Introduction à la statistique non paramétrique

poly: ©Laëtitia Comminges, Gabriel Turinici
cours G. Turinici

M1 Math Université Paris Dauphine - PSL, 2019/20

Table des matières

Introduction

Ces notes de cours font suite aux notes du cours d'introduction à la statistique non paramétrique de Catherine Mathias, Vincent Rivoirard et Laëtitia Comminges.

Chapitre 1

Introduction et rappels

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

La statistique paramétrique est le cadre classique de la statistique. Le modèle statistique est défini par un paramètre $\theta \in \mathbb{R}^k$ pour un certain entier k.

Exemple 1.1.
— *Modèle linéaire gaussien. La loi* $\mathbf{P}_\theta$ *des observations vérifie* $\mathbf{P}_\theta = \mathcal{N}(\mu, \sigma^2 I_n)$. *le paramètre* $\theta = (\mu, \sigma^2) \in \mathbb{R}^n \times \mathbb{R}_+^*$ *suffit à déterminer la loi des observations.*

— *Observation du nombre d'arrivées à un guichet :* $Y \sim P(\lambda)$ *(Poisson).*

Par opposition, en statistique non-paramétrique, le modèle n'est pas décrit par un nombre fini de paramètres (ou de manière équivalente par un paramètre de dimension finie).

Exemple 1.2. *Un constructeur automobile étudie le comportement d'achat de ses clients. Il a la conviction que la somme qu'ils sont prêts à débourser est une fonction de leur revenu et de la distance parcourue quotidiennement et à partir de* n *observations recueillies par sondage, il postule le modèle statistique suivant :*

$$Y_i = f(X_i) + \epsilon_i, \ \ i = 1, ..., n$$

où les ϵ_i *sont iid de loi* $\mathcal{N}(0, \sigma^2)$ *et* $X_i = (X_{i1}, X_{i2})$ *=(revenu,distance) et* Y_i *=somme à débourser.*

On peut faire différentes hypothèses a priori sur la fonction f *(selon l'expérience, les connaissances a priori sur les données, ou après une représentation graphique des données)*

— *on peut supposer que* f *est une fonction affine des variables explicatives, on obtient alors un modèle linéaire (ici gaussien puisqu'on a supposé les erreurs gaussiennes) :* $f(X_i) = \theta_1 + \theta_2 X_{i1} + \theta_3 X_{i2}$

— *On peut aussi ne faire aucune hypothèse sur la forme de la fonction* f*, et faire juste une hypothèse de régularité minimum. On obtient alors un modèle non-paramétrique.*

1.2 Quelques problèmes de statistique non-paramétrique

1.2.1 Estimation de la fonction de répartition

On observe $X_1, \ldots, X_n$ n variables réelles de loi P. On cherche à estimer la loi P. Or P est entièrement décrite par sa fonction de répartition

$$F : \begin{array}{ccc} \mathbb{R} & \to & [0,1] \\ x & \to & P(]-\infty, x]) \end{array}$$

On construit un estimateur $\hat{F}_n$ de F à l'aide des n observations $X_1, \ldots, X_n$.

1.2.2 Estimation de densité

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

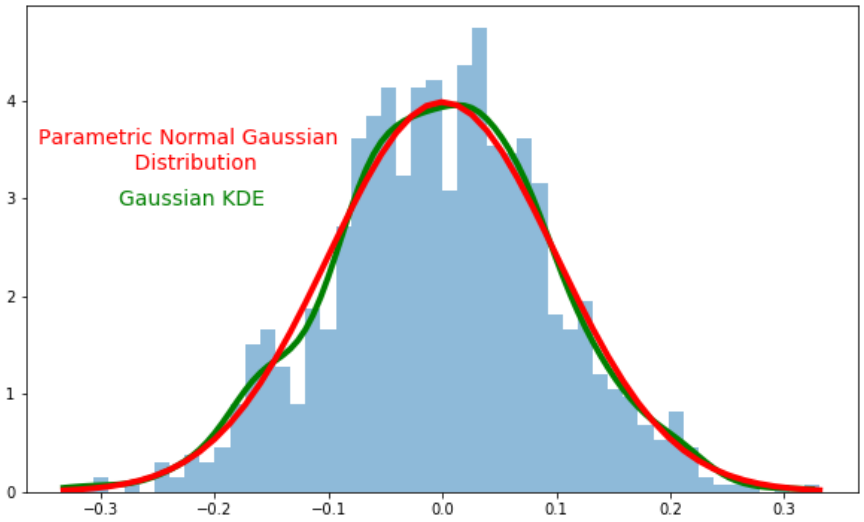


FIGURE 1.1 – Estimation de densité avec python, fonction "gausian_kde" du package "scipy.stats.kde".

1.2.3 Régression non-paramétrique

On observe une suite de couples $\big((X_i, Y_i)\big)_{1 \leq i \leq n}$ obéissant au modèle

$$Y_i = f(X_i) + \epsilon_i, \ i = 1, \ldots, n$$

On cherche à estimer la fonction de régression f.

On peut aussi considérer d'autres problèmes de statistique non-paramétrique qui ne sont pas directement de l'estimation.

1.2.4 Tests non-paramétriques

Deux exemples de problèmes possibles :

— Soit X une v.a. et P une distribution donnée. A l'aide de $X_1, \ldots, X_n$ iid de même loi que X, tester :

$$H_0 : X \sim P, \quad \text{contre} \quad X \not\sim P$$

— Soient X et Y deux v.a. et $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ des échantillons de mêmes lois respectivement que X et Y. A l'aide des deux échantillons on peut :

  — tester s'il s'agit de la même loi : $H_0 : X \sim Y$ contre $H_1 : X \not\sim Y$

  — tester l'indépendance entre X et Y : $H_0 : X \perp\!\!\!\perp Y$ contre $H_1 : X$ et Y sont non indépendants.

1.2.5 Classification supervisée

On observe n couples $\Big((X_i, Y_i)\Big)_{1 \leq i \leq n}$ où $Y_i \in \{0, 1, ..., L\}$. Y_i est l'étiquette associée à X_i. On veut trouver la fonction de classification g à valeurs dans $\{0, 1, ..., L\}$ telle que $\mathbf{P}(Y \neq g(X))$ soit la plus petite possible où $(X, Y) \sim (X_1, Y_1)$.

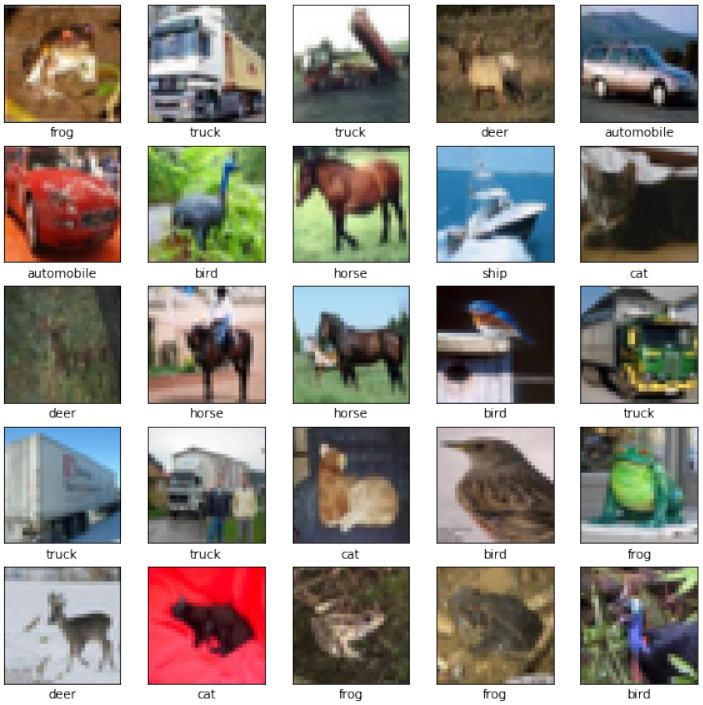


FIGURE 1.2 – Classification supervisée du dataset CIFAR10 (60000 images 32×32 format RGB = trois couleurs) avec $L = 10$ étiquettes) avec l'environnement Tensorflow v2.0 (pris du site de la librairie).

1.2.6 Classification non-supervisée, exemple génération

Idée : ayant quelques réalisations X_1, ..., X_n i.i.d. de la loi P comment générer d'autres instances de Y_1, Y_2 suivant la même loi ? Exemple : peintures de paysages. Ce sont des algorithmes de type GAN (Generative Adversarial Networks), VAE (Variational auto-encoder), ...

FIGURE 1.3 – Génération non-supervisée de paysages. Image : github.com/robbiebarrat/art-DCGAN

1.3 Rappels d'inégalités classiques

1.3.1 Inégalité de Markov

Soit X une v.a.r. positive telle que $\mathbf{E}(X) < \infty$. Alors $\forall t > 0$

$$\mathbf{P}(X \geq t) \leq \frac{\mathbf{E}(X)}{t}$$

1.3.2 Inégalité de Bienaymé-Tchebycheff (B-T)

Soit X une v.a.r. telle que $\mathbf{E}(X^2) < \infty$. Alors pour tout $t > 0$,

$$\mathbf{P}(|X - \mathbf{E}(X)| > t) \leq \frac{\mathbf{Var}(X)}{t^2}$$

1.3.3 Inégalité de Hoeffding

Soient $Y_1, \dots, Y_n$ des v.a.r. indépendantes centrées et telles que

$$a_i \leq Y_i \leq b_i \text{ p.s. pour tout i}$$

Alors

$$\forall t > 0, \quad \mathbf{P}(\sum_{i=1}^{n} Y_i \geq t) \leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big)$$

Remarque 1.3. *Sous les mêmes hypothèses, on a aussi*

$$\forall t > 0, \quad \mathbf{P}(\Big|\sum_{i=1}^{n} Y_i\Big| \geq t) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big)$$

Remarque 1.4. *Comparaison entre Hoeffding et B-T : soit $X_1, \dots, X_n \overset{iid}{\sim} Be(p)$ avec $p \in (0,1)$. On cherche un intervalle de confiance bilatéral à gauche de niveau $1-\alpha$ pour p avec l'une des inégalités ci-dessus :*

- *Si on utilise l'inégalité B-T, $\mathbf{P}(|\bar{X} - p| > c) \leq \frac{p(1-p)}{nc^2} \leq \frac{1}{4c^2 n} := \alpha$. Donc $\mathbf{P}(p \in [\bar{X} - \frac{1}{2\sqrt{n\alpha}}, \bar{X} + \frac{1}{2\sqrt{n\alpha}}]) \geq 1-\alpha$. Pour $\alpha = 5\%$ et $n = 100$, la précision, i.e. la longueur, de cet intervalle est $\frac{1}{\sqrt{n\alpha}} = 0.22$.*
- *Si on utilise l'inégalité de Hoeffding, $\mathbf{P}(|\bar{X} - p| > c) \leq 2\exp(-2nc^2)$, Donc $\mathbf{P}(p \in [\bar{X} - \sqrt{\frac{\log(\frac{2}{\alpha})}{2n}}, \bar{X} + \sqrt{\frac{\log(\frac{2}{\alpha})}{2n}}]) \geq 1-\alpha$. Pour $\alpha = 5\%$ et $n = 100$, la précision, i.e. la longueur, de cet intervalle est $\sqrt{2\frac{\log(\frac{2}{\alpha})}{n}} = 0.14$.*

1.4 Théorèmes de convergence classique

1.4.1 Lemme de Slutsky

Soient $(X_n)_{n\geq 0}$ et $(Y_n)_{n\geq 0}$ deux suites de vecteurs aléatoires tels que

- $X_n \overset{loi}{\to} X$ où X est un vecteur aléatoire quelconque.
- $Y_n \overset{proba}{\to} c$ où c est un vecteur constant.

alors $(X_n, Y_n) \overset{loi}{\to} (X, c)$.

conséquence : $X_n + Y_n \overset{loi}{\to} X + c$, $X_n Y_n \overset{loi}{\to} cX$, et de manière générale, pour toute fonction continue f (ou continue là où les variables prennent leurs valeurs) $f(X_n, Y_n) \overset{loi}{\to} f(X, c)$.

1.4.2 Delta-méthode

On se donne une suite $(U_n)_n$ de vecteurs aléatoires de $\mathbb{R}^m$, une suite déterministe $(a_n)_n$ et une application $\ell : \mathbb{R}^m \to \mathbb{R}^p$ telles que

- $a_n \to +\infty$
- $\exists U \in \mathbb{R}^m$ un vecteur déterministe (=constant) et V un vecteur aléatoire tels que $a_n(U_n - U) \overset{loi}{\to} V$.

— ℓ est une fonction différentiable en U de différentielle $D\ell(U) \in M_{pm}(\mathbb{R})$.

Alors on a la convergence en loi

$$a_n(\ell(U_n) - \ell(U)) \xrightarrow{loi} D\ell(U)V.$$

Exemple 1.5. *Soit $X_1, \ldots, X_n \overset{iid}{\sim} P(\lambda)$ avec $\lambda > 0$. Alors d'après le TCL on a $\sqrt{n}(\bar{X} - \lambda) \xrightarrow{loi} N(0, \lambda)$. Donc on a aussi, d'après le théorème ci-dessus,*

$$\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \xrightarrow{loi} N(0, \frac{1}{4})$$

En effet ici $\mathbb{R}^m = \mathbb{R}^p = \mathbb{R}$, $U_n = \bar{X}$, $U = \lambda$, $a_n = \sqrt{n}$, $V \sim N(0, \lambda)$, $\ell(u) = \sqrt{u}$, donc $D\ell(U) = \ell'(\lambda) = \frac{1}{2\sqrt{\lambda}}$ et $D\ell(U)V \sim N(0, \frac{\lambda}{4\lambda})$.

1.5 Petits rappels sur l'espérance conditionnelle

Soit X et Y deux variables aléatoires à valeurs dans $\mathbb{R}^k$ et $\mathbb{R}^p$. Pour $x \in \mathbb{R}^k$, on note $\mathbf{P}_Y^{X=x}$ la loi conditionnelle de Y sachant $X = x$.

1.5.1 Calcul d'espérance conditionnelle

On rappelle que l'espérance conditionnelle de Y sachant X, que l'on note ici $\mathbf{E}(Y \mid X)$, est une variables aléatoire qui peut s'écrire comme une fonction $g(X)$. Cette fonction est donnée par

$$\mathbf{E}(Y \mid X) = g(X) \quad \text{où} \quad g(x) = \mathbf{E}(Y \mid X = x).$$

Exemples :

1. soient Z et T deux variables aléatoires indépendantes de loi exponentielle de paramètre λ. On note $S = Z + T$ et on cherche à calculer la variables aléatoire $\mathbf{E}(Z \mid S)$. Soit $s > 0$. On trouve facilement (car tout le monde a une densité ...) que la densité conditionnelle de Z sachant $S = s$ est donnée par $f_Z^{S=s}(z) = \frac{1}{s}\mathbb{1}_{[0,s]}(z)$. On a alors immédiatement $\mathbf{E}(Z \mid S = s) = \int_0^s z f_Z^{S=s}(z)dz = \frac{s}{2}$. Et en utilisant la propriété que l'on vient de rappeler, on a finalement $\mathbf{E}(Z \mid S) = \frac{S}{2}$.
2. Soit U et V deux v.a. réelles. On rappelle la définition de la variance conditionnelle : $\mathbf{Var}(U \mid V) = \mathbf{E}\Big[(U - \mathbf{E}(U \mid V))^2 \mid V\Big] = \mathbf{E}[U^2 \mid V] - \Big(\mathbf{E}[U \mid V]\Big)^2$. On a
$$\mathbf{E}\Big[\mathbf{Var}(U \mid V)\Big] = \mathbf{E}[g(V)] \text{ avec } g(v) = \mathbf{Var}(U \mid V = v).$$
En effet $\mathbf{Var}(U \mid V) = \mathbf{E}[U^2 \mid V] - \Big(\mathbf{E}[U \mid V]\Big)^2 = \ell(V) - (h(V))^2$ avec $\ell(v) = \mathbf{E}[U^2 \mid V = v]$ et $h(v) = \mathbf{E}[U \mid V = v]$. Donc $\ell(v) - h(v)^2 = \mathbf{Var}(U \mid V = v)$.

1.5.2 Propriété du transfert conditionnel

Soit $f : \mathbb{R}^{p+k} \to \mathbb{R}^q$ une fonction borélienne. Alors la loi conditionnelle de $f(X,Y)$ sachant $X = x$ vérifie

$$\mathbf{P}^{X=x}_{f(X,Y)} = \mathbf{P}^{X=x}_{f(x,Y)}$$

et donc

$$\mathbf{E}\Big[f(X,Y) \mid X = x\Big] = \mathbf{E}\Big[f(x,Y) \mid X = x\Big]$$

En particulier si X et Y sont indépendantes, on a

$$\mathbf{P}^{X=x}_{f(X,Y)} = \mathbf{P}_{f(x,Y)}$$

et donc

$$\mathbf{E}\Big[f(X,Y) \mid X = x\Big] = \mathbf{E}\Big[f(x,Y)\Big].$$

Technique importante 1.5.1. *Supposons que X et Y sont des v.a. réelles indépendantes. On a*

$$\mathbf{P}(Y \leq X) = \mathbf{E}(\mathbb{1}_{Y\leq X}) = \mathbf{E}\Big(\mathbf{E}[\mathbb{1}_{Y\leq X} \mid X]\Big) = \mathbf{E}[g(X)]$$

où

$$g(x) = \mathbf{E}[\mathbb{1}_{Y\leq X} \mid X = x] = \mathbf{E}[\mathbb{1}_{Y\leq x}] = F_Y(x)$$

où on a noté F_Y la cdf de Y. Donc on a

$$\mathbf{P}(Y \leq X) = \mathbf{E}[F_Y(X)]$$

dès que X et Y sont indépendantes.

On peut aussi écrire :

$$\begin{aligned}
\mathbf{P}(Y \leq X) = \mathbf{E}[\mathbb{1}_{Y\leq X}] &= \int \mathbb{1}_{y\leq x} dP_{X,Y}(x,y) \\
&= \int \mathbb{1}_{y\leq x} dP_Y \otimes dP_X(y,x) \\
&= \int \Big[\int \mathbb{1}_{y\leq x} dP_Y(y)\Big] dP_X(x) \\
&= \int F_Y(x) dP_X(x) \\
&= \mathbf{E}[F_Y(X)].
\end{aligned}$$

Exemple : Reprenons l'exemple de la sous-section précédente : $Z, T \overset{iid}{\sim} \exp(\lambda)$. On veut calculer $\mathbf{E}(S^2 Z \mid S = s)$ pour $s > 0$. En utilisant la propriété ci-dessus on obtient $\mathbf{E}(S^2 Z \mid S = s) = \mathbf{E}(s^2 Z \mid S = s) = \frac{s^3}{2}$.

1.6 Rappels sur les quantiles et les lois symétriques

1.6.1 Quantiles

On ne donne ici que la définition dans le cas simple où la loi est de cdf F continue et strictement croissante.

Soit X une variable aléatoire réelle de cdf F continue et strictement croissante. Pour $\alpha \in (0,1)$, on appelle quantile d'ordre α de la loi F l'unique réel q_α^F tel que

$$F(q_\alpha^F) = P(X \leq q_\alpha^F) = \alpha$$

autrement dit

$$q_\alpha^F = F^{-1}(\alpha) \tag{1.1}$$

Attention, quand la cdf n'est pas continue, l'équation ci-dessus n'a pas toujours de solution. De plus si la cdf n'est pas strictement croissante, l'équation peut avoir une infinité de solutions. La définition générale d'un quantile sera vue dans le chapitre 2.

1.6.2 Loi symétrique

— Une variable réelle X a une loi symétrique (par rapport à 0) si $X \sim -X$.

— Si la cdf F est continue, cela se traduit par $F(x) = 1 - F(-x)$.

— Si la cdf F est continue et strictement croissante, cela se traduit, en terme de quantile, par $q_{1-\alpha}^F = -q_\alpha^F$ pour tout $\alpha \in (0,1)$.

— Si la loi a une densité f, cela se traduit par $f(-x) = f(x)$ pour presque tout $x \in \mathbb{R}$.

— Une v.a. réelle X a une distribution symétrique par rapport à b ssi $X - b$ a une distribution symétrique par rapport à 0, autrement dit ssi $X - b \sim -X + b$, autrement dit

$$X \sim 2b - X$$

.

— Si X a une loi symétrique alors $\mathbf{P}(|X| > c) = \mathbf{P}(X > c) + \mathbf{P}(-X > c) = 2\mathbf{P}(X > c)$.

— Si la loi de X est symétrique et si $\mathbf{P}(X = 0) = 0$ alors la variable aléatoire $|X|$ est indépendante de la variable aléatoire $\mathbb{1}_{X>0}$. En effet, soit A mesurable, la symétrie de la loi de X implique

$$\mathbf{P}(|X| \in A, X > 0) = \mathbf{P}(|-X| \in A, -X > 0) \tag{1.2}$$

et

$$\mathbf{P}(X > 0) = \mathbf{P}(X < 0) \tag{1.3}$$

(1.2) se réécrit

$$\mathbf{P}(|X| \in A, X > 0) = \mathbf{P}(|X| \in A, -X > 0)$$

ce qui implique

$$\mathbf{P}(|X| \in A) = \mathbf{P}(|X| \in A, X > 0) + \mathbf{P}(|X| \in A, X < 0) = 2\mathbf{P}(|X| \in A, X > 0) \tag{1.4}$$

(1.3) combinée avec la propriété $\mathbf{P}(X = 0) = 0$ impliquent

$$\mathbf{P}(X > 0) = \frac{1}{2} \tag{1.5}$$

(1.4) combiné avec (1.5) impliquent

$$\mathbf{P}(|X| \in A, X > 0) = \frac{1}{2}\mathbf{P}(|X| \in A) = \mathbf{P}(X > 0)\mathbf{P}(|X| \in A).$$

Exemples : la loi normale standard et la loi de Student sont des distributions symétriques (par rapport à 0). La loi $Be(1/2)$ est symétrique par rapport à $1/2$. La loi $B(n, 1/2)$ est symétrique par rapport à $n/2$.

