

# Introduction à la statistique non paramétrique

Catherine MATIAS

CNRS, Laboratoire Statistique & Génome, Évry

<http://stat.genopole.cnrs.fr/~cmatias>

ENSIIE - 2013/2014



# Première partie I

## Introduction à la statistique non paramétrique

# Statistique non paramétrique : c'est quoi ?

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

- ▶ Observations réelles avec un seul mode :

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

- ▶ Observations réelles avec plusieurs modes :

$\mathcal{M}_K = \{\sum_{i=1}^K p_i \mathcal{N}(\mu_i, \sigma^2), (p_1, \dots, p_K) \in (0, 1)^K, \sum_i p_i = 1, (\mu_1, \dots, \mu_K) \in \mathbb{R}^K, \sigma^2 \in \mathbb{R}^{+\star}\}$ , modèle de mélange Gaussien.

- ▶ Observations de comptage :  $\mathcal{M} = \{\mathcal{P}(\lambda); \lambda \in \mathbb{R}^{+\star}\}$ , modèle loi Poisson.

- ▶  $\mathcal{M} = \{\mathbb{P} \text{ à support dans } \mathcal{S} \text{ fini}\} \simeq [0, 1]^{|\mathcal{S}|-1}$ .

# Statistique non paramétrique : c'est quoi ?

Par opposition, en statistique **non paramétrique**, le modèle n'est pas décrit par un nombre fini de paramètres.

Divers cas de figures 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.
- ▶ On travaille sur des **espaces fonctionnels**, de dimension infinie. Exemple : les densités continues sur  $[0, 1]$ , ou les densités monotones sur  $\mathbb{R}$ .
- ▶ Le nombre de paramètres du modèle n'est pas fixé et **varie** (augmente) avec le nombre d'observations.
- ▶ Le support de la distribution est discret et **varie** (augmente) avec le nombre d'observations.

# Motivations I

## Observations rangées

Situation :

- ▶ On dispose des résultats de questionnaires où des échantillons de consommateurs ont classé un ensemble de produits par ordre de préférence,
- ▶ Les questionnaires proviennent de supermarchés situés dans des zones socio-économiques différentes,
- ▶ On se demande si un produit  $P$  obtient un classement significativement différent d'un supermarché à l'autre.

Questions :

- ▶ Comment modéliser la distribution des observations ?
- ▶ Quel test utiliser ?

Réponse : **Tests de rang.**

# Motivations II

## Observations mesurées

Situation : On observe des données quantitatives.

Questions :

- ▶ Peut-on raisonnablement supposer que les observations suivent une loi normale ? (par exemple pour faire des tests sur la moyenne). Rép : **Tests de normalité**.
- ▶ Combien de modes possède cette distribution ? Rép : **Estimation de densité**.

# Statistique non paramétrique : Quand l'utiliser ?

## Exemples de contextes d'utilisation

- ▶ Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique,
- ▶ Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle,
- ▶ Quand on ne sait pas combien de composantes on veut mettre dans un mélange,
- ▶ Quand le nombre de variables est trop grand (problème de grande dimension) et qu'un modèle paramétrique est non utilisable car il aurait de toutes façons trop de paramètres,
- ▶ ...

# Avantages/Inconvénients

## Avantages

- ▶ Moins d'a priori sur les observations,
- ▶ Modèles plus généraux, donc plus robustes au modèle.

## Inconvénients

- ▶ Vitesses de convergence **plus lentes** = il faut **plus de données** pour obtenir une précision équivalente.



## Quelques références bibliographiques pour ce cours



L. Wasserman.

*All of nonparametric statistics.*

Springer Texts in Statistics. Springer-Verlag, 2006.



E.L. Lehmann.

*Elements of large sample theory.*

Springer Texts in Statistics. Springer-Verlag, 1999.



A. B. Tsybakov.

*Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*.

Springer-Verlag, Berlin, 2004.



D. Bosq.

*Nonparametric statistics for stochastic processes*,

Springer-Verlag, 1996.

## Deuxième partie II

### Fonctions de répartition et fonctionnelles de la distribution

# Sommaire Deuxième partie

Fonction de répartition empirique

Fonctionnelles de la distribution

Fonction d'influence

# Estimer une fonction de répartition

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  $F$ .

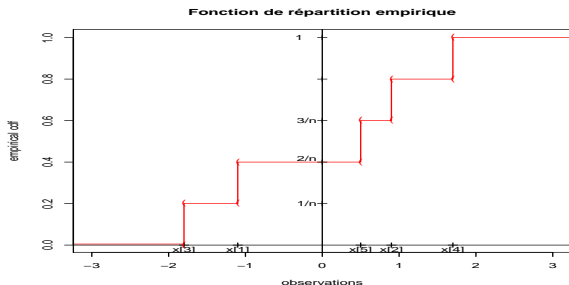


Figure : Fonction de répartition empirique.

→ Qualité de cet estimateur ?

# Propriétés ponctuelles de $\hat{F}_n(x)$ (i.e. $x$ fixé) I

- Biais

$$\mathbb{E}\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{P}(X_i \leq x) = F(x),$$

i.e. estimateur **sans biais**.

- Variance

$$\begin{aligned} \text{Var}(\hat{F}_n(x)) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(1_{X_i \leq x}) = \frac{1}{n} \text{Var}(1_{X_1 \leq x}) \\ &= \frac{F(x)(1 - F(x))}{n} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

- Erreur en moyenne quadratique (ou MSE pour "mean square error")

$$\mathbb{E}[(\hat{F}_n(x) - F(x))^2] = \text{biais}^2 + \text{variance} = \text{Var}(\hat{F}_n(x)) \xrightarrow{n \rightarrow \infty} 0.$$

## Propriétés ponctuelles de $\hat{F}_n(x)$ (i.e. $x$ fixé) II

- Convergence en probabilité

$$\hat{F}_n(x) \xrightarrow[n \rightarrow \infty]{\text{proba}} F(x).$$

En effet, d'après l'inégalité de Markov, la convergence en moyenne quadratique implique la convergence en probabilité.

$$\forall \epsilon > 0, \quad \mathbb{P}(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq \frac{\text{Var}(\hat{F}_n(x))}{\epsilon^2} \xrightarrow[n \rightarrow \infty]{} 0.$$

- LGN :

$$\hat{F}_n(x) \xrightarrow[n \rightarrow \infty]{\text{p.s.}} F(x).$$

- TCL :

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

## Propriétés ponctuelles de $\hat{F}_n(x)$ (i.e. $x$ fixé) III

- **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

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

# Propriétés uniformes de $\hat{F}_n$

- ▶ Théorème de Glivenko Cantelli

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

- ▶ Inégalité de Dvoretzky-Kiefer-Wolfowitz (DKW)

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



# Exemple d'application de l'inégalité de DKW I

Construction d'intervalles de confiance (IC) exacts sur  $F(x)$

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

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

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

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

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  $F(x)$ .

# Exemple d'application de l'inégalité de DKW II

## Remarques

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

# Sommaire Deuxième partie

Fonction de répartition empirique

Fonctionnelles de la distribution

Fonction d'influence

# Notations I

- ▶ On note  $\mathcal{F}$  l'ensemble des fonctions de répartition (fdr)

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

- ▶ Si  $F \in \mathcal{F}$ , on note  $dF$  l'**unique mesure de proba** associée et on peut définir ainsi la notation  $\int h(x)dF(x) = \mathbb{E}_F(h(X))$  pour toute fonction  $h : \mathbb{R} \rightarrow \mathbb{R}$ .
- ▶ Exemples :
  - ▶ si  **$F$  continue** (*i.e.* si  $dF$  est une mesure absolument continue) alors en notant  $f = F'$  la densité on obtient 
$$\int h(x)dF(x) = \int h(x)f(x)dx$$
  - ▶ si  **$F$  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 à chaque point du support.

## Notations II

### FDR empirique $\hat{F}_n$

Soient  $X_1, \dots, X_n$  v.a.i.i.d. réelles

- ▶  $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n 1_{X_i \leq t}, \quad \forall t \in \mathbb{R}.$
- ▶ Constante par morceaux (croissante, càd-làg)
- ▶ Associée à la mesure empirique

$$\mathbb{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(\cdot)$$

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  $X_i$ .

- ▶ Pour toute fonction  $h$ , on a  $\int h(x) d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n h(X_i).$

# Fonctionnelles de la distribution

Une fonctionnelle est une application  $T : \mathcal{F} \rightarrow \mathbb{R}$ .

## Exemples

- ▶ **Moyenne** :  $F \rightarrow \mu(F) = \int x dF(x)$  ,
- ▶ **Variance** :  $F \rightarrow \sigma^2(F) = \int (x - \mu(F))^2 dF(x) = \int x^2 dF(x) - (\int x dF(x))^2$ ,
- ▶ **Médiane** :  $F \rightarrow m(F) = F^{-1}(1/2)$  et **Quantiles** :  
 $F \rightarrow q_\alpha(F) = F^{-1}(\alpha)$ ,
- ▶ **Skewness** (ou coefficient d'asymétrie) :  
 $F \rightarrow \{\int (x - \mu(F))^3 dF(x)\} / \sigma(F)^{3/2}$ .
- ▶  $\mathbb{E}(|X_1 - X_2|)$ ,  $\mathbb{P}((X_1, X_2) \in S)$ , ...

# Cas particuliers : fonctionnelles linéaires et fonctionnelles de moment

## Définitions

- ▶ Une fonctionnelle  $T$  est dite **linéaire** s'il existe  $a : \mathbb{R} \rightarrow \mathbb{R}$  telle que  $T : F \rightarrow T(F) = \int a(x) dF(x)$ .
- ▶ Une fonctionnelle  $T$  est dite **de moment** s'il existe un entier  $k \geq 1$  et une fonction  $\phi : \mathbb{R}^k \rightarrow \mathbb{R}$  telle que  $T : F \rightarrow T(F) = \mathbb{E}_F(\phi(X_1, \dots, X_k)) = \int \phi(x_1, \dots, x_k) dF(x_1) \dots dF(x_k)$ .

## Exemples

- ▶ Les fonctionnelles linéaires sont des fonctionnelles de moment.
- ▶ Moyenne : linéaire et de moment
- ▶  $\mathbb{E}(|X_1 - X_2|)$ ,  $\mathbb{P}((X_1, X_2) \in S)$  : fonctionnelles de moment.
- ▶ Variance, médiane, quantiles, skewness : NON

# Estimateurs par substitution I

## Principe des estimateurs par substitution (ou "plug-in")

Si  $T : F \rightarrow T(F)$  est une fonctionnelle alors un estimateur naturel de  $T(F)$  est obtenu en substituant 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

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

► Variance empirique :

$$\begin{aligned}\hat{\sigma}_n^2 &= \int x^2 d\hat{F}_n(x) - \left( \int x d\hat{F}_n(x) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.\end{aligned}$$



## Estimateurs par substitution II

- ▶ **Variance empirique** (suite) : Estimateur biaisé auquel on peut préférer

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

qui est sans biais.

- ▶ **Médiane empirique**  $\hat{m} = \hat{F}_n^{-1}(1/2)$ ,

- ▶ **Quantile empirique**  $\hat{q}_\alpha = \hat{F}_n^{-1}(\alpha)$ .

- ▶ **Statistiques d'adéquation**

- ▶ **Kolmogorov** :  $\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$ ,

- ▶ **Cramer von Mises** :  $\int (\hat{F}_n(x) - F_0(x))^2 dF_0(x)$ ,

- ▶ **Pearson** :  $\sum_{j=1}^r \frac{(\hat{p}_j - p_j^0)^2}{p_j^0}$  où  $p_j^0 = F_0((a_j; a_{j+1}])$  et

$\hat{p}_j = \hat{F}_n((a_j; a_{j+1}])$ . Correspond à

$$T(F) = \sum_{j=1}^r \frac{(F(a_{j+1}) - F(a_j) - p_j^0)^2}{p_j^0}.$$

# $U$ et $V$ statistiques I

Estimateurs des fonctionnelles de moment

## $U$ et $V$ statistiques II

Soit  $T = \mathbb{E}(\phi(X_1, \dots, X_k))$  une fonctionnelle de moment.  
(On peut supposer  $\phi$  symétrique en les coordonnées).

- ▶ Son estimateur **de substitution** est la  $V$ -statistique

$$V = T(\hat{F}_n) = \frac{1}{n^k} \sum_{i_1=1}^n \dots \sum_{i_k=1}^n \phi(X_{i_1}, \dots, X_{i_k}).$$

- ▶ Un autre estimateur **sans biais** de  $T$  est la  $U$ -statistique

$$U = \binom{n}{k}^{-1} \sum_{1 \leq i_1 < i_2 < \dots < i_k \leq n} \phi(X_{i_1}, \dots, X_{i_k}).$$

- ▶ La  $U$ -stat et la  $V$ -stat correspondante ont le même comportement **asymptotique** et ne diffèrent que par des extra-termes dans la  $V$ -stat et des facteurs de normalisations différents.

# $U$ et $V$ statistiques III

## Exemples

- ▶ Ex :  $k = 2$ ,

$$U = \frac{2}{n(n-1)} \sum \sum_{i < j} \phi(X_i, X_j) = \frac{1}{n(n-1)} \sum \sum_{i \neq j} \phi(X_i, X_j)$$

et  $V = \frac{1}{n^2} \sum \sum_{i \neq j} \phi(X_i, X_j) + \frac{1}{n^2} \sum_{i=1}^n \phi(X_i, X_i)$ .

- ▶ **Différence en moyenne de Gini** :  $U = \binom{n}{2}^{-1} \sum_{i < j} |X_i - X_j|$ ,  
(mesure la dispersion de la distribution).

## Propriétés des $U$ -statistiques

- ▶ Estimateurs **sans biais**.

- ▶ **Variance** :  $\text{Var}(\sqrt{n}U) \rightarrow k^2\sigma_1^2$  où  
 $\sigma_1^2 = \text{Cov}(\phi(X, X_2, \dots, X_k); \phi(X, X'_2, \dots, X'_k))$  et  
 $X, X_2, \dots, X_k, X'_2, \dots, X'_k$  i.i.d. de loi  $F$ .

- ▶ Si  $\sigma_1^2 \in ]0, +\infty[$ , alors  $\sqrt{n}(U - T(F)) \rightsquigarrow_{n \rightarrow \infty}^{\mathcal{L}} \mathcal{N}(0, k^2\sigma_1^2)$  et  
 $\sqrt{n}(V - T(F)) \rightsquigarrow_{n \rightarrow \infty}^{\mathcal{L}} \mathcal{N}(0, k^2\sigma_1^2)$ , *i.e. asympt. gaussiens*.

# Consistance des estimateurs par substitution I

## Principe

- ▶ On a vu que  $\hat{F}_n$  converge (de diverses façons) vers  $F$ .
- ▶ Si  $T$  est assez **régulière**, les propriétés de  $\hat{F}_n$  se transmettent à  $\hat{T}_n = T(\hat{F}_n)$ .
- ▶ Attention :  $T$  est définie sur  $\mathcal{F}$  : notion de régularité à préciser.

## Rappel : Lemme de Slutsky

Soit  $(X_n)_{n \geq 0}$  suite de v.a. dans  $\mathbb{R}^d$  qui converge en loi vers  $X$  et  $h : \mathbb{R}^d \rightarrow \mathbb{R}^s$  continue. Alors  $(h(X_n))_{n \geq 0}$  suite de v.a. dans  $\mathbb{R}^s$  qui converge en loi vers  $h(X)$ .

## Continuité d'une fonctionnelle

Un fonctionnelle  $T$  est **continue** au point  $F$  si

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \rightarrow 0 \Rightarrow T(F_n) - T(F) \rightarrow 0.$$

# Consistance des estimateurs par substitution II

## Exemples de fonctionnelles continues

- ▶ Fdr en un point  $T : F \mapsto F(x_0)$
- ▶ Cramer von Mises  $T : F \mapsto \int (F - F_0)^2 dF_0$
- ▶ Quantiles  $T : F \mapsto F^{-1}(\alpha)$
- ▶ **Contre-exemple** : La moyenne n'est pas continue. En général, les fonctionnelles de moment ne sont pas continues.

## Convergence de l'estimateur plug-in (Condition suffisante)

Si  $T : \mathcal{F} \rightarrow \mathbb{R}$  est continue alors  $\hat{T}_n = T(\hat{F}_n)$  converge en proba vers  $T(F)$ .

# Consistance des estimateurs par substitution III

## Exemples d'estimateurs par substitution consistants

- ▶ Moyenne empirique :  $T : F \mapsto \int x dF(x)$  n'est pas continue en tout point mais on a quand même  $\bar{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}(X)$ ,
- ▶ Variance empirique :  $\hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \text{Var}(X)$
- ▶ Médiane empirique  $\hat{m} = \hat{F}_n^{-1}(1/2) \xrightarrow{\mathbb{P}} F^{-1}(1/2)$ ,
- ▶ Quantile empirique  $\hat{q}_\alpha = \hat{F}_n^{-1}(\alpha) \xrightarrow{\mathbb{P}} F^{-1}(\alpha)$ .

# Normalité asymptotique des estimateurs par substitution I

## Rappel : 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  $a_n(X_n - \mu) \underset{n \rightarrow \infty}{\rightsquigarrow}^{\mathcal{L}} \mathcal{N}_d(0, \Sigma)$  et si  $g : \mathbb{R}^d \rightarrow \mathbb{R}^s$  est différentiable au voisinage de  $\mu$ , alors

$$a_n(g(X_n) - g(\mu)) \underset{n \rightarrow \infty}{\rightsquigarrow}^{\mathcal{L}} \mathcal{N}_s(0, \nabla g(\mu)^\top \cdot \Sigma \cdot \nabla g(\mu)).$$

## Exemple d'application directe : variance empirique

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

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

## Dérivabilité d'une fonctionnelle

C'est la notion de **fonction d'influence**.



# Sommaire Deuxième partie

Fonction de répartition empirique

Fonctionnelles de la distribution

Fonction d'influence

# Fonction d'influence I

## Principe

- ▶ C'est une dérivée de la fonctionnelle  $T$ .
- ▶ 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  $\mathcal{F}$ .

## Accroissement élémentaire

$\forall x_0 \in \mathbb{R}$ , on note  $\delta_{x_0}$  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  $t \in \mathbb{R}$ .

## Fonction d'influence II

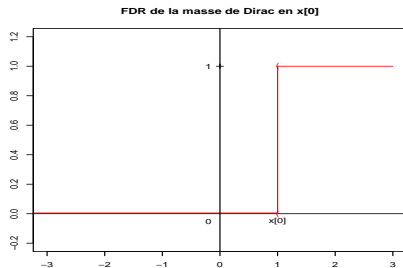


Figure : Fonction de répartition  $G_{\delta_{x_0}}$  de la masse de Dirac en  $x_0$ .

# Fonction d'influence III

## Définition

Soit  $T : F \rightarrow T(F)$  une fonctionnelle. La fonction d'influence de  $T$  en  $F$  au point  $x_0$  est définie par la limite suivante, si elle existe pour tout  $x \in \mathbb{R}$ ,

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

## 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  $+\infty$ .

## Heuristique

- ▶  $IF_{T,F}(x_0)$  mesure la variation limite subie par la fonctionnelle  $T$  en la distribution  $F$  lors d'une **perturbation infinitésimale**.
- ▶ Notion de contamination, liée à la **robustesse** statistique.

# Fonction d'influence IV

## Exemple de la moyenne (fonctionnelle linéaire)

Moyenne  $\mu : F \rightarrow \mu(F) = \int x dF(x)$ .

$\forall \epsilon > 0, \forall x_0 \in \mathbb{R}$ , 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,  $IF_{\mu, F}(x_0) = x_0 - \mu(F)$ .

**Plus généralement**, on peut facilement voir que pour une fonctionnelle linéaire  $T(F) = \int a(x) dF(x)$  on a

$$IF_{T, F}(x) = a(x) - T(F).$$

# Fonction d'influence empirique

## Définition

Soit  $T : F \rightarrow T(F)$  une fonctionnelle. La **fonction d'influence empirique** de  $T$  en  $F$  au point  $x_0$  est

$$\hat{IF}_n(x_0) = IF_{T, \hat{F}_n}(x_0).$$

## Exemple (suite)

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

## Heuristique

La quantité  $\hat{IF}_n(X_i)$  mesure la contribution de l'observation  $X_i$  à la variation de la statistique  $\hat{T}_n$ .

# Calcul de fonctions d'influence

## Proposition

1. Si  $T, S : \mathcal{F} \rightarrow \mathbb{R}$  sont deux fonctionnelles de fonctions d'influence respectives  $IF_{T,F}$  et  $IF_{S,F}$  au point  $F \in \mathcal{F}$  et si  $\lambda \in \mathbb{R}$  alors  $\lambda T + S$  est une fonctionnelle de fonction d'influence  $\lambda IF_{T,F} + IF_{S,F}$  au point  $F$  et  $T \times S$  est une fonctionnelle de fonction d'influence  $T(F) \times IF_{S,F} + S(F) \times IF_{T,F}$  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  $IF_{T,F}$  au point  $F$  alors la fonctionnelle  $S = \psi \circ T$  a pour fonction d'influence au point  $F$  la fonction  $IF_{S,F} = \psi' \circ T \times IF_{T,F}$ .

