

**Notes de cours : Introduction à la statistique non  
paramétrique**

**Catherine Matias**

# Chapitre 1

## Introduction

Quelques références bibliographiques :

- ★ Wasserman, 'All of nonparametric statistics', Springer.
- ★ Tsybakov, 'Introduction à l'estimation non paramétrique', Springer.
- ★ Lehmann & D'Abbrera, 'Nonparametrics : statistical methods based on ranks', Springer.

### 1.1 Qu'est-ce que la statistique non paramétrique ?

La statistique paramétrique est le cadre « classique » de la statistique. Le modèle statistique  $y$  est décrit par un nombre fini de paramètres. Typiquement  $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \mathbb{R}^p\}$  est le modèle statistique qui décrit la distribution des variables aléatoires observées.

**Exemples.** ★  $\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^{+\ast}\}$ , modèle Gaussien.

★  $\mathcal{M} = \{\Gamma(\alpha, \beta); (\alpha, \beta) \in \mathbb{R}^{+\ast}\}$ , modèle loi Gamma.

★  $\mathcal{M} = \{f(x; \theta) = h(x) \exp[\eta(\theta)T(x) - A(\theta)]; \theta \in \mathbb{R}^p\}$ , modèle des familles exponentielles.

Par opposition, en statistique **non paramétrique**, le modèle n'est pas décrit par un nombre fini de paramètres. Divers cas de figure peuvent se présenter, comme par exemple :

- ★ On s'autorise *toutes* les distributions possibles, *i.e.* on ne fait aucune hypothèse sur la forme/nature/type de la distribution des variables aléatoires.
- ★ Le nombre de paramètres du modèle n'est pas fixé et varie (augmente) avec le nombre d'observations.

## 1.2 Quelques exemples de problèmes de statistique non paramétrique

### 1.2.1 Estimer une fonction de répartition et des fonctionnelles de la distribution

On observe  $X_1, \dots, X_n$  variables aléatoires (v.a.) réelles, i.i.d. de fonction de répartition (fdr)  $F : x \rightarrow F(x) = \mathbb{P}(X_1 \leq x)$ . L'estimateur naturel de la fdr  $F$  est la fdr empirique  $\hat{F}_n$  définie par  $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{X_i \leq x}$ . C'est un estimateur non paramétrique de la fdr  $F$ .

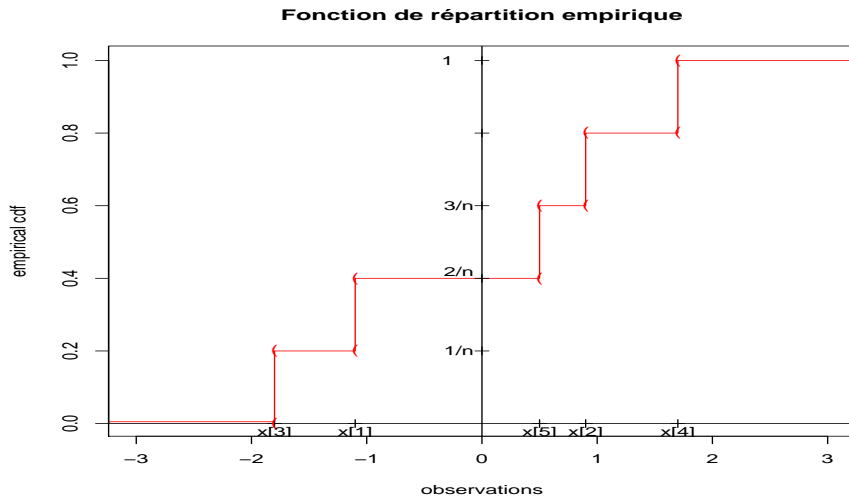


FIG. 1.1 – Fonction de répartition empirique.

#### Qualité de cet estimateur ?

##### 1) Propriétés ponctuelles (*i.e.* $x$ fixé)

- ★  $\mathbb{E}\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{P}(X_i \leq x) = F(x)$ , *i.e.* c'est un estimateur sans biais.
- ★  $\text{Var}(\hat{F}_n(x)) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(1_{X_i \leq x}) = \frac{1}{n} \text{Var}(1_{X_1 \leq x}) = F(x)(1 - F(x))/n$ . Donc  $\text{Var}(\hat{F}_n(x)) \rightarrow_{n \rightarrow \infty} 0$ .
- ★ Erreur en moyenne quadratique (ou MSE pour « mean square error ») :  $\mathbb{E}[(\hat{F}_n(x) - F(x))^2] = \text{biais}^2 + \text{variance} = \text{Var}(\hat{F}_n(x)) \rightarrow_{n \rightarrow \infty} 0$ .
- ★  $\hat{F}_n(x) \xrightarrow[n \rightarrow \infty]{\text{proba}} F(x)$ . En effet, d'après l'inégalité de Markov, la convergence en moyenne quadratique implique la convergence en probabilité.
- ★ (LGN) :  $\hat{F}_n(x) \xrightarrow[n \rightarrow \infty]{\text{p.s.}} F(x)$ .
- ★ (TCL) :  $\sqrt{n}(\hat{F}_n(x) - F(x)) \rightsquigarrow_{n \rightarrow \infty}^{\mathcal{L}} \mathcal{N}(0, F(x)(1 - F(x)))$ .

★ (Loi du logarithme itéré LIL). Rappel : si  $\{X_i\}_{i \geq 0}$  suite de v.a. i.i.d., centrées, de variance  $\sigma^2 < +\infty$  et  $S_n = \sum_{i=1}^n X_i$ . Alors

$$\limsup_{n \rightarrow \infty} \frac{|S_n|}{\sigma \sqrt{2n \log \log n}} = 1 \quad \text{p.s.}$$

En particulier

$$\limsup_{n \rightarrow \infty} \frac{\sqrt{n} |\hat{F}_n(x) - F(x)|}{\sqrt{F(x)(1-F(x))} 2 \log \log n} = 1 \quad \text{p.s.}$$

## 2) Propriétés uniformes

★ Théorème de Glivenko Cantelli (voir preuve en TD) :

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\text{p.s.}} 0.$$

★ Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW, admis) :

$$\forall n \in \mathbb{N}, \forall \epsilon > 0, \quad \mathbb{P}(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \epsilon) \leq 2 \exp(-2n\epsilon^2).$$

**Exemple d'application de l'inégalité de DKW :** Construction d'intervalles de confiance (IC) exacts sur  $F(x)$ .

En effet,  $\forall x \in \mathbb{R}$ , on a

$$\begin{aligned} \mathbb{P}(F(x) \in [\hat{F}_n(x) - \epsilon; \hat{F}_n(x) + \epsilon]) &= 1 - \mathbb{P}(|\hat{F}_n(x) - F(x)| > \epsilon) \\ &\geq 1 - \mathbb{P}(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon) \geq 1 - 2 \exp(-2n\epsilon^2). \end{aligned}$$

Pour tout  $\alpha > 0$ , on choisit alors  $\epsilon > 0$  tel que  $2 \exp(-2n\epsilon^2) = \alpha$ , *i.e.* on prend  $\epsilon = \sqrt{\log(2/\alpha)/(2n)}$  et on obtient

$$\mathbb{P}(F(x) \in [\hat{F}_n(x) - \sqrt{\log(2/\alpha)/(2n)}; \hat{F}_n(x) + \sqrt{\log(2/\alpha)/(2n)}]) \geq 1 - \alpha,$$

donc  $[\hat{F}_n(x) - \sqrt{\log(2/\alpha)/(2n)}; \hat{F}_n(x) + \sqrt{\log(2/\alpha)/(2n)}]$  est un IC au niveau  $1 - \alpha$  pour  $F(x)$ .

**Remarques.** 1) Comme  $F(x) \in [0, 1]$ , si  $n$  est petit on peut souvent raffiner cet IC en prenant plutôt  $[\hat{F}_n(x) - \sqrt{\log(2/\alpha)/(2n)}; \hat{F}_n(x) + \sqrt{\log(2/\alpha)/(2n)}] \cap [0, 1]$ .

2) Le TCL permet également d'obtenir un IC pour  $F(x)$ , à condition d'estimer la variance  $F(x)(1-F(x))$ . Mais cet intervalle est **asymptotique** uniquement. Il peut s'avérer meilleur que l'intervalle exact ci-dessus car ce dernier est fondé sur une borne **uniforme** qui peut être mauvaise pour certaines valeurs de  $x$ .

Dans le Chapitre 2, nous nous intéresserons également au cas des fonctionnelles régulières de la distribution, comme la moyenne, la variance, la médiane, *etc.*

## 1.2.2 Tests non paramétriques

Principe : faire un test statistique, sans spécifier la distribution des variables aléatoires.

**Exemples.** 1) Test d'adéquation de Kolmogorov Smirnov (KS test).

À partir d'un échantillon de v.a. réelles  $X_1, \dots, X_n$  et d'une fdr  $F_0$  fixée, on veut tester  $H_0 : F = F_0$  contre  $H_1 : F \neq F_0$ . On utilise la statistique

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|.$$

À partir de deux échantillons de v.a. réelles  $X_1, \dots, X_n$  et  $Y_1, \dots, Y_m$  on peut aussi tester  $H_0 : F_X = F_Y$  contre  $H_1 : F_X \neq F_Y$ , via la statistique

$$D_{n,m} = \sup_{t \in \mathbb{R}} |\hat{F}_{n,X}(t) - \hat{F}_{m,Y}(t)|.$$

C'est un test **asymptotique**, fondé sur le fait que la distribution limite de  $\sqrt{n}D_n$  (resp.  $\sqrt{nm/(n+m)}D_{n,m}$ ) ne dépend pas de la distribution de l'échantillon initial. Cette distribution est tabulée. Le test est restreint au cas où l'on suppose que la fdr de l'échantillon est continue (variables diffuses). C'est un test sensible à des différences à la fois dans la forme et dans la localisation des distributions.

2) Test d'adéquation du  $\chi^2$  de Pearson.

À partir d'un échantillon de v.a. discrètes (qualitatives ou quantitatives)  $X_1, \dots, X_n$  et d'une fdr  $F_0$  fixée, on teste  $H_0 : F = F_0$  contre  $H_1 : F \neq F_0$ . C'est un test **asymptotique**, fondé sur la loi limite de la statistique de test qui suit un  $\chi^2$ . On peut l'utiliser dans le cas où  $F_0$  est définie à un paramètre près  $\theta_0$  qui est alors estimé à partir des observations. Il existe une version exacte du test : le test de Fisher mais qui est coûteux en temps de calcul s'il y a beaucoup de catégories.

NB : Les tests d'adéquation ci-dessus sont non paramétriques car la distribution des variables n'est pas spécifiée sous l'alternative.

3) Test du  $\chi^2$  d'indépendance.

À partir de deux échantillons de v.a. discrètes (qualitatives ou quantitatives)  $X_1, \dots, X_n$  à valeurs dans  $\mathcal{A}$  et  $Y_1, \dots, Y_m$  à valeurs dans  $\mathcal{B}$ , on fabrique une table de contingence constituée par le nombre de couples  $(X_i, Y_j)$  qui prennent les valeurs  $(a, b)$ ,  $a \in \mathcal{A}$  et  $b \in \mathcal{B}$ . On teste  $H_0$  : « les  $\{X_i\}$  sont indépendants des  $\{Y_j\}$  » contre  $H_1$  : «  $\{X_i\}$  et  $\{Y_j\}$  sont non indépendants ». C'est aussi un test **asymptotique**.

#### 4) Tests de normalité.

Test de  $H_0$  : « la distribution des variables est gaussienne » contre  $H_1$  : « la distribution des variables n'est pas gaussienne ». Ces tests sont utiles par exemple avant de faire un test de Student, qui est peu robuste à la non-normalité des variables. Les deux tests d'adéquation mentionnés ci-dessus (KS-test et  $\chi^2$  de Pearson) peuvent être utilisés comme tests de normalité (avec  $F_0$  la fdr de la  $\mathcal{N}(0, 1)$ ).

Dans le Chapitre 3, nous verrons les tests de signe et de rang.

### 1.2.3 Estimation de densité

On observe  $X_1, \dots, X_n$  v.a. i.i.d., diffuses, de densité  $f$ . Le but est d'estimer cette densité  $f$ . Pour cela, on fait des hypothèses sur sa régularité.

Principe : on suppose que  $f$  appartient à une classe de fonctions arbitrairement grande. Alors, l'estimateur du maximum de vraisemblance n'existe pas. On construit des estimateurs dits non paramétriques.

Qualité de l'estimation : il y a deux types d'erreur

- ★ Biais = il est induit par le choix du modèle. Il traduit la « distance » de la vraie densité au modèle. Cette erreur diminue lorsqu'on passe d'un modèle paramétrique à un modèle non paramétrique.
- ★ Variance = induite par l'approximation dans un espace plus ou moins grand. Cette erreur augmente lorsqu'on passe d'un modèle paramétrique à un modèle non paramétrique.

Nous aborderons ceci dans le Chapitre 4.

### 1.2.4 Régression non paramétrique

On observe une suite de couples  $\{(X_i, Y_i)\}_{1 \leq i \leq n}$  avec  $Y_i = r(X_i) + \epsilon_i$  où  $r$  est une fonction quelconque (régulière) que l'on cherche à estimer.

Nous mettrons en œuvre les mêmes techniques d'estimation non paramétriques que pour la densité. La différence majeure réside dans la classe de fonctions  $r$  qui n'est pas contrainte à avoir une intégrale finie.

Nous aborderons ceci dans le Chapitre 5.

### 1.2.5 Estimation en grande dimension

Le problème de la grande dimension se pose lorsque le nombre de paramètres  $p$  est de l'ordre de, voire plus grand que le nombre d'observations  $n$ . Dans ce cas, plusieurs approches statistiques sont possibles

- ★ Modèles creux (ou *sparses* en anglais) : on fait l'hypothèse qu'un grand nombre de paramètres est en fait nul, mais on ne sait pas lesquels.
- ★ Approches de type sélection de variables : là encore, on fait l'hypothèse qu'un grand nombre de covariables sont non pertinentes et on cherche à sélectionner celles qui le sont.

Nous n'aborderons pas ces problèmes dans ce cours.

### **1.2.6 Autres**

On peut raconter le bootstrap aussi (c'est ce que fait Wasserman).

# Chapitre 2

## Fonction de répartition et fonctionnelles de la distribution

### 2.1 Rappels sur les fonctions de répartition

On note  $\mathcal{F}$  l'ensemble des fonctions de répartition, c'est-à-dire

$$\mathcal{F} = \{F : \mathbb{R} \rightarrow [0, 1]; F \text{ croissante, càdlàg, } \lim_{t \rightarrow -\infty} F(t) = 0, \lim_{t \rightarrow +\infty} F(t) = 1\}.$$

**Proposition 1.** *À toute mesure de probabilité  $\mathbb{P}$  sur  $\mathcal{A} \subset \mathbb{R}$  peut être associée une fonction de répartition  $F$  définie par  $F(t) = \mathbb{P}(]-\infty; t] \cap \mathcal{A})$ ,  $\forall t \in \mathbb{R}$ . Réciproquement, si  $F \in \mathcal{F}$  alors il existe une unique mesure de probabilité sur  $\mathbb{R}$  dont  $F$  soit la fdr. Cette mesure est définie à partir de sa valeur sur tous les intervalles  $]a; b] \subset \mathbb{R}$  et en posant  $\mathbb{P}(]a; b]) = F(b) - F(a)$ .*

Si  $F \in \mathcal{F}$ , on note  $dF$  l'unique mesure de proba associée et on peut définir ainsi la notation  $\int h(x)dF(x)$  pour toute fonction  $h$ .

En particulier, si  $F$  est continue (*i.e.* si  $dF$  est une mesure absolument continue) alors en notant  $f = F'$  la densité on obtient  $\int h(x)dF(x) = \int h(x)f(x)dx$  et si  $F$  est constante par morceaux (*i.e.* si  $dF$  est une mesure discrète) alors  $\int h(x)dF(x) = \sum_{a \in \mathcal{A}} h(a)w_a$  où  $\mathcal{A}$  est le support de la mesure et  $\{w_a\}_{a \in \mathcal{A}}$  l'ensemble des poids associés.

NB : La fdr empirique  $\hat{F}_n$  est une fonction constante par morceaux. Elle est associée à la mesure empirique  $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$  où  $\delta_x$  est la masse de Dirac au point  $x$ , *i.e.*  $\mathbb{P}_n$  est une mesure discrète qui associe le poids  $1/n$  à chacune des observations  $X_i$ . Alors, pour toute fonction  $h$ , on a  $\int h(x)d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n h(X_i)$ .

## 2.2 Fonctionnelles de la distribution

### 2.2.1 Estimation des fonctionnelles

Dans ce cours, une fonctionnelle est une fonction régulière  $T$  définie sur l'ensemble des fonctions de répartition  $\mathcal{F}$ .

**Exemples.** Quelques exemples de fonctionnelles régulières :

$$\text{Moyenne : } F \rightarrow \mu(F) = \int x dF(x),$$

$$\text{Variance : } F \rightarrow \sigma^2(F) = \int (x - \mu(F))^2 dF(x) = \int x^2 dF(x) - (\int x dF(x))^2,$$

$$\text{Médiane : } F \rightarrow m(F) = F^{-1}(1/2) \text{ et quantiles } F \rightarrow q_\alpha(F) = F^{-1}(\alpha),$$

$$\text{Skewness (ou coefficient d'asymétrie) : } F \rightarrow \left\{ \int (x - \mu(F))^3 dF(x) \right\} / \sigma(F)^{3/2}.$$

NB : La densité  $f = F'$  n'est pas une fonctionnelle régulière de la densité.

**Rappel.** La fdr  $F$  est croissante (pas nécessairement strictement) de  $\mathbb{R}$  dans  $[0; 1]$ . On peut définir la notion d'*inverse généralisée* de  $F$  de la façon suivante :

$$\forall y \in [0; 1], \quad F^{-1}(y) = \inf\{x \in \mathbb{R}; F(x) \geq y\}.$$

L'inverse de la fdr est la fonction quantile.

Plus généralement, une classe de fonctionnelles intéressantes est la classe des fonctionnelles linéaires, *i.e.* telles qu'il existe  $a : \mathbb{R} \rightarrow \mathbb{R}$  telle que  $T : F \rightarrow T(F) = \int a(x) dF(x)$ .

**Remarque.** La moyenne est une fonctionnelle linéaire, mais pas la variance, ni la médiane, ni les quantiles, ni le coefficient d'asymétrie.

**Méthode « plug-in » pour l'estimation de fonctionnelles.** Si  $T : F \rightarrow T(F)$  est une fonctionnelle alors un estimateur naturel de  $T(F)$  est obtenu en « injectant » l'estimateur  $\hat{F}_n$  de  $F$  dans l'expression de  $T$ , *i.e.*  $\hat{T}_n = T(\hat{F}_n)$  est un estimateur naturel de  $T(F)$ .

**Exemples.** Les estimateurs suivants sont obtenus par la méthode plug-in

$$\star \text{ Moyenne empirique : } \bar{X}_n = \int x d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n F(X_i),$$

$$\star \text{ Variance empirique : } \hat{\sigma}_n^2 = \int x^2 d\hat{F}_n(x) - (\int x d\hat{F}_n(x))^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \text{ Cet estimateur est biaisé, puisque}$$

$$\begin{aligned} \mathbb{E}(\hat{\sigma}_n^2) &= \mathbb{E}X_1^2 - \frac{1}{n^2} \left( \sum_{i=1}^n \mathbb{E}X_i^2 + \sum_{i \neq j} \mathbb{E}X_i X_j \right) \\ &= \mathbb{E}X_1^2 - \frac{1}{n} \mathbb{E}X_1^2 - \frac{n-1}{n} (\mathbb{E}X_1)^2 = \left(1 - \frac{1}{n}\right) (\mathbb{E}X_1^2 - (\mathbb{E}X_1)^2) = \left(1 - \frac{1}{n}\right) \sigma^2. \end{aligned}$$

On lui préfère souvent  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  qui est sans biais.