1.7 Rappels sur les tests (cadre paramétrique)

Test et erreur de test

Situation

On considère une expérience statistique engendrée par une observation X à valeurs dans $\left(\mathcal{X}, \mathcal{A}\right)$ et associée à la famille de lois de probabilités

$$\left\{\mathbf{P}_\theta,\ \theta \in \Theta\right\}.$$

L'ensemble des paramètres Θ est un sous-ensemble de $\mathcal{R}^d$, avec $d \geq 1$.

Principe du test statistique

On veut « décider » à partir de l'observation de X si une propriété de la loi de X est vérifiée ou non. Cette propriété se traduit mathématiquement par un sous-ensemble $\Theta_0 \subset \Theta$ de l'ensemble des paramètres, et la propriété signifie que $\theta \in \Theta_0$.

Définition 1.6 (Terminologie de test)**.** *On teste « l'hypothèse nulle »*

$$H_0\ :\ \theta \in \Theta_0$$

contre « l'alternative »

$$H_1\ :\ \theta \in \Theta_1,$$

avec $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$ et $\Theta_0 \cap \Theta_1 = \emptyset$. Construire un test signifie construire une procédure $\phi = \phi(X)$ de la forme

$$\phi(X) = 1_{\{X \in \mathcal{R}\}} = \begin{cases} 0 & \text{si } X \notin \mathcal{R}. \quad \text{« on accepte l'hypothèse nulle »} \\ 1 & \text{si } X \in \mathcal{R}. \quad \text{« on rejette l'hypothèse nulle »} \end{cases} \tag{1.6}$$

avec $\mathcal{R}$ mesurable.

Définition 1.7. *On désigne indifféremment l'ensemble $\mathcal{R} \subset \mathcal{A}$ ou bien l'événement $\Big\{X \in \mathcal{R}\Big\}$ comme zone de rejet ou encore zone critique du test ϕ.*

Définition 1.8. *L'hypothèse H_j ($j = 0$ ou $j = 1$) est dite simple si Θ_j est réduit à un singleton, sinon H_j est dite composite.*

Par exemple, le test de la forme $H_0 : \theta = 1$ contre $H_1 : \theta > 1$ a une hypothèse nulle simple et une alternative composite.

Erreur de test

Lorsque l'on effectue un test, il y a quatre possibilités. Deux sont anecdotiques et correspondent à une bonne décision :

— Accepter l'hypothèse H_0 alors que $\theta \in \Theta_0$ (c'est-à-dire l'hypothèse H_0 est vraie).

— Rejeter l'hypothèse H_0 alors que $\theta \in \Theta_1$ (c'est-à-dire l'hypothèse H_0 est fausse).

Les deux autres possibilités sont celles qui vont nous occuper, et correspondent à une erreur de décision :

— Rejeter l'hypothèse H_0 alors que $\theta \in \Theta_0$ (c'est-à-dire l'hypothèse H_0 est vraie).

— Accepter l'hypothèse H_0 alors que $\theta \in \Theta_1$ (c'est-à-dire l'hypothèse H_0 est fausse).

Définition 1.9. ***[Erreur de première et seconde espèce] L'erreur de première espèce, ou encore de "type I" correspond à la probabilité maximale de rejeter l'hypothèse alors qu'elle est vraie :***

$$\sup_{\theta \in \Theta_0} \mathbf{E}_\theta\Big[\phi(X)\Big] = \sup_{\theta \in \Theta_0} \mathbf{P}_\theta\Big[X \in \mathcal{R}\Big].$$

L'erreur de seconde espèce ("type II") correspond à la probabilité maximale d'accepter l'hypothèse alors qu'elle est fausse :

$$\sup_{\theta \in \Theta_1} \mathbf{E}_\theta\Big[1 - \phi(X)\Big] = \sup_{\theta \in \Theta_1} \mathbf{P}_\theta\Big[X \notin \mathcal{R}\Big]. \tag{1.7}$$

Intuition 1.7.1. *Sur Θ_0 et Θ_1 il n'y a pas de préférence (entre paramètres) exprimée sous la forme de loi de probabilités. Tous les éléments sont aussi importants ce qui explique les "sup" (= "pire cas") dans la définition.*

Intuition 1.7.2. *D'après cette terminologie, l'erreur de première espèce mesure la probabilité (maximale) de rejeter à tort, et l'erreur de seconde espèce d'accepter à tort. Dans le langage courant, commettre une erreur*

de première espèce revient à faire un « faux négatif », et commettre une erreur de seconde espèce revient à faire un « faux positif ».

Mise en garde 1.7.1. *Dans la plupart des situations, Θ_0 est « plus petit » que Θ_1 et le contrôle de l'erreur de seconde espèce (1.7) est difficile, surtout si Θ_1 contient des points « très proches » de Θ_0. On peut imaginer que pour des points de Θ_1 qui convergent vers un point de Θ_0 l'erreur de seconde espèce est de 100% moins l'erreur de première espèce. Elle donne alors peu d'informations nouvelles sur le test en question car elle est trop agrégée (à cause du "sup"). Pour le cas typique d'un Θ_0 singleton et Θ_1 son complémentaire, l'erreur de type II n'apporte pas d'information utile pour discriminer des tests statistiques ayant la même erreur de type I.*
Pour des informations plus précises, on introduit alors la fonction de puissance d'un test, qui mesure sa performance locale (= en tout point) sur l'alternative.

Définition 1.10. ***La fonction de puissance du test ϕ est l'application***

$$\underline{\beta} : \Theta_1 \to [0,1]$$

définie par

$$\theta \in \Theta_1 \rightsquigarrow \underline{\beta}(\theta) = \mathbf{P}_\theta\Big[X \in \mathcal{R}\Big].$$

Une illustration intuitive des erreurs et paramètres α et β est donnée en figure 1.4.

	Hypothèse H_0 vraie	Hypothèse H_1 vraie
Hypothèse H_0 acceptée	Bonne décision (1-α)	Risque β
Hypothèse H_1 acceptée	Risque α	Bonne décision (1-β)

FIGURE 1.4 – Erreurs de première et deuxième espèce, α et β. Attention : c'est une "vue d'artiste", les définitions précises sont dans le texte. Crédits : wikipedia section "Test statistique", 29 Jan. 2020.

1.7.1 Comparaison de test, principe de Neyman

Idéalement, on souhaite que l'erreur de première espèce et l'erreur de seconde espèce soient toutes deux simultanément petites. Les deux tests triviaux

$$\phi_1 = 1_\emptyset, \quad \text{et} \quad \phi_2 = 1_\mathcal{X}$$

qui consistent respectivement à accepter systématiquement l'hypothèse et à la rejeter systématiquement, sans utiliser l'observation X, ont respectivement une erreur de première espèce nulle et une erreur de seconde espèce nulle. Malheureusement la puissance de ϕ_1 est catastrophique : $\beta(\theta) = 0$ en tout point θ de toute alternative Θ_1. De même l'erreur de première espèce de ϕ_2 est égale à 1, même si l'hypothèse est réduite à un point, quelle que soit l'hypothèse.

Une méthodologie, proposée historiquement par Neyman, consiste à imposer une dissymétrie dans la problématique de test : on décide que le contrôle de l'erreur de première espèce est crucial. La démarche de construction de test sera alors, parmi les tests qui ont une erreur de première espèce contrôlée, de choisir le (ou les) test(s) le(s) plus puissant(s), c'est-à-dire ayant une erreur de seconde espèce la plus petite possible.

Définition 1.11. ***Soit*** $\alpha \in [0,1]$ ***un niveau de risque. Un test*** ϕ ***est de niveau*** α ***si son erreur de première espèce est inférieure ou égale à*** α***.***

Remarque 1.12. *On ne peut pas toujours faire en sorte que l'erreur de première espèce soit égale à* α *(problème de non continuité d'une fonction de répartition par exemple, cf chapitre 2 en particulier). C'est pourquoi on se contente d'exiger que l'erreur de première espèce soit* ***plus petite*** *que* α*.*

Définition 1.13. *On dit qu'un test est de* ***taille*** α *si l'erreur de première espèce est* ***égale*** *à* α*.*

Un test veut mesurer l'adéquation de l'hypothèse H_0 avec les observations. Pour cela il détermine les valeurs typiques de X sous H_0. Si la réalisation x de X n'est pas l'une des valeurs typiques, il rejette H_0. Sinon, faute de mieux, il conserve H_0.

Le niveau α peut être vu comme le risque maximal que l'on accepte de prendre en rejetant à tort H_0.

On prend pour H_0 :

— une hypothèse communément admise

— une hypothèse de prudence (critère de coût, de sécurité etc)

— la seule hypothèse sous laquelle on peut travailler mathématiquement.

En pratique, 2 groupes avec des visées et intérêts différents auront des couples (H_0, H_1) inversés (ex : industriels et consommateurs).

Donnons un exemple concret de ce cas : la limite légale d'un polluant contenu dans les déchets d'une usine est de 6mg/kg. On effectue un dosage sur 20 prélèvements sur lesquels on observe une moyenne empirique de 7mg/kg avec un écart-type empirique de 2.4mg/kg. On admet que la loi de dosage est gaussienne.

On observe donc $X_1, \ldots, X_{20} \overset{iid}{\sim} N(\mu, \sigma^2)$ avec μ et σ^2 inconnus. Pour le directeur de l'usine, l'erreur la plus grave serait de conclure que le niveau de polluant est trop élevé alors qu'il ne l'est pas. Il choisit donc comme hypothèses

$$H_0 : \mu \leq 6 \quad \text{contre} \quad H_1 : \mu > 6.$$

Prenons maintenant le point de vue de l'écologiste. Si la limite est supérieure à 8mg/kg, il y a danger. Contrairement au directeur d'usine, l'écologiste considère que l'erreur la plus grave serait de conclure que le niveau de polluant n'est pas trop élevé alors qu'en réalité il l'est. Il effectue donc le test suivant

$$H_0 : \mu \geq 8 \quad \text{contre} \quad \mu < 8.$$

La mise en oeuvre de ces tests sera faite en exercice (cf TD1 exercice 2).

1.7.2 Explications sur des exemples

Exemple : $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ connu.

$$H_0 : \mu = 3 \quad \text{contre } \mu \neq 3$$

Dans la pratique : nous observons $x_1, \ldots, x_n$. Comme nous voulons savoir si la moyenne est égale à 3 ou plus grande que 3, naturellement nous regardons la moyenne empirique $\bar{x}$. Imaginons que $\bar{x} = 3.5$. Alors que conclure ? Et bien ça dépend...ça dépend de plusieurs facteurs, plus exactement, ça dépend ici de n et de σ.

En effet, le problème est que, évidemment, on ne tombera jamais sur 3 exactement. Imaginons que la vraie moyenne est 3. Alors comme les X_i sont aléatoires et qu'on n'en a qu'une quantité finie n, on n'a jamais l'information exacte sur μ en utilisant l'échantillon, mais seulement une information approchée et aléatoire.

Donc si la moyenne empirique vaut 3.5, la question est : est-ce que la vraie moyenne est 3 et que je tombe sur 3.5 parce que c'est aléatoire ? Ou bien est-ce que c'est parce que ce n'est pas 3 la vraie moyenne ?

Pour répondre à ces questions, il faut utiliser les tests, et surtout utiliser toutes les informations que l'on a à notre disposition (ou que l'on peut déduire des données), en particulier la taille de l'échantillon et la variance σ^2. En effet ce sont ces deux informations qui vont nous aider à savoir si c'est "normal" de tomber sur 3.5 en ayant une vraie moyenne de 3, ou bien si c'est "anormal" (ou "atypique").

Ici regardons ce qui se passe sous H_0, c'est-à-dire quand $\mu = 3$ (c'est toujours ce qu'on fait en fréquentiste, on regarde ce qu'il est censé se passer sous H_0, donc asymétrie des deux hypothèses). Si on est vraiment sous H_0, alors la question est : qu'est-ce qu'une valeur "normale" (ou "usuelle" ou "typique") de $\bar{X}$ quand $\mu = 3$? Il suffit pour cela de standardiser pour se ramener à une variable normale standard et utiliser les quantiles de $N(0, 1)$. En effet, si $\mu = 3$ on a

$$\sqrt{n}\frac{\bar{X} - 3}{\sigma} \sim N(0, 1)$$

Donc, comme le quantile d'ordre 97.5% de la loi normale standard vaut 1.96 (environ) on a

$$\mathbf{P}_3\left(\sqrt{n}\frac{|\bar{X} - 3|}{\sigma} \leq 1.96\right) = 95\%$$

Autrement dit, on peut dire que, avec une très grande probabilité, ici plus précisément avec une probabilité de 95%, la variable aléatoire

$$T = \sqrt{n}\frac{\bar{X} - 3}{\sigma}$$

se trouve dans l'intervalle $[-1.96, 1.96]$. Autrement dit, une valeur "typique" de la statistique T, si on est vraiment sous H_0, est une valeur entre -1.96 et 1.96.

Ainsi si on tombe sur une valeur qui sort de cette intervalle, on se dit que ça n'est pas une valeur "normale" pour T sous H_0 et donc on rejette H_0.

Il est évidemment toujours possible que, tout en étant sous H_0, c'est-à-dire ici, tout en ayant une vraie valeur de μ égale à 3, on tombe sur une valeur observée de T qui sorte de l'intervalle $[-1.96, 1.96]$, puisque la loi normale a son support sur $\mathbb{R}$. Mais ceci se produit "rarement" et donc la possibilité de se tromper en rejetant à tort H_0 est faible : ici 5% (on prend toujours α petit). C'est l'erreur de type I.

Maintenant illustrons cette dépendance par rapport à σ et n dans notre exemple (donc on suppose toujours $\bar{x} = 3.5$) sur notre décision finale .

1. Imaginons d'abord que $\sigma = 1$ et $n = 100$. Alors la valeur observée de T est $t = \sqrt{100}\frac{3.5-3}{1} = 5$. Comme 5 est en dehors de l'intervalle $[-1.96, 1.96]$, on conclut que c'est une valeur "anormale" pour H_0 et donc on rejette H_0.
2. Imaginons que $\sigma = 5$ et $n = 100$. Alors la valeur observée de T est $t = \sqrt{100}\frac{3.5-3}{5} = 1$. Alors on accepte H_0. L'idée est que c'est très possible que la vraie valeur de μ soit 3 et de tomber sur une valeur aussi grande que 3.5 ici, car les données ont une grande variance.
3. Imaginons que $\sigma = 1$ et $n = 9$. Alors la valeur observée de T est $t = \sqrt{9}\frac{3.5-3}{1} = 1.5$. Alors on accepte à nouveau H_0. L'idée est qu'une valeur de $\bar{x} = 3.5$ n'est pas "anormale" pour H_0 si on n'a pas beaucoup de données (le résultat est peu précis si on n'a très peu de données donc il n'est pas "anormal" d'avoir vraiment $\mu = 3$ tout en ayant une valeur $\bar{x}$ un peu "éloignée" de 3).

Une autre alternative H_1

Dans l'exemple précédent, nous avons choisi $H_1 : \mu \neq 3$. Comment faire si $H_1 : \mu > 3$?

On va alors juste modifier la région de rejet. Il faut en fait toujours regarder H_1 pour savoir quand rejeter. On part donc de la statistique T, qui suit une loi normale standard sous H_0.

Quand on est sous H_1, cette statistique a tendance à prendre de grandes valeurs, car $\bar{X}$ est un estimateur de μ et donc $\bar{X}-3$ est proche de $\mu-3$ qui est strictement positif sous H_1. Ensuite cette quantité, $\bar{X}-3$, qui sera donc probablement strictement positive si on est sous H_1, est multipliée par $\sqrt{n}$ (et divisée par σ) pour obtenir T. Donc on se dit, au moins si n est suffisamment grand et si μ est suffisamment éloigné de 3, que la statistique T va être "grande" sous H_1, donc on rejette H_0 quand T est "trop grand". Donc la forme de la région de rejet est $T > c$ où c est une constante à déterminer en fonction, à nouveau, du comportement typique sous H_0 de T. Ici on a donc un encadrement unilatéral de T sous H_0. Le quantile d'ordre 95% de $N(0, 1)$ vaut environ 1.64. On peut alors dire que

$$\mathbf{P}_3(T \leq 1.64) = 95\%$$

C'est-à-dire que, avec une grande probabilité, plus précisément ici 95%, et si on est vraiment sous H_0, la statistique T doit être plus petite que 1.64. Donc on rejette H_0 si ce n'est pas le cas.

Fonction puissance

Pour en savoir plus 1.7.1. *Un test fréquentiste est toujours basé sur une statistique dont on connait le comportement sous H_0 et on a toujours borné l'erreur de première espèce par α. On sait donc, par construction, que si on est vraiment sous H_0 et si on rejette H_0 (à tort donc), la probabilité de se tromper est faible. Dans la construction, on regarde quand même H_1 mais c'est uniquement au moment de savoir la forme de la région de rejet. En réalité on est quand même censé dès le départ choisir une statistique qui aura un comportement différent sous H_0 et sous H_1, de façon à pouvoir faire la différence entre les deux hypothèses.*

Maintenant, après avoir construit le test, on est intéressé par l'erreur de seconde espèce et par la fonction puissance, c'est-à-dire, on est intéressé par ce qui se passe sous H_1. On veut que la probabilité de rejeter H_0, quand on est sous H_1, soit grande, c'est-à-dire qu'on veut que la puissance soit grande. Éventuellement la fonction puissance nous permet de comparer différents tests. Une des propriétés souhaitées est alors que, si on a suffisamment de données, on puisse dire qu'on est sous H_1 quand on l'est bien, avec une très grande probabilité. C'est le cas quand le test est "convergent" (ou "consistant") : la fonction puissance tend vers 1 quand n tend vers l'infini.

Évidemment, comme son nom l'indique, la puissance est une fonction, car elle dépend de l'alternative exacte. En effet en général, Θ_1 est une hypothèse composite, c'est-à-dire que Θ_1 n'est pas un singleton et on a souvent une infinité de cas possibles (exemple : $\Theta_1 = \mathbb{R} \setminus \{3\}$ ou $\Theta_1 =]3, +\infty[$). Il est évidemment plus facile de voir qu'on est sous H_1 quand le vrai μ vaut 10 que quand il vaut 3.5 (toutes choses étant égales par ailleurs). De plus, la puissance dépend également de la taille de l'échantillon et de sigma.

Exemples concrets de calculs de puissance : reprenons l'exemple des données gaussiennes ci-dessus $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Et calculons la puissance dans différents cas. On appelle α le niveau dans les 3 exemples ci-dessous.