## Application : calcul de 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 prop. précédente  $IF_{\sigma^2, F} = IF_{T, F} - 2\mu(F)IF_{\mu, F}$  où  $T : F \rightarrow T(F) = \int x^2 dF(x)$ .
- ▶ Or  $IF_{\mu, F}(x) = x - \mu(F)$  et  $T$  est aussi une fonctionnelle linéaire donc  $IF_{T, F}(x) = x^2 - T(F) = x^2 - \int u^2 dF(u)$ .
- ▶ Donc on obtient

$$\begin{aligned}IF_{\sigma^2, F}(x) &= x^2 - \int u^2 dF(u) - 2\mu(F)(x - \mu(F)) \\ &= (x - \mu(F))^2 - \sigma^2(F).\end{aligned}$$

- ▶ **NB** : On retrouve la forme  $IF_{\sigma^2, F}(x) = a_F(x) - \sigma^2(F)$  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  $F$ .



# Normalité asymptotique de $\hat{T}_n$ et construction d'IC. I

Exemple de la moyenne  $\mu(F) = \int x dF(x)$

- ▶ D'après le TCL

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

- ▶ On remarque que comme  $IF_{\mu, F}(X) = X - \mu(F)$ , on a  $\sigma^2(F) = \text{Var}(X) = \text{Var}(IF_{\mu, F}(X))$ .
- ▶ Mais  $\sigma^2(F)$  est inconnue. On l'estime par  $\hat{\sigma}_n^2 = \sigma^2(\hat{F}_n)$ , ce qui revient à estimer  $\sigma^2(F)$  par  $\text{Var}(\hat{IF}_n(X))$ .
- ▶ Au final, le Lemme de Slutsky combiné au TCL donne

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

## Normalité asymptotique de $\hat{T}_n$ et construction d'IC. II

Théorème (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) La fonction d'influence s'obtient simplement via

$$IF_{T,F}(x_0) = a(x_0) - T(F) \text{ et s'estime par}$$

$$\hat{IF}_n(x_0) = a(x_0) - T(\hat{F}_n) = a(x_0) - \frac{1}{n} \sum_{i=1}^n a(X_i),$$

ii) On a  $\mathbb{E}(IF_{T,F}(X)) = \int IF_{T,F}(x) dF(x) = 0$ , et

$\text{Var}(IF_{T,F}(X)) = \mathbb{E}(IF_{T,F}(X)^2) := \tau^2$  s'obtient via

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

De plus, si  $\tau^2 < +\infty$ , alors

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

## Normalité asymptotique de $\hat{T}_n$ et construction d'IC. III

iii) On estime  $\tau^2$  via

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

$$\hat{\tau}_n^2 \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \tau^2,$$

et par conséquent

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

# Normalité asymptotique de $\hat{T}_n$ et construction d'IC. IV

## Autres cas

- ▶ Pour les fonctionnelles de moment, on a vu la normalité asymptotique des  $V$ -statistiques (estimateurs plug-in) mais aussi des  $U$ -statistiques.
- ▶ En général (fonctionnelles ni linéaires, ni de moment), il faut démontrer la normalité asymptotique **à la main**.

## Conclusion

- ▶ **Souvent**, la variance asymptotique de l'estimateur  $T(\hat{F}_n)$  vaut l'espérance du carré de la fonction d'influence (c'est le cas pour les fonctionnelles de moment).
- ▶ Calculer la fonction d'influence permet donc d'avoir la variance de l'estimateur par substitution  $T(\hat{F}_n)$ .

# Liens avec les statistiques robustes [Hampel 74] I

## Statistiques résumées de la fonction d'influence

Estimateur par substitution  $\hat{T}_n = T(\hat{F}_n)$ .

- ▶ **Variance asymptotique** :  $\mathbb{E}(IF_{T,F}(X)^2)$ ,
- ▶ **Gross-error sensitivity** :  $\gamma^* = \sup_{x \in \mathbb{R}} |IF_{T,F}(x)|$ .
  - ▶ Mesure la pire influence approchée qu'un niveau de contamination fixé peut avoir sur la valeur de l'estimateur. On peut le voir comme une borne approchée du biais de l'estimateur.
  - ▶ Si  $\gamma^*$  est borné, alors la fonctionnelle  $T$  est **robuste aux valeurs aberrantes**. Ex : la médiane, mais pas la moyenne.
  - ▶ **Rem** : Les méthodes de robustification d'un estimateur cherchent souvent à mettre une borne sur  $\gamma^*$ . Le prix à payer est une augmentation de la variance limite.

# Liens avec les statistiques robustes [Hampel 74] II

- ▶ Local shift sensitivity

$$\lambda^* = \sup_{x \neq y} \frac{|IF_{T,F}(x) - IF_{T,F}(y)|}{|x - y|}$$

Mesure par exemple les effets locaux de l'arrondi ou du regroupement de valeurs sur la fonctionnelle  $T$ .

# Liens avec estimateur Jackknife I

## Principe du Jackknife

- ▶ On observe un  $n$ -échantillon. Estimateur initial  $\hat{\theta}_n^0$  de  $\theta$ .
- ▶ Pour  $1 \leq i \leq n$ , on construit  $\hat{\theta}_{n-1}^{(i)}$  le même estimateur de  $\theta$  sur les observations privées de la  $i$ ème.
- ▶ On forme les **pseudo-valeurs**  $\hat{\theta}^{*,i} = n\hat{\theta}_n^0 - (n-1)\hat{\theta}_{n-1}^{(i)}$ .
- ▶ Estimateur Jackknife  $\hat{\theta} = n^{-1} \sum_{i=1}^n \hat{\theta}^{*,i}$  a un **biais en général réduit**.
- ▶ Le jackknife fait plus globalement partie des méthodes de ré-échantillonnage, comme le bootstrap.
- ▶ On peut le généraliser au cas où au lieu d'enlever une valeur, on en retire  $k$ .

# Liens avec estimateur Jackknife II

## Exemples

- ▶ Si  $\theta = \mathbb{E}(X)$  et on applique cette stratégie sur  $\hat{\theta}_n^0 = \bar{X}_n$  moyenne empirique, alors la procédure ne change pas l'estimateur.
- ▶ Si  $\theta = \text{Var}(X)$  et on applique cette stratégie sur  $\hat{\theta}_n^0 = n^{-1} \sum_i (X_i - \bar{X}_n)^2$  variance empirique, alors on obtient

$$\hat{\theta} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$



## Liens avec estimateur Jackknife III

### Jackknife et fonction d'influence empirique

En prenant  $\epsilon = -1/(n - 1)$  on a

$$\begin{aligned}\hat{IF}_n(X_i) &= IF_{T, \hat{F}_n}(X_i) \simeq \frac{T((1 - \epsilon)\hat{F}_n + \epsilon G_{\delta_{X_i}}) - T(\hat{F}_n)}{\epsilon} \\ &= (n - 1)[T(\hat{F}_n) - T(\hat{F}_{n-1}^{(i)})] = \hat{T}^{*,i} - T(\hat{F}_n).\end{aligned}$$

Le Jackknife peut-être vu comme une version à taille d'échantillon finie d'une fonction d'influence empirique

[Quenouille 56, Huber 72, Miller 64].

## Références partie fdr



[Hampel 74] Frank R. Hampel.

The influence curve and its role in robust estimation.

*J. Amer. Statist. Assoc.*, 69 :383–393, 1974.



[Huber 72] P. J. Huber.

The 1972 Wald lecture. Robust statistics : A review.

*Ann. Math. Statist.*, 43 :1041–1067, 1972. Miller,



[Miller 64] Miller.

A trustworthy jackknife.

*Ann. Math. Statist.*, 35 :1594–1605, 1964.



[Quenouille 56] M.H. Quenouille.

Notes on bias in estimation.

*Biometrika*, 43 :353–360, 1956.

# Troisième partie III

## Tests non paramétriques

# Sommaire Troisième partie

## Introduction, rappels et généralités

### Tests sur une population

Tests d'adéquation à une distribution fixée

Tests d'adéquation à une famille de distributions

Tests de médiane (ou de symétrie)

### Tests sur deux populations

Tests de comparaison (ou homogénéité) de deux populations

Tests de corrélation sur variables appariées

# Introduction I

## Contexte

- ▶ 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  $Y$ .
- ▶ Les tests sont **non paramétriques** lorsque la distribution des variables aléatoires n'est pas spécifiée sous au moins une des deux hypothèses (nulle ou alternative).

# Introduction II

## Exemples

- ▶ Tests d'adéquation à une loi :  $H_0$  : "X suit la loi  $F_0$ " contre  $H_1$  : "X ne suit pas la loi  $F_0$ ".
- ▶ Tests d'adéquation à une famille de lois :  $H_0$  : "X est gaussienne" (paramètres non spécifiés) contre  $H_1$  : "X n'est pas gaussienne".
- ▶ Tests de comparaison (ou homogénéité) :  $H_0$  : "X et Y ont la même loi" contre  $H_1$  : "X et Y n'ont pas la même loi".
- ▶ Tests d'indépendance :  $H_0$  :  $\{X_i\} \amalg \{Y_i\}$  contre  $H_1$  : " $\{X_i\}$  ne sont pas indépendants des  $\{Y_i\}$ ".

# Principe

## Principe général des tests

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**.

Deux types de tests

- ▶ **bilatères** : lorsque sous l'alternative  $H_1$ , la statistique  $T$  n'est ni systématiquement "plus grande" ni "plus petite" que sous  $H_0$ .
- ▶ **unilatères** : lorsque sous l'alternative  $H_1$ , la statistique  $T$  est soit systématiquement "plus grande", soit "plus petite" que sous  $H_0$ .

Donner un sens à "plus grande" ou "plus petite" pour des variables aléatoires : notion d'**ordre stochastique**.

# Ordre stochastique I

## 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  $H_1$ .



# Ordre stochastique II

## Exemple

$T \sim \mathcal{N}(\theta, 1)$ ,  $H_0 : \theta = 0$  et  $H_1 : \theta = 1$ .

$T$  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  $Z \sim \mathcal{N}(0, 1)$ . 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.

## Propriété

Si  $X_1 \prec Y_1, X_2 \prec Y_2$  et  $X_1 \amalg X_2, Y_1 \amalg Y_2$  alors

$X_1 + X_2 \prec Y_1 + Y_2$ .

# Choix de la région de rejet I

- ▶ C'est 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  $H_0$  :  $\mathcal{R}_{H_0}$  choisie t.q. la densité de la stat. de test  $T$  sur  $\mathcal{R}_{H_0}$  est plus grande sous  $H_1$  que sous  $H_0$ .

Deux approches sont possibles :

- 1) 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  $\alpha$ .

## Choix de la région de rejet II

- 2) 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 (degré de significativité)

- ▶  $\mathcal{R}_{H_0} = \{S_n \geq s\}$
- ▶ 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  $H_0$ .

# Choix de la région de rejet III

## Cas des tests bilatères

- ▶  $\mathcal{R}_{H_0} = \{T_n \geq b\} \cup \{T_n \leq a\}$ , avec  $a \leq b$ , seuils à déterminer.
- ▶ En pratique, si on se fixe un niveau  $\alpha$  positif, alors on choisira  $a, b$  tels que  $\mathbb{P}_{H_0}(T_n \geq b) = \mathbb{P}_{H_0}(T_n \leq a) \leq \alpha/2$ .
- ▶ Si  $T_n$  a une distribution symétrique par rapport à  $m_0$  sous l'hypothèse nulle  $H_0$ , on peut écrire de façon équivalente  $\mathcal{R}_{H_0} = \{|T_n - m_0| \geq s\}$ .
- ▶ 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 !

# 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 I

## Contexte

- ▶ Stat. de test  $T_n$  **discrète** dont la **loi approchée est continue**.
- ▶ Ex : cadre asymptotique avec

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

- ▶ Si  $\mathcal{R}_{H_0} = \{T_n \geq t\}$  et  $\forall \alpha > 0$  on cherche le seuil  $t$  tel que  $\mathbb{P}_{H_0}(T_n \geq t) \leq \alpha$ .
- ▶ Or,  $\forall u \in [0, 1[$ , comme  $T_n$  est discrète,

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

- ▶ De même, si  $\mathcal{R}_{H_0} = \{T_n \leq t\}$ , alors  $\forall u \in [0; 1[$  on a  $\mathbb{P}_{H_0}(T_n \leq t) = \mathbb{P}_{H_0}(T_n \leq t + u)$ .

# Correction du continu II

## Mise en œuvre

- ▶ La correction du continu consiste à remplacer la valeur par défaut  $u = 0$  par  $u = 1/2$ .
- ▶ Ex : si  $\mathcal{R}_{H_0} = \{T_n \geq t\}$ , on cherche le seuil  $t$  tel que

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

# Test d'une hypothèse induite

## Remarques

- ▶ Il arrive que pour tester  $H_0$ , on teste en fait  $H'_0$  telle que  $H_0 \Rightarrow H'_0$ .
- ▶ Exemple :  $H_0$  : "Les variables sont gaussiennes" et  $H'_0$  : "le moment recentré d'ordre 3 est nul".
- ▶ Si on rejette  $H'_0$  alors on rejette nécessairement  $H_0$ . Par contre, si on accepte  $H'_0$ , on n'accepte pas nécessairement  $H_0$  !
- ▶ **N.B.** Lorsque  $H'_0$  est une hypothèse **paramétrique**, on sort du cadre des tests non paramétriques.
- ▶ Exemple : voir plus loin le test du signe.



# Tests de permutation

## Pincipe

- ▶ Il s'agit d'une technique générale pour échantillonner la loi de la statistique de test sous  $H_0$ .
- ▶ Requiert que les individus soient **i.i.d** (ou plus généralement échangeables).
- ▶ Permet d'obtenir un **test exact** (par opposition à asymptotique) ;
- ▶ Peut être difficile à mettre en œuvre d'un point de vue puissance de calcul (si trop de catégories par exemple).

## Exemple

- ▶ Pour tester que 2 échantillons ont la même loi, on peut permutationner les individus entre les échantillons pour échantillonner les valeurs de la stat de test (différence entre moyennes) sous  $H_0$ .

# Sommaire Troisième partie

Introduction, rappels et généralités

## Tests sur une population

Tests d'adéquation à une distribution fixée

Tests d'adéquation à une famille de distributions

Tests de médiane (ou de symétrie)

## Tests sur deux populations

Tests de comparaison (ou homogénéité) de deux populations

Tests de corrélation sur variables appariées

# Sommaire Troisième partie

Introduction, rappels et généralités

## Tests sur une population

Tests d'adéquation à une distribution fixée

Tests d'adéquation à une famille de distributions

Tests de médiane (ou de symétrie)

## Tests sur deux populations

Tests de comparaison (ou homogénéité) de deux populations

Tests de corrélation sur variables appariées

# Cas discret fini : Test d'adéquation du $\chi^2$ de Pearson I

## Description

Pour un échantillon de v.a. discrètes avec  $r$  modalités (qualitatives ou quantitatives)  $X_1, \dots, X_n$  et une distribution  $p_0$  fixée, on teste  $H_0 : "p = p_0"$  contre  $H_1 : "p \neq p_0"$ .

## Statistique de Pearson

$$\chi^2 = n \sum_{k=1}^r \frac{(\hat{p}_k - p_0(k))^2}{p_0(k)}, \quad (1)$$

où  $\hat{p}_k = \sum_{i=1}^n 1\{X_i = k\}/n$ . On rejette  $H_0$  pour les grandes valeurs de  $\chi^2$ .

# Cas discret fini : Test d'adéquation du $\chi^2$ de Pearson II

## Propriétés

- ▶ Test **asymptotique**, fondé sur la loi limite de la statistique de test qui suit un  $\chi^2(r - 1)$ .
- ▶ C'est un test **consistant** : pour toute alternative  $p \neq p_0$ , la puissance  $\beta_n(p) \rightarrow 1$ .
- ▶ **Remarque : C'est en fait un test paramétrique !** puisque la loi discrète des  $X_i$  dépend d'un nombre **fini** de paramètres.

# Cas continu : test d'adéquation du $\chi^2$ de Pearson avec catégories I

## Description

- ▶ Pour un échantillon de v.a. continues  $X_1, \dots, X_n$  et une fdr  $F_0$  fixée, on veut tester  $H_0 : "F = F_0"$  contre  $H_1 : "F \neq F_0"$ .
- ▶ On découpe  $\mathbb{R} = (-\infty, a_1) \cup (a_1, a_2) \cup \dots \cup (a_{r-2}, a_{r-1}) \cup (a_{r-1}, +\infty)$  en intervalles fixés pour obtenir  $r$  modalités :  
 $\hat{p}_k = \sum_{i=1}^n 1\{X_i \in (a_{k-1}, a_k)\} / n$  et  
 $p_0(k) = F_0(a_k) - F_0(a_{k-1})$ .
- ▶ On teste alors l'hypothèse induite  $H'_0 : "Les versions discrétisées de  $F$  et  $F_0$  sont identiques"$ .
- ▶ Même statistique qu'en (1) et même type de zone de rejet.

# Cas continu : test d'adéquation du $\chi^2$ de Pearson avec catégories II

## Propriétés

- ▶ Test **asymptotique**, fondé sur la loi limite de la statistique de test qui suit un  $\chi^2(r - 1)$ .
- ▶ Sous certaines hyps, utilisable lorsque  $F_0$  est **définie à un paramètre près**  $\theta_0 \in \mathbb{R}^s$  qui est **estimé** (loi limite  $\chi^2(r - s - 1)$ ).
- ▶ Ce test **n'est pas consistant** pour toute alternative  $F \neq F_0$  (en fait  $H'_1$  est plus "petite" que  $H_1$ ).
- ▶ Peut-être généralisé au cas où les  $a_k$  **sont aléatoires** et tels que **le nombre de points** dans chaque intervalle est **fixé**.
- ▶ Le test induit est en fait encore **paramétrique**.

# Cas continu : test de Kolmogorov Smirnov (KS)

## Description

Pour un échantillon de v.a. continues  $X_1, \dots, X_n$  et une fdr  $F_0$  fixée, on teste  $H_0 : "F = F_0"$  contre  $H_1 : "F \neq F_0"$ .

## Statistique de KS

$$\begin{aligned} D_n &= \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)| \\ &= \max_{1 \leq i \leq n} \{|F_0(X_{(i)}) - i/n|, |F_0(X_{(i)}) - (i-1)/n|\}. \end{aligned}$$

On rejette  $H_0$  pour les grandes valeurs de  $D_n$ .

## Propriétés

- ▶ Statistique libre en loi sous  $H_0$ . Cette loi est tabulée.
- ▶ Approximation pour  $n$  grand

$$\mathbb{P}_{H_0}(\sqrt{n}D_n > z) \xrightarrow{n \rightarrow \infty} 2 \sum_{j \geq 1} (-1)^{j-1} \exp(-2j^2 z^2).$$



# Autres tests : Cramér von Mises et Anderson-Darling

Dans le même contexte que KS, on peut utiliser

## Statistiques de test

- ▶  $CVM_n = n \int (\hat{F}_n(x) - F_0(x))^2 dF_0(x)$
- ▶  $A_n = n \int \frac{(\hat{F}_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x)$

On rejette  $H_0$  pour les grandes valeurs de  $CVM_n$  ou  $A_n$ .

## Propriétés

- ▶ Statistiques **libres en loi** sous  $H_0$ .
- ▶ Statistiques plus sensibles à **l'ensemble des valeurs** (pas seulement le sup). En particulier,  $A_n$  sensible aux écarts à  $F_0$  dans la queue de distribution.
- ▶ CVM et AD considérés comme plus puissants que KS (mais pas de preuve théorique).

# Conclusions sur tests d'adéquation à une loi fixée

## Adéquation à une famille de lois

- ▶  $\chi^2$  se généralise au cas de  $H'_0 : F = F_0(\theta_1, \dots, \theta_s)$  où  $\theta_1, \dots, \theta_s$  inconnus.
- ▶ Les tests KS, CVM et AD **ne s'appliquent pas directement** à ce cadre. En effet, leurs statistiques ne sont plus libres en loi (même asymptotiquement) sous  $H'_0$ . Il faut **adapter** ces tests pour chaque famille de loi considérée.

# Sommaire Troisième partie

Introduction, rappels et généralités

## Tests sur une population

Tests d'adéquation à une distribution fixée

Tests d'adéquation à une famille de distributions

Tests de médiane (ou de symétrie)

## Tests sur deux populations

Tests de comparaison (ou homogénéité) de deux populations

Tests de corrélation sur variables appariées

# Tests d'adéquation à une famille de lois

- ▶ Le test du  $\chi^2$  se généralise au cas  $H_0' : F = F_0(\theta_1, \dots, \theta_s)$  où  $\theta_1, \dots, \theta_s$  inconnus.
- ▶ Les tests KS, CVM et AD **doivent être modifiés** dans ce cadre. Voir [De Wet and Randles 87] pour plus de détails.