★ Médiane empirique  $\hat{m} = \hat{F}_n^{-1}(1/2)$  ou quantile empirique  $\hat{q}_\alpha = \hat{F}_n^{-1}(\alpha)$

**Rappel : la méthode Delta.** Si  $(X_n)_{n \geq 0}$  suite de v.a. dans  $\mathbb{R}^d$  telles qu'il existe  $\mu \in \mathbb{R}^d$  et  $(a_n)_{n \geq 0}$  suite de réelles avec  $a_n(X_n - \mu) \overset{\mathcal{L}}{\rightsquigarrow}_{n \rightarrow \infty} \mathcal{N}_d(0, \Sigma)$  et si  $g : \mathbb{R}^d \rightarrow \mathbb{R}^s$  est différentiable au voisinage de  $\mu$ , alors

$$a_n(g(X_n) - g(\mu)) \overset{\mathcal{L}}{\rightsquigarrow}_{n \rightarrow \infty} \mathcal{N}_s(0, \nabla g(\mu)^t \times \Sigma \times \nabla g(\mu)).$$

Exemple d'application : Montrer que  $\hat{\sigma}_n^2$  satisfait

$$\sqrt{n}(\hat{\sigma}_n^2 - m_2) \overset{\mathcal{L}}{\rightsquigarrow}_{n \rightarrow \infty} \mathcal{N}(0, m_4 - m_2^2),$$

où  $m_i = \mathbb{E}[(X_1 - \mathbb{E}X_1)^i]$ . (Indication : écrire un TCL sur le vecteur  $(\bar{X}_n, \overline{X_n^2})$ ).

## 2.2.2 Fonction d'influence

La fonction d'influence est un équivalent de la fonction de score en statistique paramétrique. C'est une dérivée de la fonctionnelle. Pour définir une dérivée, il faut définir un taux d'accroissement. Comme une fonctionnelle  $T$  a pour argument  $F \in \mathcal{F}$ , il faut définir un accroissement élémentaire dans  $\mathcal{F}$ .

Pour tout  $x_0 \in \mathbb{R}$ , on notera  $\delta_{x_0}$  la masse de Dirac en  $x_0$  et  $G_{\delta_{x_0}}$  la f.d.r. associée à  $\delta_{x_0}$ . Plus précisément, on a  $G_{\delta_{x_0}}(t) = 1_{x_0 \leq t}$  pour tout  $t \in \mathbb{R}$  (voir Figure 2.1).

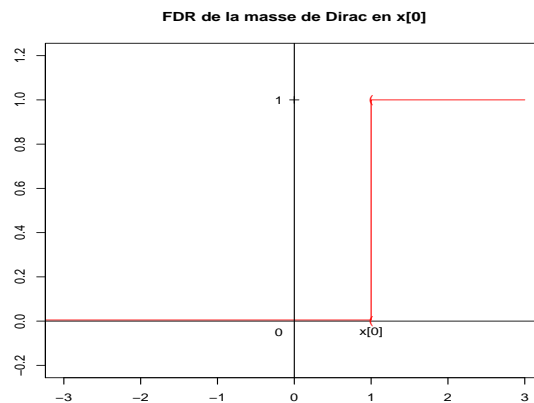


FIG. 2.1 – Fonction de répartition  $G_{\delta_{x_0}}$  de la masse de Dirac en  $x_0$ .

**Définition. (Fonction d'influence)**

Soit  $T : F \rightarrow T(F)$  une fonctionnelle régulière. La fonction d'influence de  $T$  en  $F$  au point  $x_0$  est définie par la limite suivante, si elle existe

$$L_{T,F}(x_0) = \lim_{\epsilon \rightarrow 0} \frac{T((1-\epsilon)F + \epsilon G_{\delta_{x_0}}) - T(F)}{\epsilon}.$$

**Remarque.** Si  $F \in \mathcal{F}$  alors pour tout  $\epsilon > 0$ , on a  $(1-\epsilon)F + \epsilon G_{\delta_{x_0}} \in \mathcal{F}$ . En effet, c'est une fonction croissante, càdlàg, qui tend vers 0 en  $-\infty$  et vers 1 en  $+\infty$ .

**Exemple.** Soit  $\mu : F \rightarrow \mu(F) = \int x dF(x)$ . Alors, pour tout  $\epsilon > 0$ , et tout  $x_0 \in \mathbb{R}$ , on a  $\mu((1-\epsilon)F + \epsilon G_{\delta_{x_0}}) = (1-\epsilon)\mu(F) + \epsilon\mu(G_{\delta_{x_0}})$  car  $\mu$  est linéaire!! De plus,  $\mu(G_{\delta_{x_0}}) = x_0$ . Donc on obtient

$$\frac{\mu((1-\epsilon)F + \epsilon G_{\delta_{x_0}}) - \mu(F)}{\epsilon} = \frac{(1-\epsilon)\mu(F) + \epsilon x_0 - \mu(F)}{\epsilon} = x_0 - \mu(F).$$

Ainsi,  $L_{\mu,F}(x_0) = x_0 - \mu(F)$ .

**Définition. (Fonction d'influence empirique)**

Soit  $T : F \rightarrow T(F)$  une fonctionnelle régulière. La fonction d'influence empirique de  $T$  en  $F$  au point  $x_0$  est définie par la limite suivante, si elle existe

$$\hat{L}_n(x_0) = \lim_{\epsilon \rightarrow 0} \frac{T((1-\epsilon)\hat{F}_n + \epsilon G_{\delta_{x_0}}) - T(\hat{F}_n)}{\epsilon}.$$

**Exemple. (suite).** La fonction d'influence empirique associée à la moyenne  $\mu$  en  $F$  au point  $x_0$  est  $\hat{L}_n(x_0) = x_0 - \bar{X}_n$ .

**Construction d'intervalles de confiance pour  $T(F)$ .** Nous allons d'abord détailler l'exemple de la moyenne  $\mu(F) = \int x dF(x)$ . Le théorème central limite nous donne

$$\sqrt{n} \frac{(\mu(F) - \bar{X}_n)}{\sigma(F)} \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, 1),$$

où  $\sigma^2(F) = \text{Var}(X) = \int (x - \mu(F))^2 dF(x)$ . On peut remarquer au passage que comme  $L_{\mu,F}(X) = X - \mu(F)$  et que  $\mu(F)$  est une constante, on a  $\sigma^2(F) = \text{Var}(X) = \text{Var}(L_{\mu,F}(X))$ .

Le TCL précédent ne permet pas de construire un IC (asymptotique) pour  $\mu(F)$  puisque la variance  $\sigma^2(F)$  est inconnue. Il faut donc l'estimer, par exemple par  $\hat{\sigma}_n^2 = \sigma^2(\hat{F}_n)$ . Comme on a remarqué que  $\sigma^2(F) = \text{Var}(L_{\mu,F}(X))$ , ceci revient exactement

à estimer  $\sigma^2(F)$  par  $\text{Var}(\hat{L}_n(X))$ . Au final, on peut obtenir (via le Lemme de Slutsky combiné au TCL précédent), la convergence

$$\sqrt{n} \frac{(\mu(F) - \bar{X}_n)}{\sqrt{\text{Var}(\hat{L}_n)}} \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, 1),$$

qui permet de construire un IC (asymptotique) pour  $\mu(F)$ .

Cette démarche est applicable plus généralement pour toutes les fonctionnelles linéaires, comme explicité dans le résultat suivant.

**Théorème 2. (Cas des fonctionnelles linéaires).** *Si  $T : F \rightarrow T(F)$  est une fonctionnelle linéaire, i.e. de la forme  $T(F) = \int a(x)dF(x)$ , alors*

- i)  $L_{T,F}(x_0) = a(x_0) - T(F)$  et  $\hat{L}_n(x_0) = a(x_0) - T(\hat{F}_n) = a(x_0) - \frac{1}{n} \sum_{i=1}^n a(X_i)$ ,
- ii)  $\forall H \in \mathcal{F}$ , on a  $T(H) = T(F) + \int L_{T,F}(x)dH(x)$ ,
- iii)  $\mathbb{E}(L_{T,F}(X)) = \int L_{T,F}(x)dF(x) = 0$ ,
- iv) Soit  $\tau^2 = \int L_{T,F}^2(x)dF(x) = \mathbb{E}(L_{T,F}(X)^2) = \text{Var}(L_{T,F}(X))$  alors on a

$$\tau^2 = \int (a(x) - T(F))^2 dF(x) = \int a^2(x)dF(x) - T(F)^2.$$

De plus, si  $\tau^2 < +\infty$ , alors  $\sqrt{n}(T(F) - T(\hat{F}_n)) \underset{n \rightarrow \infty}{\rightsquigarrow} \mathcal{N}(0, \tau^2)$ .

- v) On définit  $\hat{\tau}_n^2 = \frac{1}{n} \sum_{i=1}^n \hat{L}_n^2(X_i) = \frac{1}{n} \sum_{i=1}^n [a(X_i) - T(\hat{F}_n)]^2$  estimateur de  $\tau^2$ , alors on a la convergence

$$\hat{\tau}_n^2 \underset{n \rightarrow \infty}{\overset{\mathbb{P}}{\rightarrow}} \tau^2,$$

et par conséquent

$$\sqrt{n} \frac{(T(F) - T(\hat{F}_n))}{\hat{\tau}_n} \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, 1).$$

*Démonstration.* i) est immédiat ; i)  $\Rightarrow$  ii) aussi et ii)  $\Rightarrow$  iii) également.

Montrons à présent iv). D'après ii) appliqué à  $H = \hat{F}_n$  on a  $T(\hat{F}_n) = T(F) + \int L_{T,F}(x)d\hat{F}_n(x)$ , i.e.  $T(\hat{F}_n) - T(F) = \frac{1}{n} \sum_{i=1}^n L_{T,F}(X_i)$ . En appliquant le TCL avec  $\mathbb{E}L_{T,F}(X) = 0$  et  $\tau^2 = \text{Var}(L_{T,F}(X)) < +\infty$ , on a donc

$$\sqrt{n}(T(\hat{F}_n) - T(F)) \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, \tau^2).$$

v) La LGN donne la convergence en probabilité de  $\hat{\tau}_n^2$  vers  $\tau^2$ . La conclusion de la preuve provient du lemme suivant.

**Lemme 3. (Slutsky)** *Si  $X_n \underset{n \rightarrow \infty}{\rightsquigarrow} X$  et  $Y_n \underset{n \rightarrow \infty}{\rightsquigarrow} c$  constante alors  $(X_n, Y_n) \underset{n \rightarrow \infty}{\rightsquigarrow} (X, c)$ . En particulier, pour toute fonction  $h$  de 2 variables, on a  $h(X_n, Y_n) \underset{n \rightarrow \infty}{\rightsquigarrow} h(X, c)$ .*

□

Terminons à présent en donnant quelques propriétés élémentaires des fonctions d'influence.

**Proposition 4.** 1) Si  $T, S : \mathcal{F} \rightarrow \mathbb{R}$  sont deux fonctionnelles de fonctions d'influence respectives  $L_{T,F}$  et  $L_{S,F}$  au point  $F \in \mathcal{F}$  et si  $\lambda \in \mathbb{R}$  alors  $\lambda T + S$  est une fonctionnelle de fonction d'influence  $\lambda L_{T,F} + L_{S,F}$  au point  $F$  et  $T \times S$  est une fonctionnelle de fonction d'influence  $T(F) \times L_{S,F} + S(F) \times L_{T,F}$  au point  $F$ .

2) Si  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  est une fonction dérivable et si  $T$  est une fonctionnelle de fonction d'influence  $L_{T,F}$  au point  $F$  alors la fonctionnelle  $S = \psi \circ T$  a pour fonction d'influence au point  $F$  la fonction  $L_{S,F} = \psi' \circ T \times L_{T,F}$ .

*Démonstration.* Immédiat. □

**Application.** Nous allons utiliser cette proposition pour calculer la fonction d'influence de la variance. On a  $\sigma^2(F) = \int (x - \mu(F))^2 dF(x) = \int x^2 dF(x) - \mu(F)^2$ . D'après la proposition précédente  $L_{\sigma^2,F} = L_{T,F} - 2\mu(F)L_{\mu,F}$  où  $T : F \rightarrow T(F) = \int x^2 dF(x)$ . Or  $L_{\mu,F}(x) = x - \mu(F)$  et  $T$  est également une fonctionnelle linéaire donc  $L_{T,F}(x) = x^2 - T(F) = x^2 - \int u^2 dF(u)$ . Donc on obtient

$$L_{\sigma^2,F}(x) = x^2 - \int u^2 dF(u) - 2\mu(F)(x - \mu(F)) = (x - \mu(F))^2 - \sigma^2(F).$$

Il faut noter ici qu'on retrouve la forme  $L_{\sigma^2,F}(x) = a_F(x) - \sigma^2(F)$  avec  $\sigma^2(F) = \int a_F(x) dF(x)$  et  $a_F(x) = (x - \mu(F))^2$ , pourtant,  $\sigma^2$  n'est pas une fonctionnelle linéaire car la fonction  $a$  dépend de  $F$ .

**Remarque.** Il existe des liens entre la notion de robustesse et le comportement de la fonction d'influence en l'infini, qu'il serait intéressant de raconter. En effet, si

$$\exists A > 0, \forall x \in \mathbb{R}, |L_{T,F}(x)| \leq A,$$

alors la fonctionnelle  $T$  est robuste aux valeurs aberrantes. En particulier, ce n'est pas le cas de la moyenne, mais c'est le cas de la médiane.

# Chapitre 3

## Tests non paramétriques : tests de signe et de rang

### 3.1 Introduction, rappels et généralités

#### 3.1.1 Introduction

Dans la suite, on observe soit un échantillon  $X_1, \dots, X_n$  de v.a. réelles i.i.d de même loi que  $X$  ou bien deux échantillons  $X_1, \dots, X_n$  de même loi que  $X$  et  $Y_1, \dots, Y_m$  de même loi que  $Y$ .

**Exemples.** Quelques exemples de tests non paramétriques :

**Test de la médiane**  $H_0 : \mathbb{P}(X \leq 0) = 1/2$  *i.e.* « la médiane est nulle » contre  $H_1 : \mathbb{P}(X \leq 0) < 1/2$ , « la médiane est positive ».

**Test de normalité**  $H_0$  : «  $X$  est Gaussienne » contre  $H_1$  : «  $X$  n'est pas Gaussienne ».

**Tests d'adéquation**  $H_0$  : «  $X$  et  $Y$  ont la même loi » contre  $H_1$  : «  $X$  et  $Y$  n'ont pas la même loi ».

**Tests de comparaison**  $H_0 : \mathbb{P}(X \geq Y) \geq 1/2$  contre  $H_1 : \mathbb{P}(X \geq Y) < 1/2$ .

**Tests d'indépendance**  $H_0 : \{X_i\} \amalg \{Y_i\}$  contre  $H_1 : \{X_i\}$  ne sont pas indépendants des  $\{Y_i\}$ .

Ces tests sont non paramétriques car la distribution des variables aléatoires n'est pas spécifiée sous au moins une des deux hypothèses (nulle ou alternative).

**Principe général.** Trouver une statistique (de test)  $T(X_1, \dots, X_n)$  (ou bien  $T(X_1, \dots, X_n, Y_1, \dots, Y_m)$ ) dont la distribution sous  $H_0$  ne dépend pas de la distribution des v.a. observées. On parle de statistique *libre en loi*.

Comme en statistique paramétrique, on considère deux types de tests

**bilatères** : lorsqu'on sous l'alternative  $H_1$ , la statistique  $T$  n'est ni systématiquement « plus grande » ni « plus petite » que sous  $H_0$ .

**unilatères** : lorsqu'on sous l'alternative  $H_1$ , la statistique  $T$  est soit systématiquement « plus grande », soit « plus petite » que sous  $H_0$ .

### 3.1.2 Notion d'ordre stochastique

On introduit à présent des définitions qui permettent de donner un sens à la notion de variables aléatoire « plus grande » ou « plus petite ».

**Définition.** Si  $X$  v.a. réelle de fdr  $F$  et  $Y$  v.a. réelle de fdr  $G$  et si  $\forall x \in \mathbb{R}$ , on a  $G(x) \leq F(x)$  avec inégalité stricte pour au moins un  $x \in \mathbb{R}$ , alors on dit que  $Y$  est stochastiquement plus grande que  $X$  et on note  $Y \succ X$ .

En particulier, si  $T$  est une v.a.r. de fdr  $F_0$  sous l'hypothèse  $H_0$  et de fdr  $F_1$  sous l'hypothèse  $H_1$  et si  $\forall x \in \mathbb{R}$ ,  $F_0(x) \leq F_1(x)$  avec inégalité stricte en au moins un point, alors  $T$  est stochastiquement plus grande sous  $H_0$  que sous  $H_1$ .

**Exemple.** Soit  $T \sim \mathcal{N}(0, \theta)$ , et on considère  $H_0 : \theta = 0$  et  $H_1 : \theta = 1$ . Alors  $T$  est stochastiquement plus petite sous  $H_0$  que sous  $H_1$ . En effet,  $F_1(x) = \mathbb{P}_{H_1}(T \leq x) = \mathbb{P}_{H_1}(T - 1 \leq x - 1) = \mathbb{P}(Z \leq x - 1)$  où  $Z \sim \mathcal{N}(0, 1)$ . Et de même  $F_0(x) = \mathbb{P}_{H_0}(T \leq x) = \mathbb{P}(Z \leq x)$ . Donc on obtient  $F_1(x) = F_0(x - 1) < F_0(x)$ , car  $F_0$  strictement croissante.

**Proposition 5.** Si  $X_1 \prec Y_1, X_2 \prec Y_2$  et  $X_1 \amalg X_2, Y_1 \amalg Y_2$  alors  $X_1 + X_2 \prec Y_1 + Y_2$ .

*Démonstration.* On note  $F_{X_1}, F_{X_1+X_2}, F_{Y_1}$ , etc, les fdr associées aux variables aléatoires. Sans perte de généralité, on peut choisir des représentants  $Y_1 \amalg X_2$ . En effet, la propriété  $F_U \leq F_V$  ne dépend pas du choix des variables  $U, V$ . Soit  $t \in \mathbb{R}$ , on a

$$\begin{aligned} F_{X_1+X_2}(t) &= \mathbb{E}[\mathbb{E}(1\{X_1 + X_2 \leq t\} | X_2)] \stackrel{(a)}{=} \mathbb{E}[F_{X_1}(t - X_2)] \stackrel{(b)}{\geq} \mathbb{E}[F_{Y_1}(t - X_2)] \\ &\stackrel{(c)}{=} F_{Y_1+X_2}(t) \stackrel{(d)}{=} \mathbb{E}[F_{X_2}(t - Y_1)] \stackrel{(e)}{\geq} \mathbb{E}[F_{Y_2}(t - Y_1)] = F_{Y_1+Y_2}(t), \end{aligned}$$

où (a) découle de  $X_1 \amalg X_2$ , (b) découle de  $X_1 \prec Y_1$ , (c) et (d) du choix des représentants  $Y_1 \amalg X_2$ , (e) de l'hypothèse  $X_2 \prec Y_2$  et enfin (f) de l'hypothèse  $Y_1 \amalg Y_2$ . Pour finir, il faut remarquer que l'inégalité est stricte en au moins un point  $t \in \mathbb{R}$ .  $\square$

Une conséquence de la proposition précédente est la suivante.

