}{j!}(u-x)^j = \sum_{j=0}^{p} \beta_j (u-x)^j. \tag{3.34}$$

   Quand u est dans un voisinage du point x.

2. Nous ajustons localement le polynôme (3.33) aux données : $\{(X_i, Y_i) : 1 \leq i \leq n\}$ par la méthode des moindres carrés pondérés avec comme fonction de poids $K\{(.-x)/h_n\}$. La méthode consiste à minimiser par rapport au vecteur $\beta = (\beta_0, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$ la quantité suivante

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right\}^2 K\Big(\frac{X_i - x}{h_n}\Big). \tag{3.35}$$

   Les paramètres K et h_n déterminent la forme et la taille du voisinage autour du point x. On note $\widehat{\beta} = \big(\widehat{\beta}_0, \ldots, \widehat{\beta}_p\big)^T$, le vecteur qui minimise l'expression (3.35). En utilisant (3.33), la dérivée k-ième $r^{(k)}(x)$ peut donc être estimée par $\widehat{\beta}_k \times k!$, pour $k = 1, \ldots, p$.

On a donc la définition suivante

Définition 3.4.2 *La statistique*

$$\widehat{r}_n^{(k)}(x; p) = \widehat{\beta}_k \times k!, \quad 0 \leq k \leq p \tag{3.36}$$

est l'estimateur localement polynomial d'ordre p de la dérivée k-ième de la régression $\widehat{r}_n^{(k)}(x)$, et noté estimateur $[LP](p)$ de $\widehat{r}_n^{(k)}(x)$.

Si $k = p = 0$, on retrouve l'estimateur de Nadaraya-Watson noté $[NW]$. L'estimateur $\widehat{r}_n(x; 1)$ de la fonction de régression est appelé l'estimateur localement linéaire et noté $\widehat{r}_n^{LL}(x)$. En utilisant les Équations (3.35) et (3.36), il est égal à $\widehat{\beta}_0$ lorsque $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ désigne le vecteur solution de l'équation des moindres carrés suivante :

$$Arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} \{Y_i - \beta_0 - \beta_1 (X_i - x)\}^2 K\Big(\frac{X_i - x}{h_n}\Big)$$

Si on note par

$$\widehat{f}_{n,j}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \left\{ \frac{X_i - x}{h_n} \right\}^j K\Big(\frac{X_i - x}{h_n}\Big), \quad j = 0, 1, 2,$$

et

$$\widehat{g}_{n,j}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i \left\{ \frac{X_i - x}{h_n} \right\}^j K\Big(\frac{X_i - x}{h_n}\Big), \quad j = 0, 1,$$

alors l'estimateur $[LL]$ est défini par :

$$\widehat{r}_n^{LL}(x) = \frac{\widehat{g}_{n,0}(x)\widehat{f}_{n;2}(x) - \widehat{g}_{n,1}(x)\widehat{f}_{n,1}(x)}{\widehat{f}_{n,0}(x)\widehat{f}_{n,2}(x) - \widehat{f}_{n,1}(x)\widehat{f}_{n,1}(x)} \tag{3.37}$$

Remarque 3.4.2 *Dans la suite (cf. [9] (1992)), on constatera que l'estimateur* $[LP]$ *a un meilleur biais que l'estimateur de Nadaraya-Watson (cf. [25]). De plus, l'estimateur* $[LL]$ *a de bonnes propriétés minimax, il est le meilleur estimateur sur la classe des fonctions de régression à dérivée seconde bornée, parmi tous les estimateurs linéaires (cf. Fan (1993)). Pour de plus amples développements on se référe à Wand et Jones (1995) et à Fan et Gijbels (1996). C'est les résultats qui sont importants et non les démonstrations en elles mêmes.*

3.4.2 Biais et variance des estimateurs localement polynomiaux

Les estimateurs localement polynomiaux sont issus d'un problème de moindres carrés. Pour une notation matricielle on notera la matrice X_x par :

$$X_x = X = \begin{pmatrix} 1 & (X_1 - x) & \dots & (X_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x) & \dots & (X_n - x)^p \end{pmatrix}_{n \times (p+1)} \tag{3.38}$$

Nous posons

$$y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1} \qquad et \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1) \times 1}$$

La matrice diagonale $n \times n$ dénotée W_x :

$$W_x = W = diag\left\{K\Big(\frac{X_i - x}{h_n}\Big)\right\}.$$

Le problème de (3.35) se résume :

$$\min_{\beta \in \mathbb{R}^{p+1}} (y - X\beta)^T W (y - X\beta),$$

où le signe ^{T} désigne la transposition, pour une matrice ou un vecteur. On suppose dans la suite que la matrice carrée $X^T W X$ est inversible et appartenant à $\mathscr{M}_{p+1}(\mathbb{R})$ (espace des matrices carrées $p+1$ à $p+1$).

Remarque 3.4.3 *Si la matrice* $X^T W X \in \mathscr{M}_{p+1}(\mathbb{R})$ *est définie positive, l'estimateur* $[LP](p)$ *appartient à la classe des estimateurs linéaires (cf. Eq(3.39))*

D'après la théorie des moindres carrés, le vecteur solution de léquation des moindres carrés est donné par

$$\widehat{\beta} = \{X^T W X\}^{-1} X^T W y. \tag{3.39}$$

Le vecteur $\beta = \theta$*, formulé dans la Définition (3.4.1)*

$$\beta = \left\{r(x), \dots, \frac{r^{(p)}(x)}{p!}\right\}^T,$$

en utilisant (3.33).

On définit r *le vecteur* $r = \Big\{r(X_1), \dots, r(X_n)\Big\}^T$ *et* $rg = r - X\beta$*, le vecteur des résidus.*

D'après (3.39), si on note $\mathscr{X}$ *l'ensemble des variables* X_i*,* $1 \le i \le n$*, alors nous déduisons*

$$E\Big[\widehat{\beta}|\mathscr{X}\Big] = \{X^T W X\}^{-1} X^T W r \tag{3.40}$$

$$E\Big[\widehat{\beta}|\mathscr{X}\Big] = \beta + \{X^T W X\}^{-1} X^T W rg. \tag{3.41}$$

Soit

$$\sum = diag\left\{K^2\Big(\frac{X_i - x}{h_n}\Big)\right\}\sigma^2(X_i) \in \mathscr{M}_n(\mathbb{R})$$

où $\sigma^2(x) = Var\big[Y|X = x\big]$. *La matrice variance-covariance conditionnelle est*

$$Var\,[\widehat{\beta}|\mathscr{X}] = \{X^T W X\}^{-1}\{X^T \sum X\}\{X^T W X\}^{-1} \tag{3.42}$$