## Cas particulier : famille gaussienne

- ▶ Les tests de normalité permettent de tester  $H_0$  : "  $F$  suit une loi gaussienne (de **paramètres non spécifiés**)" contre  $H_1$  : "  $F$  n'est pas gaussienne".
- ▶ En pratique, on teste  $H_0' : " F = \mathcal{N}(\hat{\theta}_n, \hat{\sigma}_n^2)"$  (**paramètres estimés**).
- ▶ Il existe des généralisations de KS (test de Lilliefors) et CVM à ce cadre, implémentées sous R (Paquet *nortest*, fonctions *lillie.test* et *cvm.test*. Contient également *Pearson.test*).

# Sommaire Troisième partie

Introduction, rappels et généralités

## Tests sur une population

Tests d'adéquation à une distribution fixée

Tests d'adéquation à une famille de distributions

Tests de médiane (ou de symétrie)

## Tests sur deux populations

Tests de comparaison (ou homogénéité) de deux populations

Tests de corrélation sur variables appariées

# Rappel : test de signe I

## Contexte

On observe un échantillon  $X_1, \dots, X_n$  de v.a. réelles i.i.d. On teste  $H_0 : \mathbb{P}(X \leq 0) = 1/2$  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".

## Statistique de signe

$$S_n = \sum_{i=1}^n 1\{X_i \leq 0\} \sim \mathcal{B}(n, m),$$

où  $m = \mathbb{P}(X \leq 0)$ . Sous  $H_0 : m = 1/2$ , on a  $S_n \sim \mathcal{B}(n, 1/2)$  et sous  $H_1 : \mathbb{P}(X \leq 0) > 1/2$ , la statistique  $S_n$  est stochastiquement plus grande que sous  $H_0$ . On rejette donc  $H_0$  pour les grandes valeurs de  $S_n$ .

# Rappel : test de signe II

## Propriétés

- ▶ 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.
- ▶ 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.
- ▶ Le test de **signe et rang** utilise plus d'information sur les variables.
- ▶ **Remarque : c'est en fait un test paramétrique !** puisque la loi de  $S_n$  sous  $H_0$  et sous l'alternative est **paramétrique** ( $\mathcal{B}(n, m)$ ).

# Statistiques d'ordre et de rang I

## Définitions

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  $X_i = X_{R_X(i)}^* = X_{(R_X(i))}$ .



# Statistiques d'ordre et de rang II

## Exemple

$n = 7$ ,  $x = (4, 2, 1, 1, 2, 0, 1)$ . 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

- ▶ Cette notion est dépendante de  $n$  qui doit être fixé.
- ▶ S'il y a des ex-æquos, le vecteur des rangs n'est pas unique.

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

## Contexte

Soit  $X_1, \dots, X_n$  un échantillon de v.a. réelles de loi supposée **diffuse et symétrique par rapport à (la médiane)  $m$** . On veut tester  $H_0 : m = 0$  contre  $H_1 : m \neq 0$ .

## Statistique de Wilcoxon

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

où  $R_{|X|}$  le vecteur des rangs associé à l'échantillon.

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

## Exemple

$n = 5$  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  $W_5^- = 2 + 4 + 1 = 7$ .

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

## Remarques

- ▶  $W_n^+ + W_n^- = n(n+1)/2$  p.s.  
En effet, comme  $X$  est diffuse, on a  $X_i \neq 0$  p.s. et donc  
 $W_n^+ + W_n^- = \sum_{i=1}^n R_{|X|}(i) = \sum_{i=1}^n i = n(n+1)/2$ .
- ▶  $0 \leq W_n^+ \leq n(n+1)/2$  p.s.  
Le cas  $W_n^+ = 0$  correspond à tous les  $X_i < 0$  et le cas  
 $W_n^+ = n(n+1)/2$  à tous les  $X_i > 0$ .

## Cas des ex-æquos

- ▶ Normalement le test s'applique à des variables diffuses, donc pas d'ex-æquos en théorie.
- ▶ En pratique, on affecte des **rangs moyens** en cas d'égalité et il existe des **corrections des lois limites** dans ce cas.

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

## Théorème

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

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

et

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

## Conséquences

- ▶ Test de  $H_0$  : " $m = 0$ " contre  $H_1$  :  $m \neq 0$  en rejetant  $H_0$  pour les grandes valeurs de  $|W_n^+ - n(n+1)/4|$ .
- ▶ Test **exact** via table pour  $n \leq 20$ , **asymptotique** (via l'approximation gaussienne) sinon.

# Sommaire Troisième partie

Introduction, rappels et généralités

Tests sur une population

Tests d'adéquation à une distribution fixée

Tests d'adéquation à une famille de distributions

Tests de médiane (ou de symétrie)

Tests sur deux populations

Tests de comparaison (ou homogénéité) de deux populations

Tests de corrélation sur variables appariées

# Sommaire Troisième partie

Introduction, rappels et généralités

Tests sur une population

Tests d'adéquation à une distribution fixée

Tests d'adéquation à une famille de distributions

Tests de médiane (ou de symétrie)

Tests sur deux populations

Tests de comparaison (ou homogénéité) de deux populations

Tests de corrélation sur variables appariées

# Tests de comparaison de deux populations

Dans la suite, on distinguera :

- ▶ Le cas de deux populations **appariées**, qui se ramène au cas des tests sur **une** population.
- ▶ Le cas de deux populations **non appariées** : il s'agit là vraiment de tests sur **deux** populations.



# Échantillons appariés I

On considère  $X_1, \dots, X_n$  et  $Y_1, \dots, Y_n$  deux **échantillons indépendants** de v.a. **diffuses** de lois respectives  $F$  et  $G$ .

## Appariement

- ▶ Soit il s'agit des mêmes individus, par exemple 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).

# Échantillons appariés II

## Tests d'homogénéité

- ▶ On veut tester  $H_0 : "F = G"$  contre  $H_1 : "F \neq G"$ .
- ▶ Après appariement des variables, on construit  $Z_i = X_i - Y_i$ .
- ▶ Sous l'hypothèse  $H_0$ , la loi de  $Z$  est symétrique par rapport à sa médiane  $m$ , qui vaut 0. D'où le **test induit**  $H'_0 : "m = 0"$  contre  $H'_1 : "m \neq 0"$ .
- ▶ On applique le **test de signe et rang de Wilcoxon**.

# Échantillons non appariés I

## Contexte

- ▶ On considère  $X_1, \dots, X_n$  et  $Y_1, \dots, Y_m$  deux **échantillons indépendants** de v.a. **diffuses** de lois respectives  $F$  et  $G$ .
- ▶ On veut tester  $H_0 : "F = G"$  contre  $H_1 : "F \neq G"$ .
- ▶ Remarque : Les échantillons n'ont a priori pas la même taille et il n'existe pas d'appariement naturel entre les variables.

# Échantillons non appariés II

## Exemple

- ▶ Population de  $N$  individus sur lesquels on veut tester un nouveau traitement.
- ▶ On forme un groupe de  $n$  individus qui reçoivent le nouveau traitement et  $m = N - n$  forment le groupe "contrôle", recevant un placebo.
- ▶ On mesure une quantité relative au traitement.
- ▶ L'hypothèse nulle  $H_0 : "F = G"$  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.

# Échantillons non appariés III

## Approches possibles

- ▶ Test de Kolmogorov Smirnov de comparaison de 2 échantillons.
- ▶ Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney).

# Test de Kolmogorov Smirnov de comparaison de 2 échantillons

## Statistique de KS pour deux échantillons

$$D_{n,m} = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \hat{G}_m(x)|.$$

On rejette  $H_0$  pour les grandes valeurs de  $D_{n,m}$ .

## Propriétés

- ▶ Statistique **libre en loi** sous  $H_0$ . Cette loi est **tabulée**.

# Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney) I

## Procédure

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  $R_1, R_2, \dots, R_n$  les rangs associés au premier échantillon (i.e. les  $X_i$ ) et  $N = n + m$ .

## Exemple

$X_1 = 3.5; X_2 = 4.7; X_3 = 1.2; Y_1 = 0.7; Y_2 = 3.9$  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  $R_1 = 3, R_2 = 5, R_3 = 2$ .

## 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.

# Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney) II

## Statistique de Mann-Whitney $W_{YX}$

On note  $\Sigma_1 = R_1 + \dots + R_n$  la somme des rangs du premier échantillon. On a p.s.

$$\frac{n(n+1)}{2} \leq \Sigma_1 \leq \frac{(n+m)(n+m+1)}{2} - \frac{m(m+1)}{2} = nm + \frac{n(n+1)}{2},$$

On définit

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

On a  $0 \leq W_{YX} \leq nm$  p.s.. On définit de façon symétrique  $W_{XY} = \Sigma_2 - m(m+1)/2$  où  $\Sigma_2$  est la somme des rangs du second échantillon.



# Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney) III

## Proposition

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

# Test de la somme des rangs de Wilcoxon (ou test de Mann-Whitney) IV

## Théorème

La loi de  $W_{XY}$  (ou  $W_{YX}$ ) est libre sous  $H_0$ . Cette loi ne dépend que de  $n$  et  $m$ . De plus,

$$\mathbb{E}_{H_0}(W_{XY}) = \frac{nm}{2}, \quad \text{Var}_{H_0}(W_{XY}) = \frac{nm(N+1)}{12}$$

et

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

## Test exact ou test asymptotique

- ▶ Le test rejette  $H_0 : "F = G"$  pour les grandes valeurs de  $|W_{YX} - nm/2|$ .
- ▶ Loi tabulée pour les petites valeurs de  $n$  et  $m$  ( $\leq 10$ ).
- ▶ Pour les grandes valeurs, on utilise l'approximation gaussienne.

## Lien entre Mann-Whitney et Wilcoxon

La stat. signe et rang de Wilcoxon peut être vue comme un cas particulier de la stat. somme des rangs de Wilcoxon. En effet,

- ▶ Soit  $Z_1, \dots, Z_N$  un échantillon,
- ▶  $U_1, \dots, U_n$  sous-échantillon correspondant aux valeurs de  $Z_i$  telles que  $Z_i > 0$ ,
- ▶  $V_1, \dots, V_m$  sous-échantillon correspondant aux valeurs  $-Z_i$  pour les  $Z_i < 0$ .
- ▶ Ordonner les  $\{U_i, V_j\}$  revient à ordonner  $\{|Z_i|\}$ .
- ▶ 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.

# Remarques sur les tests de comparaison

## Remarques

- ▶ Le test de Mann-Whitney est très général et n'utilise que les **valeurs relatives** des variables entre elles.
- ▶ Le test d'homogénéité de KS pour 2 échantillons est assez différent car il prend en compte la **forme** des distributions et pas seulement des phénomènes de **translation**.

# Sommaire Troisième partie

Introduction, rappels et généralités

Tests sur une population

Tests d'adéquation à une distribution fixée

Tests d'adéquation à une famille de distributions

Tests de médiane (ou de symétrie)

Tests sur deux populations

Tests de comparaison (ou homogénéité) de deux populations

Tests de corrélation sur variables appariées

# Contexte des tests de corrélation

## Contexte

- ▶ On dispose de deux échantillons  $X_1, \dots, X_n$  et  $Y_1, \dots, Y_n$  de v.a. réelles et **appariées**.
- ▶ Exemple : on mesure deux quantités  $X$  et  $Y$  sur un ensemble d'individus.
- ▶ On veut tester  $H_0$  : "X et Y sont non corrélées" contre  $H_1$  : "X et Y sont corrélées".
- ▶ NB : si les variables ne sont pas gaussiennes, "**non corrélation**"  $\neq$  "**indépendance**".
- ▶ NB : Un test de **permutation** va tester  $H'_0$  : "X et Y sont indépendantes" et pas  $H_0$  : "X et Y sont non corrélées".

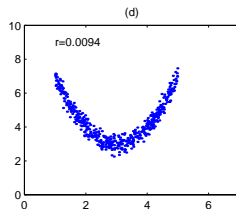
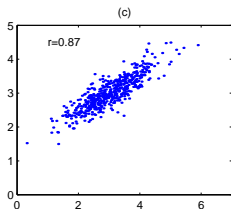
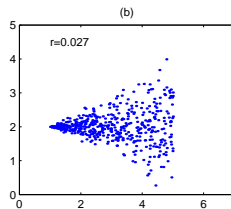
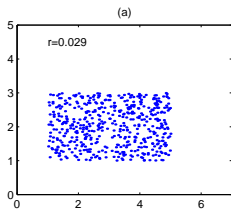
## Remarque

Le test du  $\chi^2$  d'indépendance pour variables discrètes est un test **paramétrique** car les supports des variables sont **finis** !

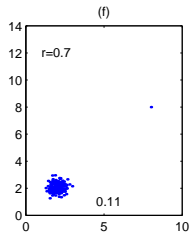
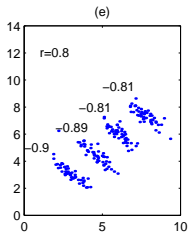
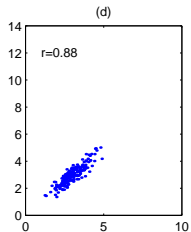
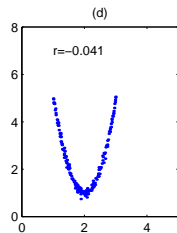
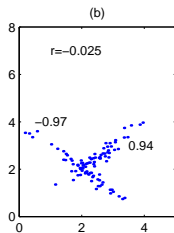
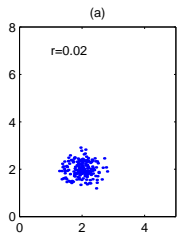
# Corrélation de Pearson I

## Rappels

Le coefficient de corrélation de Pearson mesure la dépendance **linéaire** entre deux variables **réelles**.



# Corrélation de Pearson II





# Corrélation de Pearson III

## Coefficient de corrélation

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_i (X_i - \bar{X}_n)^2} \sqrt{\sum_i (Y_i - \bar{Y}_n)^2}}.$$

## Propriétés

- ▶  $-1 \leq r \leq 1$  avec égalité lorsque la relation entre  $X$  et  $Y$  est linéaire.
- ▶ La distribution **exacte** de  $r$  sous  $H_0$  **dépend** des distributions de  $X$  et  $Y$ . Ce n'est pas une statistique libre en loi sous  $H_0$ .
- ▶ On peut utiliser un **test de permutation** dans le cas de  $H'_0$  : "X et Y sont indépendantes".
- ▶ **Asymptotiquement**, sous  $H_0$ ,  $r$  suit une loi gaussienne centrée.

# Corrélation de Pearson IV

## Remarques

- ▶ La fonction *cor.test* de R implémente différents calculs et tests de corrélation. La méthode "pearson" de cette fonction implémente le test du coefficient de corrélation de Pearson lorsque les variables sont **gaussiennes** uniquement. À manipuler avec précaution dans le cas de petits échantillons !
- ▶ Pour obtenir une statistique libre en loi sous  $H_0$ , il faut se "débarrasser" des valeurs prises par les variables. Une possibilité est de passer par les **rangs** des variables.

# Test de corrélation des rangs de Spearman I

## Contexte

- ▶ Le coefficient de corrélation des rangs de Spearman mesure la dépendance **monotone** entre deux variables **réelles et diffuses**.
- ▶ C'est le coefficient de corrélation de Pearson entre les **rangs** des variables de deux échantillons.
- ▶ Soient  $X_1, \dots, X_n$  et  $Y_1, \dots, Y_n$  deux échantillons **appariés** et  $R_1, \dots, R_n, S_1, \dots, S_n$  les rangs respectifs des variables.
- ▶ On teste donc  $H_0$  "X, Y non corrélées" contre  $H_1$  : "X, Y sont en relation monotone".

## Statistique de corrélation des rangs de Spearman

$$\rho = \frac{\sum_i (r_i - \bar{r}_n)(s_i - \bar{s}_n)}{\sqrt{\sum_i (r_i - \bar{r}_n)^2 \sum_i (s_i - \bar{s}_n)^2}}$$

On rejette  $H_0$  pour des grandes valeurs de  $|\rho|$ .

# Test de corrélation des rangs de Spearman II

## Propriétés

- ▶  $-1 \leq \rho \leq 1$  et si  $X = f(Y)$  avec  $f$  croissante (resp. décroissante) alors  $\rho = +1$  (resp.  $-1$ ).
- ▶ Sous  $H_0$ , la stat  $\rho$  est libre en loi.
- ▶ Test exact : en utilisant un test de permutation pour l'hyp  $H'_0$  (et pas  $H_0$ ). Possible si  $n$  pas trop grand.
- ▶ Tests asymptotiques de  $H_0$  : à partir de transformations de  $\rho$ .

# Test de corrélation des rangs de Spearman III

## Cas des ex-æquos

- ▶ Normalement le test s'applique à des variables diffuses, donc pas d'ex-æquos en théorie.
- ▶ En pratique, on affecte des rangs moyens en cas d'égalité.
- ▶ **S'il n'y a pas d'ex-æquos**, l'expression de  $\rho$  se simplifie et devient

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)},$$

où  $d_i = r_i - s_i$  est la différence des rangs de l'individu  $i$ .

# Test de corrélation des rangs de Kendall I

## Contexte

- ▶ Le coefficient de corrélation des rangs de Kendall mesure la dépendance entre deux variables **réelles et diffuses**.
- ▶ Pour tout couple d'individus  $(i, j)$ , on dit que les paires  $(x_i, y_i)$  et  $(x_j, y_j)$  sont **concordantes** si
  - ▶ soit  $x_i < x_j$  et  $y_i < y_j$ ,
  - ▶ soit  $x_i > x_j$  et  $y_i > y_j$ ,et **discordante** sinon.
- ▶ Cette fois encore, on teste  $H_0$  "X, Y non corrélées" contre  $H_1$  : "X, Y sont en relation monotone".

# Test de corrélation des rangs de Kendall II

## Statistique de corrélation des rangs de Kendall

$$\tau_n = \frac{(\text{Nb de paires concordantes}) - (\text{Nb de paires discordantes})}{n(n-1)/2}.$$

### Cas des ex-æquos

- ▶ Normalement le test s'applique à des variables diffuses, donc pas d'ex-æquos en théorie.
- ▶ Il existe des variantes de la définition de  $\tau_n$  qui prennent en compte le cas des ex-æquos.

# Test de corrélation des rangs de Kendall III

## Propriétés



- ▶  $-1 \leq \tau_n \leq 1$ , et  $\tau_n = 1$  lorsque la concordance entre les paires est parfaite,  $-1$  lorsque la discordance est parfaite.
- ▶ Sous  $H_0$ ,  $\tau$  est une stat **libre en loi**, de moyenne nulle ( $\mathbb{E}_{H_0}(\tau_n) = 0$ ). Loi **tabulée** pour  $n$  petit, qui donne un test **exact**.
- ▶ Test **asymptotique** pour  $n$  grand, fondé sur l'approximation gaussienne

$$\mathbb{E}_{H_0}(\tau_n) = 0, \quad \text{Var}_{H_0}(\tau_n) = \frac{2(2n + 5)}{9n(n - 1)}$$

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



## Références partie tests

-  [De Wet and Randles 87] T. De Wet and R. Randles.  
On the effect of substituting parameter estimators in limiting  $\chi^2$ ,  $u$  and  $v$  statistics.  
*Annals of Statistics*, 15 :398–412, 1987.
-  [Morris 75] C. Morris.  
Central limit theorems for multinomial sums.  
*The Annals of Statistics*, 3 :165–188, 1975.

## Quatrième partie IV

Estimation de densité : histogrammes,  
noyaux, projections et estimateurs sous  
contraintes

# Sommaire Quatrième partie

## Introduction

### Estimateur par histogramme

- Construction et risque quadratique

- Choix de la partition par validation croisée

### Estimateur à noyau

- Construction et risque quadratique

- Choix de la fenêtre par validation croisée

- Estimateur à noyau des  $k$  plus proches voisins

### Estimateur par projection

- Construction

- Propriétés

### Cas des densités multivariées

- Fléau de la dimension

- Généralisations des estimateurs précédents

### Cas des densités monotones ou unimodales ou convexes ...

- Densités monotones

- Densités unimodales

- Autres contraintes

# Contexte de l'estimation de densité (univariée)

- ▶ **Observations** :  $X_1, \dots, X_n$  v.a. i.i.d. **réelles** de fdr  $F$  et admettant une densité  $f = F'$ .
- ▶ **But** : estimer (à partir des observations)  $f$  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 de  $f$ .

## Objectifs

Obtenir des informations de **nature géométrique** sur la distribution des variables. Ex :

- ▶ Combien de modes ?
- ▶ Zones peu denses ? très denses ?

## Notation

- ▶ On notera  $\mathbb{E}_f$  l'espérance quand la densité des  $X_i$  vaut  $f$ .