**Proposition 6.** Si  $T \sim \mathcal{B}(n; \theta)$  et  $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$  avec  $\theta_0 < \theta_1$  alors  $T$  est stochastiquement plus petite sous  $H_0$  que sous  $H_1$ .

*Démonstration.* Si  $Y \sim \mathcal{B}(\theta)$  alors

$$F_Y(t) = \begin{cases} 0 & \text{si } t < 0, \\ 1 - \theta & \text{si } t \in [0; 1[, \\ 1 & \text{si } t \geq 1. \end{cases}$$

(voir la Figure 3.1). Comme  $\theta_0 < \theta_1$  on a donc que  $Y$  est plus petite sous  $H_0$  que

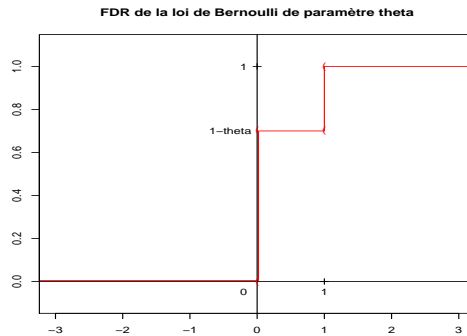


FIG. 3.1 – Fonction de répartition de la loi  $\mathcal{B}(\theta)$ .

sous  $H_1$ . De plus,  $T = \sum_{i=1}^n Y_i$  où les  $Y_i$  sont des variables indépendantes, de loi  $\mathcal{B}(\theta)$ .  $\square$

### 3.1.3 Rappels

**Choix de la région de rejet.** C'est toujours l'hypothèse alternative  $H_1$  qui détermine si le test est bilatère ou unilatère. C'est aussi l'alternative  $H_1$  qui détermine la région de rejet de l'hypothèse nulle  $H_0$ . En effet,  $\mathcal{R}_{H_0}$  est choisie comme une région où la densité de la statistique de test  $T_n$  est plus grande sous  $H_1$  que sous  $H_0$ .

**Exemple.** Si on observe  $X_1, \dots, X_n$  iid de loi  $\mathcal{B}(\theta)$  et on teste  $H_0 : \theta = 1/2$  contre  $H_1 : \theta > 1/2$ . Alors la statistique  $S_n = \sum_{i=1}^n X_i$  vérifie  $S_n \sim \mathcal{B}(n, \theta)$  et d'après la section précédente,  $S_n$  est stochastiquement plus grande sous  $H_1$  que sous  $H_0$ . Donc on rejette  $H_0$  si on observe des valeurs grandes de  $S_n$ , *i.e.*  $\mathcal{R}_{H_0} = \{S_n \geq s\}$ , où  $s$  est un seuil à déterminer.

Deux approches sont possibles

- On fixe un niveau  $\alpha$  (erreur maximale de première espèce) et on cherche le seuil (donc la zone de rejet) tel que  $\mathbb{P}_{H_0}(\text{Rejeter } H_0) \leq \alpha$ . Ce test pourra être appliqué à tout jeu de données observées ultérieurement, et l'hypothèse testée au niveau  $\alpha$ .

- On observe une réalisation  $x_1, \dots, x_n$  et on calcule le degré de significativité (ou  $p$ -value) correspondant à cette réalisation, *i.e.* le plus petit niveau  $\gamma$  tel qu'on rejette le test à ce niveau avec les valeurs observées. Ce test est spécifique à l'observation mais permet de répondre au test pour toutes les valeurs de  $\alpha$ , sur ce jeu de données.

**Exemple. (suite).** La zone de rejet de  $H_0$  est de la forme  $\mathcal{R}_{H_0} = \{S_n \geq s\}$ . Si on observe la valeur de la statistique  $s^{\text{obs}}$ , alors  $\gamma = \mathbb{P}_{H_0}(S_n \geq s^{\text{obs}})$  est le degré de significativité du test pour la valeur observée. Tout test de niveau  $\alpha < \gamma$  accepte  $H_0$  et tout test de niveau  $\alpha \geq \gamma$  rejette  $H_0$ .

Dans le cas d'un test bilatère, la zone de rejet peut souvent s'écrire de 2 façons différentes. Si  $T_n$  est la statistique de test, sous l'alternative  $H_1$  on observera généralement des valeurs plus grandes ou plus petites de  $T_n$  que sous  $H_0$ , *i.e.*,  $T_n$  est stochastiquement plus grande ou plus petite que sous  $H_0$ . La zone de rejet prend donc la forme  $\mathcal{R}_{H_0} = \{T_n \geq b\} \cup \{T_n \leq a\}$ , avec  $a \leq b$ , seuils à déterminer. En pratique, si on se fixe un niveau  $\alpha$  positif, alors on choisira  $a, b$  tels que  $\mathbb{P}_{H_0}(T_n \geq b) = \mathbb{P}_{H_0}(T_n \leq a) \leq \alpha/2$ .

Une autre formulation, lorsque  $T_n$  a une distribution symétrique par rapport à  $m_0$  sous l'hypothèse nulle  $H_0$ , consiste à choisir la région de rejet sous la forme  $\mathcal{R}_{H_0} = \{|T_n - m_0| \geq s\}$ . Ce qui revient exactement à la même chose.

Enfin, remarquons que le degré de significativité n'a pas de sens pour un test bilatère. Une fois que les données sont observées, le test rejette pour une des deux alternatives, jamais les deux en même temps !

**Exemple. (suite).** Supposons  $n = 10$  et  $S_n = 3$  alors la valeur observée 3 étant inférieure à la valeur moyenne sous  $H_0$  (*i.e.*  $n/2 = 5$ ), on définit la  $p$ -value pour le test unilatère de  $H_0 : \theta = 1/2$  contre  $H_1 : \theta < 1/2$  comme  $\gamma = \mathbb{P}_{H_0}(S_n \leq 3) = 0.171875 \simeq 17,2\%$ .

### Puissance de test.

- ★ La fonction puissance est difficile à évaluer pour un test non paramétrique car l'ensemble des alternatives est très grand et contient des distributions très différentes.
- ★ En particulier, il est difficile de comparer des tests de même niveau. On pourra plutôt en considérer plusieurs. Certains tests sont mieux adaptés à certaines alternatives que d'autres.
- ★ Par construction, ces tests ne dépendent pas de la distribution des v.a. et ont les mêmes qualités quelle que soit cette distribution. En ce sens, ils sont dits robustes.

**Correction du continu.** Supposons que la statistique de test  $T_n$  prend des valeurs discrètes, mais  $n$  étant grand, sa distribution est approché par une Gaussienne, *i.e.* on suppose

$$\frac{T_n - \mathbb{E}_{H_0}(T_n)}{\sqrt{\text{Var}_{H_0}(T_n)}} \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0; 1) \text{ sous } H_0.$$

On souhaite effectuer un test asymptotique en utilisant la loi limite de la statistique  $T_n$ . Supposons que la région de rejet prenne la forme  $\mathcal{R}_{H_0} = \{T_n \geq t\}$ . Alors, pour un niveau  $\alpha$  fixé, on souhaite déterminer le seuil  $t$  tel que  $\mathbb{P}_{H_0}(T_n \geq t) \leq \alpha$ . Or, pour toute valeur  $u \in [0, 1[$ , on note que comme  $T_n$  est une variable discrète, on a

$$\mathbb{P}_{H_0}(T_n \geq t) = \mathbb{P}_{H_0}(T_n \geq t - u).$$

De la même façon, si la région de rejet est de la forme  $T_n \leq t$  alors pour toute valeur  $u \in [0; 1[$  on a  $\mathbb{P}_{H_0}(T_n \leq t) = \mathbb{P}_{H_0}(T_n \leq t + u)$ .

La correction du continu consiste à remplacer la valeur par défaut  $u = 0$  par  $u = 1/2$ . Donc par exemple dans le cas où  $\mathcal{R}_{H_0} = \{T_n \geq t\}$ , on cherche le seuil  $t$  tel que

$$\mathbb{P}_{H_0}(T_n \geq t - 0.5) \leq \alpha \iff \mathbb{P}_{H_0} \left( \frac{T_n - \mathbb{E}_{H_0}(T_n)}{\sqrt{\text{Var}_{H_0}(T_n)}} \geq \frac{t - 0.5 - \mathbb{E}_{H_0}(T_n)}{\sqrt{\text{Var}_{H_0}(T_n)}} \right) \leq \alpha.$$

## 3.2 Tests sur une population

### 3.2.1 Test de signe

**Exemple.** On veut comparer les effets de deux traitements sur 2 populations d'individus que l'on peut appairer. Soient  $U_1, \dots, U_n$  mesures sur la première population et  $V_1, \dots, V_n$  mesures sur la seconde population. Les échantillons sont nécessairement de même taille. Après appariement, on peut considérer  $X_i = U_i - V_i, 1 \leq i \leq n$ . Le test de  $H'_0$  : « pas de différence entre les traitements » *i.e.* «  $U$  et  $V$  ont la même distribution » peut se faire à travers le test de  $H_0$  : « La médiane de  $X$  est nulle ». Plus précisément, on a  $H'_0 \Rightarrow H_0$  donc si on rejette  $H_0$ , on rejette également  $H'_0$ . Une alternative  $H_1$  de type unilatère traduit que l'un des traitements est meilleur que l'autre.

**Remarque. Sur l'appariement des variables :** soit il s'agit des mêmes individus, sur lesquels on applique des traitements à 2 temps différents, soit les individus sont différents et alors pour que l'appariement soit valable, il faut avoir collecté puis regroupé les individus en fonction de covariables (sexe, âge, etc).

Dans la suite, on observe un seul échantillon  $X_1, \dots, X_n$  de v.a. réelles i.i.d de même loi que  $X$  (éventuellement obtenu à partir de 2 échantillons appariés).

On veut tester  $H_0 : \mathbb{P}(X \leq 0) = 1/2$  i.e. « la médiane de la distribution est nulle » contre  $H_1 : \mathbb{P}(X \leq 0) > 1/2$  i.e. « la médiane de la distribution est négative » ou  $H'_1 : \mathbb{P}(X \leq 0) < 1/2$  i.e. « la médiane de la distribution est positive ».

À chaque variable  $X_i$  on associe  $Y_i = 1\{X_i \leq 0\}$  (on peut aussi considérer  $Y_i = 1\{X_i \geq 0\}$  si on veut. Ça ne change rien, il faut juste tout inverser).

**Proposition 7.** *Sous l'hypothèse  $H_0 : \mathbb{P}(X_i \leq 0) = 1/2$ , on a  $Y_i \sim \mathcal{B}(1/2)$  et  $S_n = \sum_{i=1}^n Y_i \sim \mathcal{B}(n, 1/2)$ . De plus, sous  $H_1 : \mathbb{P}(X \leq 0) > 1/2$ , la statistique  $S_n$  est stochastiquement plus grande que sous  $H_0$ .*

*Démonstration.* La première partie est évidente. La seconde provient des propriétés énoncées dans la section précédente.  $\square$

Cette proposition permet de faire le test de  $H_0 : \mathbb{P}(X_i \leq 0) = 1/2$  contre  $H_1 : \mathbb{P}(X \leq 0) > 1/2$ . On rejette l'hypothèse nulle si  $S_n \geq s$  où  $s$  est un seuil à déterminer. Par exemple, si on se fixe un niveau de test  $\alpha$ , alors  $s$  doit satisfaire

$$\mathbb{P}_{H_0}(S_n \geq s) = \sum_{k=0}^n C_n^k \frac{1}{2^n} \leq \alpha.$$

**Remarques.**  $\star$  Dans la pratique, on ne fait pas le calcul de  $\sum_{k=0}^n C_n^k$ . Pour les petites valeurs de  $n$ , la distribution  $\mathcal{B}(n; 1/2)$  est tabulée. Pour les grandes valeurs de  $n$ , on a recours à une approximation Gaussienne.

$\star$  Ce test est très général, mais il utilise très peu d'information sur les variables (uniquement leur signe, pas leurs valeurs relatives). C'est donc un test peu puissant.

### 3.2.2 Statistiques d'ordre et de rang

**Définition.** Soient  $X_1, \dots, X_n$  v.a. réelles.

i) La statistique d'ordre  $X^* = (X_{(1)}, \dots, X_{(n)})$  est obtenue par réarrangement croissant des  $X_i$ . Ainsi :  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  et  $\forall a \in \mathbb{R}, |\{i; X_i = a\}| = |\{i; X_{(i)} = a\}|$ .

ii) Le vecteur des rangs  $R_X$  est une permutation de  $\{1, \dots, n\}$  telle que  $\forall i \in \{1, \dots, n\}$ , on a  $X_i = X_{R_X(i)}^* = X_{(R_X(i))}$ .

**Exemple.** On considère  $n = 7, x = (4, 2, 1, 1, 2, 0, 1)$ . Alors  $x^* = (0, 1, 1, 1, 2, 2, 4)$  et par exemple

$X$	4	2	1	1	2	0	1
$R_X$	7	5	2	3	6	1	4

Ici on a  $x_1 = 4 = x_{(7)}$  et  $R_x(1) = 7$ .

**Remarques.**  $\star$  Cette notion est dépendante de  $n$  qui doit être fixé.

$\star$  S'il y a des ex-æquos, le vecteur des rangs n'est pas unique.

$\star$  Si toutes les observations sont distinctes, on note  $\omega_X$  la permutation inverse de  $R_X$ , i.e.  $X_i = X_{(R_X(i))} \iff X_{\omega_X(j)} = X_{(j)}$  où  $j = R_X(i)$ .

**Théorème 8.** Soient  $X_1, \dots, X_n$  v.a. réelles, i.i.d. de statistique d'ordre  $X^*$  et vecteur des rangs  $R_X$ . Alors on a  $X^* \perp R_X$  et de plus  $R_X$  est distribué uniformément sur l'ensemble  $\mathcal{S}_n$  des permutations de  $\{1, \dots, n\}$ .

*Démonstration.* On fait la preuve uniquement dans le cas où les  $X_i$  sont des variables diffuses, en particulier  $\mathbb{P}(X_i = X_j) = 0$  pour  $i \neq j$  et  $R_X$  est unique.

Conditionnellement à  $X^* = y$ , les  $n!$  valeurs possibles de  $(X_1, \dots, X_n)$  ont toutes la même probabilité d'occurrence car  $\prod_{j=1}^n \mathbb{P}(y_{\omega(j)})$  ne dépend pas de la permutation  $\omega$ . Autrement dit, conditionnellement à  $X^* = y$ , la v.a.  $(X_1, \dots, X_n)$  est distribuée uniformément sur l'ensemble  $\{(y_{\sigma(i)})_{1 \leq i \leq n}, \sigma \in \mathcal{S}_n\}$  qui contient  $n!$  éléments. Ainsi,  $\forall r \in \mathcal{S}_n$ , on a  $\mathbb{P}(R_X = r | X^* = y) = 1/(n!)$  et est indépendant de  $y$ . Donc  $R_X \perp X^*$  et  $\forall r \in \mathcal{S}_n$ , on a  $\mathbb{P}(R_X = r) = \sum_y \mathbb{P}(X^* = y) \mathbb{P}(R_X = r | X^* = y) = 1/(n!)$ .  $\square$

### 3.2.3 Test de signe et rang (ou Wilcoxon signed rank test)

**Exemple.** Comme précédemment, si on a 2 populations que l'on peut appairer, et on observe  $X_i = U_i - V_i$  des valeurs symétriques par rapport à  $m$  (qui est nécessairement la médiane), alors le test de  $H'_0$  : «  $U$  et  $V$  ont même distribution » peut se faire à travers le test de  $H_0$  : « La loi de  $X$  est symétrique par rapport à 0 ». Plus précisément, on a  $H'_0 \Rightarrow H_0$  donc si on rejette  $H_0$ , on rejette également  $H'_0$ .

Soit  $X_1, \dots, X_n$  un échantillon de v.a. réelles de même loi que  $X$  supposée **diffuse et symétrique par rapport à (la médiane)  $m$** . On veut tester  $m = 0$  contre  $m \neq 0$ , mais cette fois, on veut utiliser la valeur (relative) des  $X_i$  et pas seulement leur signe.

**Principe.** Comme pour le test de signe, on veut compter le nombre de  $X_i > 0$  (ou ce qui revient au même, le nombre de  $X_i < 0$ ), mais on leur attribue un poids d'autant plus grand que la valeur prise par  $|X_i|$  est grande.

On considère donc la statistique d'ordre associée aux  $\{|X_i|\}_{1 \leq i \leq n}$  : on a donc  $|X|_{(1)} \leq |X|_{(2)} \leq \dots \leq |X|_{(n)}$  et soit  $R_{|X|}$  le vecteur des rangs associé.

On définit la statistique de test

$$W_n^+ = \sum_{i=1}^n R_{|X|}(i) 1\{X_i > 0\}.$$

On pourrait aussi considérer de façon identique la statistique  $W_n^- = \sum_{i=1}^n R_{|X|}(i)1\{X_i < 0\}$ .

**Exemple.** On a  $n = 5$  et on observe  $\{-0.15; -0.42; 0.22; 0.6; -0.1\}$ . Alors,

$X_i$	-0.15	-0.42	0.22	0.6	-0.1
$ X_i $	0.15	0.42	0.22	0.6	0.1
$R_{ X }(i)$	2	4	3	5	1
$X_i > 0$	-	-	+	+	-

et  $W_5^+ = 3 + 5 = 8$  tandis que  $W_5^- = 2 + 4 + 1 = 7$ .

**Remarques.**  $\star$  On a  $W_n^+ + W_n^- = n(n+1)/2$  presque sûrement. En effet, comme  $X$  est diffuse, on a  $X_i \neq 0$  p.s. et donc  $W_n^+ + W_n^- = \sum_{i=1}^n R_{|X|}(i) = \sum_{i=1}^n i = n(n+1)/2$ .

$\star$  On a également  $0 \leq W_n^+ \leq n(n+1)/2$ , presque sûrement. Le cas  $W_n^+ = 0$  correspond à tous les  $X_i < 0$  et le cas  $W_n^+ = n(n+1)/2$  à tous les  $X_i > 0$ .

**Théorème 9.** *Sous l'hypothèse  $H_0$  : « La loi de  $X$  est symétrique par rapport à 0 »,  $W_n^+$  et  $W_n^-$  ont la même distribution et sont des statistiques libres en loi de la loi de  $X$ . De plus, on a*

$$\mathbb{E}_{H_0}(W_n^+) = \frac{n(n+1)}{4}, \quad \text{Var}_{H_0}(W_n^+) = \frac{n(n+1)(2n+1)}{24}.$$

Enfin, asymptotiquement,

$$\frac{W_n^+ - \mathbb{E}_{H_0}(W_n^+)}{\sqrt{\text{Var}_{H_0}(W_n^+)}} \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, 1) \text{ sous } H_0.$$

*Démonstration.* Notons d'abord que sous l'hypothèse  $H_0$ , on a en particulier  $\mathbb{P}_{H_0}(1\{X_i > 0\} = 1) = 1/2$  mais aussi  $\mathbb{P}_{H_0}(1\{X_{\sigma(i)} > 0\} = 1) = 1/2$  pour toute permutation  $\sigma \in \mathcal{S}_n$  et si  $i \neq j$  alors  $X_{\sigma(i)} \amalg X_{\sigma(j)}$ .

On note à présent  $\sigma \in \mathcal{S}_n$  la permutation telle que

$$|X_{\sigma(1)}| \leq |X_{\sigma(2)}| \leq \dots \leq |X_{\sigma(n)}|.$$

Alors on a  $W_n^+ = \sum_{j=1}^n j1\{X_{\sigma(j)} > 0\}$ . Donc d'après ce qui précède (la suite  $\{1\{X_{\sigma(j)} > 0\}\}_{1 \leq j \leq n}$  est une suite iid de variables de Bernoulli de paramètre  $1/2$  sous  $H_0$ ), la statistique  $W_n^+$  est, sous  $H_0$ , libre en loi de la loi de  $X$ . La même démarche s'applique pour montrer que  $W_n^+ = \sum_{j=1}^n j1\{X_{\sigma(j)} < 0\} \sim W_n^+$  sous  $H_0$ .