Les expressions (3.40) ne sont pas directement utilisables, elles dépendent de quantités inconnues, le vecteur des résidus rg *et de la matrice* $\sum$. *Ruppert et Wand (1994) (cf [30]) ont obtenu des développements asymptotiques pour le biais et la variance de l'estimateur localement polynomial* $\widehat{r}_n^{(k)}(x;p)$ *défini en (3.36). Pour le théorème cité plus loin il est nécessaire d'introduire les notations suivantes :*
On note les moments de K *et de* K^2 *par :*

$$\big[\mu_j(K)\big] = \int_{\mathbb{R}} u^j K(u)du \quad et \quad \big[\mu_j(K^2)\big] = \int_{\mathbb{R}} u^j K^2(u)du$$

avec $j \in \mathbb{N}$. *Soient*

$$S = \Big(\big[\mu_{j+l}(K)\big]\Big)_{0\le j,l\le p} \in \mathscr{M}_{p+1}(\mathbb{R}) \tag{3.43}$$

$$\widetilde{s} = \Big(\big[\mu_{j+l+1}(K)\big]\Big)_{0\le j,l\le p} \in \mathscr{M}_{p+1}(\mathbb{R}) \tag{3.44}$$

$$\bar{S} = \Big(\big[\mu_{j+l}(K^2)\big]\Big)_{0\le j,l\le p} \in \mathscr{M}_{p+1}(\mathbb{R}) \tag{3.45}$$

$$c_p = \Big(\big[\mu_{p+l}(K)\big],\ldots,\big[\mu_{2p+2}(K)\big]\Big)^T \in \mathbb{R}^{p+1} \tag{3.46}$$

$$\widetilde{c}_p = \Big(\big[\mu_{p+2}(K)\big],\ldots,\big[\mu_{2p+2}(K)\big]\Big)^T \in \mathbb{R}^{p+1} \tag{3.47}$$

On désigne par $e_{k+1} = (0,\ldots,0,1,0,\ldots,0)^T$ *le* $(k+1)$*-ième vecteur unité dans* $\mathbb{R}^{p+1}$.

Théorème 3.4.1 (Ruppert et Wand (1994)) *Nous supposons* $f_X(x) > 0$ *et les fonctions* $f_X(.)$, $r^{p+1}(.)$ *et* $\sigma^2(.)$ *continues dans un voisinage du point* x. *La fenêtre* h *vérifie* $h \longrightarrow 0$ *et* $nh \longrightarrow \infty$. *Alors,*

$$Var\Big[\widehat{r}_n^{(k)}(x;p)|\mathscr{X}\Big] = (k!)^2 \times e_{k+1}^T S^{-1}\overline{S}S^{-1}e_{k+1}\frac{\sigma^2(x)}{nh^{1+2k}f_X(x)} + o\Big(\frac{1}{nh^{1+2k}}\Big) \tag{3.48}$$

Lorsque $p-k$ *est impair,*

$$Biais\Big[\widehat{r}_n^{(k)}(x;p)|\mathscr{X}\Big] = k! \times e_{k+1}^T S^{-1}\frac{c_p}{(p+1)!}r^{(p+1)}(x)h^{p+1-k} + o(h^{p+1-k}) \tag{3.49}$$

Lorsque $p-k$ *est pair, en supposant* $f_X'(.)$ *et* $r^{(p+2)}(.)$ *sont continues dans un voisinage du point* x *ainsi que* $nh^3 \longrightarrow \infty$, *le biais conditionnel asymptotique est donné par,*

$$k! \times e_{k+1}^T \widetilde{S}^{-1}\frac{\widetilde{c}_p}{(p+2)!}\left\{r^{(p+2)}(x) + (p+2)r^{(p+1)}(x)\frac{f_X'(x)}{f_X(x)}\right\}h^{p+2-k} + o\Big(h^{p+2-k}\Big)$$

Démonstration Si $p = k = 0$ alors on retrouve le biais asymptotique de l'estimateur de Nadaraya-Watson. Pour les détails de la démonstration se référer à Ruppert et Wand (1996) ([30]). La meilleure représentation des estimateurs $[LP]$ est obtenue par la méthode des "noyaux équivalents ", c'est à dire en réécrivant asymptotiquement les estimateurs $[LP]$ sous une forme plus classique proche de l'estimateur de Nadaraya-Watson ([25]). □

Dans l'investigation d'autres développements, on peut se référer à la thèse de David Blondin ([1]). La variance et le biais conditionnels de l'estimateur $\widehat{r}_n^{(k)}(x;p)$ peuvent être exprimés en fonction du noyau équivalent ([20], [24]).

Autre approche

L'estimateur par polynômes locaux est une généralisation de l'estimateur de Nadaraya-Watson associée à sa caractérisation. On fixe x, on doit calculer un estimateur $r(x)$ mais on veut aussi $r(y)$.
On reprend le problème de minimisation (cf. (3.31)), mais au lieu d'utiliser une constante θ, on utilise un polynôme. Donc si r est régulière autour de x alors

$$r(u) \approx P_{l,x}(u)$$

l est un entier naturel, avec

$$P_{l,x}(u) = \sum_{k=0}^{l} \frac{r^k(x)}{k!}(u-x)^k$$

$r^k(x), k = 1\ldots, l$ sont les coefficients du développement de Taylor. $P_{l,x}$ est évidemment inconnu. On va en fait estimer le polynôme $P_{l,x}$

$$P_{l,x} = \mu_0 + \mu_1(u-x) + \ldots + \mu_l(u-x)^l.$$

On cherche donc à estimer les estimateurs $\mu_0,\ldots,\mu_l$ de ce polynôme par les estimateurs $\widehat{\mu}_0, \widehat{\mu}_1, \ldots, \widehat{\mu}_l$. L'estimateur $\widehat{\mu}_0$ est donc l'estimateur $\widehat{r}(x)$ recherché. En particulier on a

$$r(X_i) \approx P_{l,x}(X_i) \quad \text{si } X_i \text{ est proche de } x$$

Définition 3.4.3 *Un estimateur $\widehat{r}$ de la fonction de régression r est linéaire s'il s'écrit*

$$\widehat{r}(x) = \sum_{i=1}^{n} \omega_i(x) Y_i, \quad \forall\, x \in \mathbb{R} \tag{3.50}$$