# Mesure de la qualité d'un estimateur : risque I

## Ingrédients de la définition du risque

- 1) **Distance** sur  $\mathcal{F}$  pour mesurer l'écart entre  $\hat{f}_n$  et  $f$ . Ex :
  - ▶  $d(f, g) = \|f - g\|_p = [\int |f - g|^p]^{1/p}$ , pour  $p \geq 1$ . Par exemple  $p = 1$  ou  $2$ .
  - ▶  $d(f, g) = \|f - g\|_\infty = \sup_x |f(x) - g(x)|$ .
  - ▶  $d(f, g) = |f(x_0) - g(x_0)|$  où  $x_0$  fixé.
- 2) Définition d'une **fonction de perte**  $\omega : \mathbb{R} \mapsto \mathbb{R}^+$  convexe, telle que  $\omega(0) = 0$ . Ex :  $\omega : u \mapsto u^2$  fonction de perte quadratique.
- 3) L'**erreur**  $w(d(\hat{f}_n, f))$  (par ex  $d(\hat{f}_n, f)^2$ ) dépend de l'**échantillon observé**. On définit donc une **fonction de risque**

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

C'est **en moyenne**, l'erreur que l'on commet en estimant  $f$  par  $\hat{f}_n$ , pour la distance  $d$  et la perte  $\omega$ .

# Mesure de la qualité d'un estimateur : risque II

## Exemples de fonctions de risque

- ▶ En prenant la **distance**  $\mathbb{L}_2$  et la **perte quadratique**, on obtient le risque quadratique intégré : **MISE** = mean integrated squared error

$$R(\hat{f}_n, f) = \mathbb{E}_f \|\hat{f}_n - f\|_2^2 = \mathbb{E}_f \int_x (\hat{f}_n(x) - f(x))^2 dx.$$

- ▶ En prenant la **distance ponctuelle** en  $x_0$  et la **perte quadratique**, on obtient le risque quadratique ponctuel en  $x_0$  : **MSE** = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}_f |\hat{f}_n(x_0) - f(x_0)|^2.$$

# Décomposition biais-variance du risque quadratique I

## Décomposition "biais-variance" du MSE

$$\begin{aligned}R_{x_0}(\hat{f}_n, f) &= \mathbb{E}_f[\hat{f}_n(x_0) - f(x_0)]^2 \\ &= \mathbb{E}_f[\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)]^2 + \mathbb{E}_f[\hat{f}_n(x_0) - \mathbb{E}_f\hat{f}_n(x_0)]^2 + \text{dp}.\end{aligned}$$

Or  $[\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)]^2$  est déterministe donc  $\mathbb{E}_f$  disparaît, et le double produit vérifie

$$\text{dp} = 2\mathbb{E}_f\left([\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)][\hat{f}_n(x_0) - \mathbb{E}_f\hat{f}_n(x_0)]\right) = 0.$$

Donc

$$\begin{aligned}R_{x_0}(\hat{f}_n, f) &= |\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)|^2 + \mathbb{E}_f|\hat{f}_n(x_0) - \mathbb{E}_f\hat{f}_n(x_0)|^2 \\ &= \text{Biais}^2 + \text{Var}(\hat{f}_n(x_0)).\end{aligned}$$

# Décomposition biais-variance du risque quadratique II

## Décomposition "biais-variance" du MISE

$$\begin{aligned}R(\hat{f}_n, f) &= \mathbb{E}_f \|\hat{f}_n - f\|_2^2 = \mathbb{E}_f \int_x |\hat{f}_n(x) - f(x)|^2 dx \\ &= \mathbb{E}_f \|\mathbb{E}_f(\hat{f}_n) - f\|_2^2 + \mathbb{E}_f \|\hat{f}_n - \mathbb{E}_f(\hat{f}_n)\|_2^2 + dp.\end{aligned}$$

Or  $\|\mathbb{E}_f(\hat{f}_n) - f\|_2^2$  est déterministe donc  $\mathbb{E}_f$  disparaît, et le double produit vérifie

$$\begin{aligned}dp &= 2\mathbb{E}_f \left( \langle \hat{f}_n - \mathbb{E}_f(\hat{f}_n), \mathbb{E}_f(\hat{f}_n) - f \rangle_{L_2(\mathbb{R})} \right) \\ &= 2 \langle 0, \mathbb{E}_f(\hat{f}_n) - f \rangle_{L_2(\mathbb{R})} = 0.\end{aligned}$$

Donc

$$R(\hat{f}_n, f) = \|\mathbb{E}_f(\hat{f}_n) - f\|_2^2 + \mathbb{E}_f \|\hat{f}_n - \mathbb{E}_f(\hat{f}_n)\|_2^2 = \text{Biais}^2 + \text{"Var}(\hat{f}_n)\text{"}.$$



# Décomposition biais-variance du risque quadratique III

## Compromis biais/variance

- ▶ L'étude du **risque quadratique** de l'estimateur se ramène donc à l'étude de son **biais** et de sa **variance**.
- ▶ On pourra accepter des estimateurs biaisés mais de variance petite, tels que le **risque quadratique** soit **contrôlé**.

# Oracle

Idéalement, le meilleur estimateur, au sens du risque, est

$$f_n^* = \underset{\hat{f}_n}{\operatorname{Argmin}} R(\hat{f}_n, f).$$

- ▶ Pbm :  $R(\hat{f}_n, f)$  dépend de la densité  $f$  inconnue et n'est donc pas calculable. L'argmin  $f_n^*$  n'est pas un estimateur, c'est un oracle.
- ▶ Souvent, on dispose d'une famille d'estimateurs  $\hat{f}_{\lambda, n}$  dépendants d'un paramètre  $\lambda$  (partition, fenêtre, etc ...). L'oracle est donné par

$$\lambda^* = \underset{\lambda}{\operatorname{Argmin}} R(\hat{f}_{\lambda, n}, f).$$

mais  $\hat{f}_{\lambda^*, n}$  n'est pas un estimateur. On veut sélectionner le meilleur  $\lambda$  à partir de l'étude du risque de  $\hat{f}_{\lambda, n}$  et de sa dépendance en  $\lambda$ .

# Contrôles du risque

Puisque  $f$  est inconnue, le risque  $R(\hat{f}_n, f)$  au point  $f$  n'est pas calculable. Alternatives :

- ▶ Avoir recours à une méthode de **validation croisée** pour **estimer ce risque** au point  $f$ .
- ▶ 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".

- ▶ En général, dans le second cas, on prend un point de vue **asymptotique**.

# Contrôle asymptotiques du risque maximal I

## Vitesses de convergence

- ▶ On veut construire un estimateur  $\hat{f}_n$  tel que

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

- ▶ et exhiber la **vitesse de convergence** de  $\hat{f}_n$  pour le risque  $R$  : la **plus petite suite**  $(\phi_n)_{n \geq 0} \rightarrow 0$  telle que  $\{\phi_n^{-1} R(\hat{f}_n, \mathcal{F})\}_n$  bornée :

$$\exists C > 0, \forall n \in \mathbb{N}, \forall f \in \mathcal{F}, \quad R(\hat{f}_n, f) \leq C\phi_n.$$

- ▶ On dit alors que  $(\hat{f}_n)_n$  atteint la **vitesse de convergence**  $(\phi_n)_n$  sur la classe  $\mathcal{F}$  pour la distance  $d$  et la perte  $\omega$ .

# Contrôle asymptotiques du risque maximal II

## Point de vue minimax

- ▶ **Minimax** : la recherche du **meilleur estimateur**  $f_n$  pour le risque maximal, à classe  $\mathcal{F}$  fixée. Le **risque minimax** est défini par

$$\inf_{f_n} \sup_{f \in \mathcal{F}} R(f_n, f).$$

S'il existe une suite  $(\phi_n)_n$  telle que  $\exists c, C > 0$ , et une suite d'estimateurs  $(\hat{f}_n)_n$  pour lesquels

$$c\phi_n \leq \inf_{f_n} \sup_{f \in \mathcal{F}} R(f_n, f) \leq \sup_{f \in \mathcal{F}} R(\hat{f}_n, f) \leq C\phi_n,$$

alors  $(\phi_n)_n$  est la **vitesse minimax**.

- ▶ Typiquement, les classes de fonctions  $\mathcal{F}$  pour lesquelles on sait contrôler le risque minimax sont des classes de fonctions **régulières**. Comme par exemple : Lipschitz, classe  $\mathcal{C}^k$ , etc.

# Contrôle asymptotiques du risque maximal III

## Point de vue maxiset

- **Maxiset** : la recherche de la plus grande classe de fonctions  $\mathcal{F}$  telle que le risque maximal sur  $\mathcal{F}$  d'un estimateur  $\hat{f}_n$  fixé décroît à une certaine vitesse

$$\sup_{\mathcal{F}} \{ \mathcal{F}; \sup_{f \in \mathcal{F}} R(\hat{f}_n, f) \leq C\phi_n \}.$$

Voir [Cohen *et al.* 01, Kerkycharian & Picard 02].

# Remarque préliminaire à la construction d'estimateurs

## Approche plug-in naïve

- ▶ On a vu que pour estimer  $T(F)$ , on pouvait utiliser  $\hat{T}_n = T(\hat{F}_n)$  où  $\hat{F}_n$  fdr empirique,
- ▶ Ici, la densité est la **dérivée** de la fdr :  $f = F'$  d'où l'idée naïve de prendre  $\hat{f}_n = \hat{F}'_n$ .
- ▶ Pbm :  $\hat{F}_n$  n'est **pas dérivable**. Cet estimateur plug-in n'est pas défini !

# Sommaire Quatrième partie

## Introduction

### Estimateur par histogramme

- Construction et risque quadratique

- Choix de la partition par validation croisée

### Estimateur à noyau

- Construction et risque quadratique

- Choix de la fenêtre par validation croisée

- Estimateur à noyau des  $k$  plus proches voisins

### Estimateur par projection

- Construction

- Propriétés

### Cas des densités multivariées

- Fléau de la dimension

- Généralisations des estimateurs précédents

### Cas des densités monotones ou unimodales ou convexes ...

- Densités monotones

- Densités unimodales

- Autres contraintes



# Sommaire Quatrième partie

## Introduction

### Estimateur par histogramme

#### Construction et risque quadratique

Choix de la partition par validation croisée

### Estimateur à noyau

#### Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

### Estimateur par projection

#### Construction

#### Propriétés

### Cas des densités multivariées

#### Fléau de la dimension

Généralisations des estimateurs précédents

### Cas des densités monotones ou unimodales ou convexes ...

#### Densités monotones

#### Densités unimodales

#### Autres contraintes

# Histogramme I

On suppose que la densité  $f$  est définie sur un **intervalle borné**  $[a, b] \subset \mathbb{R}$  et  $f \in \mathbb{L}_2([a, b])$ .

## Définition

- ▶ Soit  $I = (I_k)_{1 \leq k \leq D}$  une **partition** de  $[a, b]$  (i.e. intervalles disjoints dont l'union est  $[a, b]$ ),
- ▶ On note  $n_k = \text{Card}\{i; X_i \in I_k\}$  le nombre d'observations dans  $I_k$ , et  $|I_k|$  la longueur de l'intervalle  $I_k$ .
- ▶ L'**estimateur par histogramme** de  $f$  est défini par

$$\hat{f}_{I,n}(x) = \sum_{k=1}^D \frac{n_k}{n|I_k|} 1_{I_k}(x).$$

- ▶ Il affecte à chaque intervalle une valeur égale à la **fréquence des observations** dans cet intervalle, **renormalisée** par la longueur de l'intervalle.

# Histogramme II

## Histogrammes dits "réguliers"

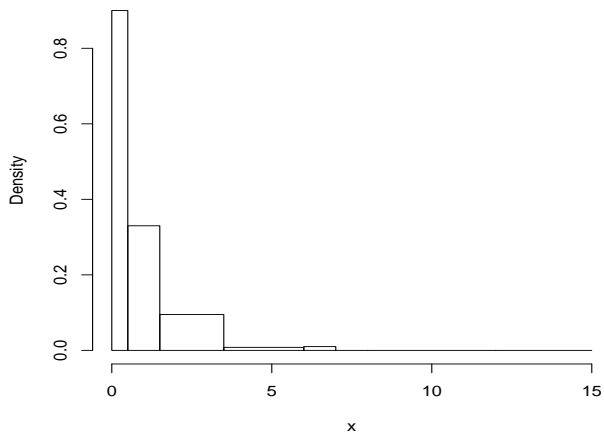
- ▶ Un histogramme est dit régulier lorsque tous les intervalles  $I_k$  de la partition ont la même longueur.
- ▶ Dans ce cas,  $|I|$  est aussi appelé fenêtré.
- ▶ Un histogramme régulier prend des valeurs proportionnelles à la fréquence des observations dans chaque intervalle.

## Remarque

- ▶ L'histogramme est une fonction constante par morceaux. C'est donc une fonction très irrégulière. Cette notion de régularité n'a rien à voir avec la précédente ...

# Histogramme III

Histogramme non régulier



Rem : La hauteur n'est pas proportionnelle à la fréquence

# Risque quadratique des histogrammes I

## Risque quadratique [Freedman and Diaconis 81]

- ▶ On suppose que  $f$  est deux fois dérivable, dans  $\mathbb{L}_2([a, b])$ , avec  $f' \in \mathbb{L}_2([a, b])$  et  $f'' \in \mathbb{L}_p([a, b])$  pour un certain  $p \in [1, 2]$ ,
- ▶ Alors on montre que pour des histogrammes réguliers, la **fenêtre**  $|I|$  qui minimise le risque quadratique est de l'ordre de  $O(n^{-1/3})$ .
- ▶ De plus, pour ce choix de fenêtre, le risque quadratique de l'estimateur par histogramme décroît en  $O(n^{-2/3})$ .
- ▶ En particulier, **asymptotiquement**, si la fenêtre  $|I|$  décroît comme  $O(n^{-1/3})$ , on obtient un estimateur **consistant**.

# Risque quadratique des histogrammes II

## Remarque

- ▶ La densité  $f$  est supposée à support **borné**  $[a, b]$ . En pratique, on observe des valeurs dans l'intervalle  $[X_{(1)}, X_{(n)}]$  et on ne peut pas estimer  $f$  en dehors de ces bornes sans **hypothèses de régularité** supplémentaires.

# Risque quadratique des histogrammes III

## Considérations pratiques

- ▶ Ce résultat ne permet pas de choisir en pratique la fenêtre  $|I|$ .
- ▶ Règle empirique de Sturges :
  - ▶ Choisir le nombre de segments de la partition  $D = 1 + \log_2 n$ .
  - ▶ Règle empirique, fondée sur la loi normale, le TCL et le triangle de Pascal.
  - ▶ C'est la règle par défaut de la fonction *hist* dans R.
- ▶ La librairie *MASS* contient la fonction *truehist* qui est plus évoluée que *hist*. En particulier, on peut
  - ▶ contrôler la taille des intervalles ou bien leur nombre,
  - ▶ sélectionner automatiquement la partition avec des considérations de type [Freedman and Diaconis 81].
- ▶ Cependant, peut-on construire une règle moins empirique que Sturges et utilisable en pratique ?

# Sommaire Quatrième partie

## Introduction

### Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

### Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

### Estimateur par projection

Construction

Propriétés

### Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

### Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes



# Choix de la partition optimale

## Minimisation du risque MISE et estimateur oracle

- ▶ On veut choisir la partition  $I$  qui **minimise** le risque quadratique intégré (MISE)  $R(I, n, f) := \mathbb{E}_f \|\hat{f}_{I,n} - f\|_2^2$ . Ainsi

$$I^* = \underset{I \in \mathcal{I}}{\operatorname{Argmin}} R(I, n, f),$$

où  $\mathcal{I}$  est l'ensemble des partitions de  $[a, b]$ .

- ▶ **Pbm** : Le MISE dépend de la densité inconnue  $f$ .

$$\underset{I \in \mathcal{I}}{\operatorname{Argmin}} R(I, n, f) = \underset{I \in \mathcal{I}}{\operatorname{Argmin}} \mathbb{E}_f \left[ \|\hat{f}_{I,n}\|_2^2 - 2 \int_x \hat{f}_{I,n}(x) f(x) dx \right].$$

Donc  $I^*$  n'est pas un estimateur. On dit que c'est un **oracle**.

- ▶ On va donc **estimer ce risque** pour sélectionner une partition  $I$ .

# Méthodes d'estimation du risque I

## Validation croisée ( $V$ -fold cross validation)

- ▶ On découpe l'échantillon de départ  $X_1, \dots, X_n$  en  $V$  paquets  $C_1, \dots, C_V$  de même taille  $n/V$ ,
- ▶ Pour  $1 \leq v \leq V$ ,
  - ▶ on construit l'estimateur  $\hat{f}_I^v$  à partir de **toutes les observations sauf** le paquet  $C_v$
  - ▶ **Lorsque c'est possible** : on construit  $\hat{R}^v(I)$  estimateur du risque  $R(\hat{f}_I^v, f)$  de  $\hat{f}_I^v$ , à partir des observations du paquet  $C_v$  qui sont **indépendantes** des précédentes.
- ▶ On construit un estimateur du risque global  $R(I, n, f)$  via

$$\hat{R}^{CV}(I) = \frac{1}{V} \sum_{v=1}^V \hat{R}^v(I).$$

# Méthodes d'estimation du risque II

## Variantes : Leave-one-out et leave-p-out

- ▶ Leave-one-out : validation croisée avec  $V = n$ , *i.e.* à chaque étape, on utilise  $n - 1$  observations pour construire l'estimateur et l'observation restante permet d'estimer son risque.
- ▶ Leave-p-out : même principe que CV appliqué à tous les paquets possibles de  $p$  variables parmi  $n$ , *i.e.* à chaque étape, on utilise  $n - p$  observations pour construire l'estimateur et les  $p$  observations restantes permettent d'estimer son risque.

## Différence entre $V$ -fold CV et leave-p-out

- ▶ Dans la  $V$ -fold, chaque variable appartient à un paquet et un seul. En particulier, chaque variable n'est utilisée qu'une seule fois pour estimer le risque.

# Estimation du MISE de l'estimateur par histogramme I

## Mise en œuvre pour l'estimateur par histogramme

- ▶ Il reste donc à **construire** les estimateurs  $\hat{R}^v(I)$  des risques  $R(\hat{f}_I^v, f)$  de chaque estimateur  $\hat{f}^v$ . Or (on note  $p = n/V$ )

$$\begin{aligned} R(\hat{f}_I^v, f) &= \mathbb{E}_f \left[ \|\hat{f}_I^v\|_2^2 - 2 \int_x \hat{f}_I^v(x) f(x) dx \right] + \text{cte} \\ &= \int_x \sum_{k=1}^D \mathbb{E}_f \left( \frac{n_k^v}{(n-p)|I_k|} \right)^2 1_{I_k}(x) dx \\ &\quad - 2 \int_x \sum_{k=1}^D \mathbb{E}_f \left( \frac{n_k^v}{(n-p)|I_k|} \right) 1_{I_k}(x) f(x) dx + \text{cte}, \end{aligned}$$

où  $n_k^v = \text{Card}\{i \notin C_v; X_i \in I_k\}$ .

- ▶ Il faut donc **estimer**  $\mathbb{E}_f(n_k^v)$  et  $\mathbb{E}_f[(n_k^v)^2]$ .

## Estimation du MISE de l'estimateur par histogramme II

- ▶ Or  $\mathbb{E}_f(n_k^v) = (n - p)\mathbb{P}(X \in I_k)$  s'estime sur les observations dans  $C_v$  par :  $(n - p)(n_k - n_k^v)/p$ ,
- ▶ Formule plus compliquée mais analogue pour  $\mathbb{E}_f[(n_k^v)^2]$ .
- ▶ Au final, on peut montrer par exemple pour l'estimateur leave-p-out [Celisse and Robin 08]

$$\hat{R}^{\text{lpo}}(I) = \frac{2n - p}{(n - 1)(n - p)} \sum_k \frac{n_k}{n|I_k|} - \frac{n(n - p + 1)}{(n - 1)(n - p)} \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n}\right)^2.$$

# Estimation par histogramme avec partition LpO optimale

## Procédure

- ▶ On se donne un ensemble de partitions  $\mathcal{I}$  de  $[a, b]$ 
  - ▶ Ex :  $\mathcal{I} = \{I^{2^m}, m_{\min} \leq m \leq m_{\max}\}$  où  $I^N$  est la partition régulière de  $[a, b]$  en intervalles de longueur  $(b - a)/N$ .
  - ▶ En pratique,  $\text{Card}(\mathcal{I})$  doit rester raisonnable pour pouvoir explorer toutes les partitions.
- ▶ Pour chaque  $I \in \mathcal{I}$ , on calcule l'estimateur LpO  $\hat{R}^{\text{lpO}}(I)$ , on sélectionne

$$\hat{I} = \underset{I \in \mathcal{I}}{\text{Argmin}} \hat{R}^{\text{lpO}}(I).$$

- ▶ On estime  $f$  par l'histogramme  $\hat{f}_{\hat{I}}$ .

## Remarques

- ▶ Cet estimateur dépend encore du choix de  $p \in \{1, \dots, n - 1\}$  utilisé pour la procédure LpO.
- ▶ Pourquoi s'arrêter là et ne pas sélectionner le meilleur  $p$  ?