De plus, on en déduit également  $\mathbb{E}_{H_0}(W_n^+) = \sum_{j=1}^n j \times 1/2 = n(n+1)/4$  et comme  $W_n^+$  est une somme de variables indépendantes (non i.d.) on a

$$\text{Var}_{H_0}[(W_n^+)^2] = \sum_{j=1}^n j^2 \text{Var}_{H_0}(1\{X_{\sigma(j)} > 0\}) = \sum_{j=1}^n \frac{j^2}{4} = \frac{1}{4} \times \frac{n(n+1)(2n+1)}{6}.$$

La normalité asymptotique est **admise** (TCL pour variables indépendantes mais non i.d.).  $\square$

**Test exact ou asymptotique.** Sous  $H_0$ , la statistique  $W_n^+$  (ou  $W_n^-$ ) est distribuée sur  $\{0, 1, 2, \dots, n(n+1)/2\}$  et de distribution symétrique par rapport à sa moyenne  $n(n+1)/4$ . Pour chaque valeur de  $n$ , on peut construire une table de la distribution de  $W_n^+$  (en énumérant toutes les configurations possibles). Dans la pratique, on utilise une table statistique pour les petites valeurs de  $n$  ( $\leq 20$ ). Pour  $n > 20$ , on utilise l'approximation Gaussienne et un test asymptotique.

### 3.3 Tests sur deux populations : Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney)

On se donne deux échantillons  $X_1, \dots, X_{n_1}$  et  $Y_1, \dots, Y_{n_2}$  indépendants de variables aléatoires réelles i.i.d et diffuses. On note  $F$  la f.d.r. de  $X_1, \dots, X_{n_1}$  et  $G$  celle de  $Y_1, \dots, Y_{n_2}$ . On veut tester  $H_0 : F = G$  contre  $H_1 : F \neq G$ .

**Remarque.** Les échantillons n'ont a priori pas même taille et il n'existe pas d'appariement naturel entre les variables.

**Exemple.** On a une population de  $N$  individus sur lesquels on veut tester un nouveau traitement. On forme un groupe de  $n_1$  individus qui reçoivent le nouveau traitement et  $n_2 = N - n_1$  forment le groupe « contrôle ». On mesure une quantité relative au traitement. On veut savoir si le nouveau traitement est efficace ( $F \neq G$ ). L'hypothèse nulle est privilégiée : si on la rejette, le nouveau traitement est déclaré efficace. On ne veut pas d'un nouveau médicament si on n'est pas sûr qu'il a un effet supérieur.

On classe les variables  $\{X_i, Y_j\}$  par leur rang *global* (i.e. on considère le vecteur des rangs  $R_{X,Y}$ ) et on note  $R_1, R_2, \dots, R_{n_1}$  les rangs associés au premier échantillon (i.e. les  $X_i$ ).

**Exemple.** Soient  $X_1 = 3.5; X_2 = 4.7; X_3 = 1.2; Y_1 = 0.7; Y_2 = 3.9$  alors  $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$  et les rangs associés à l'échantillon des  $X_i$  sont  $R_1 = 3, R_2 = 5, R_3 = 2$ .

Le principe du test est le suivant : Si les rangs dans chaque échantillon sont significativement différents, on conclura  $F \neq G$ .

**Remarque.** Suivant le contexte,  $X$  et  $Y$  peuvent mesurer des choses très différentes. Par contre, le rang relatif de ces variables est une quantité qui ne dépend pas de la nature (de la loi) des variables de départ.

On note alors

$$\Sigma_1 = R_1 + \dots + R_{n_1}$$

la somme des rangs du premier échantillon. On peut remarquer que

$$\frac{n_1(n_1 + 1)}{2} \leq \Sigma_1 \leq \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - \frac{n_2(n_2 + 1)}{2} = n_1n_2 + \frac{n_1(n_1 + 1)}{2},$$

presque sûrement (cas où les rangs sont tous les plus petits ou bien tous les plus grands). Il est donc naturel de considérer plutôt la variable

$$W_{YX} = \Sigma_1 - \frac{n_1(n_1 + 1)}{2},$$

qui varie presque sûrement entre 0 et  $n_1n_2$ . On définit de façon symétrique  $W_{XY} = \Sigma_2 - n_2(n_2 + 1)/2$  où  $\Sigma_2$  est la somme des rangs du second échantillon.

**Proposition 10.** *Sous l'hypothèse que les variables sont diffuses, on a les résultats suivants :*

- i)  $W_{XY}$  est égal au nombre de paires  $(X_i, Y_j)$  (parmi les  $n_1n_2$  paires possibles) telles que  $X_i < Y_j$ .
- ii)  $W_{XY} + W_{YX} = n_1n_2$ , p.s..
- iii) Sous l'hypothèse  $H_0 : F = G$ , la loi de  $\Sigma_1$  est symétrique par rapport à  $n_1(N+1)/2$ . Autrement dit, sous  $H_0$ , la loi de  $W_{YX}$  est symétrique par rapport à  $n_1n_2/2$ .
- iv) Sous l'hypothèse  $H_0 : F = G$ , les variables  $W_{XY}$  et  $W_{YX}$  ont la même loi.

*Démonstration.* Supposons pour commencer que les variables  $Y_i$  sont strictement ordonnées, i.e.  $Y_1 < Y_2 < \dots < Y_{n_2}$ . Notons  $S_1, S_2, \dots, S_{n_2}$  les rangs associés au second échantillon :  $S_j$  est le rang de  $Y_j$  parmi les variables  $\{X_i, Y_k\}$ . Comme  $Y_1$  est la plus petite variable parmi les  $Y_k$  et que  $S_1$  est son rang dans l'échantillon global, il y a exactement  $S_1 - 1$  variables parmi les  $X_k$  qui sont strictement plus petites que  $Y_1$ . De même pour  $Y_2$ , son rang étant  $S_2$ , il y a  $S_2 - 1$  variables de l'échantillon total qui sont strictement plus petites, et on a supposé  $Y_1 \leq Y_2 \leq Y_k$ , pour tout  $k \geq 3$ . Donc il y a exactement  $S_2 - 2$  variables  $X_k$  qui sont strictement plus petites que  $Y_2$ .

Plus généralement, on peut voir qu'il y a exactement  $S_j - j$  variables  $X_k$  qui sont strictement inférieures à  $Y_j$ . Le nombre de paires  $(X_i, Y_j)$  telles que  $X_i < Y_j$  et donc égal à

$$S_1 - 1 + S_2 - 2 + \cdots + S_{n_2} - n_2 = \Sigma_2 - (1 + \cdots + n_2) = \Sigma_2 - \frac{n_2(n_2 + 1)}{2} = W_{XY}.$$

Pour conclure, il suffit de remarquer que le résultat ne dépend pas de la numérotation des variables  $Y_j$  de départ, et que si les variables sont diffuses, alors la probabilité que deux variables soient égales est nulle.

L'assertion *ii*) est alors une conséquence immédiate de *i*). Pour montrer *iii*), on utilise le lemme suivant

**Lemme 11.** *Soient  $Z_1, \dots, Z_N$  une suite de v.a.r. i.i.d. Pour tout  $s \leq N$  et toute suite d'entiers distincts  $r_1, \dots, r_s$  dans  $\{1, \dots, N\}$ , on a*

$$\mathbb{P}((R_1, \dots, R_s) = (r_1, \dots, r_s)) = \frac{1}{N(N-1) \cdots (N-s+1)}.$$

En effet, toutes les suites de rangs possibles ont la même probabilité. (Noter en particulier que  $\mathbb{P}((R_1, \dots, R_n) = (r_1, \dots, r_n)) = 1/n!$ ).

On introduit  $S'_i = N - S_i + 1$ , i.e. le rang des variables  $Y_i$  dans l'échantillon global, lorsque les variables sont ordonnées de façon décroissante. Sous  $H_0$ , en utilisant le même argument que pour le lemme précédent, on a

$$\mathbb{P}_{H_0}((S'_1, \dots, S'_{n_2}) = (s_1, \dots, s_{n_2})) = \frac{1}{N(N-1) \cdots (N-n_2+1)},$$

et donc  $\Sigma'_2 = S'_1 + \cdots + S'_{n_2}$  a la même loi que  $\Sigma_2$  sous  $H_0$ . De plus, on a

$$\Sigma'_2 = N - S_1 + 1 + N - S_2 + 1 + \cdots + N - S_{n_2} + 1 = n_2(N + 1) - \Sigma_2$$

ce qui implique

$$\Sigma'_2 - \frac{n_2(N + 1)}{2} = \frac{n_2(N + 1)}{2} - \Sigma_2$$

i.e., pour tout  $k$ ,

$$\mathbb{P}(\Sigma'_2 - \frac{n_2(N + 1)}{2} = k) = \mathbb{P}(\frac{n_2(N + 1)}{2} - \Sigma_2 = k)$$

et comme  $\Sigma_2$  et  $\Sigma'_2$  ont même loi sous  $H_0$ ,

$$\mathbb{P}_{H_0}(\Sigma_2 - \frac{n_2(N + 1)}{2} = k) = \mathbb{P}_{H_0}(\Sigma_2 - \frac{n_2(N + 1)}{2} = -k),$$

ce qui achève la preuve de la symétrie de  $\Sigma_2$  par rapport à  $n_2(N+1)/2$ . La preuve de *iv*) découle de

$$\begin{aligned}\Sigma_1 + \Sigma_2 &= N(N+1)/2 = (n_1 + n_2)(N+1)/2 \\ W_{YX} = \Sigma_1 - n_1(N+1)/2 &= -\Sigma_2 + n_2(N+1)/2 \stackrel{\text{loi}}{=} \Sigma_2 - n_2(N+1)/2 = W_{XY}.\end{aligned}$$

□

On peut montrer plus généralement le théorème suivant (admis).

**Théorème 12.** *La loi de  $W_{XY}$  (ou  $W_{YX}$ ) est libre sous  $H_0$  (i.e. elle ne dépend pas de la f.d.r.  $F = G$ ). Cette loi ne dépend que de  $n_1$  et  $n_2$ . De plus, asymptotiquement, on a*

$$\frac{W_{XY} - \mathbb{E}_{H_0}(W_{XY})}{\sqrt{\text{Var}_{H_0}(W_{XY})}} \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, 1) \text{ sous } H_0.$$

**Test exact ou test asymptotique.** Cette loi est tabulée pour les petites valeurs de  $n_1$  et  $n_2$  ( $\leq 10$ ). Pour les grandes valeurs, on utilise une approximation Gaussienne. Pour cela, il nous faut calculer l'espérance et la variance de  $W_{XY}$  ou de façon équivalente de  $\Sigma_1$ , sous  $H_0$ .

**Calcul de l'espérance de  $\Sigma_1$  sous  $H_0$ .**

**Proposition 13.** *On a  $\mathbb{E}_{H_0}(\Sigma_1) = \frac{n_1(N+1)}{2}$ .*

*Démonstration.* On a  $\Sigma_1 + \Sigma_2 = N(N+1)/2$ , p.s. et

$$\mathbb{E}_{H_0}(W_{YX}) = \mathbb{E}_{H_0}(\Sigma_1 - n_1(n_1+1)/2) = \mathbb{E}_{H_0}(W_{XY}) = \mathbb{E}_{H_0}(\Sigma_2 - n_2(n_2+1)/2).$$

On en déduit

$$\mathbb{E}_{H_0}(\Sigma_1) + \mathbb{E}_{H_0}(\Sigma_2) = \frac{N(N+1)}{2} \tag{3.1}$$

$$\mathbb{E}_{H_0}(\Sigma_2) = \mathbb{E}_{H_0}(\Sigma_1) - \frac{n_1(n_1+1)}{2} + \frac{n_2(n_2+1)}{2}. \tag{3.2}$$

En combinant ces deux expressions, on obtient

$$2\mathbb{E}_{H_0}(\Sigma_1) - \frac{n_1(n_1+1)}{2} + \frac{n_2(n_2+1)}{2} = \frac{N(N+1)}{2},$$

ce qui donne après calculs (en utilisant  $N = n_1 + n_2$ ), le résultat  $\mathbb{E}_{H_0}(\Sigma_1) = \frac{n_1(N+1)}{2}$ . □

**Proposition 14.** *On a  $\text{Var}_{H_0}(\Sigma_1) = \frac{n_1 n_2 (N+1)}{12}$ .*

*Démonstration.* On écrit

$$\mathbb{V}\text{ar}_{H_0}(\Sigma_1) = \mathbb{V}\text{ar}_{H_0}(R_1 + R_2 + \dots + R_{n_1}) = \sum_{i=1}^{n_1} \mathbb{V}\text{ar}_{H_0}(R_i) + \sum_{\substack{1 \leq i, j \leq n_1 \\ i \neq j}} \mathbb{C}\text{ov}_{H_0}(R_i, R_j).$$

Or on a

$$\mathbb{V}\text{ar}_{H_0}(R_i) = \mathbb{E}_{H_0}[(R_i - \mathbb{E}_{H_0}(R_i))^2] = \sum_{k=1}^n (k - \mathbb{E}_{H_0}(R_i))^2 \frac{1}{N},$$

car sous  $H_0$ , les rangs  $R_i$  ont une distribution uniforme sur  $\{1, \dots, N\}$ , *i.e.*  $\mathbb{P}_{H_0}(R_i = k) = 1/N, \forall k \in \{1, \dots, N\}$ . De plus,

$$\mathbb{E}_{H_0}(R_i) = \sum_{k=1}^N \frac{k}{N} = \frac{1}{N} \times \frac{N(N+1)}{2} = \frac{N+1}{2},$$

donc on obtient

$$\mathbb{V}\text{ar}_{H_0}(R_i) = \sum_{k=1}^n \left(k - \frac{N+1}{2}\right)^2 \times \frac{1}{N}.$$

Enfin, sous  $H_0$ , pour tout  $i \neq j$  et  $k \neq l$ , on a  $\mathbb{P}_{H_0}(R_i = k, R_j = l) = 1/(N(N-1))$ . Cela implique

$$\begin{aligned} \mathbb{C}\text{ov}_{H_0}(R_i, R_j) &= \mathbb{E}_{H_0}[(R_i - \mathbb{E}_{H_0}(R_i))(R_j - \mathbb{E}_{H_0}(R_j))] \\ &= \sum_{\substack{1 \leq k, l \leq N \\ k \neq l}} \frac{1}{N(N-1)} \times \left(k - \frac{N+1}{2}\right) \left(l - \frac{N+1}{2}\right). \end{aligned}$$

On obtient donc

$$\begin{aligned} \mathbb{V}\text{ar}_{H_0}(\Sigma_1) &= n_1 \mathbb{V}\text{ar}_{H_0}(R_1) + n_1(n_1 - 1) \mathbb{C}\text{ov}_{H_0}(R_1, R_2) \\ &= \frac{n_1}{N} \sum_{k=1}^n \left(k - \frac{N+1}{2}\right)^2 + \frac{n_1(n_1 - 1)}{N(N-1)} \sum_{\substack{1 \leq k, l \leq N \\ k \neq l}} \left(k - \frac{N+1}{2}\right) \left(l - \frac{N+1}{2}\right). \end{aligned}$$

Or,

$$\left(\sum_{k=1}^N \left(k - \frac{N+1}{2}\right)\right)^2 = \sum_{k=1}^N \left(k - \frac{N+1}{2}\right)^2 + \sum_{\substack{1 \leq k, l \leq N \\ k \neq l}} \left(k - \frac{N+1}{2}\right) \left(l - \frac{N+1}{2}\right).$$

Le membre de gauche de l'équation étant nul, on récupère

$$\sum_{\substack{1 \leq k, l \leq N \\ k \neq l}} \left(k - \frac{N+1}{2}\right) \left(l - \frac{N+1}{2}\right) = - \sum_{k=1}^N \left(k - \frac{N+1}{2}\right)^2,$$

et ainsi

$$\begin{aligned}
\mathbb{V}\text{ar}_{H_0}(\Sigma_1) &= \frac{n_1}{N} \left(1 - \frac{n_1 - 1}{N - 1}\right) \sum_{k=1}^N \left(k - \frac{N + 1}{2}\right)^2 \\
&= \frac{n_1}{N} \frac{n_2}{N - 1} \left[ \sum_{k=1}^N k^2 - (N + 1) \sum_{k=1}^N k + \frac{N(N + 1)^2}{4} \right] \\
&= \dots = \frac{N(N - 1)(N + 1)}{12}.
\end{aligned}$$

□

**Remarques.**   ★ Ce test est très général et n'utilise que les valeurs relatives des variables entre elles.

- ★ Le test d'adéquation de KS pour 2 échantillons est assez différent car il prend en compte la *forme* des distributions et pas seulement des phénomènes de *translation*.
- ★ La statistique de test signe et rang de Wilcoxon peut être vue comme un cas particulier de la statistique de la somme des rangs de Wilcoxon. En effet, si  $Z_1, \dots, Z_N$  est un échantillon, on considère un premier sous-échantillon  $U_1, \dots, U_{n_1}$  correspondant aux valeurs de  $Z_i$  telles que  $Z_i > 0$ , et un second sous-échantillon  $V_1, \dots, V_{n_2}$  correspondant aux valeurs  $-Z_i$  pour les  $Z_i < 0$ . Ordonner les  $\{U_i, V_j\}$  revient à ordonner  $\{|Z_i|\}$  et la somme des rangs de l'échantillon des  $U_i$  est donc égale à la somme des rangs des  $Z_i > 0$ . Sous  $H_0$ , chacun des deux échantillons devrait être de taille environ  $N/2$ , mais il faut tenir compte de l'aléa dans la répartition des signes pour pouvoir faire un parallèle exact.

# Chapitre 4

## Estimation non paramétrique de la densité d'un échantillon

### 4.1 Introduction

Dans tout le chapitre,  $X_1, \dots, X_n$  sont des variables aléatoires i.i.d. de fdr  $F$  et admettant une densité  $f = F'$ . Le but est d'estimer (à partir des observations) la densité  $f$  en faisant le moins d'hypothèses possibles sur cette densité. Typiquement, on supposera que  $f \in \mathcal{F}$  espace fonctionnel et on notera  $\hat{f}_n$  un estimateur.

**Mesure de la qualité d'un estimateur.** Pour construire de « bons » estimateurs, il faut se demander quelles qualités on en attend. Plusieurs choses sont nécessaires :

1) Définition d'une distance sur  $\mathcal{F}$ .

**Exemples.**  $\star d(f, g) = \|f - g\|_p = [\int |f - g|^p]^{1/p}$ , pour  $p \geq 1$ . Par exemple  $p = 1$  ou  $2$ .

$\star d(f, g) = \|f - g\|_\infty = \sup_{x \in \mathbb{R}} |f(x) - g(x)|$ .

$\star d(f, g) = |f(x_0) - g(x_0)|$  où  $x_0$  fixé.

2) Définition d'une fonction de perte  $\omega : \mathbb{R} \mapsto \mathbb{R}^+$  convexe, telle que  $\omega(0) = 0$ .

**Exemple.**  $\omega : u \mapsto u^2$  fonction de perte quadratique.

3) Définition d'une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}_f(\omega(d(\hat{f}_n, f))),$$

où  $\mathbb{E}_f$  désigne l'espérance quand la densité des  $X_i$  vaut  $f$ .