Les $\omega_i(x)$ ne dépendent pas de Y_i. Si on pose $Y = (Y_1,\ldots,Y_n)^T$ et $\omega(x) = (\omega_1(x),\ldots,\omega_n(x))^T$, on a

$$\widehat{r}(x) = \omega(x)^T Y$$

Dans le cas d'un estimateur des moindres carrés ordinaires $\widehat{r}(x) = \widehat{\beta}^T x$ est linéaire en x et c'est également un estimateur linéaire :

$$\widehat{r}(x) = x^T\widehat{\beta} = x^T(X^TX)^{-1}X^TY = \omega(x)^TY$$

où $\omega(x) = \left[x^T(X^TX)^{-1}X^T\right]^T$ est un vecteur qui ne dépend pas de Y. Avant d'introduire la définition suivante, soient les notations :

$$\forall\, i = 1,\ldots,n;\ \forall\, u \in \mathbb{R}, \qquad Z_i = \frac{X_i - x}{h}, \quad V_l = \begin{pmatrix} 1 \\ u \\ \vdots \\ \frac{u^l}{l!} \end{pmatrix}$$

On pose également

$$B_{n,x} = \sum_{i=1}^{n} K(Z_i) V_l(Z_i) V_l(Z_i)^T$$

Proposition 3.4.2 *Si la matrice $B_{n,x}$ est définie positive alors l'estimateur par polynômes locaux $\widehat{r}_{n,l}(x)$ est un estimateur linéaire.*

Démonstration On a

$$\widehat{r}_n^l(x) = \widehat{\theta}(x) = e_1^T \widehat{\theta}(x) \tag{3.51}$$

avec

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\widehat{\theta}(x) = Arg \min_{\theta \in \mathbb{R}^{l+1}} \tau(\theta),$$

où

$$\tau(\theta) = \sum_{i=1}^{n} K(Z_i) \left(Y_i - \theta^T V_l(Z_i)\right)^2$$

On a donc

$$\begin{aligned} \tau(\theta) &= \sum_{i=1}^{n} K(Z_i) \left[Y_i^2 + (\theta^T V_l(Z_i))^2 - 2Y_i \theta^T V_l(Z_i)\right] \\ &= \sum_{i=1}^{n} K(Z_i) Y_i^2 + \sum_{i=1}^{n} K(Z_i) \theta^T V_l(Z_i) V_l(Z_i)^T \theta \\ &\qquad -2\theta^T \sum_{i=1}^{n} K(Z_i) Y_i V_l(Z_i) \\ &= a + \theta^T B_{n,x} \theta - 2\theta^T b \end{aligned}$$

avec $a = \sum_{i=1}^{n} K(Z_i) Y_i^2$ et $b = \sum_{i=1}^{n} K(Z_i) Y_i V_l(Z_i)$
($\widehat{r}_n^l$ est défini en (3.51)) □

Dans la suite on adoptera les notations suivantes

Notations 3.4.1
1. Si $f(x) = x^T a$ alors $\nabla f(x) = a$ et $Hf(x) = 0$ (Hf étant la hessienne de f)
2. Si $f(x) = x^T A x$ alors $\nabla f(x) = (A + A^T)x$ et $Hf(x) = A + A^T$.
3. Si A est symétrique et $f(x) = x^T A x$ alors $\nabla f(x) = 2Ax$ et $Hf(x) = 2A$.

Le point critique vérifie donc

$$\nabla \tau(\theta) = -2b + 2B_{n,x}\theta.$$

$$\nabla \tau(\theta) = 0 \Leftrightarrow B_{n,x}\theta = b.$$

Si $B_{n,x}$ est définie positive alors elle est inversible et donc il y a un seul point critique donné par

$$\widehat{\theta} = B_{n,x}^{-1} b.$$

Comme la fonction est $\tau(\theta)$ convexe, ce point critique correspond à un minimum global car

$$H\tau(\theta) = 2B_{n,x} > 0.$$

On a donc

$$\begin{aligned}\widehat{r}_{n,l}(x) &= e_1^T B_{n,x}^{-1} b \\ &= e_1^T B_{n,x}^{-1} \left[\sum_{i=1}^{n} K(Z_i) Y_i V_l(Z_i) \right] \\ &= \sum_{i=1}^{n} \omega_i(x) Y_i\end{aligned}$$

avec $\omega_i(x) = K(Z_i) e_1^T B_{n,x}^{-1} V_l(Z_i)$. $\omega_i(x)$ ne dépend que de x, K, l, h et des X_i et pas des Y_i. Donc $\widehat{r}_{n,l}$ est bien un estimateur linéaire.

Remarque 3.4.4 *Le choix de la fenêtre est crucial en comparaison avec le choix du noyau qui n'est pas important en pratique, même le degré est choisi souvent égal à* 1 *ou* 2.

Choix des paramètres de régularisation

3.4.3 Risque empirique, surajustement

On suppose dans la suite que les X_i sont aléatoires, les X_i, Y_i indépendants. On suppose en plus que $E(\epsilon_i^2) = \sigma^2$. La fonction de régression r est telle que

$$r = Arg \min_{f \in L_2(P^X)} \left[E\left(Y - f(X)\right)^2 \right]. \tag{3.52}$$

On note r_h l'estimateur utilisant la fenêtre h. Si on enlève une partie de l'échantillon $(X_i, Y_i)_{i \in I}$ avec I une partie de $\{1, \ldots, n\}$ on notera $\widehat{r}_h^{-I}$ l'estimateur calculé à partir de l'échantillon auquel on a enlevé à $(X_i, Y_i)_{i \in I}$. On veut trouver la fenêtre h qui minimise le risque

$$R(h) = E\left[(\widehat{r}_h - r)^2(X)\right] = E\left[\|\widehat{r}_h - r\|_{L_2(P^X)}\right].$$

Comme l'expression est compliquée on pourra minimiser la quantité

$$\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{r}_h(X_i) - Y_i\right)^2.$$

Cette quantité est appelée erreur d'apprentissage ou "training error".

Remarque 3.4.5 *Les mêmes données sont utilisées à la fois pour estimer* r *et estimer le risque. Il y a un manque d'indépendance, donc utiliser ce risque comme substitut du vrai risque est une mauvaise idée.*

Dans le cas de l'EMC non paramétrique, imaginons qu'on cherche à ajuster un polynôme. La question du degré M se pose de cette façon. Pour chaque M on calcule $\widehat{\beta}^M$, l'EMC associé au design $X = (X_{ij})_{1 \le j \le M, 1 \le i \le n}$ avec $X_{ij} = x_i^{j-1}$. Si M est assez grand et si les points du design sont distincts alors le risque empirique est égal à zéro. On a obtenu un polynôme qui passe par tous les points (X_i, Y_i). La variance risque fort d'être trop grande. L'erreur d'apprentissage est trop optimiste. On aura en général, $E\left[\widehat{R}_n(h)\right] < R(h)$. On a un sur-ajustement : l'estimateur associé sera trop adapté aux données particulières que nous avons et qui ne se généralisera pas bien à de nouvelles données.