# Sélection automatique de $p$ pour procédure LpO

## Risque MISE et estimateur LpO

- ▶ L'estimateur  $\hat{R}^{\text{lpO}}(I)$  dépend de  $p$ . On le note  $\hat{R}_p(I)$ .
- ▶ Le **risque quadratique de l'estimateur**  $\hat{R}_p(I)$  est donné par

$$MSE(p, I) = \mathbb{E}_f \left[ (\hat{R}_p(I) - R(I, n, f))^2 \right].$$

- ▶ Cette quantité dépend de  $f$ . [Celisse and Robin 08] ont montré que c'est une fonction  $\Phi(p, I, \alpha)$  où  $\alpha = (\alpha_1, \dots, \alpha_D)$  et  $\alpha_k = \mathbb{P}(X \in I_k)$ . On peut donc l'estimer par  $\Phi(p, I, (n_k/n)_{1 \leq k \leq D})$  et sélectionner

$$\hat{p}(I) = \underset{1 \leq p \leq n-1}{\text{Argmin}} \phi(p, I, (n_k/n)_{1 \leq k \leq D}).$$

# Estimation adaptative par histogramme

(On parle d'estimation adaptative lorsque les paramètres optimaux -ex fenêtre- sont sélectionnés automatiquement à partir des observations).

## Procédure

- ▶ On se donne un ensemble de partitions  $\mathcal{I}$  de  $[a, b]$
- ▶ Pour chaque  $I \in \mathcal{I}$ , on calcule
  - ▶ la valeur optimale de  $p$

$$\hat{p}(I) = \underset{1 \leq p \leq n-1}{\operatorname{Argmin}} \phi(p, I, (n_k/n)_{1 \leq k \leq D}),$$

- ▶ puis l'estimateur leave- $\hat{p}(I)$ -out  $\hat{R}_{\hat{p}(I)}(I)$ ,
- ▶ et on sélectionne

$$\hat{I} = \underset{I \in \mathcal{I}}{\operatorname{Argmin}} \hat{R}_{\hat{p}(I)}(I).$$

- ▶ On estime  $f$  par l'histogramme  $\hat{f}_{\hat{I}}$ .



# Résultats supplémentaires

[Celisse and Robin 08] ont montré que

- ▶ La procédure leave-p-out est meilleure que V-fold CV pour estimer le risque quadratique des estimateurs par histogramme,
- ▶ Ils ont fourni des expressions du biais et de la variance de l'estimateur du risque.

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Construction des estimateurs à noyaux I

## Retour aux estimateurs plug-in

Pour  $h$  assez petit, on a

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

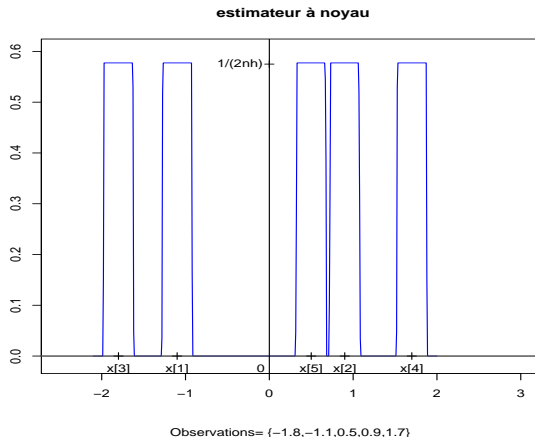
D'où l'estimateur plug-in (non naïf)

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

où  $K_0(u) = 1_{]-1;1]}(u)/2$  est le [noyau de Rosenblatt \(1956\)](#).

# Construction des estimateurs à noyaux II

## Estimateur à noyau (rectangulaire)



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

# Définition des estimateurs à noyau I

## Définition

- ▶ Soit  $K : \mathbb{R} \rightarrow \mathbb{R}$  intégrable telle que  $\int K(u) du = 1$ . Alors  $K$  est appelé **noyau**.
- ▶ Pour tout  $h > 0$  petit (en fait  $h = h_n \rightarrow_{n \rightarrow \infty} 0$ ), on peut définir

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

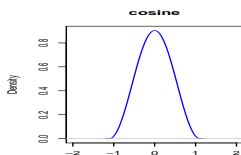
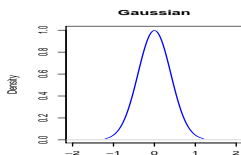
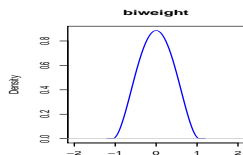
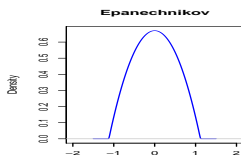
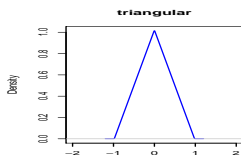
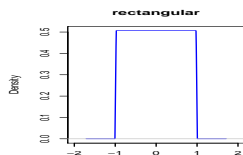
**estimateur à noyau** de  $f$ . On a  $\int \hat{f}_n(x) dx = 1$  et si  $K > 0$  alors  $\hat{f}_n$  est une densité.

- ▶ Le paramètre  $h > 0$  est appelé **fenêtre**. C'est un paramètre de lissage : **plus  $h$  est grand, plus l'estimateur est régulier**.

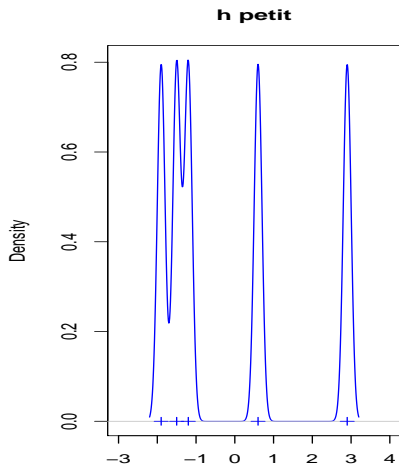
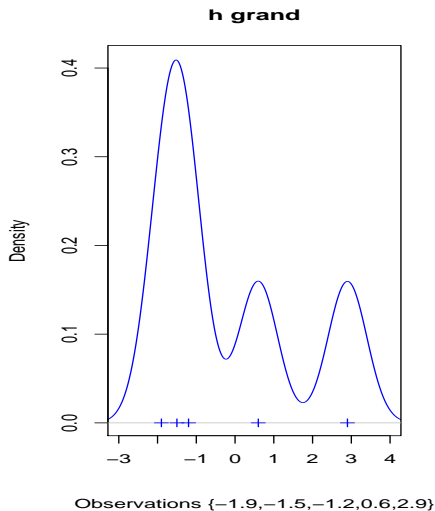
Rem : On considérera souvent des **noyaux positifs et pairs**, mais ce n'est pas obligatoire.

# Exemples de noyaux

- ▶ 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  $K(u) = \frac{\pi}{4} \cos(u\pi/2) 1_{[-1;1]}(u)$ .



# Effet de la variation de $h$ sur l'estimateur à noyau





## Mise en perspective histogrammes/noyaux

- ▶ Dans l'estimateur par histogramme, on calcule la fréquence des observations dans des intervalles **fixés à l'avance**,
- ▶ Dans l'estimateur à noyau **rectangulaire**, on calcule la fréquence des observations dans une fenêtre **glissante**.
- ▶ Dans l'estimateur à noyau **gaussien**, toutes les observations sont prises en compte : celles qui sont proches du point  $x$  où on estime la densité ont **un poids plus important** que les autres.

# Biais des estimateurs à noyau I

## Rappel sur le risque quadratique ponctuel

$$R_x(\hat{f}_n, f) = \mathbb{E}_f(\hat{f}_n(x) - f(x))^2 = \text{Biais}_f^2(\hat{f}_n(x)) + \text{Var}_f(\hat{f}_n(x)).$$

## Étude du biais : principe

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

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

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

# Biais des estimateurs à noyau II

Si de plus  $\|f'\|_\infty < +\infty$  et  $\int |vK(v)|dv < \infty$ , alors on obtient que

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

## Contrôle du biais

- ▶ 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.

# Classe de Hölder (régularité locale) I

## Définitions

- ▶ Pour tout  $\beta \in \mathbb{R}$ , on note  $\lfloor \beta \rfloor$  le plus petit entier strictement inférieur à  $\beta$ .
- ▶ 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  $\mathbb{R}$ .

# Classe de Hölder (régularité locale) II

## Remarques

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

# Noyaux d'ordre $\ell$

## 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  $\forall j \in \{1, \dots, \ell\}$ ,  $\int u^j K(u) du = 0$ .

## Remarques

- ▶ Si  $K$  est un noyau pair alors  $K$  est d'ordre au moins 1.
- ▶ Pour  $j = 0$ , on a  $\int u^j K(u) du = \int K(u) du = 1$ .
- ▶ **On sait construire** des noyaux d'ordre  $\ell$  pour tout entier  $\ell \geq 1$ .

# Biais des estimateurs à noyaux sur la classe $\Sigma_d(\beta, L)$

## Proposition

*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*

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

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

# Variance des estimateurs à noyaux

On montre que

## Proposition

*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*

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



# Compromis biais/variance

## Commentaires

- ▶ Si  $nh \rightarrow \infty$  alors on aura  $\text{Var}_f(\hat{f}_n(x)) \rightarrow 0$ . Donc on veut  $h \rightarrow 0$  (à cause du biais), 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

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

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

# Contrôle du risque quadratique ponctuel I

## Théorème

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

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

où  $C = C(c, \beta, L, K)$ .

# Contrôle du risque quadratique ponctuel II

## 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.
- ▶ En particulier, pour  $\beta = 2$  (densité dérivable avec  $f'$  Lipschitz), on obtient la vitesse  $n^{-4/5}$  pour le risque quadratique (ou  $n^{-2/5}$  pour sa racine carrée).
- ▶ 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  $\beta$** . Voir [Lepski 92].

# Risque quadratique intégré (MISE) I

## Rappel

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

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

## Contrôle de la variance

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

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

# Risque quadratique intégré (MISE) II

## Contrôle du biais

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  $f$ .

## Classe de Nikol'ski (régularité globale)

Soient  $\beta, L > 0$ , on définit la classe de fonctions

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

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

# Risque quadratique intégré (MISE) III

## Proposition

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

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

## Contrôle du risque

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

$$\text{MISE}(\hat{f}_n, \mathcal{N}_d(\beta, L)) = O(n^{-2\beta/(2\beta+1)}).$$

# Risque quadratique des histogrammes versus noyaux I

## Rappels sur les hypothèses

- ▶ Pour l'histogramme, on suppose le support de la densité **borné**, ce qui n'est pas le cas avec les noyaux.
- ▶ Les noyaux estiment des fonctions **régulières** (au sens **local** ou **global**).
- ▶ Dans la définition des classes de Hölder ou Nikol'ski, on suppose l'existence d'une **constante  $L > 0$  fixée** qui majore les "normes" des densités.
- ▶ Les estimateurs à noyau utilisent la connaissance **a priori** de la régularité  $\beta > 0$ , mais il existe des méthodes **adaptatives** pour faire sans.

# Risque quadratique des histogrammes versus noyaux II

## Comparaison des résultats

- ▶ Le risque quadratique des histogrammes décroît en  $n^{-2/3}$ ,
- ▶ Si la densité  $f$  est régulière (en fait dès que  $\beta > 1$ ), le risque quadratique des estimateurs à noyau, qui décroît en  $n^{-2\beta/(2\beta+1)}$  est plus rapide.

## Vitesses minimax

En fait la vitesse  $n^{-2\beta/(2\beta+1)}$  est une vitesse minimax

- ▶ pour le MSE sur la classe  $\Sigma_d(\beta, L)$ ,
- ▶ pour le MISE sur la classe  $\mathcal{N}_d(\beta, L)$ .

## Remarque

Les résultats sur les contrôles du risque ne donnent pas de règle pratique pour choisir la fenêtre  $h$  de l'estimateur.



# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

**Choix de la fenêtre par validation croisée**

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Choix de la fenêtre optimale pour le risque MISE I

## Rappel sur le MISE

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

$$\begin{aligned} \underset{h>0}{\text{Argmin}} MISE(h) &= \underset{h>0}{\text{Argmin}} \mathbb{E}_f \int \hat{f}_{n,h}^2(x) dx - 2\mathbb{E}_f \int f(x)\hat{f}_{n,h}(x) dx \\ &= \underset{h>0}{\text{Argmin}} J(h). \end{aligned}$$

- ▶ Comme  $J$  est inconnue (puisque dépend de  $f$  inconnue), on propose de l'estimer et de choisir la fenêtre  $h > 0$  qui minimise son estimateur.
- ▶ Par validation croisée, on calcule  $\hat{f}_{n,h}^v$  sur toutes les variables sauf le paquet  $C_v$ , sur lequel on estime le risque.

# Choix de la fenêtre optimale pour le risque MISE II

## Stratégie d'estimation de $J$

- ▶  $\mathbb{E}_f \int \hat{f}_{n,h}^2(x) dx$  est estimé sans biais par  $\int \hat{f}_{n,h}^2(x) dx$ ,
- ▶ Pour estimer sans biais  $\mathbb{E}_f \int f(x) \hat{f}_{n,h}(x) dx$  on pourrait penser prendre  $\int \hat{f}_{n,h}^2(x) dx$  mais ça ne marche pas (puisque que cette quantité est un estimateur sans biais de  $\mathbb{E}_f \int \hat{f}_{n,h}^2(x) dx$ ).
- ▶ On remarque plutôt que

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

et on introduit

$\hat{T}_n = \frac{1}{|C_v|(|C_v|-1)h} \sum_{i \in C_v} \sum_{j \in C_v, j \neq i} K\left(\frac{X_i - X_j}{h}\right)$ , estimateur sans biais de  $\frac{1}{h} \int \int f(x) K\left(\frac{u-x}{h}\right) f(u) du dx$ .

# Choix de la fenêtre optimale pour le risque MISE III

## Remarque

De façon très générale, il est important quand on estime par 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 :

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

## Choix de la fenêtre optimale pour le risque MISE IV

Ainsi, on définit

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

où  $\hat{f}_{n,h}^v$  est calculé sur toutes les variables sauf celles du paquet  $C_v$ .  
Puis

$$h^{CV} = \underset{h>0}{\text{Argmin}} \hat{J}_{n,h} \quad \text{et} \quad \hat{f}_n^{CV} \equiv \hat{f}_{n,h^{CV}}.$$

On obtient un estimateur à noyau qui est construit avec la **fenêtre aléatoire**  $h^{CV}$  (qui dépend des observations).

- ▶ On peut mq asymptotiquement,  $\hat{f}_n^{CV}$  minimise en  $h > 0$  le risque  $MISE(h) = R(\hat{f}_{n,h}, f)$  pour la densité observée.
- ▶ Rem : on ne sait pas estimer le MSE par CV.

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

**Estimateur à noyau des  $k$  plus proches voisins**

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Estimateur à noyau des $k$ plus proches voisins

## Principe

- ▶ On se donne une **distance**  $d$  sur l'ensemble des observations ( $\mathbb{R}$  ou  $\mathbb{R}^m$ ) et un noyau  $K : \mathbb{R} \rightarrow \mathbb{R}^+$  (ex : noyau gaussien),
- ▶ Estimateur à noyau de fenêtre  **$h$  variable** : en tout point  $x$ , on considère les  $k$  observations les plus proches de  $x$

$$\hat{f}_n^{knn}(x) = \frac{1}{n[V_k(x)]^m} \sum_{i=1}^n K\left(\frac{d(X_i, x)}{V_k(x)}\right),$$

où  $V_k(x)$  est le **rayon** de la plus petite boule de centre  $x$  qui contient  $k$  observations ( $m$  dimension de l'espace).

- ▶ La "fenêtre" **s'adapte**. Pbm : choisir  $k$ .

## Propriétés des $k$ plus proches voisins [Mack & Rosenblatt 79]

- ▶ Sous l'hyp.  $f$  bornée et 2 fois dérivable,
- ▶ En choisissant  $K$  tel que  $\int |x|K(x)dx = 0$  et  $\int x^2K(x)dx < +\infty$ ,

On obtient (unidimensionnel)

$$\text{Var}(\hat{f}_n^{knn}(x)) = O(k^{-1}), \quad \text{Biais}(\hat{f}_n^{knn}(x)) = O((k/n)^2).$$

Contrôles similaires à ceux des estimateurs à noyau ( $k = nh$ ).

### Remarques

- ▶ Les méthodes de  $k$  plus proches voisins sont peu utilisées en estimation de densité,
- ▶ Par contre, elles sont très populaires en **classification supervisée**.



# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

**Estimateur par projection**

**Construction**

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Construction I

Dans cette section, on suppose que  $f \in \mathbb{L}_2(I)$  où  $I = \mathbb{R}$  ou  $[a, b]$ .

## Base orthonormée (b.o.n.)

- ▶ Soit  $(\phi_j)_{j \geq 1}$  une b.o.n. de  $\mathbb{L}_2(I)$ ,
- ▶ On a  $f = \sum_{j \geq 1} \theta_j \phi_j$ , au sens d'une série convergente dans  $\mathbb{L}_2(I)$ , et où  $\theta_j = \int_I f(x) \phi_j(x) dx$  est la projection de  $f$  sur la  $j$ ème coordonnée de la base.

## Estimateur des coordonnées $\theta_j$

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

estimateur **sans biais** de  $\theta_j$ .

# Construction II

## Définition

- ▶ Soit  $\{\hat{\theta}_j\}_{j \geq 1}$  une suite d'estimateurs des coordonnées  $\{\theta_j\}_{j \geq 1}$ , on définit l'estimateur par projection de  $f$  via

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

*i.e.* on prend la projection de  $f$  sur les  $N$  premières coordonnées de la base et on estime ses coordonnées.

- ▶  $N$  joue le rôle d'un **paramètre de lissage**, comme  $h$  auparavant.
- ▶ Compromis sur la taille de la base (= la valeur de  $N$ ). Plus  $N$  est grand, plus le biais est petit, plus la variance est grande.

# Exemples de bases I

Bases (régulières dyadiques) d'histogrammes (sur  $[0, 1]$ )

On fixe  $p \geq 1$  et

$$\phi_k(x) = c_k 1_{\left\{ \left[ \frac{k-1}{2^p}, \frac{k}{2^p} \right[ \right\}}, \quad 1 \leq k \leq 2^p,$$

où  $c_k = 2^{p/2}$  constante de normalisation.

Base trigonométrique (de Fourier) de  $\mathbb{L}_2([0, 1])$

Définie par :  $\phi_1 \equiv 1$ , et

$$\begin{aligned} \forall k \geq 1, \quad \phi_{2k} : x &\rightarrow \sqrt{2} \cos(2\pi kx) \\ \text{et} \quad \phi_{2k+1} : x &\rightarrow \sqrt{2} \sin(2\pi kx). \end{aligned}$$

## Exemples de bases II

### Bases d'ondelettes de $\mathbb{L}_2(\mathbb{R})$

Soit  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  une fonction **suffisamment régulière**. On définit les fonctions **translatées en échelle et en temps**

$$\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k) \quad \forall k, j \in \mathbb{Z}.$$

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

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

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

# Exemples de bases III

## Remarques

- ▶ 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**.

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

**Propriétés**

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes



# Propriétés

## Risque quadratique

- ▶ Les estimateurs par projection ont le même type de performances que les estimateurs à noyaux (sur des classes de fonctions régulières, pour un bon choix de  $N$ )
- ▶ On peut mettre en œuvre des techniques de **sélection de modèle** pour choisir  $N$ .

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

**Cas des densités multivariées**

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

**Cas des densités multivariées**

**Fléau de la dimension**

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Estimation de densités multivariées I

## Théorie vs pratique

- ▶ Tous les estimateurs présentés dans cette partie se généralisent à la dimension supérieure,
- ▶ En pratique, l'estimation devient plus difficile quand la dimension augmente : c'est le **fléau de la dimension**.

# Estimation de densités multivariées II

## Fléau de la dimension (curse of dimensionality)

- ▶ Exemple 1 :
  - ▶ Dans  $\mathbb{R}$ 
    - ▶ Pts uniformément répartis dans  $[-1, +1]$
    - ▶ 100% de points situées à distance  $\leq 1$  de l'origine
  - ▶ Dans  $\mathbb{R}^{10}$ 
    - ▶ Pts uniformément répartis dans  $[-1, +1]^{10}$
    - ▶ % de points situées à 1 distance  $\leq 0.75$  de l'origine : 5%
- ▶ Exemple 2 : on veut construire un histogramme en s'appuyant sur au moins une moyenne de 10 points par intervalle et 10 classes par variable
  - ▶  $\mathbb{R}$  : 10 classes  $n = 100$
  - ▶  $\mathbb{R}^2$  : 100 classes  $n = 1000$
  - ▶  $\mathbb{R}^{10}$  :  $10^{10}$  classes  $n = 10^{11} = 100\text{billiards}$

# Estimation de densités multivariées III

## Fléau de la dimension (suite)

- ▶ Si  $p$  assez grand, l'espace  $\mathbb{R}^p$  est pratiquement vide : difficulté dans l'emploi des méthodes avec fenêtres,
- ▶ Les points voisins d'un point donné sont tous très loin : difficultés dans l'emploi de méthodes du type  $k$ -plus proches voisins.

## Représentations graphiques

- ▶ Problèmes pour représenter graphiquement les données,
- ▶ En dimension 2, on peut encore représenter des densités, au-delà, ça devient compliqué ...