- Exemples.**   ★ En prenant la distance  $\mathbb{L}_2$  et la perte quadratique, on obtient le risque quadratique intégré (MISE = mean integrated squared error)  $R(\hat{f}_n, f) = \mathbb{E}_f \|\hat{f}_n - f\|_2^2$ .
- ★ En prenant la distance ponctuelle en  $x_0$  et la perte quadratique, on obtient le risque quadratique ponctuel (MSE = mean squared error)  $R(\hat{f}_n, f) = \mathbb{E}_f |\hat{f}_n(x_0) - f(x_0)|^2$ .

4) Comme  $f$  est inconnue, le risque  $R(\hat{f}_n, f)$  n'est pas calculable. Plusieurs alternatives sont possibles

- ★ Avoir recours à une méthode de validation croisée pour estimer ce risque. Nous en reparlerons plus tard.
- ★ S'intéresser au risque maximal sur une classe de fonctions  $\mathcal{F}$ . On introduit alors

$$R(\hat{f}_n, \mathcal{F}) = \sup_{f \in \mathcal{F}} R(\hat{f}_n, f).$$

C'est un point de vue pessimiste, puisqu'en général les observations n'ont pas été générées sous le « pire des cas ». C'est cependant l'approche que nous suivrons principalement dans ce cours.

- ★ Avoir une approche de type minimax. Mais on en parle pas ici.

On cherche donc à construire des estimateurs tels que

$$\sup_{f \in \mathcal{F}} R(\hat{f}_n, f) \xrightarrow{n \rightarrow \infty} 0,$$

et on veut exhiber la **vitesse de convergence** de  $\hat{f}_n$  pour le risque  $R$ , *i.e.* la plus petite suite  $(\phi_n)_{n \geq 0} \rightarrow 0$  telle que  $\{\phi_n^{-1} R(\hat{f}_n, \mathcal{F})\}$  soit bornée, *i.e.* telle qu'il existe  $C > 0$  avec  $\forall n \in \mathbb{N}, \forall f \in \mathcal{F}$ , on a  $R(\hat{f}_n, f) \leq C\phi_n$ . On dit alors que  $\hat{f}_n$  atteint la vitesse de convergence  $\phi_n$  sur la classe  $\mathcal{F}$  pour la distance  $d$  et la perte  $\omega$ .

Typiquement, les classes de fonctions  $\mathcal{F}$  sont des classes de fonctions **régulières**. Comme par exemple : Lipschitz, classe  $\mathcal{C}^k$ , fonction monotones, etc.

## 4.2 Estimateurs à noyaux

**Principe.** On a, pour  $h$  assez petit,

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

La fdr  $F$  est inconnue, mais on peut la remplacer par son estimateur  $\hat{F}_n$  (fdr empirique), ce qui nous donne un estimateur de la densité  $f$ , via

$$\begin{aligned} \hat{f}_n(x) &= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{n} \sum_{i=1}^n \frac{1}{2h} 1\{X_i \in ]x-h; x+h]\} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K_0\left(\frac{X_i - x}{h}\right), \end{aligned}$$

où  $K_0(u) = 1_{]-1;1]}(u)/2$  est le noyau de Rosenblatt (1956). Voir Figure 4.1 pour une illustration.

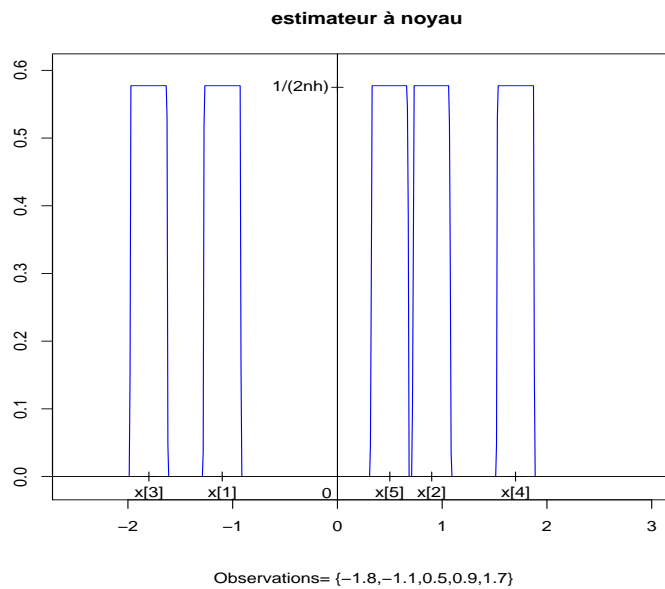


FIG. 4.1 – Estimateur à noyau (rectangulaire).

Parzen (1962), propose de remplacer  $K_0$  par un *noyau* plus général.

**Définition.** i) Soit  $K : \mathbb{R} \rightarrow \mathbb{R}$  intégrable telle que  $\int K(u)du = 1$ . Alors  $K$  est appelé noyau.

ii) Pour tout  $h > 0$  petit (en fait  $h = h_n \rightarrow_{n \rightarrow \infty} 0$ ), on peut définir

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

estimateur à noyau de  $f$ . On a  $\int \hat{f}_n(x)dx = 1$  et si  $K > 0$  alors  $\hat{f}_n$  est une densité.

iii) Le paramètre  $h > 0$  est appelé *fenêtre*. C'est un paramètre de lissage : plus  $h$  est grand, plus l'estimateur est régulier (voir Figure 4.2).

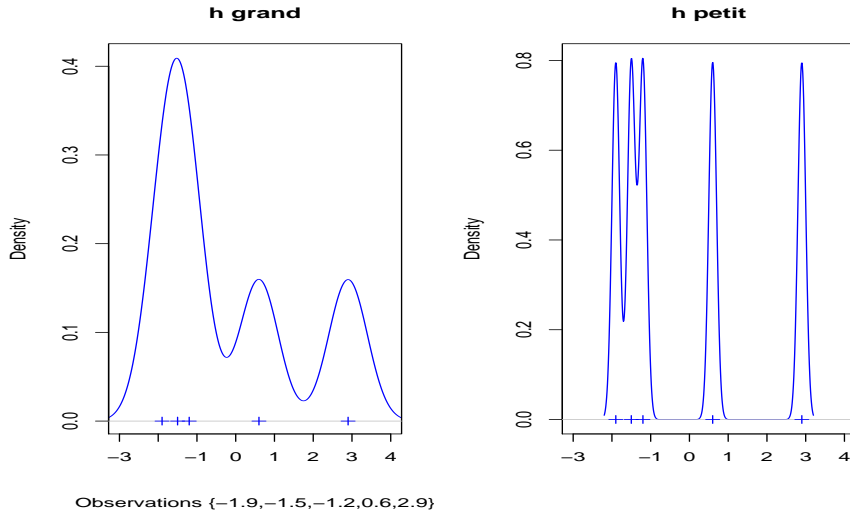


FIG. 4.2 – Effet de la variation de  $h$  sur l'estimateur à noyau.

**Remarque.** On considérera souvent des noyaux positifs et pairs, mais ce n'est pas obligatoire.

**Exemples de noyaux.** Les noyaux suivants sont représentés à la Figure 4.3.

- ★ (Rosenblatt, ou noyau rectangulaire)  $K(u) = 1_{[-1;1]}(u)/2$ .
- ★ (noyau triangle)  $K(u) = (1 - |u|)1_{[-1;1]}(u)$ .
- ★ (Epanechnikov)  $K(u) = \frac{3}{4}(1 - u^2)1_{[-1;1]}(u)$ .
- ★ (Biweight)  $K(u) = \frac{15}{16}(1 - u^2)^2 1_{[-1;1]}(u)$ .
- ★ (Gaussien)  $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$ .
- ★ (Cosine)  $K(u) = \frac{\pi}{4} \cos(u\pi/2)1_{[-1;1]}(u)$ .

### 4.3 Risque quadratique ponctuel des estimateurs à noyau

Dans cette section, on s'intéresse au risque quadratique ponctuel de  $\hat{f}_n$ , *i.e.*  $R(\hat{f}_n(x), f(x)) = \mathbb{E}_f(\hat{f}_n(x) - f(x))^2$  en tout point  $x \in \mathbb{R}$ .

On rappelle la décomposition « biais-variance » :

$$\begin{aligned} R(\hat{f}_n(x), f(x)) &= \mathbb{E}_f|\hat{f}_n(x) - f(x)|^2 \\ &= |\mathbb{E}_f(\hat{f}_n(x)) - f(x)|^2 + \mathbb{E}_f|\hat{f}_n(x) - \mathbb{E}_f\hat{f}_n(x)|^2 = \text{biais}^2 + \text{Var}(\hat{f}_n(x)). \end{aligned}$$

L'étude du risque quadratique de l'estimateur se ramène donc à l'étude de son biais et de sa variance.

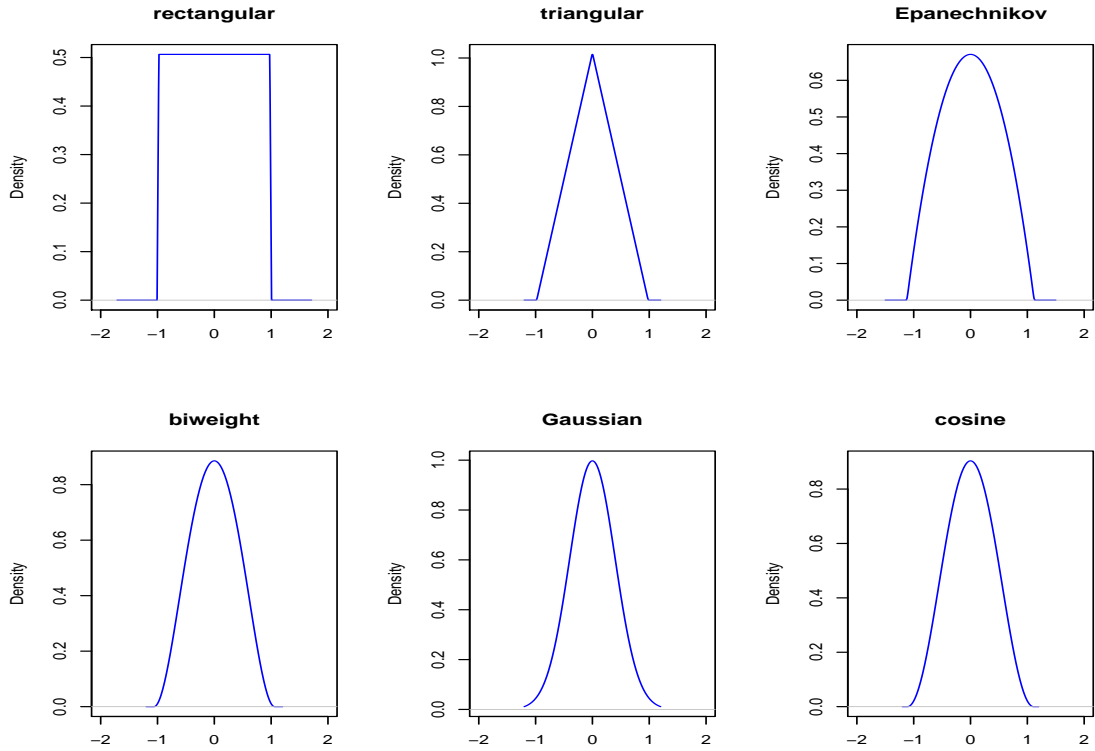


FIG. 4.3 – Exemples de noyaux.

**Biais.** On a

$$\mathbb{E}_f(\hat{f}_n(x)) = \mathbb{E}_f\left(\frac{1}{h}K\left(\frac{X-x}{h}\right)\right) = \int \frac{1}{h}K\left(\frac{u-x}{h}\right)f(u)du = \int K(v)f(x+hv)dv.$$

Si  $f$  est une fonction dérivable au voisinage de  $x$ , alors on peut écrire  $f(x+hv) = f(x) + hvf'(x + \xi hv)$ , où  $\xi \in ]0; 1[$ . On obtient alors

$$\mathbb{E}_f(\hat{f}_n(x)) = \int K(v)[f(x) + hvf'(x + \xi hv)]dv = f(x) + h \int vK(v)f'(x + \xi hv)dv,$$

car comme  $K$  est un noyau, on a  $\int K(v)dv = 1$ . Si de plus  $\|f'\|_\infty < +\infty$  et  $\int |vK(v)|dv < \infty$ , alors on obtient que

$$\mathbb{E}_f(\hat{f}_n(x)) = f(x) + O(h), \text{ lorsque } h \rightarrow 0.$$

Dans ce cas, on a montré que le biais  $|\mathbb{E}_f(\hat{f}_n(x)) - f(x)|$  converge vers 0 lorsque  $h \rightarrow 0$ . Plus généralement, si on suppose que  $f$  appartient à une classe de fonctions suffisamment régulières, on va pouvoir montrer une décroissance du terme de biais vers 0.

**Notation :** Dans toute la suite, si  $\beta \in \mathbb{R}$  alors  $\lfloor \beta \rfloor$  est le plus petit entier strictement inférieur à  $\beta$ .

**Définition.** Pour tous  $\beta > 0, L > 0$ , on définit la classe des fonctions de Hölder sur l'ensemble  $T$  par

$$\Sigma(\beta, L) = \{f : T \rightarrow \mathbb{R}; f \text{ est } \ell = \lfloor \beta \rfloor \text{ fois dérivable et} \\ \forall x, y \in T, |f^\ell(x) - f^\ell(y)| \leq L|x - y|^{\beta-\ell}\}.$$

On note également  $\Sigma_d(\beta, L)$  l'intersection entre  $\Sigma(\beta, L)$  (pour  $T = \mathbb{R}$ ) et l'ensemble des densités sur  $\mathbb{R}$ .

**Remarques.**  $\star$  Si  $\beta \in ]0; 1]$  alors  $\ell = 0$  et  $\Sigma(\beta, L)$  est la classe des fonctions contractantes (ou Hölderiennes). De plus lorsque  $\beta = 1$ , on obtient les fonctions Lipschitziennes.

$\star$  Si  $\beta \in ]1; 2]$  alors  $\ell = 1$  et  $f'$  est contractante.

**Définition.** Soit  $\ell \in \mathbb{N}^*$ . Le noyau  $K : \mathbb{R} \rightarrow \mathbb{R}$  est dit d'ordre  $\ell$  si  $\forall j \in \{1, \dots, \ell\}$ , on a  $u \rightarrow u^j K(u)$  est intégrable et  $\int u^j K(u) du = 0$ .

**Remarque.** Si  $K$  est un noyau pair alors  $K$  est d'ordre au moins 1.

**Proposition 15.** Si  $f \in \Sigma_d(\beta, L)$  avec  $\beta, L > 0$  et si  $K$  noyau d'ordre  $\ell = \lfloor \beta \rfloor$  tel que  $\int |u|^\beta |K(u)| du < +\infty$ , alors pour tout  $x \in \mathbb{R}$ , tout  $h > 0$  et tout entier  $n \geq 1$  on a

$$B_f(\hat{f}_n(x)) = |\mathbb{E}_f(\hat{f}_n(x)) - f(x)| \leq \frac{L}{\ell!} \left( \int |u|^\beta |K(u)| du \right) h^\beta.$$

En particulier, le biais tend vers 0 lorsque  $h \rightarrow 0$ .

*Démonstration.* On rappelle que

$$\mathbb{E}_f(\hat{f}_n(x)) - f(x) = \int K(v) f(x + hv) dv - f(x) = \int K(v) [f(x + vh) - f(x)] dv,$$

car  $\int K(v) dv = 1$ . On écrit un développement de Taylor à l'ordre  $\ell$  pour  $f$  au voisinage du point  $x$  :

$$f(x + hv) = f(x) + hv f'(x) + \frac{(hv)^2}{2!} f^{(2)}(x) + \dots + \frac{(hv)^{\ell-1}}{(\ell-1)!} f^{(\ell-1)}(x) + \frac{(hv)^\ell}{\ell!} f^{(\ell)}(x + \xi vh),$$

où  $\xi \in ]0; 1[$ . Ainsi,

$$B_f(\hat{f}_n(x)) = \left| \int K(v) \left[ hv f'(x) + \frac{(hv)^2}{2!} f^{(2)}(x) + \dots + \frac{(hv)^{\ell-1}}{(\ell-1)!} f^{(\ell-1)}(x) + \frac{(hv)^\ell}{\ell!} f^{(\ell)}(x + \xi vh) \right] dv \right|.$$

Comme  $K$  est un noyau d'ordre  $\ell$ , tous les termes de la forme  $\int v^j K(v)dv$ ,  $1 \leq j \leq \ell - 1$  ci-dessus s'annulent, et on obtient

$$B_f(\hat{f}_n(x)) = \left| \int K(v) \frac{(hv)^\ell}{\ell!} f^{(\ell)}(x + \xi vh) dv \right|.$$

On utilise à présent le fait que  $\int v^\ell K(v)dv = 0$  ce qui permet d'écrire

$$\begin{aligned} B_f(\hat{f}_n(x)) &= \left| \int K(v) \frac{(hv)^\ell}{\ell!} [f^{(\ell)}(x + \xi vh) - f^{(\ell)}(x)] dv \right| \\ &\leq \int |K(v)| \frac{|hv|^\ell}{\ell!} |f^{(\ell)}(x + \xi vh) - f^{(\ell)}(x)| dv. \end{aligned}$$

Puisque  $f \in \Sigma_d(\beta, L)$ , on a  $|f^{(\ell)}(x + \xi vh) - f^{(\ell)}(x)| \leq L|\xi vh|^{\beta-\ell} \leq L|vh|^{\beta-\ell}$ , puisque  $\xi \in ]0; 1[$ . Ainsi, on obtient

$$B_f(\hat{f}_n(x)) \leq \int |K(v)| L \frac{|hv|^\beta}{\ell!} dv,$$

ce qui achève la preuve. □

**Variance.** On montre la proposition suivante.

**Proposition 16.** *Si  $f$  est une densité bornée sur  $\mathbb{R}$  (i.e.  $\|f\|_\infty < \infty$ ) et si  $K$  est un noyau tel que  $\int K^2(u)du < +\infty$ , alors pour tout  $x \in \mathbb{R}$ , pour tout  $h > 0$  et tout  $n \geq 1$ , on a*

$$\text{Var}_f(\hat{f}_n(x)) \leq \frac{\|f\|_\infty (\int K^2(u)du)}{nh}.$$

*Si de plus,  $f(x) > 0$  et  $f$  continue au voisinage de  $x$  et  $\int |K(u)|du < +\infty$ , alors*

$$\text{Var}_f(\hat{f}_n(x)) = \frac{f(x)}{nh} \left( \int K^2(u)du \right) (1 + o(1)), \text{ lorsque } h \rightarrow 0.$$

*Démonstration.* On note  $Z_i = K\left(\frac{X_i - x}{h}\right)$ . Les variables  $\{Z_i\}_{1 \leq i \leq n}$  sont i.i.d. donc

$$\begin{aligned} \text{Var}_f(\hat{f}_n(x)) &= \text{Var}_f \left( \frac{1}{nh} \sum_{i=1}^n Z_i \right) = \frac{1}{(nh)^2} \sum_{i=1}^n \text{Var}_f(Z_i) \\ &\leq \frac{1}{(nh)^2} \sum_{i=1}^n \mathbb{E}_f(Z_i^2) = \frac{1}{nh^2} \int K^2 \left( \frac{u-x}{h} \right) f(u) du = \frac{1}{nh} \int K^2(v) f(x + hv) dv. \end{aligned}$$

Ainsi

$$\text{Var}_f(\hat{f}_n(x)) \leq \frac{1}{nh} \|f\|_\infty \int K^2(v) dv.$$

On procède à présent à une évaluation exacte. On a

$$\text{Var}_f(\hat{f}_n(x)) = \frac{1}{nh^2} \text{Var}_f(Z_1) = \frac{1}{nh^2} [\mathbb{E}_f(Z_1^2) - (\mathbb{E}_f Z_1)^2].$$

On utilise alors le lemme suivant.