Remarque 3.4.6 – *Si les Y_i, $i = 1,\dots,n$, indépendantes et identiquement distribuées, ont la même loi que Y alors pour estimer $E(Y)$ on utilise souvent la moyenne empirique $\frac{1}{n}\sum_{i=1}^{n} Y_i$.*

– *Si g est une fonction fixe (ne dépendant pas des données) alors $Y_i - g(X_i)$ a la même loi que $Y - g(X)$, car si g est fixe on utilise*

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n}(g(X_i)-Y_i)^2\right] &= \|g-r\|_{L_2(P^X)} + \sigma^2 \\ Var\left[\frac{1}{n}\sum_{i=1}^{n}(g(X_i)-Y_i)^2\right] &= \frac{1}{n}Var\left(g(X)-Y\right)^2. \end{aligned}$$

Quand on se donne un ensemble de fonctions déterministes $(g_h)_{h\in H}$ dépendant d'un paramètre h (g_h ne dépend pas de l'échantillon), alors minimiser le risque empirique semble un bon substitut à la minimisation du risque quadratique $\|g_h - r\|^2_{L_2(P^X)}$, pour choisir le paramètre h.

3.4.4 Validation croisée

Cette méthode est assez générale et s'applique à de nombreuses procédures d'estimation. On se donne une grille de valeurs H de fenêtres, parmi lesquelles on veut choisir une fenêtre optimale $\widehat{h}$ en se basant sur les données uniquement. Le principe est de diviser l'échantillon en un ensemble d'apprentissage (training set) et un ensemble de validation (validation set). On fabrique des estimateurs à partir de l'ensemble d'apprentissage et ensuite l'ensemble de validation est utilisé pour estimer leur risque de prédiction. Les shémas les plus connus sont :

1. **Hold-out CV**
   On divise l'échantillon en deux parties I_1 et I_2 ($I_1 \cap I_2 = \emptyset$). On calcule les estimateurs $(\widehat{r}_h^{I_1})_{h\in H}$ à partir de $(X_i, Y_i)_{i\in I_1}$, puis on calcule les estimateurs des risques associés,

   $$\widehat{R}(h) = \frac{1}{n_2}\sum_{i\in I_2}(Y_i - \widehat{r}_h^{I_1}(X_i))^2$$

   $n_2 = Card(I_2)$, où CV est une abréviation de validation croisée (¨cross validation¨, en anglais).
2. **V-fold CV**
   Les données sont divisées en V ensembles disjoints $I_1,\dots,I_V$. Chacun des V sous ensembles est utilisé à tour de rôle comme ensemble de validation, le reste est donc utilisé pour l'apprentissage : on calcule, pour chaque $j \in \{1,\dots,V\}$, l'ensemble des estimateurs $(\widehat{r}_h^{-I_j})_{h\in H}$ fabriqués avec $(X_i, Y_i)_{i\notin I_j}$. Ensuite le risque de prédiction pour une fenêtre h est estimé par

   $$\widehat{R}(h) = \frac{1}{V}\sum_{j=1}^{V}\frac{1}{n_j}\sum_{i\in I_j}\left(Y_i - \widehat{r}^{-I_j}(X_i)\right)^2 \tag{3.53}$$

   $n_j = Card(J_j)$. Dans la pratique on choisit $V = 5$, ou $V = 10$.
3. **Leave-one-out** : V-fold CV, avec $V = n$.
4. **Leave-q-out**
   Tout sous ensemble de cardinal q de l'échantillon est utilisé comme ensemble de validation

et le reste comme ensemble d'apprentissage. On choisit

$$\widehat{h} = Arg\min_{h\in H} \widehat{R}(h)$$

L'estimateur final est

$$\widehat{r} = \widehat{r}_{n,\hat{h}}$$

où $\widehat{r}_{n,h}$ est l'estimateur par polynômes locaux avec la fenêtre h et en utilisant tout l'échantillon. Le V-fold est la méthode la plus connue. Les méthodes présentées plus haut sont les plus intensives en calculs.

3.4.5 Cas du leave-one-out

Pour chaque h de la grille de valeurs H et pour chaque $i \in \{1,\dots,n\}$, on construit un estimateur $\widehat{r}_h^{(-i)}$ en utilisant toutes les observations sauf la i-ème. La i-ème observation est utilisée pour mesurer la performance de $\widehat{r}_h^{(-i)}$ par $(Y_i - \widehat{r}_h^{(-i)}(X_i))^2$. On pose donc

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{r}_h^{(-i)}(X_i)\right)^2$$

On minimise R pour trouver $\widehat{h}$.

Dépendance et risque empirique

On considère $X_1^n = (X_1,\dots,X_n)$, $Y_1^n = (Y_1,\dots,Y_n)$ et on cherche h tel que $E\left[(Y-\widehat{r}_h(X))^2\right]$ soit minimal. Plusieurs cas de figures se présentent :

1. Si $g = \widehat{r}$, g n'est plus fixe, mais ne dépend pas des données (X^n, Y^n) et on a

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{r}(X_i))^2\right] \neq E\left[(Y-\widehat{r}(X))^2\right]$$

Car si l'estimateur est symétrique en ses variables (exemple des estimateurs par polyômes locaux) alors

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{r}(X_i))^2\right] = E\left[(Y_1 - \widehat{r}(X_1))^2\right]$$

La dépendance de $\widehat{r}$ à (X^n, Y^n) en écrivant $\widehat{r}(x) = g(X_1,\dots,X_n,Y_1,\dots,Y_n,x)$. Si f est la densité du couple (X,Y) on a alors

$$\begin{aligned} E\left[(Y_1 - \widehat{r}(X_1))^2\right] &= E\left[(Y_1 - g(X_1,\dots,X_n,Y_1,\dots,Y_n,X_1)\right)^2\Big] \\ &= \int \Big(y_1 - g(x_1,\dots,x_n,y_1,\dots,y_n,x_1))^2 f(x_1,y_1)\dots f(x_n,y_n)dx_1dy_1\dots dx_ndy_n \end{aligned}$$

Tandis que

$$\begin{aligned} E\left[(Y-\widehat{r}(X))^2\right] &= E\Big[(Y - g(X_1,\dots,X_n,Y_1,\dots,Y_n,X))^2\Big] \\ &= \int \Big(y - g(x_1,\dots,x_n,y_1,\dots,y_n,x_1)\Big)^2 f(x_1,y_1)\dots f(x_n,y_n)\times \\ &\qquad f(x,y)dx_1dy_1\dots dx_ndy_ndxdy \end{aligned}$$

On déduit que le risque empirique est un mauvais estimateur du "vrai" risque $E\left[(Y-\widehat{r}(X))^2\right]$.