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

**Cas des densités multivariées**

Fléau de la dimension

**Généralisations des estimateurs précédents**

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Estimateurs à noyaux multivariés I

## Définition dans $\mathbb{R}^2$

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

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

## Généralisation dans $\mathbb{R}^p$

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



# Estimateurs à noyaux multivariés II

## Propriétés

- ▶ Contrôles similaires du biais et de la variance, pour des classes de régularité généralisées à la dimension  $p > 1$ .
- ▶ Exemple avec  $f \in \Sigma_{d,p}(1, L) =$  ensemble des densités sur  $\mathbb{R}^p$  qui sont Lipschitziennes
  - ▶ le biais ne dépend pas de la dimension de l'espace :  
Biais =  $O(h)$ ,
  - ▶ Par contre, la variance dépend de  $p$  :  
 $\text{Var}_f(\hat{f}_n(x)) = O(1/(nh^p))$
  - ▶ la fenêtre optimale pour le risque quadratique est  
 $h = cn^{-1/(2+p)}$
  - ▶ et la vitesse de convergence du risque quadratique correspondante est  $n^{-2/(2+p)}$ .
  - ▶ Quand la dimension  $p$  augmente, cette vitesse est plus lente.

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Densités monotones

On suppose  $f$  densité monotone.

## Exemples

- ▶ temps d'attente avant un évènement  $f : [0, +\infty) \rightarrow [0, +\infty)$  décroissante (ex : loi exponentielle)
- ▶ distribution des  $p$ -values sous l'hypothèse alternative  $f : [0, 1] \rightarrow [0, +\infty)$  décroissante,

# Estimateur de Grenander pour densités monotones I

## Log-vraisemblance

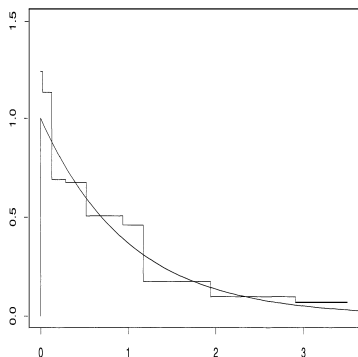
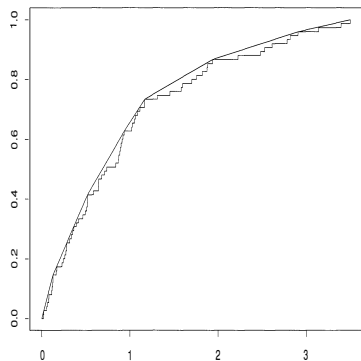
$$\ell_n(f) = \frac{1}{n} \sum_{i=1}^n \log f(X_i),$$

log-vraisemblance de la densité  $f$ .

- ▶ Si  $f$  n'est pas contraint (densité quelconque, même supposée régulière) alors  $\sup_f \ell_n(f) = +\infty$  et l'e.m.v. n'est pas défini.
- ▶ Si  $f$  est monotone, alors l'e.m.v. existe et est unique.
- ▶ Lorsque  $f$  est décroissante, c'est la **dérivée à gauche du plus petit majorant concave de la fdr empirique  $\hat{F}_n$**  (Rem : nécessairement décroissante).
- ▶ Lorsque  $f$  est croissante, c'est la **dérivée à droite du plus grand minorant convexe de la fdr empirique  $\hat{F}_n$**  (Rem : nécessairement croissante).

## Estimateur de Grenander pour densités monotones II

À gauche : plus petit majorant concave de la fdr empirique d'un échantillon de  $n = 75$  variables de loi exponentielle. À droite : dérivée à gauche et vraie densité.



# Estimateur de Grenander pour densités monotones III

## Autre expression

- ▶ On considère les variables ordonnées  
 $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ .
- ▶ On note  $\hat{F}_n$  la fdr empirique et  $\hat{\mathbb{F}}_n$  son plus petit majorant concave :  $\hat{\mathbb{F}}_n$  est linéaire par morceaux.
- ▶ Sur chaque intervalle  $(X_{(i-1)}, X_{(i)}]$ , l'estimateur  $\hat{f}_n$  vaut la pente de  $\hat{\mathbb{F}}_n$ .

## Implémentation

- ▶ Sous R, la fonction *grenander* de la bibliothèque *fdrtool* implémente l'estimateur de Grenander.

# Propriétés

- ▶ L'estimateur de Grenander est également l'estimateur **des moindres carrés** de  $f$ .
- ▶ Estimateur consistant :  $\forall x, \hat{f}_n(x) \rightarrow f(x)$  presque sûrement.
- ▶ Vitesse de convergence du risque quadratique  $n^{-1/3}$ , sans hypothèse de dérivées (à comparer à  $n^{-m/(2m+1)}$  lorsqu'on suppose  $f$  est  $m$ -fois dérivable).
- ▶ Cette vitesse est **minimax** : c'est la meilleure possible si on suppose juste la monotonie.



# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

**Densités unimodales**

Autres contraintes

# Densités unimodales I

## Définition

- ▶ S'il existe  $M_f \in \mathbb{R}$  tel que  $f$  est croissante sur  $(-\infty, M_f]$  et décroissante sur  $[M_f, +\infty)$  alors  $f$  est dite unimodale.
- ▶ Le mode  $M_f$  n'est pas nécessairement unique.

## Estimation à mode connu

- ▶ Si le mode  $M_f$  est connu a priori, alors l'estimation de  $f$  se fait sur chaque intervalle  $(-\infty, M_f]$  et  $[M_f, +\infty)$  via l'estimateur de Grenander.
- ▶ Si aucune observation ne prend la valeur du mode, alors on peut montrer que cet estimateur maximise la vraisemblance.

# Densités unimodales II

## Estimation à mode inconnu

Si  $M_f$  n'est pas a priori connu, l'e.m.v. n'existe pas.

- ▶ L'idée naïve qui consiste à estimer le mode par une autre méthode, et utiliser l'estimateur de Grenander avec le mode estimé, ne fonctionne pas (sans hyps supplémentaires sur  $f$ ),
- ▶ On peut par contre pour chaque valeur de  $M$  fixée, considérer l'estimateur de Grenander  $\hat{f}_n^M$  (de fdr associée  $\hat{\mathbb{F}}_n^M$ ), puis vouloir sélectionner **le meilleur**, par exemple au sens suivant

$$\hat{f} = \underset{\hat{f}_n^M}{\operatorname{Argmin}} \|\hat{\mathbb{F}}_n^M - \hat{F}_n\|_\infty,$$

où  $\hat{F}_n$  fdr empirique.

- ▶ On peut mq cet estimateur a les mêmes performances asymptotiques que l'estimateur de Grenander.

# Sommaire Quatrième partie

Introduction

Estimateur par histogramme

Construction et risque quadratique

Choix de la partition par validation croisée

Estimateur à noyau

Construction et risque quadratique

Choix de la fenêtre par validation croisée

Estimateur à noyau des  $k$  plus proches voisins

Estimateur par projection

Construction

Propriétés

Cas des densités multivariées

Fléau de la dimension

Généralisations des estimateurs précédents

Cas des densités monotones ou unimodales ou convexes ...

Densités monotones

Densités unimodales

Autres contraintes

# Densités convexes

On suppose  $f$  densité convexe décroissante,

## EMV et MC

- ▶ L'e.m.v. et l'estimateur des moindres carrés sont alors bien définis et uniques,
- ▶ Par contre, ils ne coïncident pas en général.
- ▶ Sous l'hyp supplémentaire  $f$  deux fois dérivable, l'e.m.v. converge à la vitesse ponctuelle  $n^{-2/5}$

Voir [Groeneboom *et al.* 01] pour plus de détails. Il existe un cadre plus général des fonctions  $k$ -monotones [Balabdaoui & Wellner 08].

# Densités log-concaves

## Définition

Une densité  $f$  est dite **log-concave** si  $-\log(f)$  est une fonction convexe sur le support de  $f$  (convention  $-\log 0 = +\infty$ ).

## Exemples (paramétriques)




Gaussienne, uniforme, Gamma, Beta, Laplace, logistique, ...

## Propriétés




- ▶ Les fonctions log-concaves sont nécessairement **unimodales**, mais la réciproque est fautive.
- ▶ L'estimateur du max de vraisemblance existe et peut être obtenu par des algos de maximisation sous contrainte.

[Rufibach 06, Rufibach 07] pour plus de détails.

# Références I





-  [Balabdaoui & Wellner 08] Balabdaoui and Wellner  
Estimation of a  $k$ -monotone density : Limit distribution theory  
and the spline connection.  
*Ann. Statist.* 35 : 2536–2564, 2008.
-  [Celisse & Robin 2008] Celisse, A. and S. Robin (2008).  
Nonparametric density estimation by exact leave- $p$ -out  
cross-validation.  
*Comput. Statist. Data Anal.*, 52(5) :2350–2368.
-  [Cohen *et al.* 01] A. Cohen and R. DeVore and  
G. Kerkycharian and D. Picard  
Maximal spaces with given rate of convergence for  
thresholding algorithms.  
*Bernoulli*, 8(2) :219–253, 2002.

## Références II

-  [Freedman & Diaconis 81] D. Freedman and P. Diaconis  
On the Histogram as a Density Estimator :  $\mathbb{L}_2$  Theory.  
*Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 57(4) :453–476,  
1981.
-  [Groeneboom *et al.* 01] Groeneboom & Jongbloed and  
Wellner  
Estimation of a convex function : Characterization and  
asymptotic theory.  
*Ann. Statist.* 29 : 1653–1698, 2001.
-  [Kerkycharian & Picard 02] G. Kerkycharian and D. Picard  
Minimax or maxisets ?  
*ACHA*, 11 :167–191, 2001.



## Références III

-  [Lepski 92] O.V. Lepski  
Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates.  
*Theory Probab. Appl.*, 37(3) :433–448, 1992.
-  [Mack & Rosenblatt 79] Y.P. Mack and M. Rosenblatt  
Multivariate k-nearest neighbour density estimates.  
*J. Multivariate Anal.*, 9 :1–15, 1979.
-  [Rufibach 06] Rufibach, K.  
Log-concave density estimation and bump hunting for I.I.D. observations.  
Ph.D. thesis, Universities of Bern and Göttingen, 2006.
-  [Rufibach 07] Rufibach, K.  
Computing maximum likelihood estimators of a log-concave density function.  
*J. Statist. Comp. Sim.* 77 : 561–574, 2007.

# Cinquième partie V

## Régression non paramétrique

## Contexte

- ▶  $(X, Y)$  couple de v.a.r. avec  $\mathbb{E}|Y| < +\infty$ .
- ▶  $r : x \rightarrow r(x) = \mathbb{E}(Y|X = x)$  la fonction de régression de  $Y$  sur  $X$ .
- ▶  $(X_1, Y_1), \dots, (X_n, Y_n)$  échantillon de même loi que  $(X, Y)$ .
- ▶ On veut 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  $\sigma^2 = \text{Var}(\xi_i)$ .

# Premières remarques

- ▶ Techniques **très similaires** à celles de la partie Estimation de densités,
- ▶ **Différence notable** : les fonctions  $r$  estimées sont différentes des densités, par ex pas intégrables sur  $\mathbb{R}$ . Par contre, on va les supposer **régulières**.
- ▶ **Validation croisée** : toujours faisable grâce à l'existence d'une mesure d'erreur très naturelle :

$$\frac{1}{V} \sum_{v=1}^V \frac{1}{|C_v|} \sum_{i \in C_v} (\hat{r}^v(x_i) - y_i)^2$$

(ex de la  $V$ -fold CV), où  $\hat{r}^v$  est un estimateur obtenu à partir de toutes les observations **sauf** celles du paquet  $C_v$ . **NB** : lors de la formation des paquets  $\{C_v\}_{1 \leq v \leq V}$ , il faut s'assurer que les variables  $x_i$  sont "uniformément réparties" dans chaque paquet (ex : ne pas faire des paquets consécutifs à partir de variables ordonnées).

# Sommaire Cinquième partie

Régressogramme

Estimateur de Nadaraya-Watson (noyaux pour la régression)

Les estimateurs par polynômes locaux

Estimateurs par projection

Splines de régression

Régression non paramétrique multivariée

# Principe

- ▶ Il s'agit de l'équivalent de l'histogramme pour le problème de régression. On suppose que la fonction de régression  $r$  est définie sur un **intervalle borné**  $[a, b] \subset \mathbb{R}$  et  $r \in \mathbb{L}_2([a, b])$ .

## Définition

- ▶ Soit  $I = (I_k)_{1 \leq k \leq D}$  une **partition** de  $[a, b]$  (i.e. intervalles disjoints dont l'union est  $[a, b]$ ),
- ▶ On note  $n_k = \text{Card}\{i; X_i \in I_k\}$  le nombre de variables  $X$  dans  $I_k$ .
- ▶ L'**estimateur par régressogramme** de  $r$  est défini par

$$\hat{r}_{I,n}(x) = \sum_{k=1}^D \left[ \sum_{i=1}^n \frac{Y_i}{n_k} 1_{X_i \in I_k} \right] 1_{I_k}(x).$$

- ▶ Il affecte à chaque intervalle  $I_k$  une valeur égale à la **moyenne des observations  $Y$**  dans cet intervalle, **renormalisée** par le nombre de variables  $X$  de cet intervalle.

# Illustration

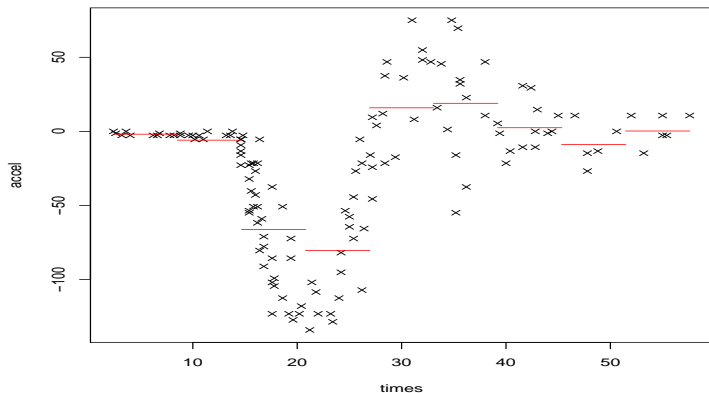


Figure : Régressogramme sur les données *mcycle* de la librairie *MASS*.

# Remarques

Comme pour les histogrammes,

- ▶ Fixer les intervalles à l'avance n'est pas la meilleure chose à faire ;
- ▶ On peut voir la moyenne sur un intervalle comme le résultat d'un lissage par un noyau rectangulaire, et donc préférer des noyaux plus réguliers.

On introduit donc les noyaux pour la régression.



# Sommaire Cinquième partie

Régressogramme

Estimateur de Nadaraya-Watson (noyaux pour la régression)

Les estimateurs par polynômes locaux

Estimateurs par projection

Splines de régression

Régression non paramétrique multivariée

## Estimateur de Nadaraya-Watson I

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,

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

Les densités  $p$  et  $p_X$  sont inconnues mais on peut les estimer via

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

puis on considère l'estimateur de la régression

$$\forall x \in \mathbb{R}, \quad \hat{r}_n(x) = \frac{\int y \hat{p}_n(x, y) dy}{\hat{p}_{n,X}(x)} \mathbf{1}_{\hat{p}_{n,X}(x) \neq 0}. \quad (2)$$

# Estimateur de Nadaraya-Watson II

## Proposition

Si  $K$  est un noyau d'ordre 1, l'estimateur défini par (2) vérifie

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

où  $K_h(\cdot) = K(\cdot/h)$ .

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

# Estimateur de Nadaraya-Watson III

## Démonstration.

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

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

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

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



# Illustration (simpliste)

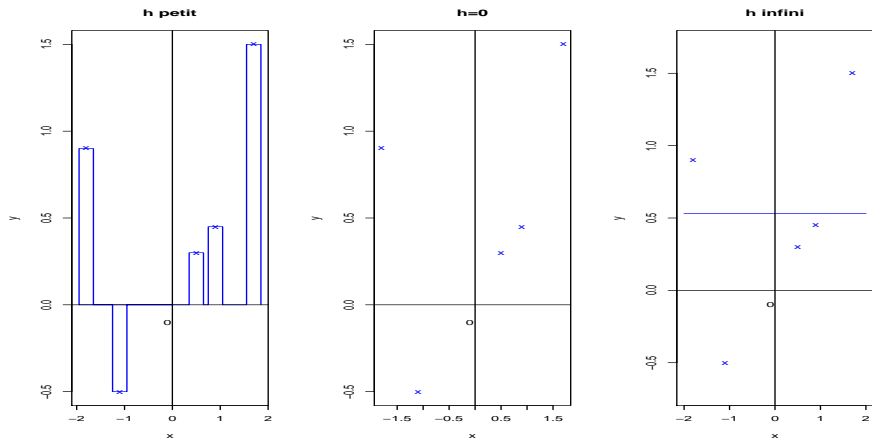


Figure : Estimateur de Nadaraya-Watson avec noyau rectangulaire ( $K : u \rightarrow 1_{|u| \leq 1/2}$ ) pour différentes valeurs de fenêtre  $h$ .

# Illustration

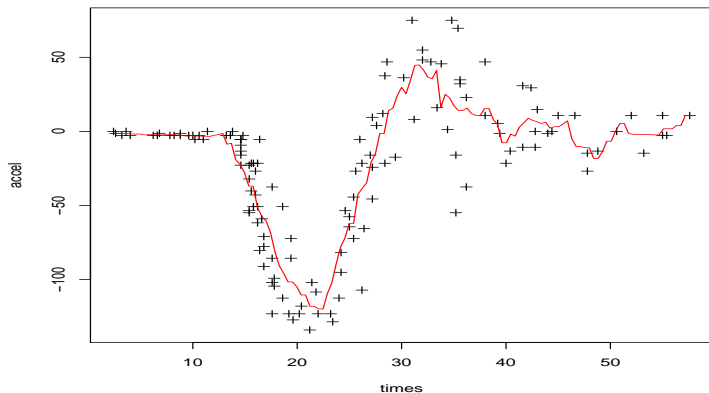


Figure : Estimateur de Nadaraya-Watson sur les données *mcycle* de la librairie *MASS*.

# Expression matricielle

On observe un ensemble de points  $(x_i, y_i)_{1 \leq i \leq n}$ . Les valeurs de  $\hat{r}_n^{NW}$  aux points  $x_i$  s'obtiennent par **lissage** des valeurs  $y_i$  de la façon suivante

$$\begin{pmatrix} \hat{r}_n^{NW}(x_1) \\ \vdots \\ \hat{r}_n^{NW}(x_n) \end{pmatrix} = \begin{pmatrix} \frac{K_h(0)}{\sum_l K_h(x_l - x_1)} & \cdots & \frac{K_h(x_1 - x_n)}{\sum_l K_h(x_l - x_n)} \\ \vdots & \frac{K_h(x_i - x_j)}{\sum_l K_h(x_l - x_j)} & \vdots \\ \frac{K_h(x_n - x_1)}{\sum_l K_h(x_l - x_1)} & \cdots & \frac{K_h(0)}{\sum_l K_h(x_l - x_n)} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

# Interprétation

- ▶ L'estimateur de Nadaraya-Watson est une **moyenne pondérée des observations**  $Y_i$

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

où les poids  $w_{n,i}(x)$  vérifient

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

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



## Remarques I

- ▶ Ne pas confondre **régression linéaire** et **estimateur linéaire** (en les obs.) 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

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

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

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

## Remarques II

- ▶ 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.

# Sommaire Cinquième partie

Régressogramme

Estimateur de Nadaraya-Watson (noyaux pour la régression)

Les estimateurs par polynômes locaux

Estimateurs par projection

Splines de régression

Régression non paramétrique multivariée

# Construction des estimateurs par polynômes locaux I

On remarque que si  $K$  est un noyau positif, alors

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

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

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

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

# Construction des estimateurs par polynômes locaux II

## Conséquences

- ▶ Au sens des **moindres carrés pondérés**, on a approché les  $Y_i$  par une constante (en  $x$ ), notée  $\theta$ .
- ▶ Plus généralement, on peut approcher  $Y_i$  par un polynôme.
- ▶ Cela revient à dire que la fonction  $r$ , au voisinage du point  $x$  peut-être approchée par un polynôme (local), et pas seulement une constante.

# Construction des estimateurs par polynômes locaux III

## Mise en œuvre

Si la fonction  $r$  est  $\ell$  fois dérivable au voisinage de  $x$ , on introduit l'approximation locale polynomiale de  $r$  au voisinage de  $x$

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

Puis on introduit artificiellement  $h$

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

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

$$V_\ell(z) = (1, z, z^2/(2!), \dots, z^\ell/(\ell!))^\top.$$

# Construction des estimateurs par polynômes locaux IV

## Définition

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

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

Alors  $\hat{\theta}_n$  est l'estimateur localement polynomial d'ordre  $\ell$  de la fonction  $\theta : x \mapsto (r(x), r'(x)h, \dots, r^{(\ell)}(x)h^\ell)$ . 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  $r$ .