1. σ connu et problème de test $H_0 : \mu = 3$ contre $H_1 : \mu \neq 3$.
   $\phi = \mathbb{1}_{|T|>q}$ avec $q = q^{N(0,1)}_{1-\frac{\alpha}{2}}$ et $T = \sqrt{n}\frac{\bar{X}-3}{\sigma}$. La fonction puissance, pour $\mu \neq 3$,

est donnée par

$$\begin{aligned}
\beta(\mu) &= \mathbf{P}_\mu(|T| > q) = \mathbf{P}_\mu(|\sqrt{n}\frac{\bar{X}-3}{\sigma}| > q) \\
&= \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-3}{\sigma} > q) + \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-3}{\sigma} < -q) \\
&= \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu+\mu-3}{\sigma} > q) + \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu+\mu-3}{\sigma} < -q) \\
&= \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} > q+\sqrt{n}\frac{3-\mu}{\sigma}) + \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} < -q+\sqrt{n}\frac{3-\mu}{\sigma}) \\
&= 1 - \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq q+\sqrt{n}\frac{3-\mu}{\sigma}) + \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} < -q+\sqrt{n}\frac{3-\mu}{\sigma}) \\
&= 1 - \Phi(q+\sqrt{n}\frac{3-\mu}{\sigma}) + \Phi(-q+\sqrt{n}\frac{3-\mu}{\sigma})
\end{aligned}$$

Quelques exemples d'applications numériques avec $\alpha = 5\%$ (arrondis à deux chiffres après la virgule) :
Code python pour calculer α

```
import scipy.stats as stat
import numpy as np

def calcul_puissance(alpha,sigma,n,mureel,muH0):
q = stat.norm.ppf(1.0-alpha/2,loc=0,scale=1)
beta=(1.0- stat.norm.cdf(q + np.sqrt(n)*(muH0-mureel)/sigma,loc=0,scale=1)
+ stat.norm.cdf(- q + np.sqrt(n)*(muH0-mureel)/sigma,loc=0,scale=1))
print("mu=",mureel," muH0=",muH0," sigma=",sigma," n=",n,
" beta(",mureel,")=",np.round(beta,2))
return beta

calcul_puissance(0.05,1,100,3.5,3.0);
calcul_puissance(0.05,1,10,3.5,3.0);
calcul_puissance(0.05,2,100,3.5,3.0);
calcul_puissance(0.05,1,100,3.1,3.0);

murange = np.linspace(0,6,100)
betan100=np.zeros_like(murange)
betan10=np.zeros_like(murange)
betan2=np.zeros_like(murange)

for index,mureel in enumerate(murange):
betan100[index]=calcul_puissance(0.05,1,100,mureel,3.0);
betan10[index]=calcul_puissance(0.05,1,10,mureel,3.0);
betan2[index]=calcul_puissance(0.05,1,2,mureel,3.0);

plt.figure(14)
plt.rc('font',size=14)
```

```
plt.plot(murange,betan100,"g",murange,betan10,"b",
murange,betan2,"r",linewidth=4)
plt.ylabel("Puissance",size=14)
plt.xlabel("$\mu$",size=14)
plt.legend(["n=100","n=10","n=2"])
plt.title("Puissance pour $\sigma=1, \mu_{H0}=3.0$")
plt.savefig("betaplot.pdf")
```

```
==============Resultats:==============================

mu= 3.5  muH0= 3.0  sigma= 1  n= 100  beta( 3.5 )= 1.0
mu= 3.5  muH0= 3.0  sigma= 1  n= 10  beta( 3.5 )= 0.35
mu= 3.5  muH0= 3.0  sigma= 2  n= 100  beta( 3.5 )= 0.71
mu= 3.1  muH0= 3.0  sigma= 1  n= 100  beta( 3.1 )= 0.17
```

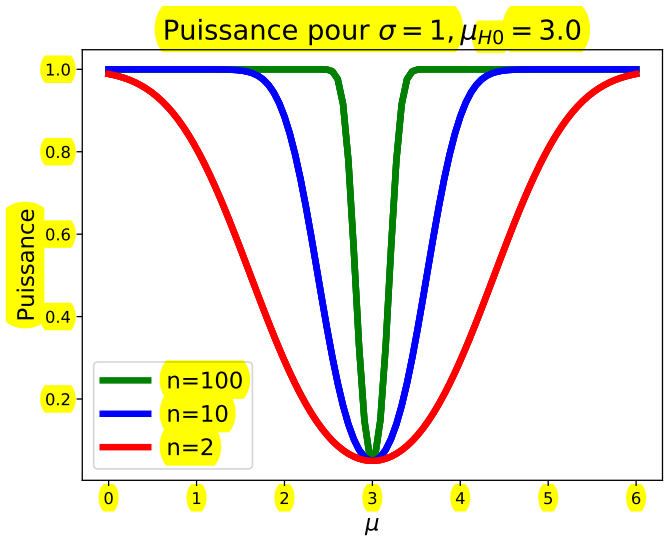


FIGURE 1.5 – Fonction puissance pour les exemples du test bilatéral $\mu = 3$ contre $\mu \neq 3$.

2. σ connu et problème de test $H_0 : \mu = 3$ contre $H_1 : \mu > 3$.
$\phi = \mathbb{1}_{T>q}$ avec $q = q_{1-\alpha}^{N(0,1)}$ et $T = \sqrt{n}\frac{\bar{X}-3}{\sigma}$. La fonction puissance, pour $\mu > 3$, est donnée par

$$\begin{aligned}\beta(\mu) &= \mathbf{P}_\mu(T > q) = \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-3}{\sigma} > q)\\ &= \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} > q + \sqrt{n}\frac{3-\mu}{\sigma})\\ &= 1 - \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq q + \sqrt{n}\frac{3-\mu}{\sigma})\\ &= 1 - \Phi(q + \sqrt{n}\frac{3-\mu}{\sigma})\end{aligned}$$

Dans ces deux premiers exemples, on voit immédiatement que la fonction puissance tend vers 1 lorsque $n \to \infty$. Cela signifie que pour tout μ de l'alternative et pour tout $\epsilon > 0$, il existe une taille d'échantillon n_0 telle que la probabilité de rejeter à tort H_1, quand on est sous $\mathbf{P}_\mu$ pour ce μ particulier, est plus petite que ϵ si $n \geq n_0$. En revanche, dans les deux cas, on peut montrer que la fonction puissance ne tend pas vers 1 uniformément, ce qui signifie que ce n_0 dépend de μ (considérer par exemple la suite $\mu_n = 3 + \frac{1}{n}$). L'erreur de seconde espèce, qui est définie par un sup, ne tend pas vers 0. Voir aussi l'encadré 1.7.1.

3. σ inconnu et problème de test $H_0 : \mu = 3$ contre $H_1 : \mu > 3$.

Si σ est inconnu, on ne peut plus baser notre test sur $T = \sqrt{n}\frac{\bar{X}-3}{\sigma}$ car T n'est plus calculable. On remplace donc σ par un estimateur, ici prenons $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$. Avec cet estimateur $\hat{\sigma}$ on définit donc $T = \sqrt{n}\frac{\bar{X}-3}{\hat{\sigma}}$. La loi de cette statistique sous H_0 est la loi de Student à $n-1$ degrés de liberté. En effet on a

$$T = \sqrt{n}\frac{\bar{X}-3}{\hat{\sigma}} = \frac{\sqrt{n}\frac{\bar{X}-3}{\sigma}}{\frac{\hat{\sigma}}{\sigma}} = \frac{\sqrt{n}\frac{\bar{X}-3}{\sigma}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \tag{1.8}$$

avec

$$\sqrt{n}\frac{\bar{X}-3}{\sigma} \sim N(0,1) \quad \text{et} \quad \frac{\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n \frac{(X_i-\bar{X})^2}{\sigma^2}}{n-1} \sim \frac{\chi^2(n-1)}{n-1}$$

et $\bar{X}$ est indépendant de $\hat{\sigma}^2$ (cf. résultat de type Cochrane). On pose donc $\phi = \mathbb{1}_{T>q}$ avec $q = q_{1-\alpha}^{T(n-1)}$. Ce test est appelé test de Student.

La fonction puissance, pour $\mu > 3$, est donnée par

$$\begin{aligned}
\underline{\beta}(\mu) &= \mathbf{P}_\mu(T > q) = \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-3}{\hat{\sigma}} > q) \\
&= \mathbf{P}_\mu(\sqrt{n}(\bar{X} - \mu + \mu - 3) > q\hat{\sigma}) \\
&= \mathbf{P}_\mu(\sqrt{n}(\bar{X} - \mu) > q\hat{\sigma} + \sqrt{n}(3-\mu)) \\
&= \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} > \frac{q\hat{\sigma} + \sqrt{n}(3-\mu)}{\sigma}) \\
&= 1 - \mathbf{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq \frac{q\hat{\sigma} + \sqrt{n}(3-\mu)}{\sigma}).
\end{aligned}$$

Si on pose $U = \sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ et $V = \frac{q\hat{\sigma}+\sqrt{n}(3-\mu)}{\sigma}$ on a $\underline{\beta}(\mu) = 1 - \mathbf{P}(U \leq V)$ avec U et V indépendantes, puisque $\bar{X}$ et $\hat{\sigma}$ sont indépendantes. Donc, d'après l'exemple 2 de la section 1.5, $\underline{\beta}(\mu) = 1 - \mathbf{E}[F_U(V)]$ où F_U est la cdf de $U = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim N(0,1)$. On obtient donc finalement

$$\underline{\beta}(\mu) = 1 - \mathbf{E}\Big(\Phi(\frac{q\hat{\sigma} + \sqrt{n}(3-\mu)}{\sigma})\Big).$$

On peut ensuite vérifier si la puissance tend bien simplement vers 1 quand n tend vers l'infini. Pour cela on peut utiliser l'expression ci-dessus. Mais on peut aussi déduire cette propriété directement, sans faire appel à cette

expression. Rappelons une méthode assez fréquemment utilisée pour montrer cette propriété : pour fixer les idées, on doit montrer que $\mathbf{P}_\mu(T > c_n)$ tend vers 1 quand n tend vers l'infini. Il suffit alors de :

— montrer que la statistique de test T_n se décompose en $T_n = T_{n,0} + T_{n,1}$ avec

— $T_{n,0} = O_P(1)$ ("grand O en probabilité", typiquement on montre que $T_{n,0}$ converge en loi)

— $T_{n,1} \overset{Proba}{\to} +\infty$,

— et $c_n = O(1)$.

On a ici : $T_n = \sqrt{n}\frac{\bar{X}-3}{\hat{\sigma}} = T_{n,0} + T_{n,1}$ avec $T_{n,0} = \sqrt{n}\frac{\bar{X}-\mu}{\hat{\sigma}}$, $T_{n,1} = \sqrt{n}\frac{\mu-3}{\hat{\sigma}}$ et $c_n = q_{1-\alpha}^{T(n-1)}$. On a, sous $\mathbf{P}_\mu$ avec $\mu > 3$ et quand $n \to +\infty$,

$$T_{n,0} \sim T(n-1) \quad \text{donc} \quad T_{n,0} = O_P(1).$$

$$T_{n,1} \overset{p.s.}{\to} +\infty$$

et enfin $q_{1-\alpha}^{T(n-1)} \to q_{1-\alpha}^{N(0,1)}$ car la loi de Student à $n-1$ degrés de liberté tend vers la loi normale standard (cf chapitre 2 théorème 2.20).

1.7.3 La p-valeur

Définition 1.14. *Supposons avoir construit une famille de tests $\phi_\alpha(X)$, chacun de niveau α, pour $\alpha \in [0,1]$. La p-valeur associée à cette famille est la variable aléatoire réelle définie par*

$$p(X) = \inf\{\alpha \in [0,1] : \phi_\alpha(X) = 1\}.$$

Intuition 1.7.3. *Interprétation de la p-valeur : plus la p-valeur observée est petite, plus on a envie de rejeter H_0 car cela signifie que la valeur observée de la statistique utilisée pour le test est atypique pour H_0.*

Remarque 1.15. *On constate que $p(X)$ est le niveau à partir duquel on se met à rejeter H_0. C'est comme si on faisait le test sans connaître le α et on tire la conclusion à la fin une fois que le α est dévoilé. Donc*

— *Si $p(x) < \alpha$ alors on rejette H_0 au niveau α.*

— *Si $p(x) > \alpha$ alors on conserve H_0 au niveau α.*

Mise en garde 1.7.2. *Une p-valeur petite ne veut pas dire que l'on a plus de chances d'être sous H_1 que sous H_0 : ça dépend en fait du comportement de la p-valeur sous H_1. On sait que, sous certaines conditions du moins (cf chapitre 2), la p-valeur suit une loi uniforme sous H_0, mais on ne sait pas forcément le comportement*

de la p-valeur sous H_1. La question de la probabilité de H_0 sachant les données est une question bayésienne à laquelle on peut répondre si on a un a priori sur l'alternative (il faut aussi parfois un a priori sur H_0). Attention donc à l'interprétation des p-valeurs, ***ne pas dire*** *"la p-valeur est petite donc la probabilité que H_0 soit fausse est grande".*

Pour autant, une p-valeur importante n'implique pas forcément que H_0 soit vraie. Il se peut que le test ne soit pas puissant. Par exemple considérons le test $\phi(X) \equiv 0$: ce test accepte toujours H_0. L'ensemble dans la définition de la p-valeur est vide, par convention on prend son sup pour définir la p-valeur, c'est-à-dire que la p-valeur est égale à 1.

Exemple 1.16. *Un exemple de cas où le calcul de la p-valeur est très simple : supposons que le test est de la forme $\phi_\alpha(X) = \mathbb{1}_{T(X)>k_\alpha}$, que $\Theta_0 = \{\theta_0\}$ et que la statistique $T(X)$ a, sous $\mathbf{P}_{\theta_0}$, une loi de cdf F_0 strictement croissante et continue. Alors on a $k_\alpha = F_0^{-1}(1-\alpha)$. Et on voit facilement que*

$$p(x) = 1 - F_0(T(x)).$$

En effet,

$$\mathbf{P}_{\theta_0}(T(X) > k_\alpha) = \alpha \iff 1 - F_0(k_\alpha) = \alpha \iff k_\alpha = F_0^{-1}(1-\alpha).$$

Et la p-valeur observée est donnée par

$$\begin{aligned} p(x) &= \inf\{\alpha \in]0,1[\colon T(x) > F_0^{-1}(1-\alpha)\} \\ &= \inf\{\alpha \in]0,1[\colon F_0(T(x)) > 1-\alpha\} \\ &= \inf\{\alpha \in]0,1[\colon \alpha > 1 - F_0(T(x))\} \\ &= 1 - F_0(T(x)). \end{aligned}$$

Dans ce cours, on supposera toujours que

— le test est conçu de façon à maximiser la région de rejet.

— $\phi_\alpha(X)$ décroit quand α décroit.

Intuition 1.7.4. *La première hypothèse est naturelle. Dans la définition d'un test de niveau α, on exige que l'erreur de première espèce soit plus petite que α. On a alors une infinité de solutions possibles : en effet si $\mathbf{P}_{\theta_0}(T > c_1) \leq \alpha$ alors $\mathbf{P}_{\theta_0}(T > c_2) \leq \alpha$ pour tout $c_2 > c_1$. Si on veut minimiser l'erreur de seconde espèce, il faut alors maximiser la région de rejet (et donc prendre c le plus petit possible).*
La seconde hypothèse est aussi très naturelle. Elle se réécrit

$$\alpha_1 \leq \alpha_2 \Longrightarrow \mathcal{R}_{\alpha_1} \subset \mathcal{R}_{\alpha_2}$$

autrement dit, si on rejette à un niveau α_1 alors on rejette aussi à tout niveau $\alpha_2 \geq \alpha_1$.

Théorème 1.17. *("théorème de Wasserman")*
On suppose que les tests que l'on fait à un niveau α donné maximisent la région de rejet.

- *Supposons qu'une famille de tests soit de la forme $\phi_\alpha(X) = \mathbb{1}_{T(X)\leq k_\alpha}$, pour $\alpha \in]0,1[$. Alors, si le test est de <u>taille</u> α, la p-valeur s'écrit*

$$p(x) = \sup_{\theta\in\Theta_0} \mathbf{P}_\theta(T(X) \leq T(x)),$$

  où x est la valeur observée de X.
- *Pour une famille de tests de <u>taille</u> α de la forme $\phi_\alpha(X) = \mathbb{1}_{T(X)\geq k_\alpha}$, on a $p(x) = \sup_{\theta\in\Theta_0} \mathbf{P}_\theta(T(X) \geq T(x))$.*
- *Si la variable $T(X)$ a une loi discrète de cdf F_0 fixe sous H_0 et si la famille de tests est de la forme $\phi_\alpha(X) = \mathbb{1}_{T(X)\leq k_\alpha}$ alors*

$$p(x) = F_0(T(x)) = \mathbf{P}_{H_0}(T(X) \leq T(x))$$

- *Si la variable $T(X)$ a une loi discrète de cdf F_0 fixe sous H_0 et si la famille de tests est de la forme $\phi_\alpha(X) = \mathbb{1}_{T(X)\geq k_\alpha}$, avec les mêmes hypothèses, on a $p(x) = \mathbf{P}_{H_0}(T(X) \geq T(x))$.*
- *Ces formules sont encore vraies s'il existe θ_0 tel que pour tout t,*

$$\sup_{\theta\in\Theta_0} \mathbf{P}_\theta(T(X) \leq t) = \mathbf{P}_{\theta_0}(T(X) \leq t)$$

  si le test s'écrit $\phi_\alpha(X) = \mathbb{1}_{T(X)\leq k_\alpha}$ ou

$$\sup_{\theta\in\Theta_0} \mathbf{P}_\theta(T(X) \geq t) = \mathbf{P}_{\theta_0}(T(X) \geq t)$$

  si le test s'écrit $\phi_\alpha(X) = \mathbb{1}_{T(X)\geq k_\alpha}$

Admis.
Ce qu'on veut dire par loi fixe : $T(X)$ a la même loi $\forall\theta \in \Theta_0$. Par exemple c'est le cas si $\Theta_0 = \{\theta_0\}$. C'est aussi le cas pour le test de Kolmogorov-Smirnov, le test du signe et les tests de Wilcoxon (cf chapitre 2).

Exemples de calculs de p-valeurs

- Soit $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ avec μ et σ^2 inconnus. On veut tester

$$H_0 : \mu = 2 \quad \text{contre} \quad \mu \leq 2.$$

  Le test utilisé est alors le test de Student (cf "exemples de calcul de puissance", item 3). Le test est alors $\phi = \mathbb{1}_{T\leq q_\alpha^{T(n-1)}}$ avec $T = \sqrt{n}\frac{\bar{X}-2}{\hat{\sigma}}$. On est alors dans les conditions d'application du théorème de Wassermann, item 1 : en effet,

puisque la loi de Student est une loi continue, on a bien un test de taille α (et pas seulement de niveau α). La p-valeur observée est donc donnée par

$$p(x_1^n) = F_{T(n-1)}(T(x_1^n))$$

où $T(x_1^n)$ est l'observation de la statistique $T(X_1^n)$.
Application numérique : $n = 20$, $\frac{1}{20}\sum_{i=1}^{20} x_i = 1.34$ et $\sqrt{\frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2} = 1.06$, ce qui donne la valeur observée $T(x_1^n) = \sqrt{20}(1.34-2)/1.06 = -2.78$. La p-valeur observée est donnée par $p(x_1^n) = 0.01$. On peut trouver cette valeur sur R en utilisant la commande `pt(-2.78,19)`. Pour obtenir directement ce résultat sans faire de calcul, on peut utiliser la commande R `t.test(-2.78,mu=2,alternative="less")`.

— Soit $X_1, \dots, X_n \overset{iid}{\sim} Be(p)$. On veut tester

$$H_0 : p = 1/2 \quad \text{contre} \quad p > 1/2.$$

On utilise la statistique $T = \sum_{i=1}^n X_i$ qui suit, sous H_0, une loi binomiale $B(n, 1/2)$. Au vu de H_1, on rejette quand T est trop grand. Donc on pose $\phi = \mathbb{1}_{T>c}$ où c est déterminé par le fait que le test est de niveau α

$$\mathbf{P}_{1/2}(T > c) \leq \alpha. \tag{1.9}$$

Attention ici la statistique T est discrète donc sa cdf $F_{B(n,1/2)}$ sous $\mathbf{P}_{1/2}$ n'est pas continue. Donc on ne peut pas toujours avoir l'égalité.
On verra au chapitre 2 que le plus petit entier c vérifiant (1.9) est donné par $c = q_{1-\alpha}^{B(n,1/2)}$, i.e. le quantile d'ordre $1-\alpha$ de la loi binomiale $B(n, 1/2)$. Cela donne le test suivant

$$\phi = \mathbb{1}_{T > q_{1-\alpha}^{B(n,1/2)}}.$$

Remarquez que, comme T est presque sûrement à valeurs entières, ce test peut aussi s'écrire $\phi = \mathbb{1}_{\{T \geq q_{1-\alpha}^{B(n,1/2)}+1\}}$. Donc il a bien l'une des formes indiquées dans le théorème de Wassermann. Nous sommes bien dans les conditions d'application de ce théorème (2ème item), donc la p-valeur observée est donnée par

$$p(x_1^n) = \mathbf{P}(Z \geq T(x_1^n)) = 1 - \mathbf{P}(Z < T(x_1^n)) = 1 - F_{B(n,1/2)}(T(x_1^n) - 1).$$

où Z désigne une variable aléatoire de loi $B(n, 1/2)$.
Par exemple, si $n = 20$, et si la valeur observée de $\sum_{i=1}^n X_i$ est 11 alors la p-valeur du test est $p(x_1^n) = 0.41$. Donc on a tendance à accepter H_0 au vu des données.
On peut obtenir cette valeur sur R avec la commande `pbinom(10,20,1/2)`. On retrouve cette p-valeur directement utilisant la commande `binom.test(11,20,1/2,alternative="greater")` .

Remarque 1.18. *Attention aux inégalités strictes versus inégalités larges, elles ont leur importance, surtout pour des variables discrètes. Dans le théorème de Wassermann, il s'agit d'inégalités larges.*

Pour une variable discrète, on définira le test à partir d'inégalités strictes comme ci-dessus. On peut toujours transformer ce type de test en un test avec égalité large (ex : $T > 3 \iff T \geq 4$ si T prend des valeurs entières).

Méthode pour construire un test

1. Choix de H_0 et H_1.
2. Détermination de $T(X)$, la statistique de test. On doit connaitre sa loi sous H_0. Evidemment on souhaite aussi que cette statistique ait un comportement différent sous H_0 et sous H_1 pour pouvoir discriminer les deux hypothèses.
3. Allure de la zone de rejet en fonction de H_1 (i.e. en fonction du comportement de $T(X)$ sous H_1).
4. Observation de la réalisation $T(x)$ de $T(X)$.
5. Calcul de la p-valeur associée $p(x)$ et comparaison à un seuil fixé par un non-statisticien.
6. Conservation ou non de H_0.

1.7.4 Interprétation des p-valeurs : d'autres exemples et détails

Comme la p-valeur est définie comme un infimum, ce n'est pas forcément un "min" donc on ne sait pas a priori si $p(x) \in \left\{\alpha \in]0,1[: \phi_\alpha(x) = 1\right\}$ ou pas, c'est-à-dire qu'on ne sait pas si on rejette H_0 pour le niveau $\alpha = p(x)$. Appelons α^* la p-valeur $p(x)$ de façon à la considérer comme un niveau. Pour fixer les idées (ça ne change rien au raisonnement), supposons que $\phi_{\alpha*}(x) = 1$, autrement dit, pour le niveau α_* on rejette H_0. On rappelle que le niveau α est choisi comme étant la probabilité de rejeter à tort H_0 (ou un majorant de cette probabilité si on ne peut pas avoir l'égalité pour tout α).