**Lemme 17. (de Bochner).** *Si  $f$  est bornée sur  $\mathbb{R}$ , continue au voisinage de  $x$  et si  $K : \mathbb{R} \rightarrow \mathbb{R}$  est une fonction telle que  $\int |K(u)|du < +\infty$ , alors*

$$\lim_{h \rightarrow 0} \frac{1}{h} \int K\left(\frac{u-x}{h}\right) = f(x) \int K(u)du.$$

La preuve du lemme de Bochner s'obtient par convergence dominée, en remarquant que

$$\frac{1}{h} \int K\left(\frac{u-x}{h}\right) = \int K(v)f(x+vh)dv.$$

En appliquant ce lemme (avec  $f$  bornée et  $K$  intégrable), on obtient

$$\begin{aligned} \frac{1}{h} \mathbb{E}_f\left(\frac{X_1-x}{h}\right) &= \frac{1}{h} \int K\left(\frac{u-x}{h}\right) f(u)du = f(x) \left(\int K(u)du\right)(1+o(1)) \\ &= f(x)(1+o(1)), \text{ lorsque } h \rightarrow 0. \end{aligned}$$

De la même façon, si on suppose  $f$  bornée et  $K^2$  intégrable, on a aussi

$$\begin{aligned} \frac{1}{h} \mathbb{E}_f\left[\left(\frac{X_1-x}{h}\right)^2\right] &= \frac{1}{h} \int K^2\left(\frac{u-x}{h}\right) f(u)du \\ &= f(x) \left(\int K^2(u)du\right)(1+o(1)), \text{ lorsque } h \rightarrow 0. \end{aligned}$$

Ainsi, on obtient

$$\begin{aligned} \text{Var}_f(\hat{f}_n(x)) &= \frac{1}{nh} [f(x) \left(\int K^2(u)du\right)(1+o(1)) - hf(x)^2(1+o(1))] \\ &= \frac{1}{nh} f(x) \left(\int K^2(u)du\right)(1+o(1)), \text{ lorsque } h \rightarrow 0. \end{aligned}$$

□

### Commentaires

★ Si  $nh \rightarrow \infty$  alors on aura  $\text{Var}_f(\hat{f}_n(x)) \rightarrow 0$ . Donc on veut  $h \rightarrow 0$ , mais pas trop vite (*i.e.*  $nh \rightarrow \infty$ ) : il ne faut pas sous-lisser.

★ Sur la classe de Hölder  $\Sigma_d(\beta, L)$ , le biais de  $\hat{f}_n(x)$  est en  $O(h^\beta)$  et si la densité  $f$  est bornée, on sait contrôler sa variance. La question naturelle qui se pose alors est : les fonctions de  $\Sigma_d(\beta, L)$  sont elles bornées ?

**Lemme 18. (admis).** Soit  $\beta, L > 0$ . Il existe une constante  $M(\beta, L) > 0$  telle que  $\forall f \in \Sigma_d(\beta, L)$ , on a

$$\sup_{x \in \mathbb{R}} \sup_{f \in \Sigma_d(\beta, L)} f(x) \leq M(\beta, L).$$

Noter que ce lemme montre non seulement que les fonctions  $f \in \Sigma_d(\beta, L)$  sont bornées, mais aussi qu'elles sont bornées uniformément dans cette classe.

Les résultats précédents nous permettent d'établir le résultat suivant.

**Théorème 19. (Contrôle du risque quadratique ponctuel sur la classe  $\Sigma_d(\beta, L)$ ).** Soit  $\beta > 0, L > 0$  et  $K$  un noyau d'ordre  $\ell = \lfloor \beta \rfloor$  tel que  $\int K^2(u)du < +\infty$  et  $\int |u|^\beta |K(u)|du < +\infty$ . Alors, en choisissant une fenêtre  $h = cn^{-1/(2\beta+1)}$ , avec  $c > 0$ , on obtient

$$\forall x \in \mathbb{R}, \quad R(\hat{f}_n(x), \Sigma_d(\beta, L)) = \sup_{f \in \Sigma_d(\beta, L)} \mathbb{E}_f[|\hat{f}_n(x) - f(x)|^2] \leq Cn^{-2\beta/(2\beta+1)},$$

où  $C = C(c, \beta, L, K)$ .

**Remarques.** ★ L'estimateur  $\hat{f}_n$  atteint la vitesse de convergence  $\phi_{n,\beta} = n^{-2\beta/(2\beta+1)}$  sur la classe  $\Sigma_d(\beta, L)$  pour le risque quadratique ponctuel maximal.  
 ★ Le choix de la fenêtre optimale  $h$  dépend de  $\beta =$  régularité maximale de la densité  $f$  inconnue. Il peut paraître artificiel de supposer qu'on connaît  $\beta$  quand on ne connaît pas  $f$ . IL existe des méthodes d'estimation dites adaptatives qui n'utilisent pas la connaissance a priori de  $\beta$ .

*Démonstration.* On a déjà vu que

$$R(\hat{f}_n(x), f(x)) = \text{biais}^2 + \text{variance}.$$

On applique alors les Propositions 15 et 16 qui donnent

$$R(\hat{f}_n(x), f(x)) \leq \frac{L^2}{(\ell!)^2} \left( \int |u|^\beta |K(u)|du \right)^2 h^{2\beta} + \frac{M(\beta, L) \int K^2(u)du}{nh}.$$

Comme la majoration est uniforme par rapport à  $f \in \Sigma_d(\beta, L)$ , on obtient

$$\begin{aligned} \sup_{f \in \Sigma_d(\beta, L)} R(\hat{f}_n(x), f(x)) &\leq \frac{L^2}{(\ell!)^2} \left( \int |u|^\beta |K(u)|du \right)^2 h^{2\beta} + \frac{M(\beta, L) \int K^2(u)du}{nh} \\ &= C_B h^{2\beta} + \frac{C_V}{nh}. \end{aligned}$$

On veut alors optimiser la borne de droite par rapport à  $h > 0$ . On définit  $\psi : h \rightarrow C_B h^{2\beta} + C_V/(nh)$ . Cette fonction admet un point singulier en  $h^* = (C_V/(2\beta C_B))^{1/(2\beta+1)} n^{-1/(2\beta+1)}$  et ce point est bien un minimum. (Faire par exemple un tableau de variation. On peut distinguer le cas  $\beta < 1/2$ .) Alors on obtient

$$\min_{h>0} C_B h^{2\beta} + \frac{C_V}{nh} = C_B (h^*)^{2\beta} + \frac{C_V}{nh^*} = C n^{-2\beta/(2\beta+1)}.$$

Ainsi, l'estimateur  $\hat{f}_n^*$  qui utilise la fenêtre optimale  $h^* = (C_V/(2\beta C_B))^{1/(2\beta+1)} n^{-1/(2\beta+1)}$  vérifie

$$\sup_{f \in \Sigma_d(\beta, L)} R(\hat{f}_n^*(x), f(x)) \leq C n^{-2\beta/(2\beta+1)}.$$

□

## 4.4 Construction de noyaux d'ordre $\ell$

Le théorème précédent repose sur l'existence de noyaux d'ordre  $\ell = \lfloor \beta \rfloor$ . Nous allons montrer que de tels noyaux existent.

Soit  $\{\phi_m\}_{m \geq 0}$  la b.o.n. des polynômes de Legendre dans  $\mathbb{L}_2([0; 1])$ , définie par

$$\phi_0 \equiv \frac{1}{\sqrt{2}} \quad \text{et } \forall m \geq 1, \quad \phi_m(x) = \sqrt{\frac{2m+1}{2}} \frac{1}{2^m m!} \frac{d^m}{dx^m} [(x^2 - 1)^m].$$

Cette base est obtenue par orthonormalisation (de Gramm-Schmidt) à partir de  $\{P_k : x \rightarrow x^k\}_{k \geq 0}$ . Elle a les propriétés suivantes

- ★  $\int_{-1}^1 \phi_m(u) \phi_k(u) du = 1_{m=k}$ , (car b.o.n.)
- ★  $\phi_m$  est un polynôme de degré  $m$ , (base orthonormalisée à partir de  $\{P_k\}_k$ )
- ★  $\phi_0 \equiv 1/\sqrt{2}$ ,  $\phi_1(x) = \sqrt{3/2}x$ ,  $\phi_2(x) = \sqrt{5/2}(3x^2 - 1)/2$ ,
- ★  $\phi_{2k}$  est une fonction paire et  $\phi_{2k+1}$  impaire.

**Proposition 20.** *Soit  $K : u \rightarrow \sum_{m=0}^{\ell} \phi_m(0) \phi_m(u) 1_{|u| \leq 1}$ . Alors  $K$  est un noyau d'ordre  $\ell$ .*

*Démonstration.* Pour tout  $j \in \mathbb{N}, j \leq \ell$ , on a  $u \rightarrow u^j K(u)$  est intégrable sur  $[-1; 1]$ . Comme  $\{\phi_m\}_{m \geq 0}$  base de  $\mathbb{L}_2([-1; 1])$ , il existe des coefficients  $\{a_m(j)\}_{m \geq 0}$  tels que

$$\forall u \in [-1; 1], \quad u^j = \sum_{m \geq 0} a_m(j) \phi_m(u).$$

De plus, comme  $\phi_m$  est un polynôme de degré  $m$ , les coefficients  $a_m(j)$  pour  $m > j$  sont nuls et on a

$$\forall u \in [-1; 1], \quad u^j = \sum_{m=0}^j a_m(j) \phi_m(u).$$

Alors on a

$$\begin{aligned}
\int_{-1}^1 u^j K(u) du &= \int_{-1}^1 \sum_{m=0}^j a_m(j) \phi_m(u) K(u) du \\
&= \int_{-1}^1 \sum_{m=0}^j a_m(j) \phi_m(u) \sum_{k=0}^{\ell} \phi_k(0) \phi_k(u) 1_{|u| \leq 1} du \\
&= \sum_{m=0}^j \sum_{k=0}^{\ell} a_m(j) \phi_k(0) \int_{-1}^1 \phi_m(u) \phi_k(u) du \stackrel{(1)}{=} \sum_{m=0}^j a_m(j) \phi_m(0) = 0^j = 1_{j=0}.
\end{aligned}$$

On remarque pour l'égalité (1) que  $\int_{-1}^1 \phi_m(u) \phi_k(u) du = 1_{k=m}$ . Ainsi,  $K$  est bien un noyau d'ordre  $\ell$ .  $\square$

**Remarque.** Comme  $\phi_{2k}$  est une fonction paire et  $\phi_{2k+1}(0) = 0$  (fonction impaire), on a pour tout  $k \geq 0$ ,

$$K_{2k}(u) = \sum_{m=0}^{2\ell} \phi_m(0) \phi_m(u) 1_{|u| \leq 1} = \sum_{m=0}^k \phi_{2m}(0) \phi_{2m}(u) 1_{|u| \leq 1}.$$

Donc  $K_{2k}$  est une fonction paire. Comme c'est un noyau d'ordre  $2k$ , c'est donc aussi un noyau d'ordre  $2k + 1$ .

## 4.5 Cas d'une densité multivariée

On se place dans  $\mathbb{R}^2$ , la généralisation étant triviale. Soit  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  une densité et  $(X_1, Y_1), \dots, (X_n, Y_n)$  un échantillon de densité  $f$ . On utilise un *noyau produit* et on construit

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

Plus généralement dans  $\mathbb{R}^p$ , on utilise

$$\hat{f}_n(x^1, \dots, x^p) = \frac{1}{nh^d} \sum_{i=1}^n \prod_{j=1}^p K\left(\frac{X_i^j - x^j}{h}\right).$$

Alors on peut montrer des contrôles du biais et de la variance par des méthodes identiques à celles utilisées ci-dessus.

Par exemple, si  $f \in \Sigma_{d,p}(1, L)$ , l'ensemble des densités sur  $\mathbb{R}^p$  qui sont Lipschitziennes (*i.e.*  $\forall x, y \in \mathbb{R}^p, |f(x) - f(y)| \leq L\|x - y\|$ ), alors on obtient  $B_f^2(\hat{f}_n(x)) = |\mathbb{E}_f \hat{f}_n(x) - f(x)|^2 = O(h^2)$ , *i.e.* le biais ne dépend pas de la dimension de l'espace.

Par contre, le terme de variance en dépend, on obtient  $\text{Var}_f(\hat{f}_n(x)) = O(1/(nh^p))$ . Si on cherche la fenêtre optimale (*i.e.* qui minimise le risque quadratique), on obtient alors  $h = cn^{-1/(2+p)}$  et la vitesse de convergence du risque quadratique correspondante est  $n^{-2/(2+p)}$ . Quand la dimension  $p$  augmente, cette vitesse est plus lente, c'est le **fléau de la dimension**.

## 4.6 Risque quadratique dans $\mathbb{L}_2(\mathbb{R})$

Dans la Section 4.3, on a considéré les performances **ponctuelles** de l'estimateur  $\hat{f}_n$ , *i.e.* on a fixé  $x \in \mathbb{R}$  et on s'est intéressés à l'estimation du paramètre  $f(x) \in \mathbb{R}$  dans un modèle non paramétrique. Mais comme  $f$  est une fonction, il peut être plus intéressant de chercher à caractériser les performances de  $\hat{f}_n$  de façon globale, par exemple à travers le risque quadratique dans  $\mathbb{L}_2(\mathbb{R})$ .

On rappelle que

$$MISE(\hat{f}_n, f) = \mathbb{E}_f \|\hat{f}_n - f\|_2^2 = \mathbb{E}_f \left[ \int (\hat{f}_n(x) - f(x))^2 dx \right] = \text{Var}_f(\hat{f}_n) + B_f^2(\hat{f}_n),$$

où  $\text{Var}_f(\hat{f}_n) = \mathbb{E}_f \|\hat{f}_n - \mathbb{E}_f \hat{f}_n\|_2^2$  et  $B_f^2 = \|\mathbb{E}_f \hat{f}_n - f\|_2^2$ .

En faisant simplement l'hypothèse  $f \in \mathbb{L}_2(\mathbb{R})$ , on peut montrer le contrôle suivant sur la variance (au sens  $\mathbb{L}_2$ ) de  $\hat{f}_n$ .

**Proposition 21.** *Si  $f \in \mathbb{L}_2(\mathbb{R})$ , et si  $K$  noyau tel que  $\int K^2(u)du < \infty$ , alors pour toute fenêtre  $h > 0$  et tout entier  $n \geq 1$ , on a*

$$\text{Var}_f(\hat{f}_n) = \mathbb{E}_f \|\hat{f}_n - \mathbb{E}_f \hat{f}_n\|_2^2 = \frac{1}{nh} \left( \int K^2(u)du \right) (1 + o(1)).$$

*Démonstration.* Voir TD. □

Pour contrôler le biais de cet estimateur, il faut introduire une classe de fonctions régulières. Ici, le contrôle souhaité étant global, on introduit une classe qui contrôle la régularité globale de la fonction  $f$ .

**Définition. (classe de Nikol'ski.)** Soient  $\beta, L > 0$ , on définit la classe de fonctions de Nikol'ski  $\mathcal{N}(\beta, L)$  par

$$\mathcal{N}(\beta, L) = \{f : \mathbb{R} \rightarrow \mathbb{R}, f \text{ est } \ell = \lfloor \beta \rfloor \text{ fois dérivable et } \forall t \in \mathbb{R}, \\ \|f^{(\ell)}(\cdot + t) - f^{(\ell)}\|_2 = \left( \int \left( f^{(\ell)}(x+t) - f^{(\ell)}(x) \right)^2 dx \right)^{1/2} \leq L|t|^{\beta-\ell}\}.$$

De plus, on note  $\mathcal{N}_d(\beta, L)$  l'ensemble des densités qui sont dans la classe  $\mathcal{N}(\beta, L)$ .

On peut alors montrer le résultat suivant

**Proposition 22.** *Si  $f \in \mathcal{N}_d(\beta, L)$  et si  $K$  est un noyau d'ordre  $\ell = \lfloor \beta \rfloor$  tel que  $\int |u|^\beta |K(u)| du < +\infty$ , alors pour tout  $h > 0$  et tout  $n \geq 1$ , on a*

$$B_f^2 = \|\mathbb{E}_f \hat{f}_n - f\|_2^2 \leq \left( \frac{L}{(\ell)!} \int |u|^\beta |K(u)| du \right)^2 h^{2\beta}.$$

*Démonstration.* Voir TD. □

La fenêtre optimale qui minimise le risque quadratique intégré est alors  $h^* = cn^{-1/(2\beta+1)}$ , et pour cette fenêtre, l'estimateur  $\hat{f}_n^*$  vérifie  $MISE(\hat{f}_n^*, \mathcal{N}(\beta, L)) = O(n^{-2\beta/(2\beta+1)})$ .

## 4.7 Choix de la fenêtre par validation croisée

Jusqu'à présent, on a considéré un risque maximal sur la classe :  $R(\hat{f}_n, \mathcal{F}) = \sup_{f \in \mathcal{F}} R(\hat{f}_n, f)$ . C'est un point de vue pessimiste et les observations n'ont aucune raison d'avoir été générées sous le pire des cas possibles sur la classe  $\mathcal{F}$ . Peut-on choisir la fenêtre optimale qui minimise le risque  $R(\hat{f}_n, f)$  pour les observations données ? La réponse est oui dans certains cas, par exemple pour le MISE, mais pas pour le MSE.

On reprend donc le cas du MISE. On a

$$MISE(h) = \mathbb{E}_f \int [\hat{f}_{n,h}(x) - f(x)]^2 dx = \mathbb{E}_f \int \hat{f}_{n,h}^2(x) dx - 2\mathbb{E}_f \int f(x) \hat{f}_{n,h}(x) dx + cte,$$

et donc

$$\operatorname{argmin}_{h>0} MISE(h) = \operatorname{argmin}_{h>0} \mathbb{E}_f \int \hat{f}_{n,h}^2(x) dx - 2\mathbb{E}_f \int f(x) \hat{f}_{n,h}(x) dx \equiv \operatorname{argmin}_{h>0} J(h).$$

Comme  $J$  est inconnu (puisque dépend de  $f$  inconnu), on propose de l'estimer et de choisir la fenêtre  $h > 0$  qui minimise son estimateur.

Pour estimer sans biais le terme  $\mathbb{E}_f \int \hat{f}_{n,h}^2(x) dx$ , il suffit de prendre  $\int \hat{f}_{n,h}^2(x) dx$ . De plus, pour estimer sans biais le terme  $\mathbb{E}_f \int f(x) \hat{f}_{n,h}(x) dx$  on pourrait penser prendre  $\int \hat{f}_{n,h}^2(x) dx$  mais ça ne marche pas du tout (puisque qu'on a vu que cette quantité était un estimateur sans biais de  $\mathbb{E}_f \int \hat{f}_{n,h}^2(x) dx$ ). On remarque plutôt que

$$\begin{aligned} \mathbb{E}_f \int f(x) \hat{f}_{n,h}(x) dx &= \stackrel{\text{Fubini}}{\int} f(x) \mathbb{E}_f[\hat{f}_{n,h}(x)] dx \\ &= \int f(x) \frac{1}{h} \int K\left(\frac{u-x}{h}\right) f(u) du dx. \end{aligned}$$

On utilise alors le résultat suivant.