## Expression matricielle

On observe un ensemble de points  $(x_i, y_i)_{1 \leq i \leq n}$ . Matriciellement, le problème à résoudre est

$$\min_{\theta} (\mathbf{y} - \mathbf{V}_\ell(x)\theta)^\top \mathbf{W}(x) (\mathbf{y} - \mathbf{V}_\ell(x)\theta),$$

où  $\mathbf{y} = (y_1, \dots, y_n)^\top$ ,  $\mathbf{W}(x) = \text{diag}(K_h(x_1 - x), \dots, K_h(x_n - x))$   
et

$$\mathbf{V}_\ell(x) = \begin{pmatrix} 1 & (x_1 - x) & \dots & (x_1 - x)^\ell \\ \vdots & & & \vdots \\ 1 & (x_n - x) & \dots & (x_n - x)^\ell \end{pmatrix}.$$

La solution est donnée par

$$\hat{\theta} = \left( \hat{r}_n(x), \dots, \frac{\hat{r}_n^{(\ell)}(x)}{\ell!} \right) = \{\mathbf{V}_\ell(x)^\top \mathbf{W}(x) \mathbf{V}_\ell(x)\}^{-1} \mathbf{V}_\ell(x)^\top \mathbf{W}(x) \mathbf{y}.$$



# Remarques

- ▶ On a vu que pour  $\ell = 0$ , on a  $\hat{r}_n^{LP(0)} = \hat{r}_n^{NW}$ .
- ▶  $\hat{\theta}_n$  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  $\ell$ .
- ▶ On peut montrer que  $\hat{r}_n^{LP(\ell)}$  est un **estimateur linéaire** (en les obs.) de la fonction de régression

# Biais et variance des estimateurs par polynômes locaux I

## Contexte

- ▶ On considère ici uniquement le modèle de régression à effets fixes sur  $[0; 1]$  *i.e.*

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

et on s'intéresse au **risque ponctuel en  $x \in [0; 1]$  fixé.**

- ▶ On note  $\hat{r}_n$  l'estimateur localement polynomial d'ordre  $\ell$ .
- ▶ On suppose que le design est **suffisamment uniforme** et que le noyau  $K$  est borné et à support compact.

# Biais et variance des estimateurs par polynômes locaux II

## Biais

On suppose que  $r \in \Sigma(\beta, L)$  sur  $[0; 1]$  (classe de Hölder sur l'intervalle  $[0; 1]$ ), pour certaines constantes  $\beta, L > 0$ . Alors il existe  $C > 0$  telle que

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

## Contrôle de la variance

On montre que

$$\sigma^2(x) \leq \frac{2\sigma^2 K_{\max}}{nh}.$$

# Biais et variance des estimateurs par polynômes locaux III

## Compromis biais-variance

On a

$$\begin{aligned}MSE(x) &= \mathbb{E}_r[(\hat{r}_n(x) - r(x))^2] = b^2(x) + \sigma^2(x) \\ &\leq C_1 h^{2\beta} + C_2 \times \frac{1}{nh}.\end{aligned}$$

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

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

# Sommaire Cinquième partie

Régressogramme

Estimateur de Nadaraya-Watson (noyaux pour la régression)

Les estimateurs par polynômes locaux

**Estimateurs par projection**

Splines de régression

Régression non paramétrique multivariée

## Contexte des estimateurs par projection

- ▶ Cadre de la régression à effets fixes sur  $[0; 1]$ .
- ▶ On suppose que  $r \in \mathbb{L}_2([0; 1])$ .
- ▶ Soit  $(\phi_j)_{j \geq 1}$  une b.o.n. de  $\mathbb{L}_2([0; 1])$ . Alors,

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

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

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

- ▶ Ici,  $N$  joue le rôle d'un paramètre de lissage, comme  $h$  auparavant.

## Exemple

Cas du dispositif fixe uniforme sur  $[0; 1]$

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**.

## Exemples de bases de $\mathbb{L}_2$ I

Bases (régulières dyadiques) de régressogrammes (sur  $[0, 1]$ )

On fixe  $p \geq 1$  et

$$\phi_k(x) = c_k 1\left\{\left[\frac{k-1}{2^p}, \frac{k}{2^p}\right[\right\}, \quad 1 \leq k \leq 2^p,$$

où  $c_k = 2^{p/2}$  constante de normalisation.

Base trigonométrique (de Fourier)

Définie par :  $\phi_1 \equiv 1$ , et

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



# Exemples de bases de $\mathbb{L}_2$ II

## Bases d'ondelettes de $\mathbb{L}_2(\mathbb{R})$

Soit  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  une fonction **suffisamment régulière**. On définit les fonctions **translatées en échelle et en temps**

$$\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k) \quad \forall k, j \in \mathbb{Z}.$$

Alors, sous certaines hypothèses sur  $\psi$ , les fonctions  $\{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$  forment une b.o.n. de  $\mathbb{L}_2(\mathbb{R})$ .

## Remarque

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 construire 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.

# Sommaire Cinquième partie

Régressogramme

Estimateur de Nadaraya-Watson (noyaux pour la régression)

Les estimateurs par polynômes locaux

Estimateurs par projection

**Splines de régression**

Régression non paramétrique multivariée

# Introduction I

## Principe des splines

- ▶ Compromis entre régression polynomiale globale et les méthodes de lissage local précédentes.
- ▶ Les splines sont des polynômes par morceaux qui se raccordent de façon lisse.
- ▶ Les points de raccordement sont appelés **nœuds**.
- ▶ **Exemple : spline cubique**. Les polynômes par morceaux sont de degré 3 et contraints à avoir des dérivées d'ordre 2 continues aux noeuds.

# Introduction II

## Représentation

- ▶ À un ensemble fixé de nœuds  $\{\xi_k\}_{1 \leq k \leq K}$ , on associe une base de fonctions.

ex : base de polynômes tronqués de degré  $\ell$ , donnée par

$$\{\phi_j\}_{1 \leq j \leq \ell+1+K} = \{1, x, x^2, x^3, \dots, x^\ell, (x - \xi_1)_+^\ell, \dots, (x - \xi_K)_+^\ell\}$$

- ▶ La fonction de régression  $r$  se décompose dans cette base  $r(x) = \sum_{j=1}^{\ell+K+1} \beta_j \phi_j(x)$ .

- ▶ Les coefficients  $\beta_j$  sont estimés par minimisation de l'erreur quadratique :

$$\min_{\beta} \sum_{i=1}^n \left( y_i - \sum_{j=1}^{\ell+K+1} \beta_j \phi_j(x_i) \right)^2$$

# Les splines vues comme un lissage

On peut montrer que l'estimateur par spline cubique est la solution d'un problème de moindres carrés pénalisés

$$\min_r \sum_{i=1}^n (y_i - r(x_i))^2 + \lambda \int [r^{(2)}(t)]^2 dt \quad (\star)$$

- ▶ Le terme de pénalité agit sur les fluctuations de  $r$ .
- ▶  $\lambda$  est un paramètre de pénalité qui fait le compromis entre l'ajustement aux observations et le contrôle des fluctuations de  $r$ .
- ▶ La solution de  $(\star)$  est unique et c'est la spline cubique dont les nœuds sont les observations  $\{x_i\}_{1 \leq i \leq n}$ .

# Sommaire Cinquième partie

Régressogramme

Estimateur de Nadaraya-Watson (noyaux pour la régression)

Les estimateurs par polynômes locaux

Estimateurs par projection

Splines de régression

Régression non paramétrique multivariée

## Cas multivarié

Ici la covariable  $X = (X^1, \dots, X^p) \in \mathbb{R}^p$ .

- ▶ Tous les estimateurs présentés ont des généralisations naturelles lorsque  $X \in \mathbb{R}^p$ ,
- ▶ mais, le **fléau de la dimension** joue là encore un rôle,
- ▶ et les représentations graphiques sont impossibles dès que  $p > 2$ .
- ▶ En pratique, il faut essayer de ne pas ajouter des variables  $X^j$  qui soient liées aux autres.

## Régularisation et Réduction de la dimension

- ▶ On peut essayer de réduire la dimension par des techniques de **sélection de variables**, ou essayer de **régulariser** le problème.
- ▶ Dans certains contextes, on peut chercher à faire de la régression **parcimonieuse** (ex : méthodes **lasso**).
- ▶ On peut aussi utiliser des **modèles contraints**.

# Quelques modèles contraints de régression multivariée

- ▶ Modèles additifs

$$r(x^1, \dots, x^p) = \sum_{j=1}^p r_j(x^j),$$

où  $(r_1, \dots, r_p)$  fonctions de  $\mathbb{R} \mapsto \mathbb{R}$  inconnues,

- ▶ Modèles à direction révélatrice (single-index model)

$$r(x^1, \dots, x^p) = \phi(\theta^\top \mathbf{x}),$$

où  $\phi : \mathbb{R} \mapsto \mathbb{R}$  et  $\theta \in \mathbb{R}^p$  inconnus.



# Sixième partie VI

## Ré-échantillonnage

# Les techniques de ré-échantillonnage

- ▶ Les méthodes statistiques fondées sur le ré-échantillonnage consistent à partir d'un échantillon  $\mathbf{X} = (X_1, \dots, X_n)$  de variables i.i.d., à **estimer la loi d'une variable  $T(\mathbf{X})$**  sur la base de l'observation de  $\mathbf{X}$ .
- ▶ Le ré-échantillonnage est un cadre générique qui permet de **mesurer des incertitudes et donc la qualité des procédures.**

## Quelques exemples

- ▶ Jackknife [Quenouille 49, Quenouille 56, Tukey 58]
- ▶ Bootstrap [Efron 79]
- ▶ Validation croisée
- ▶ Tests de permutation
- ▶ ...

# Sommaire Sixième partie

Le bootstrap

Exemples d'estimateurs bootstrap

Compléments sur le bootstrap

Jackknife

# Motivation I

## L'exemple des souris

On dispose de 16 souris malades, parmi lesquelles 7 ont été tirées au hasard pour recevoir un nouveau traitement tandis que les autres reçoivent un placebo. On mesure les durées de vie des souris (en jours) après traitement.

groupe traitement	94, 38, 23, 197, 99, 16, 141
groupe placebo	52, 10, 40, 104, 51, 27, 146, 30, 46

On obtient les moyennes de durée de vie suivantes pour chacun des 2 groupes

$$\bar{x}_{trait.} = 86.86 \quad \bar{x}_{placebo} = 56.22$$

Soit une différence  $\bar{x}_{trait.} - \bar{x}_{placebo} = 30.63$ . Est-ce significatif ?

# Motivation II

## Éléments de réponse

- ▶ **Test de Student** sur la moyenne dans chaque groupe : implique une hypothèse de normalité sur les variables,
- ▶ **Test nonparamétrique** : **KS test** ou **test de la somme des rangs de Wilcoxon** (échantillons non appariés) : perte d'information,
- ▶ **Approche naïve** : quelle est la **précision de ces estimateurs** ?

- ▶ Pour la moyenne empirique, on sait calculer et estimer la variance

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X_1)}{n} \text{ estimé par } \frac{s^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2.$$

- ▶ On obtient  $\sqrt{s_{\text{trait.}}^2/n_{\text{trait.}}} = 25.24$ ,  $\sqrt{s_{\text{pl.}}^2/n_{\text{pl.}}} = 14.14$  et l'erreur standard pour  $\bar{x}_{\text{trait.}} - \bar{x}_{\text{placebo}}$  vaut  $\sqrt{25.24^2 + 14.14^2} = 28.93$ .
- ▶ Sous l'hyp que  $\bar{X}_{\text{trait.}} - \bar{X}_{\text{placebo}}$  suit une loi normale, on conclut que **la différence observée n'est pas significative**.

# Motivation III

## Cas général

- ▶ Comment faire si on n'a pas de formule pour estimer l'erreur standard d'un estimateur ?
- ▶ **Ex** : on compare les traitements via leur médiane et non plus leur moyenne.  $med_{trait} = 94$  et  $med_{placebo} = 46$  soit  $med_{trait} - med_{placebo} = 48$ . Mais comment estimer la variance de cette quantité ?
- ▶ **Autre ex** : Si on s'intéresse au coefficient de corrélation entre deux variables, comment on estime la variabilité de ce coefficient ?
- ▶ Au-delà de sa variance, comment évaluer la distribution de l'estimateur ?

# Heuristique du bootstrap

- ▶ Si on avait plusieurs échantillons de la même loi  $\mathbb{P}$ , on pourrait répondre à ces questions.
- ▶ En non paramétrique, la meilleure approximation de  $\mathbb{P}$  est la loi empirique  $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$  (mesure aléatoire),
- ▶ En l'absence de tels échantillons, on peut **simuler des échantillons** sous cette loi empirique  $\mathbb{P}_n$  (loi conditionnelle à l'échantillon initial).

# Objectifs du bootstrap

## Contexte

- ▶ On observe un échantillon  $\mathbf{X} = (X_1, \dots, X_n)$  de variables i.i.d. de loi inconnue  $\mathbb{P}$ .
- ▶ On s'intéresse à une fonctionnelle  $\psi(\mathbb{P})$  de la distribution  $\mathbb{P}$ .

## Problèmes envisagés

- ▶ Estimation ponctuelle de  $\psi(\mathbb{P})$  par  $\hat{\psi} = T(\mathbf{X})$  et étude de la précision de cet estimateur
  - ▶ Biais de  $\hat{\psi}$ , Variance de  $\hat{\psi}$ , Risque MSE de  $\hat{\psi}$ , ...
- ▶ Construction d'une région de confiance pour  $\psi(\mathbb{P})$ ,
- ▶ Construction de tests d'hypothèses sur  $\psi(\mathbb{P})$ .

## Réponses aux pbms envisagés

- ▶ Soit on connaît (exact ou asympt.) la loi de  $T(\mathbf{X}) - \psi(\mathbb{P})$ ,
- ▶ Soit on sait **échantillonner** cette loi et on utilise des procédures de Monte Carlo pour approcher ces quantités.



## Principe général du bootstrap

À partir d'un échantillon  $\mathbf{X} = (X_1, \dots, X_n)$  de variables i.i.d., on considère

- ▶  $T(\mathbf{X})$  une statistique calculée sur l'échantillon initial,
- ▶ Pour  $1 \leq b \leq B$ ,
  - ▶ on tire **aléatoirement et avec remise**  $n$  variables dans l'échantillon de départ, notées  $\mathbf{X}^b = (X_1^b, \dots, X_n^b)$ .
  - ▶ on calcule la valeur de la statistique sur chaque échantillon bootstrap, notée  $T^b = T(\mathbf{X}^b)$ .
- ▶  $\mathbf{T} = (T^1, \dots, T^B)$  est un  $B$ -échantillon de la loi de  $T$  conditionnelle à  $\mathbf{X}$ .
- ▶ Cet échantillon peut être utilisé pour estimer  $\mathbb{E}(T|\mathbf{X})$ ,  $\text{Var}(T|\mathbf{X})$ , des IC pour  $\mathbb{E}(T|\mathbf{X})$ , ... On utilise par exemple

$$\bar{T}_{boot} = \frac{1}{B} \sum_{b=1}^B T^b, \quad \widehat{\text{Var}}_{boot}(T) = \frac{1}{B-1} \sum_{b=1}^B (T^b - \bar{T}_{boot})^2.$$

- ▶ **NB** : pour que l'approx. Monte Carlo soit bonne il faut  $B$  assez grand.

# Éléments de théorie

Soit  $X_1, \dots, X_n$  des variables i.i.d. et  $\mathbb{P}_n$  la distribution empirique de l'échantillon

$$\mathbb{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(\cdot),$$

où  $\delta_a$  est la masse de Dirac au point  $a$ . Alors  $\mathbb{P}_n$  est un estimateur non paramétrique de la vraie distribution  $\mathbb{P}$ .

## Échantillon bootstrap

- ▶ Un échantillon bootstrap est un tirage avec remise parmi les variables  $\{X_1, \dots, X_n\}$  : chaque variable est tirée indépendamment avec probabilité  $1/n$ . C'est donc un échantillon de la loi  $\mathbb{P}_n$ .
- ▶ Les répliqués bootstrap  $T^1, \dots, T^B$  ont la loi de  $T$  sachant  $\mathbf{X}$ .

# Correspondance monde réel et monde bootstrap

Efron extrapole le principe du plug-in pour construire un monde Bootstrap, miroir du monde réel dans lequel aucune quantité n'est inconnue.

Monde réel	Monde bootstrap
$\mathbb{P}$ loi de proba inconnue $\mathbf{X} = (X_1, \dots, X_n)$ éch. initial	$\mathbb{P}_n$ loi empirique $\mathbf{X}^* = (X_1^*, \dots, X_n^*)$ éch. bootstrap $\mathbb{P}(X_j^* = X_i   \mathbf{X}) = n^{-1}, \forall i, j$
$\theta = \psi(\mathbb{P})$ param. de la loi $\mathbb{P}$ $\hat{\theta} = T(\mathbf{X})$ estim. de $\theta$ Loi de $\hat{\theta}$ inconnue approchée par loi de $\hat{\theta}   \mathbf{X}$	$\theta_n = \psi(\mathbb{P}_n)$ param. empirique $\hat{\theta}^* = T(\mathbf{X}^*)$ réplikat bootstrap <b>Appr. Monte Carlo :</b> $\mathbb{P}(\hat{\theta} \in A   \mathbf{X}) \simeq \frac{1}{B} \sum_{b=1}^B 1\{\hat{\theta}^b \in A\}$

**Ex 1 :**  $\mathbb{P} = \mathcal{N}(\theta, 1); \theta = \int x d\mathbb{P}(x); \hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i; \theta_n = \hat{\theta}.$

**Ex 2 :**  $\mathbb{P} = \mathcal{N}(0, \theta); \theta = \int x^2 d\mathbb{P}(x); \hat{\theta} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2;$   
 $\theta_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$

**Remarque :** si  $\hat{\theta}$  est un estimateur par substitution, alors

$$\hat{\theta} = \psi(\mathbb{P}_n) = \theta_n.$$

# Sommaire Sixième partie

Le bootstrap

Exemples d'estimateurs bootstrap

Compléments sur le bootstrap

Jackknife

# Estimation bootstrap du biais d'un estimateur

## Contexte

- ▶  $\mathbf{X} = (X_1, \dots, X_n)$  i.i.d. de loi  $\mathbb{P}$ ,
- ▶  $\theta = \psi(\mathbb{P})$  une fonctionnelle de la loi  $\mathbb{P}$  et  $\theta_n = \psi(\mathbb{P}_n)$  son équivalent empirique,
- ▶  $\hat{\theta}$  un estimateur de  $\theta$  (éventuellement  $\theta_n$ ),
- ▶  $(\mathbf{X}^1, \dots, \mathbf{X}^B)$   $B$  réplicats bootstrap de l'éch. initial (de taille  $n$ ) et  $(\hat{\theta}^1, \dots, \hat{\theta}^B)$  estim. associés.

## Estimation du biais de $\hat{\theta}$

$$\text{biais}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta,$$

$$\widehat{\text{biais}}_{\text{boot}}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^b - \theta_n.$$

## Application : réduction du biais d'un estimateur

À partir d'un estimateur  $\hat{\theta}$  initial, on peut considérer

$$\hat{\theta}_{\text{corrigé}} = \hat{\theta} - \widehat{\text{biais}}_{\text{boot}}(\hat{\theta}).$$

Cas particulier  $\hat{\theta} = \theta_n = \psi(\mathbb{P}_n)$  alors

$$\hat{\theta}_{\text{corrigé}} = 2\hat{\theta} - \frac{1}{B} \sum_{b=1}^B \hat{\theta}^b.$$

### Précautions

- ▶ Réduire le biais peut augmenter la variance!!!
- ▶ En pratique, avoir une idée du biais d'un estimateur n'est utile que si on estime aussi sa variance.

# Estimation bootstrap de la variance d'un estimateur

## Contexte

- ▶  $\mathbf{X} = (X_1, \dots, X_n)$  i.i.d. de loi  $\mathbb{P}$ ,
- ▶  $\theta = \psi(\mathbb{P})$  une fonctionnelle de la loi  $\mathbb{P}$ ,
- ▶  $\hat{\theta}$  un estimateur de  $\theta$ ,
- ▶  $(\mathbf{X}^1, \dots, \mathbf{X}^B)$   $B$  répliquats bootstrap de l'éch. initial (de taille  $n$ ) et  $(\hat{\theta}^1, \dots, \hat{\theta}^B)$  estim. associés.

## Estimation de la variance de $\hat{\theta}$

$$\text{Var}(\hat{\theta}) = \mathbb{E}\{\hat{\theta} - \mathbb{E}(\hat{\theta})\}^2,$$

$$\widehat{\text{Var}}_{boot}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^B \{\hat{\theta}^b - \frac{1}{B} \sum_{b'} \hat{\theta}^{b'}\}^2.$$

# Application : erreur standard des estimateurs obtenus par algorithme EM

## Contexte

- ▶ Modèle à données manquantes ou à variables cachées (mélange, HMM ...).
- ▶ L'estim. du max. de vraisemblance  $\hat{\theta}^{MLE}$  du paramètre  $\theta$  n'a pas d'expression analytique simple. On l'approche via l'algorithme EM (expectation-maximization) par  $\hat{\theta}^{EM}$ .
- ▶ L'algo EM ne fournit pas d'estimateur de  $\text{Var}(\hat{\theta}^{EM})$ .