Donc si on regarde la p-valeur comme un niveau α^*, alors on rejette pour ce niveau α^* et, si α^* est très petit, alors la probabilité de rejeter à tort est très faible. En quelque sorte, plus la p-valeur observée est petite, plus on on a envie de rejeter H_0.

Supposons que l'on ait observé une p-valeur de $p = 0.001$, qui est donc très petite. Alors pour le niveau $\alpha = 5\%$ on rejette H_0 puisque $0.001 < 0.05$. Mais en plus, le fait de connaitre la p-valeur nous apporte une information supplémentaire : le fait que p soit vraiment petit ici nous donne une certaine confiance dans notre rejet. Par exemple si on avait eu $p = 0.04$ alors on aurait aussi rejeté au niveau $\alpha = 5\%$ mais on l'aurait fait avec moins d'assurance.

Les logiciels de statistique donnent toujours la p-valeur quand on leur demande de faire un test. Prenons l'exemple du test de Student. On a dans un vecteur x un échantillon de gaussiennes de moyenne et variance inconnues et on veut tester $H_0 : \mu = 1.5$ contre $\mu \neq 1.5$ où μ est la moyenne. Alors on peut utiliser la commande R suivante

```
t.test(x,mu=1.5)
```

dont la sortie est

```
One Sample t-test
```

```
data:  x
t = -1.9561, df = 19, p-value =
0.06532
alternative hypothesis: true mean is not equal to 1.5
95 percent confidence interval:
 0.6181763 1.5298237
sample estimates:
mean of x
    1.074
```

On tombe sur une p-valeur d'environ 0.06 donc on accepte (tout juste) H_0 au niveau 5%. Là encore, comme la p-valeur est proche du niveau, on n'a pas une confiance énorme en le résultat final.

Interprétation à l'aide du théorème de Wassermann

Reprenons un des exemples précédents : premier item de "exemples de calculs de p-valeurs". La p-valeur s'écrit dans cet exemple $p(x_1^n) = F_{T(n-1)}(T(x_1^n))$. Dans l'application numérique, la valeur observée de la statistique T est $t = -2.78$. Si on est vraiment sous H_0, t est alors censée être la valeur observée d'une statistique qui suit une loi de Student à 19 degrés de liberté, et la p-valeur mesure alors la probabilité qu'une variable de Student à 19 degrés de liberté soit plus petite que -2.78, c'est-à-dire la probabilité d'observer une valeur de T plus petite que -2.78 **si on est vraiment sous** H_0. Donc la p-valeur mesure en quelque sorte le côté atypique de la valeur observée, par rapport à ce qu'il est censé se passer sous H_0.

Ici, si on était vraiment sous H_0, il y aurait une probabilité de 1% d'observer une valeur inférieure ou égale à -2.78 pour la statistique T. Donc -2.78 est plutôt une valeur atypique pour H_0 et on penche donc pour le rejet de H_0.

1.8 Exercices

Exercice 1.1. *Soit X une variable aléatoire réelle, absolument continue de densité continue f, de fonction de répartition F. On observe un n-échantillon iid $(X_1, \dots, X_n)$ de même loi que X. On considère la statistique T qui ordonne l'échantillon dans le sens croissant :*

$$T(X_1, \dots, X_n) = (X_{(1)}, \dots, X_{(n)}),$$

avec $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. $(X_{(1)}, \dots, X_{(n)})$ s'appelle la statistique d'ordre.

1. *On suppose pour cette question uniquement que les X_i sont seulement indépendants et de lois continues (c'est-à-dire que les X_i sont indépendants et ont tous une fonction de répartition F_i continue, mais pas forcément absolument continue). Montrer que*

$$\mathbb{P}\left(\exists\, i \neq j : \quad X_i = X_j\right) = 0,$$

   et que dans la définition de la statistique d'ordre, on peut donc se limiter à des inégalités strictes : $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$.
2. *Déterminer la densité de la loi du n-uplet $(X_{(1)}, \dots, X_{(n)})$.*
3. *Déterminer la fonction de répartition F_k et la densité f_k de $X_{(k)}$.*
4. *Montrer que si $\mathbb{E}[|X|]$ est finie, alors il en est de même de $\mathbb{E}[|X_{(k)}|]$.*
5. *Rappeler les densités des lois de $X_{(1)}$ et $X_{(n)}$ et déterminer la densité du couple $(X_{(1)}, X_{(n)})$. Quelle est la loi de $W_n = X_{(n)} - X_{(1)}$?*
6. *On considère une suite $(U_i)_{i\in\mathbb{N}}$ de variables i.i.d. selon la loi uniforme sur $[0, 1]$, et on pose*

$$Y_n = \min_{1\leq i\leq n} U_i \qquad Z_n = \max_{1\leq i\leq n} U_i - \min_{1\leq i\leq n} U_i$$

   (a) *Montrer que nY_n converge en loi vers une loi exponentielle.*

   (b) *Étudier la convergence en loi de Z_n, puis sa convergence en probabilité et $\mathbb{L}^1$.*

   (c) *Soit $\epsilon > 0$. Calculer $\mathbb{P}[|Z_n - 1| > \epsilon]$. En déduire que Z_n converge presque sûrement.*

   (d) *Rappeler les implications logiques entre les modes de convergence étudiés : en loi, en probabilité, en norme $\mathbb{L}^1$, en norme $\mathbb{L}^2$, presque sûre.*

Exercice 1.2. *On reprend un exemple du cours. La limite légale d'un polluant contenu dans les déchets d'une usine est de 6mg/kg. On effectue un dosage sur 20 prélèvements sur lesquels on observe une moyenne empirique de 7mg/kg avec un écart-type empirique de 2.4mg/kg. On admet que la loi de dosage est gaussienne.*

On observe donc $X_1, \dots, X_{20} \overset{iid}{\sim} N(\mu, \sigma^2)$ avec μ et σ^2 inconnus.

1. *Faire un test de niveau α pour le problème de test suivant :*

$$H_0 : \mu \leq 6 \quad \textit{contre} \quad H_1 : \mu > 6.$$

2. *On calcule à partir de ces données* $\bar{x} = 7$ *et* $\hat{\sigma}^2 = \frac{1}{19}\sum_{i=1}^{20}(x_i - \bar{x})^2 = 2.4^2$. *Calculer la p-valeur observée et conclure si on choisit le niveau* $\alpha = 5\%$.

Exercice 1.3. *On dispose d'un échantillon de loi Bernoulli de paramètre p :* $X_1, \ldots, X_n \overset{iid}{\sim} Be(p)$.

1. *Proposer une procédure de test pour le problème suivant*

$$H_0 : p = 1/2 \quad \textit{contre} \quad H_1 : p > 1/2.$$

2. *Proposer une procédure de test pour le problème suivant*

$$H_0 : p = 1/2 \quad \textit{contre} \quad H_1 : p \neq 1/2$$

3. *Proposer un test asymptotique pour le problème de la question précédente.*
4. *Calculer la puissance du test asymptotique de la question 3. La puissance tend-elle simplement vers 1 quand n tend vers l'infini ?*
5. *Application numérique. On calcule à l'aide des données,* $n = 100$, $\bar{x} = 0.59$, $q_{0.95} = 58$, $q_{0.975} = 60$ *où on note* q_α *le quantile d'ordre* α *de la loi binomiale* $B(100, 1/2)$. *Quelle est la conclusion des deux premiers tests ci-dessus, au niveau* $\alpha = 0.05$, *pour ces données ?*

Chapitre 2

Estimation de la fonction de répartition

2.1 Consistance des fonctions de répartition empiriques

On considère $X_1^n = (X_1, \dots, X_n)$ un n-échantillon iid de cdf F : $\forall x \in \mathbb{R}$, $F(x) = \mathbf{P}(X_1 \leq x)$. On rappelle que :

— F est croissante

— F est continue à droite

— $\lim_{x \to +\infty} F(x) = 1$ et $\lim_{x \to -\infty} F(x) = 0$.

On peut préciser que, étant croissante, elle a une limite à gauche en tout point et elle admet au plus un nombre dénombrable de discontinuités aux points x tel que $\mathbf{P}(X_j = x) \neq 0$.

Il existe un estimateur naturel de F : la fonction de répartition empirique.

Définition 2.1. *la fonction de répartition empirique associée à $X_1^n = (X_1, \dots, X_n)$ est la fonction aléatoire définie par :* $\hat{F}_n : \begin{array}{ccc} \mathbb{R} & \to & [0,1] \\ x & \to & \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x} \end{array}$

Remarque 2.2. *Pour insister sur le caractère aléatoire de $\hat{F}_n$, on peut écrire parfois $\hat{F}_n(\omega, x)$ au lieu de $\hat{F}_n(x)$. $\hat{F}_n(\omega, x)$ désigne donc la valeur de la cdf $\hat{F}_n$ en x quand l'observation est ω.*

Remarque 2.3. *On construit facilement $\hat{F}_n$ car c'est une fonction en escalier. Fixons ω et écrivons $(x_1, \dots, x_n) = (X_1(\omega), \dots, X_n(\omega))$. Alors $\hat{F}_n(\omega, x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x_i \leq x}$ est la fonction de répartition de la variables aléatoire Z à valeurs dans $\{x_1, \dots, x_n\}$ et telle que $\mathbf{P}(Z = x_i) = \frac{k}{n}$ si la valeur x_i apparait k fois dans $\{x_1, \dots, x_n\}$. Par exemple, si tous les x_i distincts, alors $\hat{F}_n(\omega, \cdot)$ est la cdf de la loi uniforme sur $\{x_1, \dots, x_n\}$.*

Soit $(X_{(1)}, \dots, X_{(n)})$ la statistique d'ordre associée à X_1^n. On rappelle que cela signifie que $\{X_{(1)}, \dots, X_{(n)}\} = \{X_1, \dots, X_n\}$ et

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

La fonction $\hat{F}_n(\omega, \cdot)$ est discontinue aux points $X_{(j)}(\omega)$. Elle a un saut égal au nombre de fois où la valeur $X_i(\omega)$ apparait dans $\{X_1(\omega), \dots, X_n(\omega)\}$. En particulier si tous les $X_i(\omega)$ sont distincts, i.e. $X_{(j)}(\omega) < X_{(j+1)}(\omega)$ pour tout j, alors $\hat{F}_n(\omega, x) = \frac{j}{n}$ pour tout $x \in [X_{(j)}(\omega), X_{(j+1)}(\omega)[$. Dans tous les cas, elle vaut 0 sur $]-\infty, X_{(1)}(\omega)[$ et 1 sur $[X_{(n)}(\omega), +\infty[$.

Proposition 2.4. *Soit $x \in \mathbb{R}$, $\hat{F}_n(x)$ est un estimateur sans biais de $F(x)$ et $\lim_{n\to\infty} \hat{F}_n(x) = F(x)$ p.s. Par ailleurs*

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{Loi} \mathcal{N}(0, F(x)(1 - F(x)))$$

Démonstration. $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \le x}$ avec $\mathbb{1}_{X_i \le x} \overset{iid}{\sim} Be(F(x))$ donc $\lim_{n\to\infty} \hat{F}_n(x) = F(x)$ p.s. découle de la LGN. La deuxième propriété vient du théorème limite central en remarquant que $\mathbf{Var}(\mathbb{1}_{X_i \le x}) = F(x)(1 - F(x))$. □

Ce résultat est de nature paramétrique car x est fixé. On peut aller plus loin.

Théorème 2.5. *(Glivenko-Cantelli) Soit $(X_1, \dots, X_n)$ un $n-$échantillon iid de fonction de répartition F. Alors la fonction de répartition empirique est un estimateur fortement consistant de F pour la norme de la convergence uniforme :*

$$\lim_{n\to\infty} \|\hat{F}_n - F\|_\infty = \lim_{n\to\infty} \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| = 0 \quad p.s.$$

La preuve sera donnée dans la section 2.3.

Définition 2.6. *A toute fonction de répartition F on associe son inverse généralisé $F^{(-1)}$ définie comme suit :*

$$\forall q \in [0,1] \quad F^{(-1)}(q) = \inf\{x \in \mathbb{R} : F(x) \ge q\}$$

$F^{(-1)}$ est aussi appelée la fonction quantile.

Proposition 2.7. *On a $F^{(-1)} = F^{-1}$ quand F est bijective. De plus,*

1. *$F(F^{(-1)}(q)) \ge q$ pour tout $q \in [0,1]$.*
2. *$\forall x \in \mathbb{R}, \forall q \in [0,1], \quad F(x) \ge q \Leftrightarrow x \ge F^{(-1)}(q)$.*
3. *Si $U \sim U[0,1]$ alors $F^{(-1)}(U)$ est une v.a. de fonction de répartition F.*
4. *Si F est continue alors $F(F^{(-1)}(q)) = q$. pour tout $q \in]0,1[$.*
5. *Si Z admet pour fonction de répartition F continue alors $F(Z) \sim U[0,1]$.*
6. *$F^{(-1)}$ est croissante.*

Démonstration.

1. Par définition de $F^{(-1)}(q)$, il existe une suite $(u_n)_{n\ge 0}$ telle que $F(u_n) \ge q$ et $u_n \underset{n\to\infty}{\to} F^{(-1)}(q)$ en décroissant (c'est donc une limite à droite) . Comme F est continue à droite, $F(u_n) \to F(F^{(-1)}(q))$. Donc $F(F^{(-1)}(q)) \ge q$.
2.
   - Si $F(x) \ge q$ alors par définition $F^{(-1)}(q) \le x$.
   - Si $x \ge F^{(-1)}(q)$ alors par croissance de F on a $F(x) \ge F(F^{(-1)}(q))$ donc $F(x) \ge q$ par l'item 1.

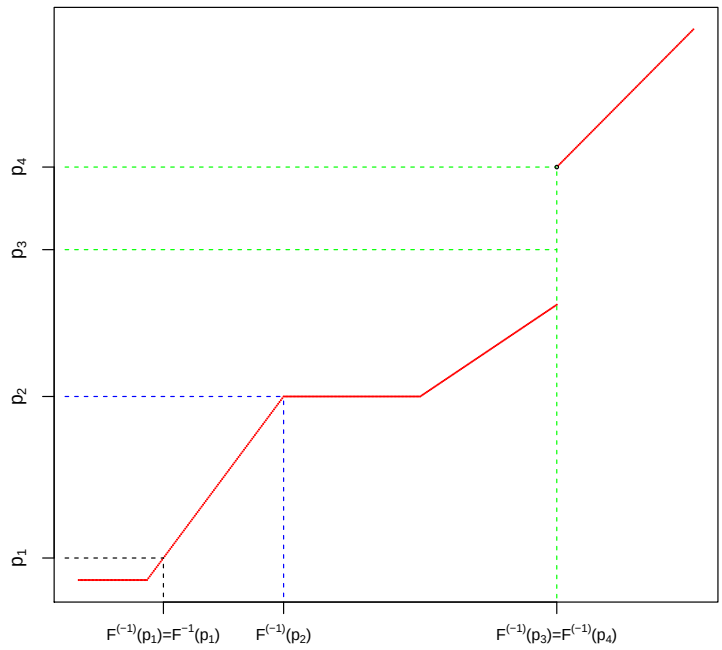


FIGURE 2.1 – Fonction de répartition (en rouge) avec palier et saut

3. D'après l'item 2 on a $\mathbf{P}(F^{(-1)}(U) \leq t) = \mathbf{P}(U \leq F(t))$ et $\mathbf{P}(U \leq F(t)) = F(t)$ car $F(t) \in [0,1]$.
4. D'après l'item 1, il suffit de montrer que $F(F^{(-1)}(q)) \leq q$. Si F est continue alors $]0,1[\subset Im(F)$, d'après le théorème des valeurs intermédiaires. Donc il existe $x_q \in \mathbb{R}$ tel que $F(x_q) = q$. Donc par définition $F^{(-1)}(q) \leq x_q$. Donc par croissance de F, $F(F^{(-1)}(q)) \leq q$.
5. Soit $t \in]0,1[$. On a

$$\begin{aligned}\mathbf{P}(F(Z) < t) &= 1 - \mathbf{P}(F(Z) \geq t) \\ &= 1 - \mathbf{P}(Z \geq F^{(-1)}(t)) \\ &= \mathbf{P}(Z < F^{(-1)}(t)) \\ &= F(F^{(-1)}(t)) \\ &= t\end{aligned}$$

où on a utilisé l'item 2 pour la 2ème ligne, le fait que F est continue pour la 4ème ligne, et l'item 4 pour la dernière ligne. Comme $]-\infty, x] = \cap_{t>x}]-\infty, t[$ on a $\mathbf{P}(F(Z) \leq x) = \lim_{t \to x, t>x} \mathbf{P}(F(Z) < t) = \lim_{t \to x, t>x} t = x$. Donc $F(Z) \sim U[0,1]$.
6. Soit $q_1, q_2 \in [0,1]$ avec $q_1 \leq q_2$. Alors $\{x \in \mathbb{R} : F(x) \geq q_2\} \subset \{x \in \mathbb{R} : F(x) \geq q_1\}$ donc $F^{(-1)}(q_1) \leq F^{(-1)}(q_2)$.

□

Remarque 2.8. *Les item 1 et 4 peuvent "se déduire" à partir d'un dessin. On "voit" également que les paliers de F correspondent à un point de discontinuité de $F^{(-1)}$ et qu'un saut de F correspond à un palier de $F^{(-1)}$.*

Remarque 2.9. *Dans un certain nombre de cas (cf exemples ci-dessous), la p-valeur $p(X)$ d'un test suit une loi uniforme sous H_0.*

Exemple 2.10. *Un échantillon $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ avec σ inconnu. Problème de test $H_0 : \mu = \mu_0$ contre $H_1 : \mu \leq \mu_0$. On utilise le test de Student $\phi = \mathbb{1}_{T \leq q_\alpha^{T(n-1)}}$. Alors d'après le théorème de Wassermann, la p-valeur observée s'écrit $p(x_1^n) = F_{T(n-1)}(T(x_1^n))$. Donc la p-valeur $p(X_1^n)$, en tant que variable aléatoire, vérifie $p(X_1^n) = F_{T(n-1)}(T(X_1^n))$. Or $T(X_1^n) \sim T(n-1)$ sous H_0. Donc, d'après l'item 5 de la proposition précédente, $p(X_1^n)$ suit une loi uniforme sous H_0.*

Exemple 2.11. *: même contexte mais avec $H_1 : \mu \geq \mu_0$. Alors le test s'écrit $\phi = \mathbb{1}_{T \geq q_{1-\alpha}^{T(n-1)}}$. Et la p-valeur observée satisfait $p(x_1^n) = \mathbf{P}_{\mu_0}(T(X_1^n) \geq T(x_1^n)) = \mathbf{P}(Z \geq T(x_1^n))$ avec $Z \sim T(n-1)$. Donc $p(x_1^n) = 1 - F_{T(n-1)}(T(x_1^n))$. Ainsi la p-valeur, en tant que variable aléatoire, satisfait $p(X_1^n) = 1 - F_{T(n-1)}(T(X_1^n))$. A nouveau, sous H_0, $F_{T(n-1)}(T(X_1^n)) \sim U[0,1]$ donc $p(X_1^n) \sim U[0,1]$.*

Exemple 2.12. *Même contexte mais avec $H_1 : \mu \neq \mu_0$. Alors le test s'écrit $\phi_\alpha = \mathbb{1}_{|T| \geq q_{1-\frac{\alpha}{2}}^{T(n-1)}}$. et la p-valeur observée s'écrit*

$$\begin{aligned} p(x_1^n) &= \inf\Big\{\alpha \in]0,1[: |T(x_1^n)| \geq F^{-1}_{T(n-1)}(1 - \frac{\alpha}{2})\Big\} \\ &= \inf\Big\{\alpha \in]0,1[: F_{T(n-1)}(|T(x_1^n)|) \geq 1 - \frac{\alpha}{2}\Big\} \\ &= \inf\Big\{\alpha \in]0,1[: \alpha \geq 2\Big[1 - F_{T(n-1)}(|T(x_1^n)|\Big]\Big\} \\ &= 2[1 - F_{T(n-1)}(|T(x_1^n)|]. \end{aligned}$$

Ainsi

$$p(X_1^n) = 2\Big[1 - F_{T(n-1)}(|T(X_1^n)|)\Big]$$

Pour simplifier, on note $T(X_1^n) = T$ et $F_{T(n-1)} = F$. On a alors si $x \in [0,1]$,

$$\begin{aligned} &\mathbf{P}\Big(2[1 - F(|T|)] \leq x\Big) \\ &= \mathbf{P}\Big(F(|T|) \geq 1 - \frac{x}{2}\Big) \\ &= \mathbf{P}\Big(F(T) \geq 1 - \frac{x}{2}, T \geq 0\Big) + \mathbf{P}\Big(F(-T) \geq 1 - \frac{x}{2}, -T \geq 0\Big) \\ &= 2\mathbf{P}\Big(F(T) \geq 1 - \frac{x}{2}, T \geq 0\Big) \\ &= 2\mathbf{P}\Big(F(T) \geq 1 - \frac{x}{2}\Big) \\ &= x \end{aligned}$$

On a utilisé

— *pour la 3ème ligne : la symétrie de la loi de T.*

— *pour la 4ème ligne :* $1-\frac{x}{2} > 1/2$ *si* $x \in]0,1[$. *Or* F *est la cdf de* $T(n-1)$, *donc* F *est continue et correspond à une loi symétrique. Donc* $F(x) = 1-F(-x)$. *Donc* $F(0) = 1/2$ *et comme* F *est strictement croissante, on a* $F(x) \geq \frac{1}{2} \Longrightarrow x \geq 0$

— *pour la dernière ligne :* $1 - F(T) \sim U[0,1]$ *et* $1 - \frac{x}{2} \in]0,1[$.