**Lemme 23.** On considère

$$\hat{T}_n = \frac{1}{n(n-1)h} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K\left(\frac{X_i - X_j}{h}\right).$$

Alors  $\hat{T}_n$  est un estimateur sans biais de

$$\frac{1}{h} \int \int f(x) K\left(\frac{u-x}{h}\right) f(u) du dx.$$

*Démonstration.* En effet,

$$\begin{aligned} \mathbb{E}_f \hat{T}_n &= \frac{1}{n(n-1)h} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \mathbb{E}_f K\left(\frac{X_i - X_j}{h}\right) =_{X_i \text{ i.i.d.}} \frac{1}{h} \mathbb{E}_f K\left(\frac{X_1 - X_2}{h}\right) \\ &= \frac{1}{h} \int \int K\left(\frac{u-x}{h}\right) f(u) f(x) du dx. \end{aligned}$$

□

**Remarque.** On peut noter que de façon très générale, il est important quand on considère une somme double de la forme  $\sum_{i,j} \phi(X_i - X_j)$ , de la priver de sa diagonale  $i \neq j$ , sinon on augmente le biais. En effet, considérons par exemple

$$\tilde{T}_n = \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n K\left(\frac{X_i - X_j}{h}\right).$$

Alors la moyenne de  $\tilde{T}_n$  fait apparaître un terme parasite :

$$\mathbb{E}_f \tilde{T}_n = \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_f K\left(\frac{X_i - X_j}{h}\right) = \frac{1}{nh} K(0) + \frac{n-1}{nh} \mathbb{E}_f K\left(\frac{X_1 - X_2}{h}\right).$$

Revenons à l'estimation de  $J$ . Ainsi, on définit

$$\hat{J}_{n,h} = \int \hat{f}_{n,h}^2(x) dx - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K\left(\frac{X_i - X_j}{h}\right),$$

et d'après ce qui précède, c'est un estimateur sans biais de  $J(h)$ . On définit la fenêtre de cross-validation

$$h^{CV} = \operatorname{argmin}_{h>0} \hat{J}_{n,h}$$

et l'estimateur à noyau correspondant  $\hat{f}_n^{CV} \equiv \hat{f}_{n,h^{CV}}$  : c'est un estimateur à noyau qui est construit avec la fenêtre aléatoire  $h^{CV}$  (qui dépend des observations).

On peut montrer (admis) que  $\hat{f}_n^{CV}$  est un estimateur qui a de bonnes propriétés asymptotiques : asymptotiquement, il minimise en  $h > 0$  le risque  $MISE(h) = R(\hat{f}_{n,h}, f)$  pour la densité observée.

# Chapitre 5

## Régression non paramétrique

Dans tout ce chapitre,  $(X, Y)$  est un couple de variables aléatoires réelles avec  $\mathbb{E}|Y| < +\infty$ . On note  $r : x \rightarrow r(x) = \mathbb{E}(Y|X = x)$  la fonction de régression de  $Y$  sur  $X$ . On observe un échantillon  $(X_1, Y_1), \dots, (X_n, Y_n)$  de variables i.i.d. de même loi que  $(X, Y)$ . On cherche à estimer  $r$  en faisant le moins d'hypothèses possibles (uniquement  $r \in \mathcal{F}$  où  $\mathcal{F}$  est une classe de fonctions). Les variables  $X_1, \dots, X_n$  constituent le dispositif expérimental, ou « design ». Elles peuvent être déterministes ou aléatoires. Dans le premier cas, on parle d'« effets fixes », dans le second, d'« effets aléatoires ».

On note  $\xi = Y - \mathbb{E}(Y|X)$  le résidu. On peut alors écrire

$$Y_i = r(X_i) + \xi_i, 1 \leq i \leq n, \quad \text{où } \xi_i \text{ i.i.d. centrés.}$$

Les  $\{\xi_i\}_{1 \leq i \leq n}$  jouent le rôle d'un bruit. On supposera que ces variables ont un moment d'ordre 2 fini et on note  $\sigma^2 = \mathbb{V}\text{ar}(\xi_i)$ .

**Remarque.** Les méthodes de ce chapitre sont très semblables à celles du chapitre précédent. Il faut cependant garder en tête que les fonctions  $r$  que l'on estime ici sont très différentes des densités du chapitre précédent. En effet, sur  $\mathbb{R}$ , il est courant (en statistique paramétrique) de considérer des régressions polynomiales (*i.e.* supposer que  $r$  est un polynôme), en particulier, on a rarement  $r$  intégrable, contrairement au cas des densités.

### 5.1 Estimateur de Nadaraya-Watson

Supposons que  $(X, Y)$  a une densité  $p : (x, y) \rightarrow p(x, y)$  sur  $\mathbb{R}^2$  et que  $p_X : x \rightarrow p_X(x) = \int p(x, y)dy > 0$  (densité de  $X$ ). Alors, on peut écrire

$$\forall x \in \mathbb{R}, \quad r(x) = \mathbb{E}(Y|X) = \frac{\int yp(x, y)dy}{p_X(x)}.$$

Comme les densités  $p$  et  $p_X$  sont inconnus, mais qu'on observe un échantillon de variables de ces densités, on peut les estimer via un estimateur à noyau. On considère donc

$$\begin{aligned}\forall (x, y) \in \mathbb{R}^2, \quad \hat{p}_n(x, y) &= \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right), \\ \hat{p}_{n,X}(x) &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right),\end{aligned}$$

puis l'estimateur de la régression,

$$\forall x \in \mathbb{R}, \quad \hat{r}_n(x) = \frac{\int y \hat{p}_n(x, y) dy}{\hat{p}_{n,X}(x)} 1_{\hat{p}_{n,X}(x) \neq 0}. \quad (5.1)$$

**Proposition 24.** *Si  $K$  est un noyau d'ordre 1, l'estimateur défini par (5.1) vérifie*

$$\forall x \in \mathbb{R}, \quad \hat{r}_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} 1_{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0}.$$

*C'est l'estimateur de Nadaraya-Watson, noté  $\hat{r}_n^{NW}$ .*

*Démonstration.* En effet, pour tout  $x \in \mathbb{R}$  tel que  $\hat{p}_{n,X}(x) \neq 0$ , on a

$$\hat{r}_n(x) = \frac{\int y \hat{p}_n(x, y) dy}{\hat{p}_{n,X}(x)} 1_{\hat{p}_{n,X}(x) \neq 0} = \frac{1}{h} \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \int y K\left(\frac{Y_i - y}{h}\right) dy}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}.$$

Or,  $\int y K\left(\frac{Y_i - y}{h}\right) dy = h \int (Y_i - uh) K(u) du = h Y_i$  si  $K$  est un noyau d'ordre 1. Donc on obtient bien

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

□

L'estimateur de Nadaraya-Watson est une moyenne pondérée des observations  $Y_i$  (voir la Figure 5.1 pour une illustration). On a

$$\forall x \in \mathbb{R}, \quad \hat{r}_n^{NW}(x) = \sum_{i=1}^n w_{n,i}(x) Y_i$$

où les poids  $w_{n,i}(x)$  vérifient

$$w_{n,i}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)} 1_{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right) \neq 0}.$$

Ces poids ne dépendent que des effets (et pas des observations  $Y_i$ ). En particulier,  $\hat{r}_n^{NW}$  est un **estimateur linéaire** de la régression non paramétrique des  $Y_i$  sur les  $X_i$ .

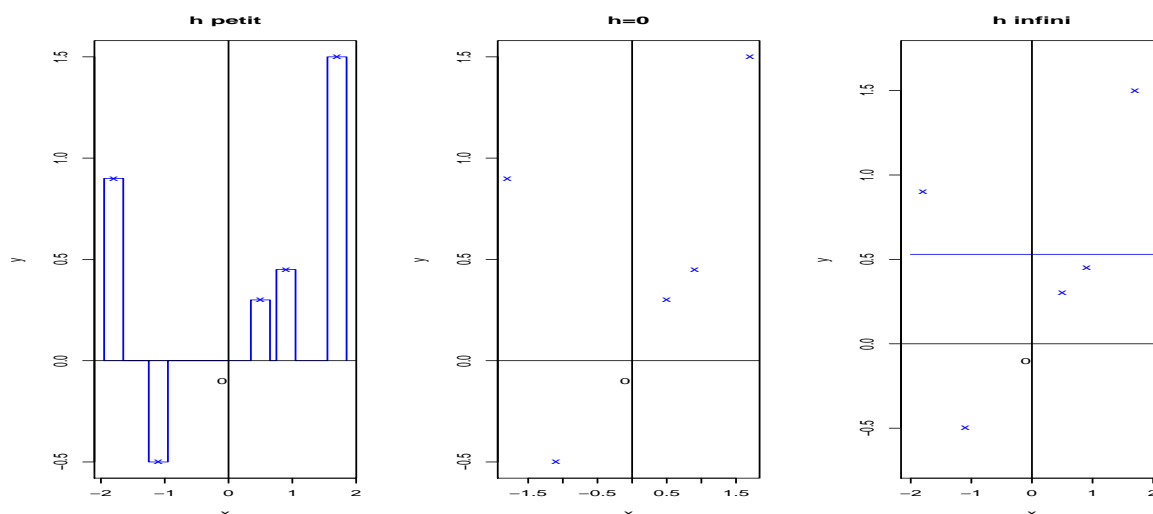


FIG. 5.1 – Estimateur de Nadaraya-Watson avec noyau rectangulaire ( $K : u \rightarrow 1_{|u| \leq 1/2}$ ) pour différentes valeurs de fenêtre  $h$ .

**Remarques.** ★ Ne pas confondre régression linéaire et estimateur linéaire de la régression.

★ Pour tout  $x \in \mathbb{R}$ , on a  $\sum_{i=1}^n w_{n,i}(x) = 1$  ou 0.

★ Si la densité marginale  $p_X$  des  $X_i$  est connue, on utilisera plutôt l'estimateur

$$\forall x \in \mathbb{R}, \quad \hat{r}_n(x) = \frac{\int y \hat{p}_n(x, y) dy}{p_X(x)} 1_{p_X(x) \neq 0} = \frac{1}{nh p_X(x)} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right) 1_{p_X(x) \neq 0}.$$

En particulier, si les effets sont uniformes sur  $[0; 1]$ , alors on utilise

$$\forall x \in \mathbb{R}, \quad \hat{r}_n(x) = \frac{1}{nh} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right) 1_{[0;1]}(x).$$

★ Dans le cas d'effets fixes réguliers, *i.e.*  $X_i = i/n, 1 \leq i \leq n$ , il n'y a pas de densité  $p_X$ . Cependant, l'estimateur précédent

$$\forall x \in \mathbb{R}, \quad \hat{r}_n(x) = \frac{1}{nh} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right) 1_{[0;1]}(x).$$

est parfaitement adapté.

★ L'estimateur de Nadaraya-Watson est un cas particulier d'une classe plus générale : les estimateurs par polynômes locaux, que nous allons introduire et étudier dans la section suivante.

## 5.2 Les estimateurs par polynômes locaux

On remarque la chose suivante : si  $K$  est un noyau positif, alors

$$\forall x \in \mathbb{R}, \quad \hat{r}_n^{NW}(x) = \operatorname{argmin}_{\theta \in \mathbb{R}} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) (Y_i - \theta)^2.$$

En effet, si on cherche les points singuliers correspondants, on obtient

$$-2 \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) (Y_i - \theta) = 0 \iff \theta = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} \mathbf{1}_{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0},$$

et il s'agit bien d'un minimum si  $K \geq 0$ .

Ainsi, au sens des moindres carrés pondérés, on a approché les  $Y_i$  par une constante (en  $x$ ), notée  $\theta$ . On pourrait plus généralement approcher  $Y_i$  par un polynôme en  $x$ . Cela revient à dire que la fonction  $r$ , au voisinage du point  $x$  peut-être approchée par un polynôme (local) en  $x$  (et pas seulement une constante). Si la fonction  $r$  est  $\ell$  fois dérivable au voisinage de  $x$ , on note

$$\forall u \in \mathbb{R}, \quad P_\ell(u) = r(x) + r'(x)(u - x) + \dots + \frac{r^{(\ell)}(x)}{\ell!} (u - x)^\ell.$$

On introduit artificiellement la fenêtre  $h$  dans cette définition

$$\begin{aligned} \forall u \in \mathbb{R}, \quad P_\ell(u) &= r(x) + r'(x)h \left(\frac{u - x}{h}\right) + \dots + \frac{r^{(\ell)}(x)h^\ell}{\ell!} \left(\frac{u - x}{h}\right)^\ell \\ &= \langle \theta(x); V_\ell \left(\frac{u - x}{h}\right) \rangle = \theta(x)^\top \cdot V_\ell \left(\frac{u - x}{h}\right), \end{aligned}$$

où  $\theta(x) = (r(x); r'(x)h; \dots; r^{(\ell)}(x)h^\ell)^\top$  est un vecteur qui contient les valeurs de  $r$  et ses dérivées au point  $x$  et  $V_\ell(z) = (1; z; z^2/(2!); \dots; z^\ell/(\ell!))^\top$ .

**Définition.** Soit  $K : \mathbb{R} \rightarrow \mathbb{R}^+$  un noyau positif,  $h > 0$  une fenêtre et  $\ell \geq 0$  un entier. On définit

$$\forall x \in \mathbb{R}, \quad \hat{\theta}_n(x) = \operatorname{argmin}_{\theta \in \mathbb{R}^{\ell+1}} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \left[ Y_i - \theta^\top \cdot V_\ell \left(\frac{X_i - x}{h}\right) \right]^2.$$

Alors  $\hat{\theta}_n$  est l'estimateur localement polynomial d'ordre  $\ell$  de la fonction  $\theta$ . De plus, la statistique

$$\forall x \in \mathbb{R}, \quad \hat{r}_n^{LP(\ell)}(x) = \hat{\theta}_n(x)^\top \cdot V_\ell(0)$$

(i.e. la première coordonnée du vecteur  $\hat{\theta}_n$ ) est l'estimateur localement polynomial d'ordre  $\ell$  de la fonction de régression  $r$ .

**Remarques.**  $\star$  On a vu que pour  $\ell = 0$ , on a  $\hat{r}_n^{LP(0)} = \hat{r}_n^{NW}$ .

$\star$   $\hat{\theta}_n$  contient plus qu'un estimateur de la régression  $r$ , puisque les coordonnées de ce vecteur contiennent en fait des estimateurs des dérivées successives de  $r$ , jusqu'à l'ordre  $\ell$ .

Dans la suite, on cherche une forme explicite de l'estimateur  $\hat{\theta}_n$ . Puisque  $\hat{\theta}_n$  est un estimateur des moindres carrés (pondérés), on a,

$$\forall x \in \mathbb{R}, \quad \hat{\theta}_n(x) = \underset{\theta \in \mathbb{R}^{\ell+1}}{\operatorname{argmin}} \{ \theta^\top B_{n,x} \theta - 2\theta^\top a_{n,x} \},$$

où  $B_{n,x}$  matrice de taille  $(\ell + 1) \times (\ell + 1)$  et  $a_{n,x}$  vecteur de  $\mathbb{R}^{\ell+1}$  définis par

$$\begin{aligned} B_{n,x} &= \frac{1}{nh} \sum_{i=1}^n V_\ell \left( \frac{X_i - x}{h} \right) V_\ell^\top \left( \frac{X_i - x}{h} \right) K \left( \frac{X_i - x}{h} \right) \\ a_{n,x} &= \frac{1}{nh} \sum_{i=1}^n Y_i K \left( \frac{X_i - x}{h} \right) V_\ell \left( \frac{X_i - x}{h} \right). \end{aligned}$$

Ainsi,  $\hat{\theta}_n(x)$  doit satisfaire  $B_{n,x} \hat{\theta}_n(x) = a_{n,x}$ . La matrice  $B_{n,x}$  est une matrice symétrique réelle et positive, puisque de la forme  $\sum_i \alpha_i U_i U_i^\top$  où les  $U_i$  sont des vecteurs réels et les  $\alpha_i > 0$ . Donc elle est inversible si et seulement si elle est définie **positive**. Dans ce qui suit, on suppose que la matrice  $B_{n,x}$  est définie positive. Alors la solution est unique et donnée par  $\hat{\theta}_n(x) = B_{n,x}^{-1} a_{n,x}$ , *i.e.*

$$\hat{\theta}_n(x) = \frac{1}{nh} \sum_{i=1}^n Y_i K \left( \frac{X_i - x}{h} \right) B_{n,x}^{-1} V_\ell \left( \frac{X_i - x}{h} \right),$$

et en considérant uniquement la première coordonnée, on note que l'on peut écrire

$$\forall x \in \mathbb{R}, \quad \hat{r}_n^{LP(\ell)}(x) = \sum_{i=1}^n w_{ni}(x) Y_i,$$

*i.e.* l'estimateur de la régression localement polynomial est un estimateur linéaire, avec poids

$$w_{n,i}(x) = \frac{1}{nh} \left[ V_\ell^\top(0) B_{n,x}^{-1} V_\ell \left( \frac{X_i - x}{h} \right) \right] K \left( \frac{X_i - x}{h} \right). \quad (5.2)$$

Ainsi, si la matrice  $B_{n,x}$  est définie positive au point  $x$ , alors l'estimateur  $\hat{r}_n^{LP(\ell)}(x)$  est un estimateur linéaire, dont les poids sont données par (5.2).

On montre à présent une propriété de la suite des poids  $w_{ni}$  (à rapprocher de la propriété de *noyau d'ordre*  $\ell$ ).

**Proposition 25.** Soit  $x \in \mathbb{R}$  tel que  $B_{nx}$  soit définie positive et  $Q$  un polynôme de degré inférieur ou égal à  $\ell$ . Alors la suite des poids  $\{w_{ni}(x)\}_{1 \leq i \leq n}$  définie par (5.2) vérifie

$$\sum_{i=1}^n Q(X_i)w_{n,i}(x) = Q(x).$$

**Remarques.**  $\star$  Cette proposition signifie que si on observe  $Y_i = Q(X_i)$ ,  $1 \leq i \leq n$  sans erreurs (résidus  $\xi_i = 0$ ) et  $Q$  est un polynôme de degré  $\leq \ell$ , alors l'estimateur localement polynomial d'ordre  $\ell$  de  $Q$  vérifie  $\hat{r}_n^{LP(\ell)} = Q$  (au point  $x \in \mathbb{R}$  satisfaisant les hypothèses).