## Erreur standard de $\hat{\theta}^{EM}$

- ▶ Pour chaque échantillon bootstrap  $\mathbf{X}^b$ , on utilise l'algo EM pour construire un estimateur  $\hat{\theta}^{EM,b}$ .
- ▶ On estime l'erreur standard de  $\hat{\theta}^{EM}$  par

$$\widehat{\text{se}}_{boot}(\hat{\theta}^{EM}) = \left\{ \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{EM,b} - \frac{1}{B} \sum_{b'} \hat{\theta}^{EM,b'})^2 \right\}^{1/2}.$$



# Estimation bootstrap du risque MSE d'un estimateur

## Contexte

- ▶  $\mathbf{X} = (X_1, \dots, X_n)$  i.i.d. de loi  $\mathbb{P}$ ,
- ▶  $\theta = \psi(\mathbb{P})$  une fonctionnelle de la loi  $\mathbb{P}$  et  $\theta_n = \psi(\mathbb{P}_n)$  son équivalent empirique,
- ▶  $\hat{\theta}$  un estimateur de  $\theta$  (éventuellement  $\theta_n$ ),
- ▶  $(\mathbf{X}^1, \dots, \mathbf{X}^B)$   $B$  répliqués bootstrap de l'éch. initial (de taille  $n$ ) et  $(\hat{\theta}^1, \dots, \hat{\theta}^B)$  estim. associés.

## Estimation du MSE (mean square error) de $\hat{\theta}$

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2],$$

$$\widehat{MSE}_{boot}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}^b - \theta_n)^2.$$

# Estimation bootstrap de l'erreur de prédiction (classification supervisée) I

## Contexte

- ▶  $\mathcal{L} = \{(X_1, Y_1), \dots, (X_L, Y_L)\}$  i.i.d. de loi  $\mathbb{P}$  sur  $\mathcal{X} \times \mathcal{Y}$ , avec  $\mathcal{Y}$  fini ; **échantillon d'apprentissage**,
- ▶  $r : \mathcal{X} \rightarrow \mathcal{Y}$  : **règle de classification** (= estimateur),
- ▶ L'erreur de prédiction de la règle  $r$  est  $\mathbb{P}(r(X) \neq Y)$ . On l'estime par son équivalent empirique

$$\widehat{err}(r) = \frac{1}{L} \sum_{i=1}^L 1\{r(X_i) \neq Y_i\}.$$

- ▶ **Pbm** : Si  $r$  est construite sur l'éch. d'apprentissage et si l'erreur de  $r$  est estimée sur **le même** échantillon, alors on a un **estimateur optimiste** (biaisé) de cette erreur.

# Estimation bootstrap de l'erreur de prédiction (classification supervisée) II

## Approches possibles

- ▶ Validation croisée ;
- ▶ Réduction bootstrap du biais de l'estimateur de l'erreur : Pour chaque  $1 \leq b \leq B$ , on tire un éch. bootstrap  $\mathcal{L}^b = \{(X_1^b, Y_1^b), \dots, (X_L^b, Y_L^b)\}$ , on construit le classifieur  $r^b$  sur  $\mathcal{L}^b$  et on calcule
  - ▶  $\widehat{err}_{\mathcal{L}^b}(r^b)$  erreur empirique calculée sur  $\mathcal{L}^b$ ,
  - ▶  $\widehat{err}_{\mathcal{L}}(r^b)$  erreur empirique calculée sur  $\mathcal{L}$ ,
  - ▶ Estimateur bootstrap du biais
$$\widehat{Biais}_{boot}(err) = B^{-1} \sum_b [\widehat{err}_{\mathcal{L}^b}(r^b) - \widehat{err}_{\mathcal{L}}(r^b)],$$
  - ▶ Estimateur bootstrap à biais réduit
$$\widehat{err}_{boot.}(r) = \widehat{err}(r) - \widehat{Biais}_{boot}(err).$$
- ▶ Pbm : l'estimateur bootstrap du biais est encore biaisé (car calculé sur l'éch. d'apprentissage).

# Estimation bootstrap de l'erreur de prédiction (classification supervisée) III

## Bootstrap 632 [Efron & Tibshirani 97]

- ▶ Pour chaque  $1 \leq b \leq B$ , on tire un éch. bootstrap  $\mathcal{L}^b = \{(X_1^b, Y_1^b), \dots, (X_L^b, Y_L^b)\}$ , on construit le classifieur  $r^b$  sur  $\mathcal{L}^b$ , on calcule l'erreur empirique  $\widehat{err}_{\mathcal{L} \setminus \mathcal{L}^b}(r^b)$  sur les variables de  $\mathcal{L} \setminus \mathcal{L}^b$ , puis

$$\varepsilon_0 = B^{-1} \sum_b \widehat{err}_{\mathcal{L} \setminus \mathcal{L}^b}(r^b)$$
$$\widehat{err}_{boot.632}(r) = 0.368 \widehat{err}(r) + 0.632 \varepsilon_0.$$

- ▶ **Explication** :  $0.632 \simeq 1 - 1/e = \text{proba}$  (quand  $L \rightarrow +\infty$ ) qu'un couple de variables  $(X_i, Y_i)$  de l'éch. initial soit présent dans l'éch. bootstrap  $(\mathbf{X}^b, \mathbf{Y}^b)$ .
- ▶ Meilleures performances que la validation croisée.

# Estimation bootstrap de la loi d'une stat. de l'échantillon

## Contexte

- ▶  $\mathbf{X} = (X_1, \dots, X_n)$  i.i.d. de loi  $\mathbb{P}$ ,
- ▶  $\theta = \psi(\mathbb{P})$  une fonctionnelle de la loi  $\mathbb{P}$  et  $\theta_n = \psi(\mathbb{P}_n)$  son équivalent empirique,
- ▶  $\hat{\theta}$  un estimateur de  $\theta$  (éventuellement  $\theta_n$ ),
- ▶  $(\mathbf{X}^1, \dots, \mathbf{X}^B)$   $B$  réplicats bootstrap de l'éch. initial (de taille  $n$ ) et  $(\hat{\theta}^1, \dots, \hat{\theta}^B)$  estim. associés.

## Estimation de la loi de $T$

$(T^1, \dots, T^B)$  est un éch. de la loi de  $T$  cond. à  $\mathbf{X}$  avec lequel on peut

- ▶ calculer des quantiles empiriques (donc par ex. faire des tests),
- ▶ estimer la densité de  $T$  (histogramme, noyau ...),
- ▶ construire des intervalles de confiance sur  $\psi(\mathbb{P})$ ,
- ▶ ...

# Application : test d'unimodalité d'une distribution I

## Contexte

- ▶ On observe  $X_1, \dots, X_n$  i.i.d. de densité  $f$  inconnue.
- ▶ On veut tester  $H_0$  : " $f$  est unimodale" contre  $H_1$  : " $f$  a au moins 2 modes".
- ▶ On estime  $f$  par un estimateur à noyau  $\hat{f}_{n,h}$  de fenêtre  $h$ .
- ▶ Plus  $h$  est grand, plus on sur-régularise et moins on a de modes pour  $\hat{f}_{n,h}$ .
- ▶ Soit  $h_{min} = \min\{h > 0; \forall h' < h, \hat{f}_{n,h'} \text{ a au moins 2 modes}\}$ .

## Construction du test

- ▶ On rejette  $H_0$  si  $h_{min}$  est plus grand qu'un seuil  $s$ .
- ▶ Pour régler le niveau du test, il faut la loi de  $h_{min}$  sous l'hypothèse  $H_0$  : **estimateur bootstrap du quantile  $q_{1-\alpha}$  de la loi de  $h_{min}$** .

# Application : test d'unimodalité d'une distribution II

## En pratique

Pour  $1 \leq b \leq B$ ,

- ▶ On tire un échantillon aléatoire avec remise  $\mathbf{X}^b$  parmi les obs. initiales,
- ▶ On identifie (via une grille sur  $h$ ), la valeur  $h_{min}^b$  pour cet échantillon,
- ▶ On obtient une distribution empirique  $\hat{F}_B^*$  pour  $h_{min}$ .

Au final, on estime le quantile  $q_{1-\alpha}$  de la variable  $h_{min}$  par

$$\hat{q}_{1-\alpha} = (\hat{F}_B^*)^{-1}(1 - \alpha).$$

Il s'agit de la stat d'ordre d'indice  $\lceil B(1 - \alpha) \rceil$  de l'échantillon  $\{h_{min}^b\}_{1 \leq b \leq B}$ .

# Construction d'intervalles de confiance I

## Contexte

- ▶  $\mathbf{X} = (X_1, \dots, X_n)$  i.i.d. de loi  $\mathbb{P}$ ,
- ▶  $\theta = \psi(\mathbb{P}) \in \mathbb{R}$  un **paramètre réel** de la loi  $\mathbb{P}$ , et  $\theta_n = \psi(\mathbb{P}_n)$  son équivalent empirique,
- ▶ Un IC pour  $\theta$  au niveau  $1 - \alpha$  est un intervalle aléatoire  $[T_1(\mathbf{X}), T_2(\mathbf{X})]$  tel que  $\mathbb{P}(\theta \in [T_1(\mathbf{X}), T_2(\mathbf{X})]) \geq 1 - \alpha$ .

## Principe du pivot

- ▶ Trouver une quantité **pivot**  $g(\mathbf{X}, \theta)$  dont la loi ne dépend pas de  $\theta$ .
- ▶ Déterminer les quantiles  $q_{\alpha/2}$  et  $q_{1-\alpha/2}$  de la loi de  $g(\mathbf{X}, \theta)$ .
- ▶ Inverser l'équation  
$$g(\mathbf{X}, \theta) \in [q_{\alpha/2}, q_{1-\alpha/2}] \iff \theta \in [T_1(\mathbf{X}), T_2(\mathbf{X})].$$



# Construction d'intervalles de confiance II

## Construction par intervalle bootstrap-t

- ▶ On tire  $B$  répliqués bootstrap  $g(\mathbf{X}^b, \theta_n)$  de la fonction pivotale, à partir desquels on estime les quantiles :  $\hat{q}_{\alpha/2}^*$  et  $\hat{q}_{1-\alpha/2}^*$  sont les stats d'ordre d'indices  $\lceil B\alpha/2 \rceil$  et  $\lceil B(1 - \alpha/2) \rceil$  de l'échantillon bootstrap.
- ▶ **Avantage** : pas besoin d'hyp. sur la loi du pivot (par ex. approximation gaussienne ou Student).
- ▶ **Inconvénient** : nécessite l'existence d'un pivot.

## Exemple

- ▶  $\theta$  paramètre de localisation, estimé par  $\hat{\theta}$ ,
- ▶ Sous l'hyp que  $(\hat{\theta} - \theta)/se(\hat{\theta})$  a une loi (asympt.) indépendante de  $\theta$ , on peut construire un intervalle de confiance bootstrap-t pour  $\theta$ .

# Construction d'intervalles de confiance III

## Construction par quantiles bootstrap

- ▶ On tire  $B$  réplicats bootstrap d'un estimateur  $\{\hat{\theta}^b\}_{1 \leq b \leq B}$ ,
- ▶  $\hat{q}_{\alpha/2}^*$  et  $\hat{q}_{1-\alpha/2}^*$  les stats d'ordre d'indices respectifs  $\lceil B\alpha/2 \rceil$  et  $\lceil B(1 - \alpha/2) \rceil$  de l'échantillon bootstrap  $\{\hat{\theta}^b\}_{1 \leq b \leq B}$ ,
- ▶ L'IC bootstrap pour  $\theta$  de niveau  $1 - \alpha$  est  $[\hat{q}_{\alpha/2}^*, \hat{q}_{1-\alpha/2}^*]$ .
- ▶ **Avantage** : Mise en œuvre possible dès qu'on a un estimateur,
- ▶ **Inconvénient** : Niveau de confiance **uniquement approximatif**.

## Exemple

- ▶ Construction d'un IC pour le coeff. de corrélation entre deux variables.

# Sommaire Sixième partie

Le bootstrap

Exemples d'estimateurs bootstrap

Compléments sur le bootstrap

Jackknife

# Bootstrap paramétrique

## Principe

- ▶  $X_1, \dots, X_n$  i.i.d. de loi paramétrique  $\mathbb{P}_\theta$ ,
- ▶  $\hat{\theta}$  un estimateur de  $\theta$ ,
- ▶ On échantillonne  $\mathbf{X}^b = (X_1^b, \dots, X_n^b)$  selon la loi  $\mathbb{P}_{\hat{\theta}}$ .

## Avantages/Inconvénients

- ▶ Les échantillons bootstrap sont plus "variables" puisqu'on peut avoir de nouvelles observations (non incluses dans l'éch. initial),
- ▶ Hypothèse paramétrique forte.

# Consistance du bootstrap

## Notations

- ▶  $\mathbf{X} = (X_1, \dots, X_n)$  iid de fdr  $F$ ,
- ▶ On s'intéresse à une racine  $R_n(\mathbf{X}, F)$  de loi  $\mu_n$ ,
- ▶ Approximée par une quantité bootstrap  $R_n(\mathbf{X}^*, \hat{F}_n)$  où  $\hat{F}_n$  fdr empirique (ou bien estimateur consistant de  $F_n$ ), de loi conditionnelle sachant  $\mathbf{X}$  notée  $\mu_n^*$ .

## Conditions de validité du bootstrap

Soit  $G$  fdr et  $\mathbf{Y} = (Y_1, \dots, Y_n)$  iid de fdr  $G$ . Si

- ▶  $R_n(\mathbf{Y}, G)$  converge faiblement vers  $\mathcal{L}_G$  pour tout  $G$  au voisinage de  $F$ ,
- ▶ Cette convergence est uniforme sur le voisinage,
- ▶ La fonction  $G \mapsto \mathcal{L}_G$  est continue,

alors, l'approximation bootstrap est valide, *i.e.* pour une certaine distance  $\rho$  sur l'ensemble des mesures de proba sur  $\mathbb{R}$ , on a  $\rho(\mu_n^*, \mu_n) \rightarrow 0$  (en proba ou p.s.).

# Le bootstrap à poids I

## Limites du bootstrap

Inconsistance des estimateurs bootstrap dans certaines situations :

- ▶ Estimation de la moyenne lorsque la variance est infinie,
- ▶ Estimation du min et du max d'une suite de variables i.i.d.
- ▶ ...

Ces pbms sont corrigés par le *m out of n bootstrap*.

## Principe des poids

- ▶ On se donne une suite de poids aléatoires  $\mathbf{w} = \{w_i\}_{1 \leq i \leq n}$  tels que  $w_i \geq 0$  et  $\sum_i w_i = 1$  (ex : poids uniformes  $w_i = 1/n$ ),
- ▶ Pour chaque échantillon bootstrap, on tire chaque variable  $X_i$  avec proba  $w_i$ .
- ▶ **Rem** : le bootstrap "naïf" d'Efron affecte en fait à chaque valeur un poids proportionnel au nombre de fois où elle apparaît dans l'échantillon initial.

# Le bootstrap à poids II

## Exemple : $m$ out of $n$ bootstrap

Tirage (avec ou sans remise) de  $m$  variables parmi les  $n$  initiales.

- ▶ Correspond à une suite de poids  $\mathbf{w} = \sigma(\mathbf{w}^0)$  où  $\sigma$  permutation de  $\{1, \dots, n\}$  et  $\mathbf{w}^0 = (1/m, \dots, 1/m, 0, \dots, 0)$
- ▶ En théorie, choisir  $m \rightarrow +\infty$  avec  $m/n \rightarrow 0$ .

# Sommaire Sixième partie

Le bootstrap

Exemples d'estimateurs bootstrap

Compléments sur le bootstrap

Jackknife



# Jackknife I

## Historique

- ▶ Introduit bien avant le bootstrap, à l'origine comme une technique de réduction de biais [Quenouille 49],
- ▶ Utilisé par [Tukey 58] pour estimer l'erreur d'un estimateur,
- ▶ De nos jours, le jackknife s'utilise en complément du bootstrap : méthodes **Jackknife-after-bootstrap**.

## Principe

On crée  $n$  nouveaux échantillons en retirant à chaque fois une seule observation de l'échantillon initial.

# Jackknife II

## Détails du Jackknife

- ▶ On observe un  $n$ -échantillon. Estimateur initial  $\hat{\theta}$  de  $\theta$ .
- ▶ Pour  $1 \leq i \leq n$ , on construit  $\hat{\theta}^{(-i)}$  le même estimateur de  $\theta$  sur les observations privées de la  $i$ ème et  $\overline{\hat{\theta}^{(-)}} = \frac{1}{n} \sum_i \hat{\theta}^{(-i)}$ .

- ▶ **Fonction d'influence Jackknife** : estime le biais de  $\hat{\theta}$

$$\widehat{Biais}_{Jackk}(\hat{\theta}) = (n - 1)(\overline{\hat{\theta}^{(-)}} - \hat{\theta}).$$

- ▶ **Estimateur Jackknife à biais réduit** :

$$\hat{\theta} - \widehat{Biais}_{Jackk}(\hat{\theta}) = n\hat{\theta} - (n - 1)\overline{\hat{\theta}^{(-)}}.$$

Rappel : réduire le biais peut cependant augmenter la variance !

- ▶ **Estimateur Jackknife de la variance** :

$$\widehat{Var}_{Jackk}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}^{(-i)} - \overline{\hat{\theta}^{(-)}})^2.$$

# Remarques et compléments

## Remarques

- ▶ Les éch. Jackknife ne sont pas tirés selon la loi empirique  $\mathbb{P}_n$ .
- ▶ En particulier : différence dans la normalisation entre Bootstrap et Jackknife. Compense le fait que les échantillons Jackknife sont bcp plus similaires à l'éch. initial que les éch. bootstrap.

## Compléments sur le Jackknife

- ▶ Utilisé pour mesurer l'importance d'une variable sur la valeur d'une statistique, **lié à la notion de fonction d'influence**.
- ▶ Typiquement, le jackknife est mauvais sur les statistiques peu robustes.
- ▶ **Variante : delete- $d$  Jackknife** : on retire  $d$  observations (version déterministe du  $m$  out of  $n$  bootstrap).

# Jackknife after Bootstrap I

Quelle est la précision des estimateurs bootstrap de la précision d'un estimateur ?

## Principe

- ▶ Estimer la variabilité de statistiques bootstrap  $\hat{S}_{boot}$ , comme par ex :  $\widehat{se}_{boot}(\hat{\theta})$ ,  $\widehat{biais}_{boot}(\hat{\theta})$ , ...
- ▶ Pour chaque valeur  $1 \leq i \leq n$ , on calcule la valeur de  $\hat{S}_{boot}^{(-i)}$  et la moyenne de ces valeurs  $\overline{\hat{S}_{boot}^{(-)}}$ , puis

$$\widehat{se}_{Jackk}(\hat{S}_{boot}) = \left\{ \frac{n-1}{n} \sum_{i=1}^n \left( \hat{S}_{boot}^{(-i)} - \overline{\hat{S}_{boot}^{(-)}} \right)^2 \right\}^{1/2}.$$

- ▶ **Implémentation naïve** : faire  $nB$  ré-échantillonnages.

# Jackknife after Bootstrap II

## Implémentation directe

- ▶ La stat. bootstrap a la forme  $\hat{S}_{boot} = B^{-1} \sum_{b=1}^B R(\mathbf{X}^b, \mathbb{P}_n)$ .
- ▶ Donc à partir de  $B$  éch. bootstrap, on peut calculer  $\hat{S}_{boot}^{(-i)}$  via la moyenne des  $R(\mathbf{X}^b, \mathbb{P}_n)$  sur les échantillons bootstrap  $b$  qui ne contiennent pas la variable  $X_i$

$$\hat{S}_{boot}^{(-i)} = \frac{1}{|\mathcal{B}^i|} \sum_{b \in \mathcal{B}^i} R(\mathbf{X}^b, \mathbb{P}_n)$$

$\mathcal{B}^i$  = ens. des éch. bootstrap qui ne contiennent pas  $X_i$ .

- ▶ Inutile de faire  $n$  fois  $B$  ré-échantillonnages bootstrap !



[Efron 79] B. Efron.

Bootstrap methods : Another look at the jackknife.

*Ann. Stat.*, 7 :1-26, 1979.



[Efron & Tibshirani 97] B. Efron & R. Tibshirani.

Improvements on Cross-Validation : The .632+ Bootstrap Method.

*Journal of the American Statistical Association*,  
92(438) :548-560, 1997.



[Quenouille 49] M.H. Quenouille.

Approximate tests of correlation in time series.

*Journal of the Royal Statistical Society, Series B*, 11 :18-44,  
1949.



[Quenouille 56] M.H. Quenouille.

Notes on bias in estimation.

*Biometrika*, 43 :353-360, 1956.



[Tukey 58] J.W. Tukey.

Bias and confidence in not quite large samples.

*Annals of Mathematical Statistics.*, 29 :614, 1958.