2. Si (X_{n+1}, Y_{n+1}) est une nouvelle donnée indépendante de (X_1^n, Y_1^n) et de même loi que (X, Y) alors

$$\begin{aligned} E\left[\left(Y_{n+1} - \widehat{r}(X_{n+1})\right)^2\right] &= E\left[\left(Y_{n+1} - g(X_1, \ldots, X_n, Y_1, \ldots, Y_n, X_{n+1})\right)^2\right] \\ &= \int (y_{n+1} - g(x_1, \ldots, x_n, y_1, \ldots, y_n, x_{n+1}))^2 f(x_1, y_1) \ldots f(x_n, y_n) \times \\ &\qquad f(x_{n+1}, y_{n+1}) dx_1 dy_1 \ldots dx_n dy_n dx_{n+1} dy_{n+1} \\ &= \int \left(y - g(x_1, \ldots, x_n, y_1, \ldots, y_n, x)\right)^2 f(x_1, y_1) \ldots f(x_n, y_n) \times \\ &\qquad f(x, y) dx_1 dy_1 \ldots dx_n dy_n dx dy \\ &= E\left[\left(Y - \widehat{r}(X)\right)^2\right]. \end{aligned}$$

On a utilisé le fait que

$$Y_{n+1} - \widehat{r}(X_{n+1}) = Y_{n+1} - g\Big(X_1, \ldots, X_n, Y_1, \ldots, Y_n, X_{n+1}\Big).$$

Ce dernier terme est équivalent à $Y - g\Big(X_1, \ldots, X_n, Y_1, \ldots, Y_n, X\Big)$ quantité qui vaut $Y - \widehat{r}(X)$.

L'idée est de séparer l'échantillon en deux si on a n assez grand. On note $n+p$ données, on sépare en prenant $(X_1, Y_1), \ldots, (X_n, Y_n)$ pour estimer $\widehat{r}$ puis $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+p}, Y_{n+p})$ pour valider l'estimateur (ou estimer le risque de cet estimateur ou faire un choix de paramètre d'ajustement comme le choix de la fenêtre h pour un estimateur par polynômes locaux). On a alors un bon estimateur du risque $E\left[(Y - \widehat{r}(X))^2\right]$ en le posant égal à :

$$\frac{1}{p} \sum_{k=1}^{p} \left[Y_{n+k} - \widehat{r}(X_{n+k}) \right]^2$$

Car on a, en conditionnant sur $(X_1, \ldots, X_n)$,

$$Y_{n+1} - \widehat{r}(X_{n+1}), \ldots, Y_{n+p} - \widehat{r}(X_{n+p}) \sim Y - \widehat{r}(X)$$

C'est l'idée du Hold-out. Une autre idée est le leave-one out : On fabrique un estimateur $\widehat{r}_{n-1}^{(-i)}$ en utilisant l'échantillon (X^n, Y^n) privé de (X_i, Y_i). On utilise ensuite (X_i, Y_i) pour valider cet estimateur

$$E\left[\left(Y_i - \widehat{r}_{n-1}^{(-i)}(X_i)\right)^2\right] = E\left[\left(Y - \widehat{r}_{n-1}(X)\right)^2\right]$$

On a noté $\widehat{r}_{n-1}$ l'estimateur fabriqué avec seulement $n-1$ données. Donc on déduit que la moyenne empirique $\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{r}_{n-1}^{(-i)}(X_i)\right)^2$ semble un bon estimateur (en particulier sans biais) de $E\left[\left(Y - \widehat{r}_{n-1}(X)\right)^2\right]$ qui est le "vrai" risque de l'estimateur $\widehat{r}_{n-1}$ fabriqué à partir de $n-1$ données (on s'attend à ce que $E\left[\left(Y - \widehat{r}_{n-1}(X)\right)^2\right]$ soit proche de $E\left[\left(Y - \widehat{r}_n(X)\right)^2\right]$ où $\widehat{r}_n$ est l'estimateur de départ n fabriqué avec n données). On admet la proposition suivante qui relie les poids associés à l'estimateur $\widehat{r}_h^{(-i)} Y_j$ alors, pour tout $j \neq i$

$$\widetilde{\omega}_{j,h}(X_i) = \frac{\omega_{j,h}(X_i)}{1 - \omega_{i,h}(X_i)}. \tag{3.54}$$

Pour le calcul de $\left(\widehat{r}_h^{(-i)}\right)_{1\leq i\leq n}$, dans le cas des polynômes locaux on utilise la proposition suivante :

Proposition 3.4.3 *Si* $\widehat{r}_h(x)=\sum_{i=1}^{n}\omega_{i,h}(x)Y_i$ *alors*

$$CV(h)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i-\widehat{r}_h(X_i)}{1-\omega_{i,h}(X_i)}\right)^2. \tag{3.55}$$

Démonstration En s'inspirant de [13], [16] et de [17] On a

$$CV(h)=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\widehat{r}_h^{(-i)}(X_i)\right)^2$$

avec

$$\begin{aligned}
Y_i-\widehat{r}_h^{(-i)}(X_i) &= Y_i-\sum_{j\neq i}\widetilde{\omega}_{j,h}(X_i)Y_j\\
&= Y_i-\sum_{j\neq i}\frac{\omega_{j,h}(X_i)}{1-\omega_{i,h}(X_i)}Y_j\\
&= \frac{(1-\omega_{i,h}(X_i))Y_i-\sum_{j\neq i}\omega_{j,h}(X_i)Y_j}{1-\omega_{i,h}(X_i)}\\
&= \frac{Y_i-\widehat{r}_h(X_i)}{1-\omega_{i,h}(X_i)}.
\end{aligned}$$

Il existe une autre alternative appelée validation croisée généralisée notée

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i-\widehat{r}_h(x_i)}{1-\Omega/n}\right)^2 \tag{3.56}$$

$$= \frac{1}{\left(1-\frac{\Omega}{n}\right)^2}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\widehat{r}_h(x_i)\right)^2 \tag{3.57}$$

on a posé $\Omega=\sum_{i=1}^{n}\omega_{i,h}(x_i)$. □

Remarque 3.4.7 *Si* $\Omega << n$ *alors* $(1-\frac{\Omega}{n})^{-2}\approx 1+2\frac{\Omega}{n}$, *donc on a*

$$GCV(h)\approx\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\widehat{r}_h(x_i)\right)^2\left(1+\frac{2\Omega}{n}\right)$$

Chapitre 4

Rééchantillonnage

Ces méthodes sont utilisées quand on a des hypothèses faibles ou aucune hypothèse n'est émise au sujet de la distribution de la population.