$\star$  Conséquence de cette proposition : en prenant  $Q \equiv 1$  puis  $Q_k : u \rightarrow (u - x)^k$ , pour toutes les valeurs  $1 \leq k \leq \ell$ , on obtient les identités suivantes

$$\sum_{i=1}^n w_{ni}(x) = 1 \text{ et } \forall 1 \leq k \leq \ell, \sum_{i=1}^n (X_i - x)^k w_{ni}(x) = Q_k(x) = 0. \quad (5.3)$$

*Démonstration.* Puisque  $Q$  est un polynôme de degré  $\leq \ell$ , on peut écrire un développement exact

$$Q(X_i) = Q(x) + Q'(x)(X_i - x) + \dots + \frac{Q^{(\ell)}(x)}{\ell!}(X_i - x)^\ell = q(x)^\top V_\ell \left( \frac{X_i - x}{h} \right),$$

où  $q(x) = (Q(x), Q'(x)h, \dots, Q^{(\ell)}(x)h^\ell)^\top \in \mathbb{R}^{\ell+1}$ . Considérons dans la suite l'observation de  $Y_i = Q(X_i)$ ,  $1 \leq i \leq n$ . Par définition, l'estimateur localement polynomial d'ordre  $\ell$  vérifie

$$\begin{aligned} \hat{\theta}_n(x) &= \operatorname{argmin}_{\theta \in \mathbb{R}^{\ell+1}} \sum_{i=1}^n K \left( \frac{X_i - x}{h} \right) \left[ Q(X_i) - \theta^\top V_\ell \left( \frac{X_i - x}{h} \right) \right]^2 \\ &= \operatorname{argmin}_{\theta \in \mathbb{R}^{\ell+1}} \sum_{i=1}^n K \left( \frac{X_i - x}{h} \right) \left[ (q(x) - \theta)^\top V_\ell \left( \frac{X_i - x}{h} \right) \right]^2 \\ &= \operatorname{argmin}_{\theta \in \mathbb{R}^{\ell+1}} \sum_{i=1}^n K \left( \frac{X_i - x}{h} \right) (q(x) - \theta)^\top V_\ell \left( \frac{X_i - x}{h} \right) V_\ell^\top \left( \frac{X_i - x}{h} \right) (q(x) - \theta). \end{aligned}$$

Or, on rappelle que  $B_{nx} = (nh)^{-1} \sum_{i=1}^n V_\ell \left( \frac{X_i - x}{h} \right) V_\ell^\top \left( \frac{X_i - x}{h} \right) K \left( \frac{X_i - x}{h} \right)$ , donc

$$\hat{\theta}_n(x) = \operatorname{argmin}_{\theta \in \mathbb{R}^{\ell+1}} (q(x) - \theta)^\top B_{nx} (q(x) - \theta).$$

Ainsi, si  $B_{nx}$  est définie positive, on obtient  $\hat{\theta}_n(x) = q(x)$  (le point singulier est donné par l'équation  $-2B_{nx}(q(x) - \theta) = 0$ ). En considérant uniquement la première

coordonnée, on obtient  $\hat{r}_n^{LP(\ell)}(x) = Q(x)$ . De plus, on a vu que l'estimateur localement polynomial d'ordre  $\ell$  est un estimateur linéaire de poids donnés par (5.2), ce qui permet de conclure

$$Q(x) = \hat{r}_n^{LP(\ell)}(x) = \sum_{i=1}^n w_{n,i}(x) Y_i = \sum_{i=1}^n w_{n,i}(x) Q(X_i).$$

□

### 5.3 Biais et variance des estimateurs par polynômes locaux

On considère ici uniquement le modèle de régression à effets fixes sur  $[0; 1]$

$$Y_i = r(x_i) + \xi_i, 1 \leq i \leq n, x_i \in [0; 1],$$

et on s'intéresse au risque ponctuel en  $x \in [0; 1]$  fixé. On note plus simplement  $\hat{r}_n$  l'estimateur localement polynomial d'ordre  $\ell$ . On utilise les hypothèses suivantes.

**Hypothèse. (H1)**  $\exists \mu_0 > 0$  et  $n_0 \in \mathbb{N}$  tels que  $\forall n \geq n_0, \forall x \in [0; 1]$ , la plus petite valeur propre  $\mu_{\min}(B_{nx})$  vérifie  $\mu_{\min}(B_{nx}) \geq \mu_0$ .

L'hypothèse **(H1)** est une condition uniforme en  $x \in \mathbb{R}$  et  $n$  assez grand, qui renforce l'hypothèse  $B_{nx}$  définie positive.

**Hypothèse. (H2)**  $\exists a_0 > 0$  telle que  $\forall A \subset [0; 1]$  et  $\forall n \geq 1$ ,

$$\frac{1}{n} \sum_{i=1}^n 1_{x_i \in A} \leq a_0 \max \left( \lambda(A); \frac{1}{n} \right),$$

où  $\lambda(A)$  est la mesure de Lebesgue de l'ensemble  $A$ .

L'hypothèse **(H2)** est une hypothèse qui porte sur la répartition des effets fixes : ils doivent être « assez uniformément » répartis sur  $[0; 1]$ .

**Hypothèse. (H3)**  $K$  est à support compact dans  $[-1; 1]$  avec  $\|K\|_\infty = K_{\max} < +\infty$ .

Les deux hypothèses **(H1)**, **(H2)** seront reformulées dans des cas particulier ci-dessous. L'hypothèse **(H3)** n'est pas contraignante.

**Contrôle du biais.** Comme d'habitude, il faut supposer que  $r$  a une certaine régularité pour contrôler le biais de  $\hat{r}_n$ . Dans la suite, on suppose donc que  $r \in \Sigma(\beta, L)$  sur  $[0; 1]$  (classe de Sobolev sur l'intervalle  $[0; 1]$ ), pour certaines constantes  $\beta, L > 0$ .

On a alors

$$b(x) = \mathbb{E}_r \hat{r}_n(x) - r(x) = \sum_{i=1}^n w_{n,i}(x) \mathbb{E}_r(Y_i) - r(x).$$

Or on sait d'après (5.3) que  $\sum_{i=1}^n w_{n,i}(x) = 1$  et  $\mathbb{E}_r(Y_i) = r(x_i)$  ce qui donne

$$b(x) = \sum_{i=1}^n w_{n,i}(x) [r(x_i) - r(x)].$$

On effectue un développement limité de  $r$  au voisinage de  $x$  et on utilise également (voir (5.3)) que pour tout  $1 \leq k \leq \ell$ , on a  $\sum_{i=1}^n (x_i - x)^k w_{n,i}(x) = 0$ , ce qui donne

$$\begin{aligned} b(x) &= \sum_{i=1}^n w_{n,i}(x) \left[ \sum_{k=1}^{\ell-1} \frac{r^{(k)}(x)}{k!} (x_i - x)^k + \frac{r^{(\ell)}(x + \eta_i(x_i - x))}{\ell!} (x_i - x)^\ell \right] \\ &= \sum_{i=1}^n w_{n,i}(x) \frac{r^{(\ell)}(x + \eta_i(x_i - x))}{\ell!} (x_i - x)^\ell \\ &= \sum_{i=1}^n w_{n,i}(x) [r^{(\ell)}(x + \eta_i(x_i - x)) - r^{(\ell)}(x)] \frac{(x_i - x)^\ell}{\ell!}, \end{aligned}$$

avec  $\eta_i \in ]0; 1[$ . En utilisant l'hypothèse  $r \in \Sigma(\beta, L)$ , on a donc

$$|b(x)| \leq \frac{L}{\ell!} \sum_{i=1}^n |w_{n,i}(x)| |x_i - x|^\beta.$$

Or, chaque poids  $w_{n,i}(x)$  est proportionnel au terme  $K(\frac{x_i - x}{h})$  et puisque  $K$  est à support sur  $[-1; 1]$ , s'annule pour  $|x_i - x| > h$ . On peut donc écrire

$$|b(x)| \leq \frac{L}{\ell!} \sum_{i=1}^n |w_{n,i}(x)| |x_i - x|^\beta \mathbf{1}_{|x_i - x| \leq h} \leq \frac{Lh^\beta}{\ell!} \sum_{i=1}^n |w_{n,i}(x)|. \quad (5.4)$$

**Lemme 26.** *Il existe une constante  $C^* > 0$  ne dépendant que de  $\mu_0, a_0, K_{max}$ , telle que pour tout  $x \in [0; 1]$ ,*

$$\sum_{i=1}^n |w_{n,i}(x)| \leq C^*.$$

*Démonstration.* En reprenant la définition (5.2) des  $w_{ni}$  et l'hypothèse **(H3)**, on a

$$|w_{n,i}(x)| \leq \frac{K_{\max}}{nh} 1_{|x_i-x| \leq h} \left| V_\ell^\top(0) B_{nx}^{-1} V_\ell \left( \frac{x_i - x}{h} \right) \right|.$$

Or, la norme euclidienne est sous-multiplicative, donc on peut écrire

$$\left| V_\ell^\top(0) B_{nx}^{-1} V_\ell \left( \frac{x_i - x}{h} \right) \right| \leq \|V_\ell^\top(0)\| \times \left\| B_{nx}^{-1} V_\ell \left( \frac{x_i - x}{h} \right) \right\| \leq \frac{1}{\mu_0} \left\| V_\ell \left( \frac{x_i - x}{h} \right) \right\|,$$

car  $\|V_\ell^\top(0)\| = 1$  et la plus grande valeur propre de  $B_{nx}^{-1}$  est majorée par  $1/\mu_0$  d'après **(H1)**. Ainsi,

$$\begin{aligned} |w_{n,i}(x)| &\leq \frac{K_{\max}}{nh\mu_0} \left\| V_\ell \left( \frac{x_i - x}{h} \right) \right\| 1_{|x_i-x| \leq h} \\ &\leq \frac{K_{\max}}{nh\mu_0} \sqrt{1 + 1 + \frac{1}{(2!)^2} + \cdots + \frac{1}{(\ell!)^2}} 1_{|x_i-x| \leq h} \\ &\leq \sqrt{1 + e} \frac{K_{\max}}{nh\mu_0} 1_{|x_i-x| \leq h} \leq 2 \frac{K_{\max}}{nh\mu_0} 1_{|x_i-x| \leq h}. \end{aligned} \quad (5.5)$$

Finalement,

$$\sum_{i=1}^n |w_{n,i}(x)| \leq 2 \frac{K_{\max}}{nh\mu_0} \sum_{i=1}^n 1_{|x_i-x| \leq h}.$$

Or, d'après l'hypothèse **(H2)**, on a

$$\frac{1}{n} \sum_{i=1}^n 1_{|x_i-x| \leq h} \leq a_0 \max \left( 2h; \frac{1}{n} \right),$$

donc

$$\sum_{i=1}^n |w_{n,i}(x)| \leq 2 \frac{K_{\max}}{nh\mu_0} a_0 \max \left( 2h; \frac{1}{n} \right) \leq 4 \frac{K_{\max}}{nh\mu_0},$$

dès que  $1/(nh) \leq 1$ . □

En conclusion, en reprenant (5.4) et le lemme précédent, on a montré le contrôle suivant sur le biais

$$|b(x)| \leq \frac{LC^* h^\beta}{\ell!}.$$

**Contrôle de la variance.** On note

$$\begin{aligned}\sigma^2(x) &= \mathbb{E}_r[\hat{r}_n(x) - \mathbb{E}_r\hat{r}_n(x)]^2 = \mathbb{E}_r\left[\left(\sum_{i=1}^n w_{ni}(x)Y_i - r(x_i)\right)^2\right] \\ &= \mathbb{E}_r\left[\left(\sum_{i=1}^n w_{ni}(x)(Y_i - r(x_i))\right)^2\right] = \sum_{i=1}^n w_{ni}(x)^2\sigma^2,\end{aligned}$$

car les résidus  $Y_i - r(x_i)$  sont i.i.d. de variance commune  $\sigma^2$ . Alors, d'après le lemme précédent et l'inégalité (5.5), on a

$$\sigma^2(x) \leq \sigma^2 \sup_{x \in \mathbb{R}, 1 \leq i \leq n} |w_{ni}(x)| \sum_{i=1}^n |w_{ni}(x)| \leq \sigma^2 \frac{2K_{\max}C^*}{nh\mu_0} = \frac{2\sigma^2 K_{\max}C^*}{\mu_0} \times \frac{1}{nh}.$$

En conclusion, on a montré le contrôle suivant sur la variance

$$\sigma^2(x) \leq \frac{2\sigma^2 K_{\max}C^*}{\mu_0} \times \frac{1}{nh}.$$

Ainsi, nous obtenons le théorème suivant.

**Théorème 27.** *Sous les hypothèses (H1), (H2) et (H3), dans le modèle de régression à effets fixes sur  $[0; 1]$ , et si  $r \in \Sigma(\beta, L)$  sur  $[0; 1]$ , l'estimateur localement polynomial d'ordre  $\ell = \lfloor \beta \rfloor$  vérifie*

$$\begin{aligned}\forall x \in \mathbb{R}, \quad MSE(x) &= \mathbb{E}_r[(\hat{r}_n(x) - r(x))^2] = b^2(x) + \sigma^2(x) \\ &\leq \left(\frac{LC^*}{\ell!}\right)^2 h^{2\beta} + \frac{2\sigma^2 K_{\max}C^*}{\mu_0} \times \frac{1}{nh}.\end{aligned}$$

De plus, en choisissant une fenêtre de la forme  $h^* = cn^{-1/(2\beta+1)}$  où  $c > 0$ , on obtient pour l'estimateur correspondant qu'il existe une constante  $C \in ]0; +\infty[$  telle que

$$\limsup_{n \rightarrow \infty} \sup_{r \in \Sigma(\beta, L)} \sup_{x \in [0; 1]} \mathbb{E}_r[n^{-2\beta/(2\beta+1)} |\hat{r}_n^*(x) - r(x)|^2] \leq C.$$

*Démonstration.* Les contrôles du biais et de la variance donnent immédiatement

$$\forall x \in \mathbb{R}, \quad MSE(x) \leq \left(\frac{LC^*}{\ell!}\right)^2 h^{2\beta} + \frac{2\sigma^2 K_{\max}C^*}{\mu_0} \times \frac{1}{nh}.$$

Puisque la borne de droite est uniforme en  $x \in [0; 1]$  et  $r \in \Sigma(\beta, L)$ , on en déduit

$$\sup_{r \in \Sigma(\beta, L)} \sup_{x \in [0; 1]} \mathbb{E}_r[|\hat{r}_n(x) - r(x)|^2] \leq \left(\frac{LC^*}{\ell!}\right)^2 h^{2\beta} + \frac{2\sigma^2 K_{\max}C^*}{\mu_0} \times \frac{1}{nh}.$$

On cherche ensuite à minimiser la borne de droite par rapport à  $h > 0$ , ce qui conduit à sélectionner  $h^* = cn^{-1/(2\beta+1)}$  avec  $c > 0$ , et qui donne ensuite le résultat voulu.  $\square$

**Retour sur les hypothèses (H1) et (H2) dans le cas d'un dispositif expérimental uniforme sur  $[0; 1]$ .** On se place dans le cas particulier d'un dispositif expérimental uniforme sur  $[0; 1]$ . Ainsi, si  $x_i = i/n, 1 \leq i \leq n$ .

On montre tout d'abord que l'hypothèse (H2) est vérifiée. En effet,  $|\{i; x_i \in A\}| \leq n\lambda(A) + 1$  donc  $\frac{1}{n} \sum_{i=1}^n 1_{x_i \in A} \leq \lambda(A) + 1/n \leq 2 \max(\lambda(A); 1/n)$ .

Concernant (H1), on peut montrer (admis) que

$$B_{nx} \xrightarrow{n \rightarrow \infty} B = \int V_\ell(u) V_\ell^\top(u) K(u) du.$$

[L'idée qui sous tend cette convergence est la suivante : si  $h$  est fixé, il est facile de voir que

$$\begin{aligned} B_{nx} &= \frac{1}{nh} \sum_{i=1}^n V_\ell \left( \frac{i/n - x}{h} \right) V_\ell^\top \left( \frac{i/n - x}{h} \right) K \left( \frac{i/n - x}{h} \right) \\ &\xrightarrow{n \rightarrow \infty} \int_0^1 V_\ell \left( \frac{u - x}{h} \right) V_\ell^\top \left( \frac{u - x}{h} \right) K(u) du, \end{aligned}$$

puis un changement de variable donne le résultat. Cependant, comme  $h \rightarrow 0$  avec  $n$ , il faut montrer une convergence uniforme.]

Alors, si  $n$  est assez grand, il suffit de vérifier que la plus petite valeur propre de la limite  $B$ , notée  $\mu_{\min}(B)$  est strictement positive, *i.e.* que  $B$  est définie positive.

Or, si on suppose que le noyau  $K$  est tel que  $\lambda(\{u; K(u) > 0\}) > 0$ , *i.e.*  $K$  est non nul sur un ouvert, alors on montre que  $B$  est définie positive. En effet, pour tout  $v \in \mathbb{R}^{\ell+1}$ , on a

$$v^\top B v = \int [v^\top V_\ell(u) V_\ell^\top(u) v] K(u) du = \int [v^\top V_\ell(u)]^2 K(u) du \geq 0,$$

et si  $v^\top B v = 0$  alors  $u \rightarrow v^\top V_\ell(u)$  est une fonction nulle sur le support  $\mathcal{S} = \{u \in \mathbb{R}, K(u) > 0\}$  du noyau  $K$ . Or, d'après la définition du vecteur  $V_\ell$ , la fonction  $u \rightarrow v^\top V_\ell(u)$  est un polynôme et  $\mathcal{S}$  contient un intervalle ouvert, donc  $u \rightarrow v^\top V_\ell(u)$  est nécessairement une fonction nulle et donc  $v = 0$ .

## 5.4 Estimateurs par projection

On se place à nouveau dans le cadre de la régression à effets fixes sur  $[0; 1]$ . On suppose à présent que la fonction de régression  $r$  vérifie  $r \in \mathbb{L}_2([0; 1])$ . On considère  $(\phi_j)_{j \geq 1}$  une b.o.n. de  $\mathbb{L}_2([0; 1])$ . On peut alors écrire

$$r = \sum_{j \geq 1} \theta_j \phi_j,$$

au sens d'une série convergente dans  $\mathbb{L}_2([0;1])$ , et où  $\theta_j = \int_0^1 r(x)\phi_j(x)dx$  est la projection de  $r$  sur la  $j$ ème coordonnée de la base.

Si  $\{\hat{\theta}_j\}_{j \geq 1}$  est une suite d'estimateurs des coordonnées  $\{\theta_j\}_{j \geq 1}$ , alors on peut définir un estimateur par projection de  $r$  via

$$\hat{r}_{n,N} = \sum_{j=1}^N \hat{\theta}_j \phi_j,$$

*i.e.* on prend la projection de  $r$  sur les  $N$  premières coordonnées de la base on estime ses coordonnées. Ici,  $N$  joue le rôle d'un paramètre de lissage, comme  $h$  auparavant.

**Exemple.** Prenons le cas du dispositif fixe uniforme sur  $[0;1]$ . Alors on observe

$$Y_i = r(i/n) + \xi_i, 1 \leq i \leq n,$$

et les coordonnées de  $r$  sur la base  $\{\phi_j\}_{j \geq 1}$  sont données par

$$\theta_j = \int_0^1 r(x)\phi_j(x)dx \simeq \frac{1}{n} \sum_{i=1}^n r(i/n)\phi_j(i/n),$$

donc un estimateur naturel de  $\theta_j$  est

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^n Y_i \phi_j(i/n),$$

ce qui donne pour estimateur de la régression

$$\hat{r}_{n,N}(x) = \frac{1}{n} \sum_{i=1}^n Y_i \left( \sum_{j=1}^N \phi_j(i/n)\phi_j(x) \right).$$

On peut noter que c'est un estimateur linéaire.

Les bases de  $\mathbb{L}_2([0;1])$  les plus utilisées sont la base trigonométrique et les bases d'ondelettes.

**Base trigonométrique (de Fourier).** Elle est donnée par

$$\phi_1 \equiv 1, \quad \phi_{2k} : x \rightarrow \sqrt{2} \cos(2\pi kx), \quad \phi_{2k+1} : x \rightarrow \sqrt{2} \sin(2\pi kx), \forall k \geq 1.$$

**Bases d'ondelettes** Soit  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  une fonction suffisamment régulière, à support compact. On définit  $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$  pour tous  $k, j \in \mathbb{Z}$ . Alors, sous certaines hypothèses sur  $\psi$ , les fonctions  $\{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$  forment une b.o.n. de  $\mathbb{L}_2(\mathbb{R})$  et pour tout  $h \in \mathbb{L}_2(\mathbb{R})$ , on a

$$h = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \theta_{j,k} \psi_{j,k},$$

où  $\theta_{j,k} = \int h \psi_{j,k}$  et la série ci-dessus converge dans  $\mathbb{L}_2(\mathbb{R})$ .

Pour une fonction de régression, on ne fera jamais l'hypothèse  $r \in \mathbb{L}_2(\mathbb{R})$  qui n'est pas raisonnable. Par contre, on peut supposer  $r \in \mathbb{L}_2([0; 1])$ , et on peut aussi une base d'ondelettes sur  $\mathbb{L}_2([0; 1])$  par le même procédé que ci-dessus, mais en faisant également des corrections aux bords.

Une base d'ondelettes est constituée de deux indices  $j$  pour l'échelle (=fréquence) et  $k$  pour la translation (=temps). La base trigonométrique localise les fonctions en fréquence tandis que les bases d'ondelettes localisent les fonctions en fréquence et en temps.