Donc, à nouveau, $p(X_1^n) \sim U[0,1]$.

2.2 Estimation de quantiles

Pour la construction de tests et de régions de confiance, on s'appuie sur la notion de quantiles. On rappelle la définition générale d'un quantile.

Définition 2.13. *Pour* $\beta \in [0,1]$, *on appelle quantile d'ordre* β *d'une loi de probabilité* P *à support dans* $\mathbb{R}$ *la quantité*

$$q_\beta = \inf\{x \in \mathbb{R} : \mathbf{P}(]-\infty, x]) \geq \beta\}$$

Autrement dit, en utilisant la fonction inverse généralisé, si P admet F pour fonction de répartition

$$q_\beta = F^{(-1)}(\beta)$$

Exemple 2.14. *Soit la loi* $\frac{\delta_0+\delta_1+\delta_2}{3}$. *La quantile de* 25% *est* 0 *et celle de* 75% *est* 3.

Proposition 2.15. 1. *quand la fonction de répartition* F *est inversible, le quantile d'ordre* β *est égale à* $F^{-1}(\beta)$ *et alors on a* $F(q_\beta) = \beta$. *Et le quantile est l'unique solution de cette équation.*

2. *Plus généralement si* F *est continue, on a* $F(q_\beta) = \beta$. *(mais la solution n'est pas unique)*

3. *On a toujours* $F(q_\beta) \geq \beta$ *et,* $F(q_\beta^-) \leq \beta$, *i.e* $P(X < q_\beta) \leq \beta$. *Autrement dit*

$$P(X \leq q_\beta) \geq \beta \quad \text{et } P(X \geq q_\beta) \geq 1-\beta.$$

Démonstration. 1. évident.

2. $F(q_\beta) \geq \beta$ est l'item 1 de la proposition 2.7.

3. $F(x^-) \equiv \lim_{t\to x, t<x} F(t) = \lim_{t\to x, t<x} \mathbf{P}(]-\infty, t]) = \mathbf{P}(]-\infty, x[)$. De plus si $x < q_\beta$ alors, par définition de q_β, on a $F(x) < \beta$. Donc $F(q_b^-) \leq \beta$.

□

Exemple 2.16. *La médiane* m *vérifie*

$$\mathbf{P}(X \leq m) \geq 1/2 \quad \text{et} \quad \mathbf{P}(X \geq m) \geq 1/2.$$

Et on a $\mathbf{P}(X \leq m) = \mathbf{P}(X \geq m) = 1/2$ *quand* F *est continue.*

Remarque 2.17. *D'autres conventions existent pour la définition d'un quantile. On peut aussi définir un quantile de manière non unique. Souvent, on appelle quantile d'ordre* β *de la loi* F *tout nombre* q_β *tel que*

$$P(X \leq q_\beta) \geq \beta \quad \text{et } P(X \geq q_\beta) \geq 1-\beta. \tag{2.1}$$

Proposition 2.18. *Soit X une variable aléatoire réelle de cdf F, et $\alpha \in]0,1[$. Le plus petit réel c tel que $\mathbf{P}(X > c) \leq \alpha$ est égal à $q^F_{1-\alpha}$.*

Démonstration. $\mathbf{P}(X > c) \leq \alpha \Leftrightarrow \mathbf{P}(X \leq c) \geq 1 - \alpha$. Par définition, le plus petit réel c vérifiant cette inégalité est $F^{(-1)}(1-\alpha)$. □

Exemple 2.19. *Soit $X_1, \dots, X_n \overset{iid}{\sim} Be(p)$. On veut tester au niveau α*

$$H_0 : p = 1/2 \quad contre \quad p > 1/2.$$

On utilise une procédure de test $\phi_\alpha = \mathbb{1}_{\sum_{i=1}^n X_i > c}$ avec c choisi de façon à ce que le niveau du test soit plus petit que α et tel que la région de rejet soit maximisée. On choisit donc $c = q^{B(n,1/2)}_{1-\alpha}$.

Si on veut tester au niveau α

$$H_0 : p = 1/2 \quad contre \quad p < 1/2.$$

On utilise une procédure de test de la forme $\phi = \mathbb{1}_{\sum_{i=1}^n X_i \leq c}$. Attention ici, la valeur $c = q_\alpha$ ne fonctionne pas (ni avec le test $\phi = \mathbb{1}_{\sum_{i=1}^n X_i \leq c}$ ni avec $\phi = \mathbb{1}_{\sum_{i=1}^n X_i < c}$). On sait en effet seulement que $\mathbf{P}(\sum_{i=1}^n X_i \leq q_\alpha) \geq \alpha$ (alors qu'on souhaite $\leq \alpha$). Dans cet exemple, on pourrait utiliser $c = -q^{-B(n,1/2)}_{1-\alpha}$.

Ce type de problème ne se pose pas pour les variables continues puisque dans ce cas on a l'égalité (et en plus le fait d'utiliser une inégalité large ou stricte n'a pas d'importance). Dans la suite, nous n'utiliserons essentiellement que des tests de la forme $\mathbb{1}_{T>c}$ ou $\mathbb{1}_{|T|>c}$, que la loi de T soit continue ou discrète.

On admet le théorème suivant. Une preuve, pour les étudiants intéressés, se trouve dans les annales de l'examen 2018.

Théorème 2.20. *Soit $(F_n)_{n\geq 0}$ une suite de fonctions de répartition sur $\mathbb{R}$ et F une fonction de répartition sur $\mathbb{R}$. Alors F_n converge vers F en tout point de continuité de F si et seulement si $F_n^{(-1)}$ converge vers $F^{(-1)}$ en tout point de continuité de $F^{(-1)}$.*

Exemple 2.21. *La loi de Student à n degrés de liberté tend vers la loi normale standard. Φ^{-1} est continue. Donc, pour tout $\alpha \in]0,1[$, $q^{T(n)}_\alpha \underset{n\to\infty}{\to} q^{N(0,1)}_\alpha$.*

On a besoin des quantiles pour les procédures de tests ainsi que pour les régions de confiance. Parfois on ne sait pas calculer les quantiles de la loi mais on sait simuler cette loi. Le quantile empirique peut alors être utilisé en remplacement du vrai quantile.

On rappelle la notation suivante pour les statistiques d'ordre :

$$X_{(1)} \leq \dots \leq X_{(n)}$$

Définition 2.22. *Le quantile empirique d'un n échantillon iid $X = (X_1, \dots, X_n)$ est défini, pour $\beta \in]0,1]$, par*

$$\hat{q}_{n,\beta} = \hat{F}_n^{(-1)}(\beta)$$

Intuition 2.2.1. *Il s'agit donc des quantiles des cdf (lois) empiriques.*

Proposition 2.23.

$$\hat{F}_n^{(-1)}(\beta) = X_{(\lceil n\beta \rceil)}$$

où on a noté $\lceil t \rceil = \min\{m \in \mathbb{N} : m \geq t\}$.

Intuition 2.2.2. *Dans la formule précédente* $X_{(\lceil n\beta \rceil)}$ *est en pratique la valeur de la* $\lceil n\beta \rceil$*-ème variable de la statistique d'ordre.*

Démonstration. On va utiliser la propriété immédiate suivante : pour tout x,

$$x \leq \lceil x \rceil < x + 1.$$

1. Il y a au moins $\lceil n\beta \rceil$ indices $i \in [n]$ tels que $X_i \leq X_{(\lceil n\beta \rceil)}$ donc

$$\hat{F}_n(X_{(\lceil n\beta \rceil)}) \geq \frac{\lceil n\beta \rceil}{n} \geq \beta. \tag{2.2}$$

2. Soit $x < X_{(\lceil n\beta \rceil)}$. Il y a au plus $\lceil n\beta \rceil - 1$ indices $i \in [n]$ tels que $X_i \leq x$ donc

$$\hat{F}_n(x) \leq \frac{\lceil n\beta \rceil - 1}{n} < \beta. \tag{2.3}$$

(2.2) et (2.3) donnent le résultat. □

Le théorème de Glivenko-Cantelli assure que $\|\hat{F}_n - F\|_\infty \to 0$ presque sûrement. On s'attend donc à ce que $\hat{q}_{n,\beta}$ soit proche de q_β quand n est grand.

Théorème 2.24. *Soit* $\beta \in]0,1[$ *tel que* $F^{(-1)}$ *est continue en* β*. Alors on a*

$$\lim_{n\to\infty} \hat{q}_{n,\beta} = q_\beta \quad p.s.$$

Démonstration. D'après le théorème de Glivenko-Cantelli, il existe un ensemble mesurable A tel que $\mathbf{P}(A) = 1$ et si $\omega \in A$, $\|\hat{F}_n(\omega, \cdot) - F(\cdot)\|_\infty \underset{n\to\infty}{\to} 0$. Soit $\omega \in A$. On a en particulier $\hat{F}_n(\omega, t) \underset{n\to\infty}{\to} F(t)$ pour tout $t \in \mathbb{R}$. Donc $\hat{F}_n^{(-1)}(\omega, t) \underset{n\to\infty}{\to} F^{(-1)}(t)$ en tout point de continuité de t de $F^{(-1)}$ d'après le théorème 2.20. □

Remarque 2.25. *Un point de continuité* β *pour* $F^{(-1)}$ *correspond à point de croissance stricte* q_β *pour* F*.*

Remarque 2.26. *On voit donc que si on ne sait pas calculer facilement le quantile d'une loi, mais si on sait simuler cette loi, on peut avoir une valeur approchée de ses quantiles en simulant un échantillon suffisamment grand et en calculant le quantile empirique. Une question associée est : quelle est la taille d'échantillon nécessaire pour avoir une précision donnée ? Le théorème suivant donne en partie une réponse à cette question. Sa preuve dépasse le cadre de ce cours donc on admettra ce théorème.*

Théorème 2.27. *Si F est dérivable en q_β avec $F'(q_\beta) > 0$ alors*

$$\sqrt{n}(\hat{q}_{n,\beta} - q_\beta) \xrightarrow{Loi} \mathcal{N}\left(0, \frac{\beta(1-\beta)}{\left(F'(q_\beta)\right)^2}\right)$$

Remarque 2.28. *Les conditions du théorème sont en particulier vérifiées si la loi F est à densité f strictement positive sur $\mathbb{R}$. Pour construire un IC pour q_β, il faut alors connaitre $f(q_\beta)$.*

2.3 Test d'ajustement à une loi ou à une famille de lois

2.3.1 Ajustement à une loi donnée

On fixe une loi de référence, de fonction de répartition F_0 et on observe un n-échantillon iid $X_1^n = (X_1, \dots, X_n)$ dont on note F la fonction de répartition commune. On veut tester

$$H_0 : F = F_0 \quad \text{contre} \quad H_1 : F \neq F_0$$

On va naturellement utiliser la statistique de test suivante

$$h_n(X_1^n, F_0) = \|\hat{F}_n - F_0\|_\infty$$

Remarque 2.29. *Il s'agit bien d'une statistique, c'est-à-dire que h_n est bien mesurable. En effet on peut montrer (grâce à la continuité à droite) que*

$$h_n(X_1^n, F_0) = \sup_{x \in \mathbb{Q}} \left|\hat{F}_n(x) - F_0(x)\right|.$$

Proposition 2.30. *On suppose F_0 et F continues. Alors*

$$h_n(X_1^n, F_0) = \max_{1 \leq j \leq n} \left\{ \max\{ \frac{j}{n} - F_0(X_{(j)}), F_0(X_{(j)}) - \frac{j-1}{n} \} \right\}$$

Démonstration. Comme F est continue, presque sûrement, tous les X_i sont distincts (cf TD). Donc $X_{(1)} < X_{(2)} < \dots < X_{(n)}$. Donc on peut décrire $\hat{F}_n$ de la manière suivante :

$$\hat{F}_n(x) = \begin{cases} 0 & \text{si } x < X_{(1)} \\ \frac{j}{n} & \text{si } x \in [X_{(j)}, X_{(j+1)}[, \ 1 \leq j \leq n-1 \\ 1 & \text{si } x \geq X_{(n)} \end{cases}$$

On va donc utiliser l'égalité suivante

$$h_n(X_1^n, F_0) = \sup_{0 \leq j \leq n} M_j$$

où, pour $1 \leq j \leq n-1$,

$$M_j = \sup_{x \in [X_{(j)}, X_{(j+1)}[} \left|\hat{F}_n(x) - F_0(x)\right|$$

et

$$M_0 = \sup_{x<X_{(1)}} \left|\hat{F}_n(x) - F_0(x)\right| \quad \text{et} \quad M_n = \sup_{x\geq X_{(n)}} \left|\hat{F}_n(x) - F_0(x)\right|$$

En utilisant la croissance de F_0 on obtient

$$M_n = \sup_{x\geq X_{(n)}} \left|1 - F_0(x)\right| = \sup_{x\geq X_{(n)}} \{1 - F_0(x)\} = 1 - F_0(X_{(n)})$$

et

$$M_0 = \sup_{x<X_{(1)}} \left|0 - F_0(x)\right| = \sup_{x<X_{(1)}} F_0(x) = F_0(X_{(1)}^-)$$

Et par la continuité de F_0,

$$M_0 = F_0(X_{(1)})$$

Considérons maintenant M_j pour $1 \leq j \leq n-1$. On a

$$M_j = \sup_{x\in[X_{(j)},X_{(j+1)}[} \left|\frac{j}{n} - F_0(x)\right|.$$

Soit f une fonction croissante et continue sur un segment $[a,b]$. On a

$$\begin{aligned}
\sup_{a\leq x<b} |f(x)| &= \sup_{a\leq x<b} \Big\{ \sup\{f(x), -f(x)\} \Big\} \\
&= \sup\Big\{ \sup_{a\leq x<b} f(x), \sup_{a\leq x<b} -f(x) \Big\} \\
&= \sup\Big\{ \sup_{a\leq x<b} f(x), -\inf_{a\leq x<b} f(x) \Big\} \\
&= \max\{(f(b), -f(a)\}
\end{aligned}$$

En appliquant cette propriété à la fonction croissante et continue $F_0 - \frac{j}{n}$, on obtient

$$M_j = \max\{F_0(X_{(j+1)}) - \frac{j}{n}, \frac{j}{n} - F_0(X_{(j)})\}.$$

En rassemblant tous les résultats on obtient finalement

$$h_n(X_1^n, F_0) = \max\left\{ \max_{1\leq j\leq n-1} \{F_0(X_{(j+1)}) - \frac{j}{n}\}, \max_{1\leq j\leq n-1} \{\frac{j}{n} - F_0(X_{(j)})\}, F_0(X_{(1)}), 1 - F_0(X_{(n)}) \right\}$$

On obtient le résultat final en remarquant que

$$\max\left\{ \max_{1\leq j\leq n-1} \{\frac{j}{n} - F_0(X_{(j)})\}, 1 - F_0(X_{(n)}) \right\} = \max_{1\leq j\leq n} \{\frac{j}{n} - F_0(X_{(j)})\}$$

et

$$\begin{aligned}
\max\left\{ \max_{1\leq j\leq n-1} \{F_0(X_{(j+1)}) - \frac{j}{n}\}, F_0(X_{(1)}) \right\} &= \max\left\{ \max_{2\leq j\leq n} \{F_0(X_{(j)}) - \frac{j-1}{n}\}, F_0(X_{(1)}) \right\} \\
&= \max_{1\leq j\leq n} \{F_0(X_{(j)}) - \frac{j-1}{n}\}
\end{aligned}$$

□

Définition 2.31. — *On dit qu'une variable Z est **diffuse** si sa cdf est continue.*
— *"La statistique $h_n(X_1^n, F_0)$ est **libre** de F_0" signifie que sa loi ne dépend pas de F_0.*

Nous faisons maintenant deux remarques utiles pour la preuve de la proposition suivante.

Remarque 2.32. *Si $F : \mathbb{R} \to [0,1]$ est une fonction de répartition alors*

$$F \text{ continue } \Leftrightarrow]0,1[\subset F(\mathbb{R})$$

En effet
— *Si F est continue alors on peut appliquer le théorème des valeurs intermédiaires.*
— *Si F n'est pas continue, alors il y a au moins un saut en un certain $x \in \mathbb{R}$, alors les valeurs entres $F(x)$ et $F(x^-)$ ne sont pas prises par F.*

Remarque 2.33. *Si $Z = \max_{j=1,\dots,k} X_j$ avec des variables X_j diffuses, alors Z est diffuse. En effet, pour tout x,*

$$\mathbf{P}(Z = x) \leq \mathbf{P}(\cup_{j=1}^k \{X_j = x\}) \leq \sum_{j=1}^k \mathbf{P}(X_j = x) = 0$$

Proposition 2.34. *Sous H_0, si F_0 est continue alors $h_n(X_1^n, F_0)$ est une statistique libre de F_0 et de loi continue.*

Démonstration. Soit $U_1^n = (U_1, \dots, U_n)$ est un n-échantillon iid de loi uniforme sur $]0,1[$ Sous H_0, comme $X_i \overset{iid}{\sim} F_0$, on a $(X_1, \dots, X_n) \sim (F_0^{(-1)}(U_1), \dots, F_0^{(-1)}(U_n))$ d'après la proposition 2.7. On a donc aussi, sous H_0,

$$h_n(X_1^n, F_0) \sim \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{F_0^{(-1)}(U_i) \leq x} - F_0(x) \right|$$

En utilisant l'item 1 de la proposition 2.7, on obtient

$$\begin{aligned} \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{F_0^{(-1)}(U_i) \leq x} - F_0(x) \right| &= \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq F_0(x)} - F_0(x) \right| \\ &= \sup_{s \in Im(F_0)} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq s} - s \right| \end{aligned}$$

En utilisant la remarque 2.32, ceci donne, presque sûrement,

$$\sup_{s \in Im(F_0)} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq s} - s \right| = \sup_{s \in]0,1[} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq s} - s \right|.$$

En effet on a $]0,1[\subset Im(F_0) \subset [0,1]$ et la valeur de la fonction $s \mapsto \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq s} - s$ en $s = 0$ et en $s = 1$ est égale à 0 presque sûrement. On a donc obtenu que, sous H_0

$$h_n(X_1^n, F_0) \sim h_n(U_1^n, G)$$

où on a noté G la fonction de répartition de la loi uniforme sur $[0,1]$, i.e. la fonction définie par $G(s) = s$ pour $s \in [0,1]$. Cela montre que la loi, sous H_0, de $h_n(X_1^n, F_0)$ est libre de F_0.

On prouve maintenant que la loi de $h_n(U_1^n, G)$ est continue. D'après la proposition 2.30, comme G est continue,

$$h_n(U_1^n, G) = \max_{1\le j\le n}\left\{\max\{\frac{j}{n} - U_{(j)}, U_{(j)} - \frac{j-1}{n}\}\right\}$$

Comme la loi des U_j est absolument continue, celle de $U_{(j)}$ aussi (fait en TD1, on a même ici $U_{(j)} \sim Beta(j, n-j+1)$, toujours d'après le TD1). Donc, d'après la remarque 2.33, $h_n(U_1^n, G)$ est bien de loi continue. □

Exemple 2.35. *Soit* $X_1, \dots, X_n \overset{iid}{\sim} N(0,1)$ *et* $Y_1 \dots, Y_n \overset{iid}{\sim} exp(1)$ *alors*

$$\sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\le x} - \Phi(x)\right| \sim \sup_{x>0}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i\le x} - (1-\exp(-x))\right| \sim \sup_{s\in]0,1[}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i\le s} - s\right|$$

Pour tout $\alpha \in]0,1[$, si on note $\xi_{n,\alpha}$ le quantile d'ordre α de la loi de la statistique $h_n(U_1^n, G)$, on a donc, par continuité de cette loi,

$$\mathbf{P}_{X_i\overset{iid}{\sim}F_0}(h_n(X_1^n, F_0) \le \xi_{n,\alpha}) = \alpha.$$

Pour les petites valeurs de n, on a tabulé les quantiles de cette statistique.
On en déduit une bande de confiance de niveau $1-\alpha$ en posant

$$\begin{aligned} B(n,\alpha) &= \{ \text{ fonctions de répartitions } G : \forall x \in \mathbb{R} \quad \hat{F}_n(x) - \xi_{n,1-\alpha} \le G(x) \le \hat{F}_n(x) + \xi_{n,1-\alpha}\} \\ &= \{G : h_n(X_1^n, G) \le \xi_{n,1-\alpha}\} \end{aligned}$$

Pour tester $H_0 : F = F_0$ contre $H_1 : F \neq F_0$, on pose

$$\phi_\alpha(X_1^n) = \mathbb{1}_{h_n(X_1^n, F_0)\ge \xi_{n,1-\alpha}}$$

On a donc obtenu le résultat suivant

Théorème 2.36. *(Test de Kolmogorov) Soit* $(X_1, \dots, X_n)$ *un n-échantillon iid de fonction de répartition* F*. Le test* $\phi_\alpha(X_1^n)$ *est de taille* α *pour tester* $H_0 : F = F_0$ *contre* $H_1 : F \neq F_0$ *quand* F_0 *est continue.*

Remarque 2.37. *Quand* F_0 *n'est pas continue, le test n'est plus de taille* α *mais il reste de niveau* α*. En effet on a, d'après la preuve de la Proposition 2.34,*

$$h_n(X_1^n, F_0) \sim \sup_{s\in Im(F_0)}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i\le s} - s\right|$$

(En effet, la continuité de F_0 *n'est pas nécessaire pour obtenir cette égalité en loi). Et comme*

$$\sup_{s\in Im(F_0)}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i\le s} - s\right| \le \sup_{s\in]0,1[}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i\le s} - s\right|,$$

on a

$$\mathbf{P}_{X_1,\dots,X_n\overset{iid}{\sim}F_0}\Big(h_n(X_1^n, F_0) \ge \xi_{n,1-\alpha}\Big) \le \mathbf{P}\Big(h_n(U_1^n, G) \ge \xi_{n,1-\alpha}\Big) = \alpha.$$

Remarque 2.38. *Quand le nombre de données n est grand, on utilise un test asymptotique.*

On a aussi l'inégalité de Dvoretzky-Kiefer-Wolfowitz, et qui est valable sans condition sur F_0. Sous $H_0 : X_1, \dots X_n \overset{iid}{\sim} F_0$,

$$\mathbf{P}(\|\hat{F}_n - F_0\|_\infty > \epsilon) \leq 2e^{-2n\epsilon^2} \text{ pour tout } n \in \mathbb{N} \text{ et tout } \epsilon > 0$$

On termine en donnant les preuves du théorème de Glivenko-Cantelli. On aura besoin du résultat d'analyse suivant, que l'on admet (niveau L1, cf par ex wikipedia).