4.1 La méthode du jackknife

La méthode du jackknife (canif, couteau de poche) est une méthode déterministe assez ancienne, généralisée par la méthode du bootstrap. Elle a été proposée par Maurice Quenille en 1950. Le but du jackknife est de réduire le biais et d'obtenir des intervalles de confiance. Pour estimer l'écart type d'une loi F inconnue on a l'estimateur bien connu

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

donc $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Soit s une réalisation correspondante pour une réalisation $x_1, x_2, \ldots, x_n$ de l'échantillon, l'estimateur du jackknife est obtenu de la façon suivante

1. On calcule la valeur notée s_{-1}, de l'écart type du sous échantillon obtenu en ne tenant pas compte de la valeur x_1

$$s_{-1} = \sqrt{\frac{1}{n-2}\sum_{\substack{i=1 \\ i\neq 1}}^{n}(x_i - \overline{x})^2} \tag{4.1}$$

   Puis la valeur $s_{*1} = ns - (n-1)s_{-1}$.

2. On répète cette opération en omettant à tour de rôle chacune des observations pour obtenir n pseudo-valeurs $s_{*1}, s_{*2}, s_{*3}, \ldots, s_{*n}$ avec donc

$$s_{-i} = \sqrt{\frac{1}{n-2}\sum_{\substack{j=1 \\ j\neq i}}^{n}(x_j - \overline{x})^2} \tag{4.2}$$

$$s_{*i} = ns - (n-1)s_{-i} \tag{4.3}$$

3. L'estimation du jackknife est alors la moyenne des pseudo-valeurs noté $\overline{s}_*$.

Un intervalle de confiance approché peut être obtenu, en utilisant la variance des pseudo-valeurs

$$s_{JK}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(s_{*i} - \overline{s}_*)^2$$

$$IC_{0.95}(\sigma) \simeq \left[\overline{s}_* - t_{0.975}^{(n-1)}\frac{s_{JK}}{\sqrt{n}}, \overline{s}_* + t_{0.975}^{(n-1)}\frac{s_{JK}}{\sqrt{n}}\right].$$

De manière générale si ω est une caractéristique de la loi et T_n un estimateur convergent de ω, l'estimateur du maximum de vraisemblance est obtenu par sa version empirique. En raison du fait que la fonction de répartition F_n est l'estimateur fonctionnel du maximum de vraisemblance pour F. Soit maintenant T^{-i} l'estimateur calculé en omettant X_i. On définit les pseudo-valeurs

$$T_n^{*i} = nT_n - (n-1)T_n^{-i},\ i = 1,\ldots,n.$$

Définition 4.1.1 *L'estimateur du jackknife fondé sur T_n est défini par $T_n^* = \frac{1}{n}\sum_{i=1}^{n} T_n^{*i}$.*

Proposition 4.1.1 *Si le biais T_n est de la forme $\frac{c}{n}$, où c est une constante, alors T_n^* l'estimateur du jackknife fondé sur T_n est sans biais.*

Démonstration Comme $E(T_n) = \omega + \frac{c}{n}$, pour tout i on a

$$E(T_n^{-i}) = \omega + \frac{c}{n-1}$$

car T_n^{-i} est le même estimateur appliqué au $(n-1)$ -échantillon aléatoire $X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n$. On a donc

$$\begin{aligned} E(T_n^{*i}) &= nT(T_n) - (n-1)E(T_n^{-i}) \\ &= n\omega + c - (n-1)\left[\omega + \frac{c}{n-1}\right] \\ &= \omega \end{aligned}$$

d'où

$$E(T_n^*) = \frac{1}{n}\sum_{i=1}^{n} E(T_n^{*i}) = \omega.$$

□

Proposition 4.1.2 *Soit T_n^* l'estimateur du jackknife de la caractéristique ω, reposant sur un estimateur T_n. Soit $S_{n,JK}^2$ la variance des pseudo-valeurs, alors sous certaines conditions concernant la forme de la statistique on a*

$$\frac{T_n^* - \omega}{S_{n,JK}/\sqrt{n}} \xrightarrow[n\longrightarrow\infty]{\mathscr{L}} N(0,1). \tag{4.4}$$

On déduit l'intervalle de confiance

$$IC_{0.975}(\omega) \simeq \left[t_n^* - t_{0.975}^{(n-1)}\frac{s_{n,JK}}{\sqrt{n}}, t_n^* + t_{0.975}^{(n-1)}\frac{s_{n,JK}}{\sqrt{n}}\right] \tag{4.5}$$

où t_n^* et $s_{n,JK}$ sont les réalisations respectives de T_n^* et de $S_{n,JK}$ avec :

$$s_{n,JK}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(s_{*i} - \overline{s}_*)^2$$

Démonstration La démonstration se déduit de ([19], Proposition 8.3, page 168-169), en vertu du théorème central limite pour les pseudo-valeurs. Nous déduisons un intervalle de confiance approché pour ω en raison du fait que

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \underset{approxi}{\rightsquigarrow} t_{(n-1)}$$

et de la convergence de S^2 vers σ^2 dans le cas de ([19] section 7.4.1 page 144).

4.2 La méthode du bootstrap

4.2.1 Introduction

Le bootstrap est une approche non paramétrique plus générale. Elle a été proposée par Bradley Efron (1979). Le nom provient de l'expression : "Se hisser en tirant sur les languettes de ses bottes". Se sortir, seul, d'une situation difficile. Situation que le baron Münchhausen a vécu en tombant dans un lac profond. Soit $\widehat{\omega}$ un estimateur d'une caractéristique ω de la loi mère. Cette loi peut être continue ou discrète, de fonction de répartition inconnue F. Un problème peut se poser. Estimer la variance $V_F(\widehat{\omega})$.