Théorème 2.39. *(2ème théorème de Dini) Soit $(f_n)_{n\geq 0}$ une suite de fonctions croissantes sur un segment $[a, b]$ dans $\mathbb{R}$, qui converge simplement vers une fonction continue f. Alors $(f_n)_{n\geq 0}$ converge uniformément vers f sur $[a, b]$.*

La propriété suivante sera aussi nécessaire à la preuve :

$$\text{Si } (X_n)_{n\geq 0} \sim (Y_n)_{n\geq 0} \text{ alors } X_n \text{ converge p.s. vers } 0 \Leftrightarrow Y_n \text{ converge p.s. vers } 0. \tag{2.4}$$

On a en effet $\{X_n \overset{p.s.}{\underset{n\to\infty}{\to}} 0\} \Leftrightarrow \mathbf{P}\big(\lim X_n = 0\big) = 1 \Leftrightarrow \mathbf{P}(\cap_{\epsilon\in\mathbb{Q}} \cup_{p\in\mathbb{N}} \cap_{n\geq p} \{X_n \leq \epsilon\}\big) = 1.$

La propriété aurait été fausse avec seulement $X_n \sim Y_n$ pour tout n. Contre exemple : $X_i \overset{iid}{\sim} N(0, 1)$ et $Y_i \equiv Y \sim N(0, 1)$.

<u>Preuve du théorème de Glivenko-Cantelli</u> :

Presque sûrement, $\hat{F}_n$ converge simplement vers F d'après la proposition 2.4. On est donc tenté d'utiliser le 2ème théorème de Dini pour obtenir la convergence uniforme. Plusieurs problèmes se posent alors :

- La fonction F n'est pas forcément continue.
- La convergence n'a pas lieu sur un segment.
- La convergence presque sûre se traduit ici par : il existe un ensemble $A(x)$ de probabilité 1 tel que, pour $\omega \in A(x)$, $\hat{F}_n(\omega, x) \underset{n\to\infty}{\to} F(x)$. Autrement dit l'ensemble sur lequel la convergence a lieu dépend de x.

Pour régler les deux premiers problèmes, on va à nouveau se ramener à des variables uniformes sur $[0, 1]$. En effet, on a, pour tout n, d'après la preuve du théorème de Kolmogorov-Smirnov,

$$\sup_{t\in\mathbb{R}} \Big|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i\leq t} - F(t)\Big| \sim \sup_{s\in Im(F)} \Big|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq s} - s\Big|$$

Posons $V_n = \sup_{t\in\mathbb{R}} \Big|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i\leq t} - F(t)\Big|$, et $W_n = \sup_{s\in Im(F)} \Big|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq s} - s\Big|$. On a même

$$(V_n)_{n\geq 0} \sim (W_n)_{n\geq 0}$$

Donc, d'après la propriété 2.4, pour prouver que V_n converge presque sûrement vers 0, il suffit de prouver que W_n converge presque sûrement vers 0. Pour cela, il suffit de prouver que $\sup_{s\in[0,1]} \Big|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq s} - s\Big|$ converge p.s. vers 0 car

$$\sup_{s\in Im(F)} \Big|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq s} - s\Big| \leq \sup_{s\in[0,1]} \Big|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i\leq s} - s\Big|.$$

On est alors ramené à prouver le résultat pour des variables uniformes sur $[0,1]$, pour lesquelles les deux premiers problèmes ne se posent pas, puisque la cdf est ici $G(x) = x$, définie sur le segment $[0,1]$, et continue. Il reste donc à prouver que, presque sûrement, $\hat{G}_n$ converge uniformément vers G. Pour cela, il reste à régler le dernier problème. Il suffirait, pour conclure à l'aide de Dini, de prouver qu'il existe un ensemble mesurable A de probabilité 1 tel que, si $\omega \in A$, alors **pour tout** x, $\hat{G}_n(\omega, x)$ tend vers $G(x)$. On pose $A = \cap_{q \in \mathbb{Q} \cap [0,1]} A(q)$. Si $\omega \in A$ on a donc

$$\hat{G}_n(\omega, q) \underset{n\to\infty}{\to} G(q), \quad \forall q \in \mathbb{Q} \cap [0,1]$$

De plus, comme $\mathbb{Q}$ est dénombrable, $\mathbf{P}(A) = 1$. Il reste à prouver que, pour tout $s \in [0,1]$, on a aussi $\hat{G}_n(\omega, s) \underset{n\to\infty}{\to} G(s)$, si $\omega \in A$. Soit donc $\omega \in A$, $s \in [0,1]$ et $\epsilon > 0$. Par densité de $\mathbb{Q}$ dans $\mathbb{R}$, Il existe des rationnels q_1 et q_2 tels que $s - \epsilon \leq q_1 \leq s \leq q_2 \leq s + \epsilon$. Par croissance de $\hat{G}_n$ on a

$$\hat{G}_n(q_1) \leq \hat{G}_n(s) \leq \hat{G}_n(q_2)$$

Donc en passant à la limite sup et la limite inf (attention à ce stade on ne sait pas encore que $\hat{G}_n(\omega, s)$ converge, donc on doit utiliser les limites supérieures et inférieures qui, elles, existent toujours), on obtient

$$s - \epsilon \leq q_1 = G(q_1) \leq \liminf_{n\to\infty} \hat{G}_n(\omega, s) \leq \limsup_{n\to\infty} \hat{G}_n(\omega, s) \leq G(q_2) = q_2 \leq s + \epsilon.$$

Ces inégalités étant vraies pour tout $\epsilon > 0$, on a bien $\lim_{n\to\infty} \hat{G}_n(\omega, s) = s$ et la preuve est terminée.

2.3.2 Ajustement à une famille paramétrique de lois : le cas des familles exponentielles

Soit $(X_1, \ldots, X_n)$ un n- échantillon iid de variables positives de fonction de répartition F. On veut tester si la loi des X_i est exponentielle, c'est-à-dire on veut tester s'il existe un $\lambda > 0$ tel que $F = F_\lambda$ avec $F_\lambda(x) = (1 - e^{-\lambda x})\mathbb{1}_{\mathbb{R}^+}(x)$ pour tout $x \in \mathbb{R}$. Cette hypothèse correspond à H_0.

Sous H_0 on va estimer le paramètre λ. L'EMV est $\hat{\lambda} = \frac{1}{\overline{X}}$. On considère alors la statistique $h'_n(X_1^n) = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_{\hat{\lambda}}(x)|$.

Proposition 2.40. *Sous H_0, la loi de $h'_n(X_1^n)$ est libre du paramètre λ. De plus cette loi est continue.*

Démonstration. On se place sous H_0. On pose $Y_i = \lambda X_i$, pour $1 \leq i \leq n$. Alors $Y_1, \ldots Y_n \overset{iid}{\sim} exp(1)$.

$$\begin{aligned} h'_n(X_1^n) &= \sup_{x>0} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x} - (1 - e^{-\frac{x}{\overline{X}}}) \right| \\ &= \sup_{x>0} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\frac{Y_i}{\lambda} \leq x} - (1 - e^{-\frac{\lambda x}{\overline{Y}}}) \right| \\ &= \sup_{t>0} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{Y_i \leq t} - (1 - e^{-\frac{t}{\overline{Y}}}) \right| \end{aligned}$$

La statistique $\sup_{t>0} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i\leq t} - (1-e^{-\frac{t}{\bar{Y}}})\right|$ a une loi indépendante de λ.
On admet la continuité de cette loi. □

On en déduit un test de taille α en posant

$$\phi_\alpha(X_1^n) = \mathbb{1}_{h'_n(X_1^n)\geq q_{n,1-\alpha}}$$

où $q_{n,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de $\sup_{t\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i\leq t} - (1-e^{-\frac{t}{\bar{Y}_n}})\right|$ avec $Y_1,\dots,Y_n \overset{iid}{\sim} exp(1)$.

Remarque 2.41. *On peut aussi faire le même style de test avec un certain nombre de familles de lois (cf TD pour un exemple avec les lois normales).*

2.4 Test d'homogénéité de Kolmogorov Smirnov

On observe deux échantillons iid $X_1^n = (X_1,\dots,X_n)$ et $Y_1^m = (Y_1,\dots,Y_m)$, indépendants entre eux, avec m qui peut être différent de n. On veut tester si les deux échantillons ont la même loi. Autrement dit, si on note F la cdf des X_i et G la cdf des Y_i, on veut tester

$$H_0 : F = G \quad \text{contre} \quad H_1 : F \neq G.$$

On note comme précédemment $\hat{F}_n$ et $\hat{G}_m$ les fonctions de répartitions empiriques respectives des échantillons $X_1^n = (X_1,\dots,X_n)$ et $Y_1^m = (Y_1,\dots,Y_m)$ et on pose

$$h_{n,m}(X_1^n, Y_1^m) = \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|$$

Proposition 2.42. *Sous $H_0 : F = G$ et si F est continue alors la loi de $h_{n,m}(X_1^n, Y_1^m)$ ne dépend pas de F.*

Démonstration. Sous H_0, $X_1,\dots,X_n,Y_1,\dots,Y_m \overset{iid}{\sim} F$ donc, si $U_1,\dots,U_n,V_1,\dots,V_m \overset{iid}{\sim} U[0,1]$, on a

$$\left(F^{(-1)}(U_1),\dots,F^{(1)}(U_n),F^{(-1)}(V_1),\dots,F^{(-1)}(V_m)\right) \sim (X_1,\dots,X_n,Y_1,\dots,Y_m)$$

Ainsi on obtient, sous H_0,

$$\begin{aligned}
h_{n,m}(X_1^n, Y_1^m) &= \sup_{t\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\leq t} - \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{Y_i\leq t}\right| \\
&\sim \sup_{t\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{F^{(-1)}(U_i)\leq t} - \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{F^{(-1)}(V_i)\leq t}\right| \\
&= \sup_{t\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i\leq F(t)} - \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{V_i\leq F(t)}\right| \\
&= \sup_{s\in Im(F)} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i\leq s} - \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{V_i\leq s}\right| \\
&\overset{p.s.}{=} \sup_{s\in]0,1[} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i\leq s} - \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{V_i\leq s}\right|
\end{aligned}$$

On a utilisé la proposition 2.7 ainsi que la continuité de F. En effet, si F est continue, $]0,1[\subset Im(F) \subset [0,1]$ et on vérifie immédiatement que, presque sûrement, la fonction $s \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{U_i \leq s} - \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{V_i \leq s}$ vaut 0 en $s=0$ et $s=1$. □

Cette loi est tabulée. On pose

$$\phi_\alpha(X_1^n, Y_1^m) = \mathbb{1}_{h_{n,m}(X_1^n, Y_1^m) > x_{n,m,1-\alpha}}$$

où $x_{n,m,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de la statistique $h_{n,m}(X_1^n, Y_1^m)$ sous H_0.

Remarque 2.43. *Le problème que nous venons de traiter concerne l'ajustement d'une distribution inconnue à une distribution théorique. Il existe un autre test pour cela et qui est encore plus connu : le test du chi-deux. Voici les différences essentielles entre le test de Kolmogorov-Smirnov et le test du chi-deux :*

- *Le test du chi-deux est plus adapté aux lois discrètes. Si on veut l'utiliser pour des lois continues, c'est possible, mais il faut discrétiser en choisissant des classes (quelles classes ? combien de classes ?).*
- *Le test de Kolmogorov-Smirnov a la particularité d'être exact pour de petits échantillons : la loi est libre à n fini. Le test du chi-deux est uniquement asymptotique (basé sur le TCL). Donc pour des échantillons de petite taille, on préférera le test de Kolmogorov-Smirnov.*

Remarque 2.44. *De façon similaire, il existe un test d'indépendance adapté à des variables ne prenant qu'un nombre fini de valeurs (des facteurs). Par exemple : tester l'indépendance entre le fait qu'une mère a fumé pendant sa grossesse et le fait que le bébé a une malformation a la naissance. On utilise le test d'indépendance du chi-deux (vu en TD en L3 et également dans le cours d'analyse de données de M1).*

Avec R : Pour tester si deux échantillons x et y ont la même loi, on peut utiliser `ks.test` du package `stat`. La formule est `ks.test(x,y)`.

Illustrons maintenant les sections précédentes.

Si on veut vérifier qu'un échantillon `x` suit bien une loi gaussienne de moyenne 3 et d'écart-type 2 :

```
ks.test(x,"pnorm",3,2)
```

Si on veut vérifier qu'un échantillon `x` suive bien une loi gamma avec 3 comme paramètre de forme et 2 pour le taux :

```
ks.test(x,"pgamma",3,2)
```

Attention, la fonction `ks.test` se comporte mal en cas d'ex æquo (dans le cas du test d'égalité des lois de deux échantillons, il ne faut pas avoir un ex æquo de type $x_i = y_j$). Normalement, en théorie, on ne peut avoir deux valeurs identiques si la loi sous-jacente est continue. Mais dans la pratique, on peut avoir des mesures pas assez précises qui donnent donc un échantillon avec des ex aequo. Voyons ce qui se passe avec la fonction `ks.test` de R sur un exemple numérique présentant des ex

aequo (tiré des documents pédagogiques de F-G Carpentier, cf biblio). L'échantillon se nomme x.

```
> x= c(8.43, 8.70, 11.27, 12.92, 13.05, 13.05, 13.17, 13.44, 13.89,
18.90)
> ks.test(x,"pnorm",mean=13, sd=3)

One-sample Kolmogorov-Smirnov testdata: x
D = 0.2834, p-value = 0.3982
alternative hypothesis: two-sided
Warning message:
cannot compute correct p-values with ties in: ks.test(x, "pnorm", mean
= 13, sd = 3)
```

On peut éviter le message d'avertissement concernant les ex aequo en modifiant légèrement l'une des valeurs 13.05 :

```
> x <- c(8.43, 8.70, 11.27, 12.92, 13.05, 13.050001, 13.17, 13.44, 13.89,
18.90)
> ks.test(X,"pnorm",mean=13, sd=3)

One-sample Kolmogorov-Smirnov test
data: x
D = 0.2834, p-value = 0.3326
alternative hypothesis: two-sided
```

On observe effectivement une valeur du niveau de significativité assez différent du précédent.

Dans le cas où on veut tester qu'un échantillon x suit bien une loi normale, sans préciser la moyenne ou la variance (cf section "ajustement à une famille paramétrique de lois", cas des familles normales, fait en TD), on peut utiliser la fonction `lillie.test` du package `nortest`.

```
> library(nortest)
> lillie.test(x)
Lilliefors (Kolmogorov-Smirnov) normality test
data: x
D = 0.2451, p-value = 0.0903
```

Le test du chi-deux peut se faire à l'aide de la procédure `chisq.test`. Par exemple, si on veut tester qu'un échantillon x est à loi discrète à valeurs dans $\{1, \dots, m\}$ représentée par le vecteur de probabilités prob$= (p_1, \dots, p_m)$, on peut utiliser `chisq.test(table(x), p=prob)`.

Remarque 2.45. *Pour les étudiants qui préfèrent Python à R, on peut appeler les commandes R depuis Python. Par exemple on peut faire*

```
from rpy2 import robjects
rks=robjects.r('ks.test')
```

Ensuite on utilise normalement la fonction qu'on a appelée `rks`*, en prenant garde de transformer aussi les entrées. Par exemple si on a un échantillon dans le vecteur* `x`*, et si on veut vérifier qu'il s'agit d'un échantillon gaussien standard :*

```
y=robjects.FloatVector(x)
z=rks(y,"pnorm")
```

On peut aussi utiliser directement les fonctions natives `stats.kstest`*, qui se comportent différemment en cas d'ex aequo.*

2.5 Exercices

Exercice 2.1. *Soit* $X_1, \dots, X_n \overset{iid}{\sim} Be(p)$. *On veut tester*

$$H_0 : p = 1/2 \quad \textit{contre} \quad p \neq 1/2.$$

1. *Proposer une procédure de test.*
2. *Donner l'expression de la p-valeur.*

Exercice 2.2. *On considère un n-échantillon i.i.d.* $X_1^n = (X_1, \dots, X_n)$. *On note* F *la fonction de répartition et* $\hat{F}_n$ *la fonction de répartition empirique associées à cet échantillon. On se donne* F_0 *une fonction de répartition.*

1. *Montrer que si* F_0 *est continue la loi, sous* H_0, *de la statistique*

$$h_n^+(X_1^n, F_0) = \sup_{t \in \mathcal{R}} \left\{ \hat{F}_n(t) - F_0(t) \right\}_+$$

   est libre de F_0.

2. *Proposer une procédure de test de*

$$H_0 : F = F_0 \quad \textit{contre} \quad H_1 : \exists\, t \in \mathcal{R}\ F(t) > F_0(t).$$

Exercice 2.3. *On s'intéresse dans cet exercice à la puissance du test de Kolmogorov-Smirnov. On considère donc un échantillon i.i.d.* $(X_1, \dots, X_n)$ *de loi de cdf* F *et de cdf empirique* $\hat{F}_n$. *On veut tester*

$$H_0 : F = F_0 \quad \textit{contre} \quad H_1 : F \neq F_0$$

où F_0 *est une loi donnée. On veut savoir si le test est capable de nous dire, avec une grande probabilité, que l'échantillon ne suit pas la loi* F_0, *quand c'est bien le cas, et du moment que la taille de l'échantillon est suffisamment grande. Autrement dit, on veut savoir si le test est puissant.*

1. *A l'aide de l'inégalité DKW vue en cours, montrer que le quantile* $\xi_{n,1-\alpha}$ *d'ordre* $1-\alpha$ *de la statistique de Kolmogorov-Smirnov, vérifie* $\xi_{n,1-\alpha} = O(\frac{1}{\sqrt{n}})$ *quand* $n \to \infty$.
2. *On suppose que* $F \neq F_0$, *c'est-à-dire que l'échantillon ne suit pas la loi* F_0. *Montrer que si on pose*

$$\underline{\beta}(F) = \mathbb{P}_{X_1,\dots,X_n \overset{iid}{\sim} F}\big(\|\hat{F}_n - F_0\|_\infty \geq \xi_{n,1-\alpha}\big)$$

   alors

$$\underline{\beta}(F) \underset{n\to\infty}{\to} 1$$

Exercice 2.4. *On considère un n-échantillon i.i.d.* $X_1^n = (X_1, \dots, X_n)$ *de variables aléatoires. On note* F *la fonction de répartition et* $\hat{F}_n$ *la fonction de répartition empirique associées à cet échantillon. Si les variables* X_i *sont de lois normales de paramètres* μ *et* σ^2, *on note également* N_{μ,σ^2} *leur fonction de répartition commune.*

1. *On suppose que* $F = N_{\mu,\sigma^2}$. *Déterminer l'estimateur du maximum de vraisemblance* $(\hat{\mu}, \hat{\sigma}^2)$ *de* (μ, σ^2).
2. *On pose*

$$\Delta_n = \sup_{t \in \mathcal{R}} |\hat{F}_n(t) - N_{\hat{\mu},\hat{\sigma}^2}(t)|.$$

   Montrer que si $F = N_{\mu,\sigma^2}$, *alors la loi de* Δ_n *ne dépend pas de* μ *et* σ^2.

3. *En déduire un test d'appartenance à la famille des lois normales, c'est-à-dire un test de*

$$H_0 : F \in \mathcal{F}N \quad \textit{contre} \quad H_1 : F \notin \mathcal{F}N,$$

   où

$$\mathcal{F}N = \left\{ G : \quad \exists\, (\mu, \sigma^2) \in (\mathcal{R} \times \mathcal{R}_+^*) \ \textit{ tel que } \ G = N_{\mu,\sigma^2} \right\}.$$

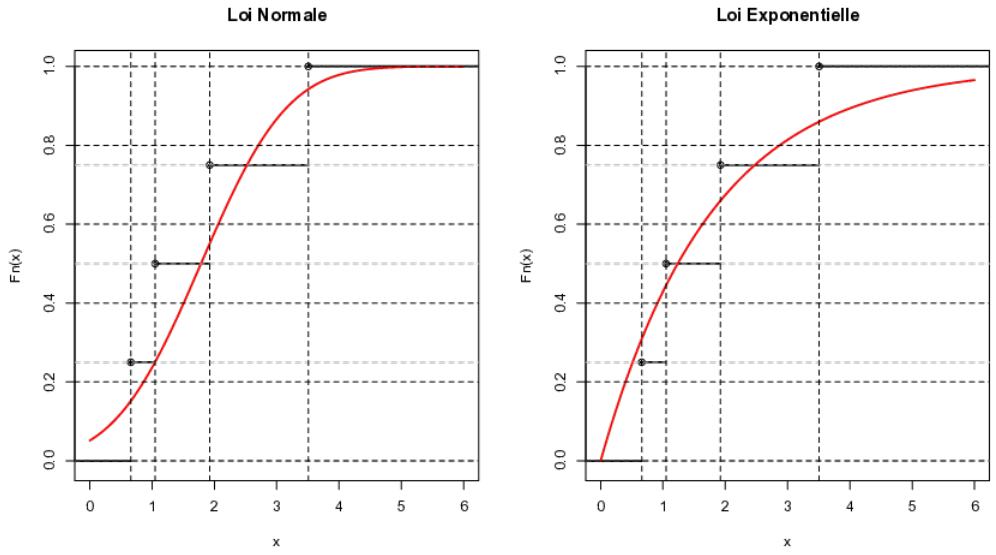


FIGURE 2.2 – Exercice 3 : fonction de répartition empirique de l'échantillon (en escalier) et fonction de répartition à tester.

4. Application (quasi indépendante du reste de l'exercice) *: La loi de la statistique de test de la question 3 a été tabulée. On s'intéresse aussi au test, vu en cours (section 2.3.2 du poly), d'appartenance à la famille exponentielle. On fournit ci-dessous quelques quantiles intéressants pour* $n = 4$ *:*

	$q_{5\%}$	$q_{10\%}$	$q_{90\%}$	$q_{95\%}$
Stat du test d'appartenance à la loi normale	*0.18*	*0.20*	*0.36*	*0.39*
Stat du test d'appartenance à la loi expo	*0.21*	*0.23*	*0.44*	*0.48*

Considérons la réalisation d'un échantillon de taille $n = 4$ *:*

$$0.66 \quad 3.51 \quad 1.92 \quad 1.05$$

Nous cherchons à tester si cet échantillon est distribué selon une loi normale et s'il est distribué selon une loi exponentielle. Pour cela nous proposons d'appliquer le test précédemment construit et le test du cours. Sur la figure 2.2 (à droite et à gauche) nous avons tracé la fonction de répartition empirique correspondant à l'échantillon donné. D'autre part, à gauche nous avons tracé la fonction $N_{\widehat{\mu},\widehat{\sigma^2}}$ *où* $\widehat{\mu}$ *et* $\widehat{\sigma^2}$ *sont les estimateurs du maximum de vraisemblance (*$\widehat{\mu} = 1.78 \quad \widehat{\sigma^2} = 1.20$*). A droite nous avons tracé la fonction de répartition de la loi exponentielle de paramètre* $\widehat{\lambda}$ *(*$\widehat{\lambda} = 0.56$*).*

   (a) *Par une lecture graphique sur la figure 2.2, donner la valeur de la statistique des 2 tests.*

   (b) *En utilisant les quantiles donnés ci-dessus, effectuer les 2 tests pour un niveau* 5%*.*

   (c) *Les deux conclusions vous semblent-elles cohérentes ?*

Exercice 2.5. *L'objectif de cet exercice est d'étudier la performance du test de Student à un seul échantillon quand il est effectué sur un échantillon non gaussien.*
On suppose que l'on dispose d'un échantillon iid $(X_1, \dots, X_n)$ *tel que* $\mathbb{E}X_1^2 < \infty$*. On note* σ^2 *la variance de* X_1 *et* $\mu = \mathbb{E}X_1$*. On veut tester* $H_0 : \mu = 0$ *contre* $H_1 : \mu > 0$ *au niveau* α *pour* $\alpha \in (0, 1)$*.*
On appelle Φ *le test de Student. On a donc* $\Phi = \mathbb{1}_{T_n > q_{1-\alpha}^{T(n-1)}}$ *où* $q_{1-\alpha}^{T(n-1)}$ *est le quantile d'ordre 1-*α *de la loi de Student à* $n-1$ *degrés de liberté et* $T_n = \frac{\sqrt{n}\bar{X}}{\hat{\sigma}}$ *avec* $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$*.*

1. *Montrer que, sous H_0, T_n tend en loi vers la loi normale standard.*
2. *Montrer que l'erreur de première espèce du test de Student appliqué à l'échantillon $(X_1, \dots, X_n)$ tend vers α quand n tend vers l'infini.*

   Pour cela, on admettra le résultat suivant (qui est une généralisation du 2ème théorème de Dini) : Si $(F_n)_{n\geq 0}$ et F des fonctions de répartition, si F est continue et si F_n converge simplement vers F alors la convergence est uniforme.
3. *Montrer que la puissance du test tend simplement vers 1 quand n tend vers l'infini.*

Exercice 2.6. *L'objectif de cet exercice est de proposer une procédure de tests multiples lorsque le nombre d'hypothèses à tester est élevé. On considère dans tout l'exercice $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$ un modèle statistique.*

Partie A. *On se place tout d'abord dans le cadre simple où on veut tester*

$$H_0 : \theta = \theta_0 \quad \textit{contre} \quad H_1 : \theta \in \Theta_1$$

où $\theta_0 \notin \Theta_1$. Pour cela, on dispose d'une observation réelle X de loi $\mathbb{P}_\theta$. Pour $\alpha \in]0,1[$ donné , on considère un test de H_0 contre H_1 de la forme $\phi_\alpha(X) = 1_{X\geq k_\alpha}$ où $k_\alpha \in \mathcal{R}$. On note F_θ la cdf de X sous $\mathbb{P}_\theta$. On suppose que F_θ est continue.

1. *Montrer que la p-valeur observée de ce test s'écrit pour tout $x \in \mathcal{R}$:*

$$p(x) = \mathbb{P}_{\theta_0}(X \geq x). \tag{2.5}$$

2. *Quelle est la loi sous H_0 de la p-valeur $p(X)$?*
3. *Montrer que ϕ peur s'écrire $\phi_\alpha(X) = \mathbf{1}_{p(X)\leq\alpha}$.*

Partie B. *Dans cette partie, indépendante de la partie A, pour $m \in \mathbb{N}^*$, on considère $2m$ sous-ensembles de Θ notés $\Theta_{01}, \Theta_{11}, \Theta_{02}, \Theta_{12}, \dots, \Theta_{0m}, \Theta_{1m}$ avec pour tout $i \in \{1, \dots, m\}$*

$$\Theta_{0i} \cap \Theta_{1i} = \emptyset$$

*et on veut réaliser **simultanément** m tests*

$$H_{0i} : \theta \in \Theta_{0i} \quad \textit{contre} \quad H_{1i} : \theta \in \Theta_{1i}, \qquad i = 1, \dots, m.$$

On suppose pour simplifier que les hypothèses nulles sont des singletons $\Theta_{0i} = \{\theta_{0i}\}$. On note I_0 l'ensemble des indices i pour lesquels H_{0i} est vraie :

$$I_0 = \{i \in \{1, \dots, m\} : \quad H_{0i} \textit{ est vraie}\}.$$

On cherche à construire une procédure de tests multiples qui retourne un ensemble $\hat{R} \subset \{1, \dots, m\}$ correspondant aux indices i pour lesquels H_{0i} est rejetée. On note FP le cardinal de l'ensemble des indices correspondant aux hypothèses nulles rejetées à tort et TP le cardinal de l'ensemble des indices correspondant aux hypothèses nulles rejetées à raison :

$$FP = \textit{card}(\hat{R} \cap I_0), \quad TP = \textit{card}(\hat{R} \setminus I_0).$$

FP est le cardinal des faux positifs et TP celui des vrais positifs. Idéalement, on cherche une procédure de tests de sorte que FP soit petit et TP soit grand. On note $\hat{p}_i$ la p-valeur du test de H_{0i} contre H_{1i}. Donc $\hat{p}_i$ est une statistique satisfaisant, pour tout $u \in]0,1[$,

$$\mathbb{P}_{\theta_{0i}}(\hat{p}_i \leq u) = u. \tag{2.6}$$

1. *On propose tout d'abord la procédure de Bonferroni qui permet le contrôle de FP en posant pour $\alpha \in]0,1[$:*

$$\hat{R} = \left\{ i \in \{1, \dots, m\} : \quad \hat{p}_i \leq \frac{\alpha}{m} \right\}.$$

   (a) *Montrer que*

$$\mathbb{P}(FP > 0) \leq \sum_{i \in I_0} \mathbb{P}_{\theta_{0i}}(\hat{p}_i \leq \alpha/m).$$

(b) En utilisant (2.6)*, en déduire que*

$$\mathbb{P}(FP > 0) \leq \alpha.$$

2. *La procédure de Bonferroni contrôle le nombre de faux positifs mais peut produire un trop petit nombre de vrais positifs. On dit que c'est une procédure trop conservative. Aussi, on propose l'alternative suivante. On se donne une fonction* $f : \{0, 1, \ldots, m\} \to [0, m]$ *supposée croissante et on ordonne les statistiques* $\hat{p}_i$ *par ordre croissant :*

$$\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \cdots \leq \hat{p}_{(m)}.$$

On cherche à contrôler le rapport FDR défini par

$$FDR = \mathbb{E}\left[\frac{FP}{FP + TP} 1_{\{FP+TP\}\geq 1}\right]$$

avec la convention $0/0 = 0$.
On pose pour $\alpha \in]0, 1[$,

$$\hat{R} = \left\{ i \in \{1, \ldots, m\} : \quad \hat{p}_i \leq \frac{\alpha f(\hat{k})}{m} \right\}$$

avec

$$\hat{k} = \max\left\{ k \in \{1, \ldots, m\} : \quad \hat{p}_{(k)} \leq \frac{\alpha f(k)}{m} \right\}.$$

En particulier, on pose $\hat{k} = 0$ *et* $\hat{R} = \emptyset$ *si pour tout entier* k, $\hat{p}_{(k)} > \frac{\alpha f(k)}{m}$.

(a) Montrer que

$$\hat{k} = card(\hat{R})$$

et que pour $j \geq \hat{k}$,

$$f(\hat{k}) \leq f(\min(j, m)).$$

(b) Établir alors que

$$FDR = \sum_{i \in I_0} \mathbb{E}\left[1_{\{\hat{p}_i \leq \alpha f(\hat{k})/m\}} \times \frac{1_{\{\hat{k}\geq 1\}}}{\hat{k}}\right].$$

(c) Montrer que si $\hat{k} \geq 1$,

$$\frac{1}{\hat{k}} = \sum_{j=1}^{+\infty} \frac{1_{\{j \geq \hat{k}\}}}{j(j+1)}.$$

(d) En déduire finalement que

$$FDR \leq \frac{\alpha\, card(I_0)}{m} \sum_{j=1}^{+\infty} \frac{f(\min(j, m))}{j(j+1)}.$$

(e) Conclure que si f *satisfait*

$$\sum_{j=1}^{+\infty} \frac{f(\min(j, m))}{j(j+1)} \leq 1 \tag{2.7}$$

alors

$$FDR \leq \alpha.$$

(f) Donner un exemple de fonction f *satisfaisant* (2.7).

Remarque : on peut aussi généraliser ces résultats au cas d'hypothèses nulles composites (cf examen 2014 ou partiel 2018).

Remarque : si les m tests sont indépendants, on peut en fait prendre f égale à l'identité : alors on a plus de vrais positifs que dans le cas précédent tout en ayant quand même un FDR borné par α. La procédure est alors la suivante :

$$\hat{R} = \left\{ i \in \{1, \dots, m\} : \quad \hat{p}_i \leq \frac{\alpha \hat{k}}{m} \right\}$$

avec

$$\hat{k} = \max \left\{ k \in \{1, \dots, m\} : \quad \hat{p}_{(k)} \leq \frac{\alpha k}{m} \right\}.$$

Elle s'appelle la procédure de Benjamini-Hochberg (cf partiel 2018).

Code R : si on a calculé les p-valeurs des m tests indépendants dans le vecteur `p` alors on peut utiliser le code suivant pour calculer $\hat{R}$, l'ensemble des indices des hypothèses rejetées par la procédure de Benjamini-Hochberg quand on, veut un FDR plus petit que 5% :

```
k<-sum(sort(p)<=0.05*(1:m)/m) # k chapeau
R<-(1:m)[p<=0.05*k/m]#
```

Il existe un certain nombre de méthodes basées sur les p-valeurs pour résoudre ce type de problème de tests multiples. Par exemple on peut citer la procédure de Berk-Jones modifiée.

Chapitre 3

Tests robustes

L'objectif de ce chapitre est de présenter des tests qui ne nécessitent aucune hypothèse sur les distributions sous-jacentes, ou alors des hypothèses très faibles. En ce sens, ces tests sont non-paramétriques. Ils sont également plus adaptés à la présence d'observations aberrantes dans l'échantillon. On parle de tests robustes.

Dans ce chapitre, tout ce qui est écrit en petits caractères est facultatif et peut être totalement ignoré.

On rappelle qu'on n'exige pas la connaissance de R à l'examen.

3.1 Un exemple

Un exemple de question à laquelle on souhaite répondre dans ce chapitre est la suivante : les hommes gagnent-ils plus que les femmes ? Pour répondre à cette question, imaginons que nous disposions d'un échantillon $X_1^{n_1} = (X_1, \dots, X_{n_1})$ de salaires de femmes et d'un échantillon $Y_1^{n_2} = (Y_1, \dots, Y_{n_2})$ de salaires d'hommes. Nous ferons des tests différents selon que

1. Les échantillons sont iid et indépendants entre eux i.e.

$$X_1, \dots, X_{n_1} \overset{iid}{\sim} X \quad \text{et} \quad Y_1, \dots, Y_{n_2} \overset{iid}{\sim} Y \quad X_1^{n_1} \perp\!\!\!\perp Y_1^{n_2}$$

2. Les données sont appariées. Nous donnerons une définition de l'appariement plus loin. Disons juste ici que si les deux échantillons sont de même taille et si on a regroupé les données selon l'âge des personnes (i.e. les individus de même numéro ont le même âge) alors les données sont appariées.

Imaginons pour l'instant que, pour notre exemple lié aux salaires, nous soyons dans le cas des données regroupées par âge. Nous pouvons considérer les différences de salaires $Y_i - X_i$. Supposons pour simplifier que les $(Y_i - X_i)_{1 \leq i \leq n}$ sont iid.

Le test que nous souhaitons faire est donc

H_0 : les femmes gagnent autant que les hommes

contre

H_1 : les hommes gagnent plus que les femmes

Il y a bien sûr plusieurs façons de modéliser le problème. On peut formuler le problème en utilisant la variable différence $Y_1 - X_1$. Nous souhaitons ici faire un

test sur un paramètre de position. Deux exemples usuels de paramètres de position sont la moyenne et la médiane. On pourrait traduire le fait que les femmes gagnent autant que les hommes par "la variable différence $Y_1 - X_1$ a une moyenne égale à 0", ou bien, si on préfère utiliser la médiane, on pourrait le traduire par "la médiane de la différence est égale à 0".

Autrement dit, si nous choisissons la moyenne comme paramètre de position,

$$H_0 : \text{la moyenne de } Y_1 - X_1 \text{ est égale à } 0$$

contre

$$H_1 : \text{la moyenne de } Y_1 - X_1 \text{ est strictement positive}$$

Et si nous choisissons la médiane comme paramètre de position on fait plutôt le test :

$$H_0 : \text{la médiane de } Y_1 - X_1 \text{ est égale à } 0$$

contre

$$H_1 : \text{la médiane } Y_1 - X_1 \text{ est strictement positive}$$

Si nous modélisons le problème à l'aide de la moyenne et si nous supposons les données gaussiennes, alors nous ferons naturellement le test de Student, qui est un test paramétrique.

3.2 Un test paramétrique : le test de Student

3.2.1 Un seul échantillon

Soit un n-échantillon iid $(X_1, \dots, X_n)$ de loi $N(\mu, \sigma^2)$ avec μ et σ inconnus. On veut tester

$$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu \neq \mu_0$$

(ou bien $H_1 : \mu > \mu_0$ ou bien $H_1 : \mu < \mu_0$.)

Le test de Student est basé sur la statistique $\hat{T} = \sqrt{n}\frac{\bar{X} - \mu_0}{\hat{\sigma}}$ qui suit une loi de Student à $n-1$ degrés de libertés sous H_0, où $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Au niveau α, Le test est $\phi_\alpha(X_1^n) = \mathbb{1}_{|\hat{T}| > q_{1-\alpha/2}^{T(n-1)}}$ pour H_1, $\phi_\alpha(X_1^n) = \mathbb{1}_{\hat{T} > q_{1-\alpha}^{T(n-1)}}$ pour $H_1 : \mu > \mu_0$ et $\phi_\alpha(X_1^n) = \mathbb{1}_{\hat{T} < q_\alpha^{T(n-1)}}$ pour $H_1 : \mu < \mu_0$.

Problèmes éventuels qu'on peut avoir pour réaliser ce test dans la pratique :

- l'échantillon n'est pas de loi normale,
- les variables sont gaussiennes mais pas de même variance : par exemple on peut avoir $X_i \sim N(\mu, \sigma_i^2)$.
- l'échantillon est contaminé par des outliers (=observations aberrantes)

Disons déjà, en simplifiant, que le problème éventuel de non-normalité n'est pas forcément grave si la taille de l'échantillon est grande (cf TD).

Toutefois, si on veut tester la normalité d'un échantillon, on suggère d'abord des représentations graphiques, en particulier un qqplot. On peut faire un des nombreux tests de normalité, par exemple Shapiro-Wilk (qui semble être le plus puissant dans de nombreux cas).

Code R

Le test de Student sur un échantillon dans R peut se faire par la procédure `t.test`.

Le test de Shapiro-Wilk peut se faire avec `shapiro.test`.

3.2.2 Deux échantillons indépendants

On dispose de deux échantillons indépendants $U_1, \dots, U_n$ et $V_1, \dots, V_p$, pas forcément de même taille et on veut tester l'égalité des moyennes. On suppose que

$$U_1, \dots, U_n \overset{iid}{\sim} N(\mu_1, \sigma_1^2), \quad V_1, \dots, V_p \overset{iid}{\sim} N(\mu_2, \sigma_2^2), \quad \sigma_1 = \sigma_2, \quad V_1^p \perp\!\!\!\perp U_1^n$$

et on veut tester :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2$$

(ou bien $H_1 : \mu_1 < \mu_2$ ou bien $H_1 : \mu_1 > \mu_2$)

On note σ^2 la variance commune et on suppose que σ est inconnu.

On utilise alors la variable

$$T = \frac{\bar{V} - \bar{U}}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{p}}}$$

où on a posé

$$\hat{\sigma}^2 = \frac{1}{n+p-2}\left[\sum_{i=1}^{n}(U_i - \bar{U})^2 + \sum_{i=1}^{p}(V_i - \bar{V})^2\right].$$

Sous H_0, la variable T suit une loi de student à $n+p-2$ degrés de liberté. En effet,

- $\sum_{i=1}^{n} \frac{(U_i-\bar{U})^2}{\sigma^2} \sim \chi^2(n-1)$ et $\sum_{i=1}^{p} \frac{(V_i-\bar{V})^2}{\sigma^2} \sim \chi^2(p-1)$
- Ces deux variables sont indépendantes donc $\sum_{i=1}^{n} \frac{(U_i-\bar{U})^2}{\sigma^2} + \sum_{i=1}^{p} \frac{(V_i-\bar{V})^2}{\sigma^2} \sim \chi^2(n+p-2)$
- $\bar{U} \sim N(\mu_1, \frac{\sigma^2}{n})$ et $\bar{V} \sim N(\mu_2, \frac{\sigma^2}{p})$ et $\bar{U} \perp\!\!\!\perp \bar{V}$.
- Donc sous H_0, $\bar{V} - \bar{U} \sim N(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{p})$.

Ainsi, on a obtenu que, sous H_0,

$$\frac{\bar{V} - \bar{U}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{p}}} \sim N(0,1)$$

et

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2(n+p-2)}{n+p-2}.$$

De plus

$$\sum_{i=1}^{n}(U_i - \bar{U})^2 \perp\!\!\!\perp \bar{U}, \quad \sum_{i=1}^{p}(V_i - \bar{V})^2 \perp\!\!\!\perp \bar{V}, \quad U_1^n \perp\!\!\!\perp V_1^p.$$

Donc

$$\hat{\sigma}^2 \perp\!\!\!\perp \bar{U} - \bar{V}$$

Et finalement

$$T = \frac{\frac{\bar{V}-\bar{U}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{p}}}}{\frac{\hat{\sigma}}{\sigma}} \sim T(n+p-2).$$

Et le test pour l'alternative H_1 est alors $\phi_\alpha = \mathbb{1}_{|T|>q_{1-\frac{\alpha}{2}}^{T(n+p-2)}}$ (respectivement $\phi_\alpha = \mathbb{1}_{T>q_\alpha^{T(n+p-2)}}$ pour $H_1 : \mu_1 < \mu_2$ et $\phi_\alpha = \mathbb{1}_{T<q_{1-\alpha}^{T(n+p-2)}}$ pour $H_1 : \mu_1 > \mu_2$).

Le même type de problème que pour le test de Student sur la moyenne d'un échantillon se pose :

1. Les données ne sont peut-être pas gaussiennes
2. Les données peuvent être gaussiennes mais pas de même variance
3. Les données peuvent être contaminées par des outliers.

Evoquons d'abord le problème des variances égales ou non : il existe un test adapté à des données gaussiennes et qui ressemble au test de Student mais adapté au cas $\sigma_1 \neq \sigma_2$ (en fait c'est surtout pour le cas $\sigma_1 \neq \sigma_2$ et $n_1 \neq n_2$). Ce test s'appelle le test de Welch. La procédure est basée sur la statistique $\frac{\bar{X}-\bar{Y}}{\hat{\sigma}}$ mais $\hat{\sigma}$ est calculé différemment puisqu'on ne suppose plus que la variance est la même. La statistique ne suit alors plus une loi de Student mais elle est bien approchée par une Student avec un degré de liberté non entier et calculé à partir de s_X, s_Y et de la taille de chaque échantillon.