4.2.2 Estimation par bootstrap

L'estimation se fait suivant les étapes suivantes

1. On dispose des valeurs observées de X un échantillon de taille n, $x_1,\ldots,x_n$. On effectue n tirages au hasard avec remise parmi les valeurs x_i, on calcule l'estimation $\widehat{\omega}_1^*$ obtenue sur la base de ce nouvel échantillon.
2. On répète l'opération précédente M fois pour obtenir une série d'estimations $\widehat{\omega}_1^*,\widehat{\omega}_2^*,\ldots,\widehat{\omega}_M^*$.
3. L'estimation de la moyenne de $\widehat{\omega}$ appelée la moyenne bootstrap est donnée par

$$\overline{\widehat{\omega}^*} = \frac{\sum_{i=1}^{M} \widehat{\omega}_i^*}{M} \tag{4.6}$$

   M est appelé le nombre total d'échantillons *bootstrap* (chacun de taille n).
4. L'estimation non paramétrique de la variance de $\widehat{\omega}$ est donnée par

$$s^{2*}(\widehat{\omega}) = \frac{1}{M-1}\sum_{k=1}^{M}(\widehat{\omega}_k^* - \overline{\widehat{\omega}_k}^*)^2 \tag{4.7}$$

   $\overline{\widehat{\omega}^*}$ désigne la moyenne de la série, $\widehat{\omega}_1^*,\widehat{\omega}_2^*,\ldots,\widehat{\omega}_M^*$.

Proposition 4.2.1 *Lorsque M tend vers l'infini, l'estimateur issu de cette procédure tend presque sûrement vers l'estimateur du maximum de vraisemblance de $V_F(\widehat{\omega})$. En pratique on choisit $M = 100$ et plus.*

Démonstration Pour la démonstration c.f [19] (page 192).

Exemple 4.2.1 Avec le logiciel **R**, la fonction boot(.) procède au tirage au sort avec remise des n indices pour former l'échantillon bootstrap et qui calcule les M valeurs $x_1^*, x_2^*, \ldots, x_M^*$. Dans le cas du calcul d'une moyenne par bootstrap, et pour estimer cette moyenne on utilise la fonction :
moyenne <-function(d,w) {
n <- length(d)
return ($sum(d[w[1:n]])/n$}
La fonction boot(.) produit un objet de classe "boot", (on prendra par exemple M=500 ; le vecteur brad est l'échantillon saisi avec le logiciel **R**) :
library(boot)
brad.boot<-boot(brad,moyenne,$R = 500$)
brad.boot$t0 donne la valeur du paramètre la moyenne dans ce cas, sur l'échantillon de départ.
L'estimation bootstrap de la moyenne et l'erreur type correspondante sont obtenues :
print(mean(brad.boot$t))
[1] 659.0232
print(sd(brad.boot$t))
print [1] 5.53912.

brad est le vecteur données représentant la longueur en dixièmes de mm de la grande plume de l'aile (3è rémige) de 40 oiseaux (pinsons) sur les arbres.

$$x<-c(720,680,705,600,625,680,630,650,635,640,675,620,625,700,620,600,690,615,720,$$

$$655,620,640,675,665,620,630,680,660,690,720,650,680,700,685,625,690,650,645,690,670).$$

La moyenne et l'écart type estimés sont donnés par :

$$mean(brad) = 659.25, \quad sd(brad)/sqrt(length(brad)) = 5.399282.$$

4.2.3 Intervalles de confiance et bootstrap

Pour déterminer un intervalle de confiance pour un paramètre ω, plusieurs méthodes existent pour le déterminer à partir des M valeurs obtenues par bootstrap. On prendra dans la suite $a = \alpha/2$.

1. Méthode de l'erreur standard (standard bootstrap confidence interval) : on utilise la moyenne $\widehat{\omega}^*$ et l'erreur type s^* du paramètre obtenues par bootstrap, et la loi normale (ou la loi de Student) pour construire l'intervalle de confiance :
   – Dans le cas normal on a un intervalle de confiance approché

   $$IC_{0.95}(\omega) \simeq \left[\widehat{\omega} - 1.96 s^*(\widehat{\omega}); \widehat{\omega} + 1.96 s^*(\widehat{\omega})\right]$$

   $s^*(\widehat{\omega})$ est défini dans (4.7).
   – Dans le cas de la méthode studentisée l'intervalle de confiance est

   $$\left[\widehat{\omega} + t^*_{0.025}\frac{s}{\sqrt{n}}; \widehat{\omega} + t^*_{0.975}\frac{s}{\sqrt{n}}\right]$$

   $t^*_{0.025}$ et $t^*_{0.975}$ sont les quantiles empiriques de la série t_k^*, $k = 1\ldots, M$, où $t_k^* = \frac{\widehat{\omega}_k^* - \widehat{\omega}}{s_k^*/\sqrt{n}}$, $k = 1, \ldots, M$. $s_k^{2*} = Var(\widehat{\omega}_k^*)$, s^2 est estimateur convergent de $Var(\omega)$.

2. Méthode des pourcentiles simples (¨simple percentile confidence interval˝) : Cette méthode est assez facile à mettre en œuvre. On considère la série $\widehat{\omega}_1^*, \widehat{\omega}_2^*, \ldots, \widehat{\omega}_n^*$, et ses quantiles d'ordre respectifs **0.025** et **0.975**, on obtient l'intervalle de confiance

$$IC_{0.95}(\omega) \simeq \left[\widehat{\omega}_{0.025}^*; \widehat{\omega}_{0.975}^*\right].$$

Avec le logiciel **R**, on construit la fonction de répartition des valeurs $\widehat{\omega}_i^*$ obtenues par bootstrap. Les bornes de l'intervalle sont le $a-i$ème et le $(1-a)$-ième percentile. La fonction **boot.ci()** de **R** permet d'obtenir les intervalles calculés selon les diverses méthodes.
Avec le logiciel **R**
>print(boot.ci(brad.boot))

Annexes

Le tableau suivant donne la table de Wilcoxon pour échantillons appariés.

FIGURE 4.1 – Table des valeurs critiques pour le test de Wilcoxon- Échantillons appariés

N	Niveau de signification, test unilatéral		
	0.025	0.01	0.005
	Niveau de signification, test bilatéral		
	0.05	0.02	0.01
6	0		
7	2	0	
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

La table de la loi Mann et Whitney :
Elle donne pour la loi de Mann Whitney (MW) ; $n_2 = 7$, $n_1 = (1, \ldots, 7)$ et $U = (0, \ldots, 25)$. La probabilité $P(MW \leq U)$.
Pour un test bilatéral, pour calculer la probabilité critique il faut multiplier par 2 les valeurs ci dessus (valables pour un test unilatéral. Mann et Whitney).

FIGURE 4.2 – La table de Mann et Whitney

$n_2 = 7$							
n_1 / U	1	2	3	4	5	6	7
0	0.125	0.028	0.008	0.003	0.001	0.001	0.000
1	0.250	0.056	0.017	0.006	0.003	0.001	0.001
2	0.375	0.111	0.033	0.012	0.005	0.002	0.001
3	0.5	0.167	0.058	0.021	0.009	0.004	0.002
4	0.625	0.25	0.092	0.036	0.015	0.007	0.003
5		0.333	0.133	0.055	0.024	0.011	0.006
6		0.444	0.192	0.082	0.037	0.017	0.009
7		0.556	0.258	0.115	0.053	0.026	0.013
8			0.333	0.158	0.074	0.037	0.019
9			0.417	0.206	0.10	0.051	0.027
10			0.50	.264	0.134	0.069	0.036
11			0.583	0.324	0.172	0.090	0.049
12				0.394	0.216	0.117	0.064
13				0.464	0.265	0.147	0.082
14				0.538	0.319	0.83	0.104
15					0.378	0.223	0.130
16					0.438	0.267	0.159
17					0.500	0.314	0.191
18					0.562	0.365	0.228
19						0.418	0.267
20						0.473	0.310
21						0.527	0.355
22							0.402
23							0.451
24							0.5
25							0.549

FIGURE 4.3 – Table des valeurs de Φ pour une loi normale $N(0,1)$.
$\Phi(Z_p) = P(Z \leq Z_p),\ Z \sim N(0,1)$.

Z_p	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53386
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99606	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

Bibliographie

[1] D. Blondin (2004). Lois limites uniformes et estimation non-paramétrique de la régression. Thèse de doctorat de l'Université Paris 6- Laboratoire de statistique théorique et appliquée. UFR 920. Hal open science.

[2] A. W. Bowman (1985). A comparative study of some kernel-based nonparametric density estimates. Journal of Statistical Computation and Simulation, Vol 21, Issue 3-4, pages 313-327.

[3] A. W. Bowman (1984). An alternative method of cross-validation for the smoothing of density estimates, Biometrika 71, 353-360.

[4] P. Capéraà, B. Van Cutsem (1988). Méthodes et modèles en statistique non paramétrique, exposé fondamental, Presses de l'université Laval, Dunod.

[5] WS. Cleveland (1979). Robust Locally Weighted Regression. Journal of the American Statistical Association 74 : 829-836.

[6] G. Collomb (1977). Quelques propriétés de la méthode du noyau pour l'estimation non-paramétrique de la régression en un point fixé. *Comptes rendus de l'Académie des Sciences de Paris tome 285, série A : 289-292.*

[7] AC. Davison, DV. Hinkley (1997). *Bootstrap Methods and their Application. Cambridge University Press.*

[8] B. Efron (1960). An introduction to the bootstrap, Springer-Science+Business Media,B.V.

[9] J. Fan (1992). Local linear regression smoothers and their minimax efficiencies. Ann. Statist., 21, 196-216.

[10] J. Fan et I. Gijbels (1996). Local polynomial modelling and its applications, 66. Chapman & Hall, London.

[11] I. Gijbels (2002-2003). chapitre 3. SyllabusSTA.pdf. Estimation non paramétrique d'une fonction de répartition.

[12] I. Gijbels (2008-2009). https ://pers.ulouvain.be/ rainer.vonsachs/STAT2150/STAT2150-Transp.pdf.

[13] P. Hall (1992). Asymptotic properties of integrated square error and cross validation for kernel estimation of a regression function. Z. Wahrsch. Verw. Gebiete, 67, 175-196.

[14] W. Härdle (1990). Applied Nonparametric Regression. Cambridge University Press, Cambridge.

[15] W. Härdle, W. Hall, J. S. Marron (1992). Regression smoothing parameters that are not far from their optimum. J. Amer. Statist. Assoc., 87, 227-233.

[16] W. Härdle, G. Kelly (1987). Nonparametric kernel regression estimation - optimal choice of bandwidth. Statistics., 12.2, 612-623

[17] W. Härdle, J. S. Marron (1985). Optimal bandwidth selection in nonparametric regression function estimation. Ann. Statist., 13.4, 1465-1481.

[18] D. V. Hinkley (1997). Bootstrap methods and their applications, Cambridge University Press.

[19] M. Lejeune (2010). Statistique. La théorie et ses applications. Deuxième édition. Springer-Verlag France, Paris.

[20] M.Lejeune (1985). Estimation non-paramétrique par noyaux : régression polynomiale mobile. Revue de Statist. Appliq., 33, 43-68.

[21] M. Lejeune (1984). Optimisation in non Parametric Regression. Compstat 84, Proceeedings in Computational Statistics Appliquée vol. XXXIII, N° 3 : 43-67.

[22] M. Lejeune (1983). Estimation non-paramétrique multivariée par noyaux. Rapport technique No 3, Projet No 2.843-0.80 du FNRS.

[23] C. Matias (2013-2014). Introduction à la statistique non paramétrique. CNRS, Laboratoire Statistique & Génome, Évry. http : //stat.genopole.cnrs.fr/~ cmatias. ENSIIE - 2013-2014.

[24] H.-G, Müller (1987). Weighted local regression and kernel methods for nonparametric curve fitting. J. Amer. Statist. Assoc., 82, 231-238.

[25] E. Nadaraya (1965). On nonparametric estimation of density function and regression, Theory of Probability and its Applications. 10, 186-190.

[26] E. Nadaraya (1964). On estimating regression. Theory of Probability and its Applications. 141-143.

[27] A. Renyi (1966). Calcul des probabilités, Dunod.

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

[29] M. Rudemo (1982). Empirical choice of histograms and density estimators, Scand. J. Statist. 9, 65-78.

[30] D. Ruppert, M. Wand (1994). Multivariate weighted least squares regression. Ann. Statst., 22, 1346-1370.

[31] G. Saporta (1990). Probabilités, analyse des données et statistique. Technip, Paris.

[32] J.Shao, D. Tu (1995). The jacknife and Bootstrap. Springer-Verlag, New York.

[33] A. B. Tsybakov (2012). Introduction à l'estimation non paramétrique, collection application of mathematics, Vol. 41.

[34] M.P. Wand et M. C. Jones (1995). Kernel Smoothing. Chapmann and Hall, London.

[35] L. Wasserman (2006). All of nonparametric statistics. Springer Texts in Statistics. Springer-Verlag.

[36] G.S. Watson (1964). Smooth Regression. The Indian Journal of Statistics, Series A, Vol 26, n° 4, pp. 359-372.