Un certain nombre d'auteurs disent qu'il est inutile de tester si les variances des deux échantillons sont égales ou pas avant de se décider à faire le test de Welch ou le test de Student, et qu'il vaut mieux utiliser directement et systématiquement le test de Welch. C'est l'opinion majoritaire. En effet, d'une part, ce test est plus fiable quand les tailles d'échantillon diffèrent nettement et quand les variances diffèrent nettement, et d'autre part il donne des résultats très similaires au test de Student dans le cas contraire.

Le problème de variances non égales pour Student n'est pas très important si les tailles d'échantillon sont approximativement égales.

Avec R

Pour faire un test de Student ou un test de Welch avec R, on peut utiliser la fonction `t.test`, il faut préciser l'argument `var.equal=T` pour avoir le test de Student car `varequal=F` par défaut, et c'est alors le test de Welch.

3.2.3 Echantillons appariés (paired data)

"Définition" de l'appariement

On veut par exemple comparer les effets de deux traitements sur deux populations d'individus que l'on peut apparier.

Expliquons d'abord ce qu'est l'appariement. Concrètement, nous avons à notre disposition deux échantillons **de même taille** : $U_1, \ldots, U_n$ et $V_1, \ldots, V_n$. On parle de données appariées quand "l'individu" i du premier échantillon est lié à "l'individu" i du second échantillon.

Donc il faut bien comprendre ici que, pour chaque i, U_i et V_i sont liés, autrement dit il n'y a pas indépendance entre U_i et V_i. En revanche, on a toujours l'indépendance entre les (U_i, V_i) pour différents i. (Concrètement, par exemple on a (U_1, V_1) indépendant de (U_2, V_2) mais U_1 et V_1 ne sont pas indépendants.)

Prenons l'exemple d'un traitement médicamenteux. Imaginons donc qu'on veuille comparer l'efficacité de deux médicaments : U_1 et V_1 vont mesurer l'efficacité respective du médicament 1 et du médicament 2 sur deux individus qui se ressemblent,

par exemple deux individus de même âge. Il peut aussi s'agir du même individu, à qui on a donné deux traitements différents à deux moments différents.

De manière générale, quand on considère des échantillons appariés, cela signifie que

— soit U_i et V_i correspondent à une mesure sur le même individu,

— soit les individus sont différents mais ils sont regroupés en fonction de covariables (sexe, âge etc).

Les tests pour données appariées sont essentiellement basés sur le fait de prendre la différence des deux mesures et ensuite de faire un test sur l' échantillon résultant.

Le test de Student pour données appariées

On dispose de deux échantillons appariés $U_1, \dots, U_n$ et $V_1, \dots, V_n$. On veut tester l'égalité des moyennes. On pose

$$X_i = U_i - V_i, \quad i = 1, \dots, n$$

On suppose que

$$\text{les } X_i \text{ sont iid de loi } N(\mu, \sigma^2)$$

On veut donc tester

$$H_0 : \mu = 0, \quad \text{contre} \quad H_1 : \mu \neq 0$$

(ou bien $H_1 : \mu > 0$ ou bien $H_1 : \mu < 0$.)

On fait alors le test de Student (de la section 3.1.1) pour l'échantillon des X_i. Plus précisément cela donne :

$$\phi_\alpha(U_1^n, V_1^n) = \mathbb{1}_{|\hat{T}| > q_{1-\alpha/2}^{T(n-1)}}$$

où

$$T = \sqrt{n}\frac{\bar{U} - \bar{V}}{\hat{\sigma}} \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(U_i - V_i - \bar{U} - \bar{V})^2$$

.

Une des hypothèses faites est que la distribution des X_i est la même pour tout i. En particulier la variance doit être la même pour tout i. Certains auteurs recommandent de faire une vérification graphique de cela avec un graphe de "Bland-Altman". Fréquemment la dispersion est proportionnelle au niveau et une transformation logarithmique est utile pour remédier à ce problème.

Avec R

On peut utiliser `t.test` et il faut préciser `paired=T` pour dire que les données sont appariées : `t.test(x,y,paired=T,var.equal=T)`.

De façon équivalente on peut utiliser `t.test(x-y,var.equal=T)`.

3.2.4 Importance des conditions d'application

Le test de Student

A retenir : les tests non paramétriques s'appliquent dans des situations plus générales et sont donc plus robustes. On les utilise en général quand les conditions d'application des tests paramétriques ne sont pas vérifiées (ou pas vérifiables). Toutefois, un test paramétrique peut devenir performant avec une grande taille d'échantillon même si les conditions théoriques d'application du test paramétrique ne sont pas exactement vérifiées. En particulier, pour le test de Student de comparaison de deux populations, quand les tailles des échantillons sont importantes et sous des conditions assez faibles sur la loi des échantillons, le test de Student est valide, même si les échantillons ne sont pas gaussiens. Ce résultat est à rapprocher de ce qui se produit en modèle linéaire quand les erreurs ne sont pas gaussiennes.

De manière générale, on peut cependant préférer utiliser systématiquement les tests de Wilcoxon quand on ne sait pas si échantillon sont gaussiens. En effet, la performance des tests de Wilcoxon effectué sur des échantillons gaussiens n'est pas tellement moins bonne que la performance des tests de Student. De plus, les tests de Wilcoxon ont souvent une meilleure performance que celle du test de Student quand l'échantillon n'est pas gaussien (même avec une grande taille d'échantillon). .

Le reste de la section 3.2.4 est facultatif

Revenons à notre exemple lié aux salaires. Supposons ici que les échantillons sont iid et indépendants entre eux. Nous avons donc un échantillon iid de salaires de femmes, de taille n_1, et un échantillon iid de salaires d'hommes, de taille n_2, et ces deux échantillons sont supposés indépendants. Supposons que les échelles des distributions de salaires soient les mêmes (même dispersion).

Si les données normales alors nous choisirons le test de Student. Que se passe-t-il si nous nous trompons et que nous appliquons le test de Student à deux échantillons de loi non gaussienne par exemple ? Ou s'il y a des outliers dans les échantillons ?

Dans les simulations qui suivent, nous utilisons aussi le test de Wilcoxon de la somme des rangs pour comparer deux échantillons indépendants (appelé aussi "test de Mann-Whitney"). Ce test non paramétrique peut aussi être utilisé pour comparer les positions de deux populations. Comme tout test non paramétrique, il a des conditions d'application beaucoup plus générales que les tests paramétriques. En particulier pour l'appliquer, il n'est pas nécessaire d'avoir des échantillons gaussiens. Ce test sera étudié dans la suite.

Nous voulons donc illustrer ici ce qui se produit quand on n'est pas dans les conditions d'application du test de Student (et ensuite comparer sa performance avec le test de Mann-Whitney) pour montrer l'intérêt des tests non paramétriques. Plus précisément nous allons simuler des variables de lois normales et aussi des lois non normales : nous regardons ce qui se passe pour des échantillons de loi de student à 3 degrés de liberté et de loi de Cauchy . Comme la loi de Cauchy n'a pas de moyenne, ce que nous utilisons comme paramètre de position pour la loi de Cauchy est sa médiane.

```r
sim=function(type,n,a)
# fonction qui simule un   échantillon de taille n
{
    return(switch(type,norm=rnorm(n,mean=a),cauchy=rcauchy(n,a), student3=rt(n,df=3,a)))
}

test=function(n,type,a,b,n1,n2,outliers=F)
  #simulation, calcul de la p-valeur de chaque test
{
    u=rep(0,n); v=rep(0,n); w=rep(0,n);
    #fait avec lapply par principe mais  boucle for pas plus lente ici
lapply(1:n,function(i)
 # on fait n simulations et  les 3 tests sur chaque simulation
  {
    x=sim(type,n1,a); y=sim(type,n2,b);# simulation des deux échantillons
    if (outliers) {x[1:10]=rnorm(10,3)}
    #calcul de la p-valeur de chaque test
        v[i]<<-t.test(x,y,var.equal=T)$p.value; # Student
    w[i]<<-wilcox.test(x,y)$p.value;    # Mann-Whitney
    })
    # on regarde le taux d'erreurs si on est sous H0 et la puissance si on est sous H1
return(list( "Student"=sum(v<0.05)/n,
            "Mann-Whitney"=sum(w<0.05)/n))

}

# on s'attend à avoir un taux proche de 5/% sous H0 si tout va bien,
# et un taux important si on est sous H1 (puissance)


 #sur des lois normales  et même variance
#sous H0
test(10000,"norm",1,1,30,30)
```

```
## $Student
## [1] 0.0513
##
## $`Mann-Whitney`
## [1] 0.0503
```

```r
#sous H1
 test(10000,"norm",1,2,30,30)
```

```
## $Student
## [1] 0.9686
##
## $`Mann-Whitney`
## [1] 0.9613
```

```r
 #loi de Student à 3 degrés de liberté  (même variance)
 #sous H0
 test(10000,"student3",1,1,50,50)
```

```
## $Student
## [1] 0.0443
```

```
##
## $`Mann-Whitney`
## [1] 0.0466
```

```
 #sous H1
 test(10000,"student3",1,2,50,50)
```

```
## $Student
## [1] 0.8839
##
## $`Mann-Whitney`
## [1] 0.979
```

```
#loi de Cauchy, petits échantillons
 #sous H0
  test(10000,"cauchy",1,1,15,15)
```

```
## $Student
## [1] 0.0218
##
## $`Mann-Whitney`
## [1] 0.0446
```

```
  #sous H1
 test(10000,"cauchy",1,2,15,15)
```

```
## $Student
## [1] 0.0728
##
## $`Mann-Whitney`
## [1] 0.2701
```

```
#loi de Cauchy, grands échantillons
 #sous H0
 test(10000,"cauchy",1,1,100,100)
```

```
## $Student
## [1] 0.0201
##
## $`Mann-Whitney`
## [1] 0.0495
```

```
 #sous H1
 test(10000,"cauchy",1,2,100,100)
```

```
## $Student
## [1] 0.0762
##
## $`Mann-Whitney`
## [1] 0.9548
```

```
# lois normales avec  10\% d'outliers "assez gros"
# sous H0
 test(10000,"norm",1,1,100,100,outliers=T)
```

```
## $Student
## [1] 0.2278
##
## $`Mann-Whitney`
```

```
## [1] 0.1572
```

```
 #sous H1
test(10000,"norm",1,2,100,100,outliers=T)
```

```
## $Student
## [1] 0.9998
##
## $`Mann-Whitney`
## [1] 0.9998
```

```
#la même chose mais sans outliers

# sous H0
 test(10000,"norm",1,1,100,100)
```

```
## $Student
## [1] 0.0456
##
## $`Mann-Whitney`
## [1] 0.0441
```

```
 #sous H1
test(10000,"norm",1,2,100,100)
```

```
## $Student
## [1] 1
##
## $`Mann-Whitney`
## [1] 1
```

- Le test de Mann-Whitney est souvent moins performant que le test de Student quand on est dans les conditions d'application du test de Student, mais la différence est souvent faible.
- Avec des lois de Student à 3 degrés de liberté, le test de Mann-Whitney est plus performant que le test de Student. Le test de Student est valide dans le cas où les échantillons sont de taille suffisante. Ce comportement peut grossièrement s'expliquer ainsi : la variance est finie, donc quand n_1 et n_2 sont suffisamment grands, le test fonctionne assez bien (cf aussi TD2 exo 5) .
- Le test de Student se comporte mal avec des échantillons de loi de Cauchy, même si les tailles des échantillons sont grandes. Ce comportement peut grossièrement s'expliquer ainsi : les queues de la loi de Cauchy sont si lourdes que la variance est infinie (même la moyenne est infinie dans ce cas).
- La performance du test de Student est plus affectée par la présence d'outliers que le test de Mann-Whitney.

Dans ces deux dernières situations, c'est-à-dire quand il y a présence d'un grand nombre d'outliers ou quand loi est à queues lourdes, le test de Mann-Whitney que l'on va introduire dans la suite se comporte mieux que le test de Student. On dit qu'il est plus robuste.

3.3 Test du signe

Nous venons de voir un test paramétrique, le test de Student, qui peut être utilisé pour comparer deux populations indépendantes ou bien comparer deux traitements sur des données appariées. Ce test repose sur le caractère gaussien des données.

On va construire maintenant des tests reposant sur des hypothèses beaucoup plus faibles sur les données.

On commence par le test du signe et le test de Wilcoxon des rangs signés, qu'on peut plus ou moins voir comme des versions non-paramétriques du test de Student.

Définition 3.1. *On dit qu'une variable aléatoire U est diffuse si*

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

Cela revient à dire que sa distribution est continue, c'est-à-dire que sa cdf est continue

3.3.1 Test du signe sur un seul échantillon

Objectif : Faire un test sur un paramètre de position, qui n'est ici pas la moyenne, mais la médiane.

Intérêt : ne nécessite justement même pas l'existence d'une moyenne, plus robuste.

Données : $X_1, \ldots, X_n$.

Les conditions :

1. les X_i sont indépendantes
2. Les X_i ont une médiane commune m.
3. $P(X_i = m) = 0$

Remarquez que les X_i ne sont pas nécessairement identiquement distribuées.

L'hypothèse nulle est :

$$H_0 : m = 0$$

Remarquons que $m = 0$ implique ici que $\mathbf{P}(X_i \leq 0) = 1/2$.

En effet si 0 est la médiane commune des X_i alors (cf chapitre 2)

$$\mathbf{P}(X_i \leq 0) \geq 1/2 \quad \text{et} \quad \mathbf{P}(X_i \geq 0) \geq 1/2$$

Ici on suppose en plus que $\mathbf{P}(X_i = 0) = 0$ donc la propriété ci-dessus se réécrit

$$\mathbf{P}(X_i < 0) \geq 1/2 \quad \text{et} \quad \mathbf{P}(X_i > 0) \geq 1/2$$

et comme on a alors aussi $\mathbf{P}(X_i < 0) + \mathbf{P}(X_i > 0) = 1$, on a forcément

$$\mathbf{P}(X_i < 0) = \mathbf{P}(X_i > 0) = 1/2$$

et donc $\mathbf{P}(X_i \leq 0) = \mathbf{P}(X_i < 0) = \mathbf{P}(X_i \geq 0) = \mathbf{P}(X_i > 0) = 1/2$.

(c'est donc ici qu'intervient la condition 3.)

On pose

$$Y_i = \mathbb{1}_{X_i \leq 0}.$$

Faisons d'abord, pour simplifier l'exposition, l'hypothèse que les X_i sont de même loi.

On a

$$Y_i \overset{iid}{\sim} Be(p), \quad \text{avec} \quad p = \mathbf{P}(X_i \leq 0).$$

Donc H_0 se réécrit

$$H_0 : p = 1/2.$$

Donc on se ramène à un test d'égalité sur le paramètre p d'un échantillon iid de v.a. de Bernoulli Y_i.

Imaginons que l'alternative soit la suivante :

$$H_1 : m \neq 0$$

alors, cette alternative peut aussi s'écrire,

$$H_1 : p \neq 1/2.$$

Il s'agit donc du test de l'exercice 1 TD2. On utilise donc

$$\phi_\alpha(Y_1, \ldots, Y_n) = \mathbb{1}_{\left\{ |\sum_{i=1}^n Y_i - n/2| > q_{1-\frac{\alpha}{2}}^{B(n,1/2)} - \frac{n}{2} \right\}}$$

Avec l'échantillon initial, cela donne

$$\phi_\alpha(X_1, \ldots, X_n) = \mathbb{1}_{\left\{ |\sum_{i=1}^n \mathbb{1}_{X_i \leq 0} - n/2| > q_{1-\frac{\alpha}{2}}^{B(n,1/2)} - \frac{n}{2} \right\}}$$

Si l'hypothèse alternative est

$$H_1 : m < 0$$

Dans ce cas, le test est de la forme

$$\phi_\alpha(X_1, \ldots, X_n) = \mathbb{1}_{\left\{ \sum_{i=1}^n \mathbb{1}_{\{X_i \leq 0\}} > q_{1-\alpha}^{B(n,1/2} \right\}}$$

Si l'hypothèse alternative est $H_1 : m > 0$, on remplace $\sum_{i=1}^n \mathbb{1}_{X_i \leq 0}$ par $\sum_{i=1}^n \mathbb{1}_{X_i \geq 0}$ dans la formule ci-dessus.

Maintenant, que se passe-t-il si les X_i ne sont pas de même loi ? Pour simplifier, supposons que H_1 corresponde au fait que la médiane commune est strictement négative :

$$H_1 : m < 0$$

Alors "ça marche quand même " : en effet, sous H_0 on a bien, du fait que 0 est la médiane commune,

$$Y_i \overset{iid}{\sim} Be(1/2)$$

donc si on pose

$$\phi_\alpha(X_1,\dots,X_n) = \mathbb{1}_{\left\{\sum_{i=1}^n \mathbb{1}_{X_i\leq 0} > q_{1-\alpha}^{B(n,1/2)}\right\}},$$

on a bien un test de niveau $1-\alpha$.

Est-ce un test adapté au problème ? Cela revient à savoir si la statistique $\sum_{i=1}^n \mathbb{1}_{X_i\leq 0}$ prend bien de grandes valeurs sous H_1. C'est bien le cas ici car si on est sous H_1, les X_i ont tendance à prendre des valeurs négatives et donc le nombre de X_i négatifs va être grand.

Ce test, qui utilise donc uniquement le signe des X_i, est appelé test du signe.

Avec R

On peut utiliser la procédure `binom.test` qui fait un test sur le paramètre p d'un échantillon de va de Bernoulli. Si l'échantillon se trouve dans un vecteur `x`, et si a une alternative bilatéral $H_1 : m \neq 0$, on peut utiliser la commande suivante

```
binom.test(sum(x>0),n=length(x),p=0.5,alternative="two.sided")
```

Pour l'alternative $H_1 : m < 0$ on met `alternative="less"`

Pour $H_1 : m > 0$ on met `alternative="greater"`.

3.3.2 Test du signe sur deux échantillons

On dispose de deux échantillons appariés $U_1,\dots,U_n$ et $V_1,\dots,V_n$. Comme d'habitude avec les données appariées, on se ramène à un test sur l'échantillon des différences

$$X_i = V_i - U_i.$$

Pour fixer les idées, imaginons le cas de deux traitements que l'on veut comparer. On prend deux populations de même taille n. On les classe par âge. On donne à un individu i un premier traitement dont on mesure l'efficacité par U_i et on donne l'autre traitement à un individu du même âge, dont on mesure l'efficacité par V_i.

On veut par exemple savoir si le second traitement est plus efficace que le premier. On peut modéliser le fait que les deux traitements ont la même efficacité par l'égalité $\mathbf{P}(U_i \leq V_i) = \mathbf{P}(V_i \leq U_i)$, ce qui donne, en termes des X_i, $\mathbf{P}(X_i \leq 0) = \mathbf{P}(X_i \geq 0)$. En supposant que, presque sûrement, les X_i ne prennent jamais la valeur 0, ceci se traduit encore par "la médiane commune des X_i est égale à 0". En effet on a vu dans la sous-section 3.3.1 que $m = 0$ signifie, si la condition $\mathbf{P}(X_i = 0) = 0$ est satisfaite, que $\mathbf{P}(X_i \leq 0) = \mathbf{P}(X_i \geq 0) = 1/2$. On va donc faire un test du signe sur l'échantillon différence X_1^n.

On suppose donc que les conditions suivantes, qui sont les conditions du test du signe sur l'échantillon des X_i, sont vérifiées :

— Les X_i sont indépendants entre eux.

— Les X_i ne sont pas forcément de même loi mais ont une médiane commune m.

— $P(X_i = m) = 0$.

Le test est donc

$$\phi_\alpha(U_1,\dots,U_n,V_1,\dots,V_n) = \mathbb{1}_{\left\{\sum_{i=1}^n \mathbb{1}_{\{U_i\le V_i\}} > q_{1-\alpha}^{B(n,1/2)}\right\}}$$

Si au contraire, on pense que soit les deux médicaments ont la même efficacité, soit le premier est plus efficace, alors on échange juste les rôle de U_i et V_i, ce qui donne le test

$$\phi_\alpha(U_1,\dots,U_n,V_1,\dots,V_n) = \mathbb{1}_{\left\{\sum_{i=1}^n \mathbb{1}_{\{V_i\le U_i\}} > q_{1-\alpha}^{B(n,1/2)}\right\}}$$

Si on n'a pas d'a priori sur les médicaments, c'est-à-dire si on ne sait pas quel médicament est susceptible d'être plus efficace, l'alternative est alors $H_1 : m \neq 0$ et on fait le test

$$\phi_\alpha(X_1,\dots,X_n) = \mathbb{1}_{\left\{|\sum_{i=1}^n \mathbb{1}_{U_i\le V_i} - n/2| > q_{1-\frac{\alpha}{2}}^{B(n,1/2)} - \frac{n}{2}\right\}}$$

Remarque 3.2. *Le test du signe n'utilise que très peu d'information sur les variables $U_i - V_i$ (uniquement leur signe, pas leurs valeurs absolues). C'est donc un test peu puissant. Quel est alors l'intérêt de parler du test du signe ? Il se peut que les signes des $U_i - V_i$ soit la seule donnée disponible : c'est en effet le cas si la question posée aux patients qui ont testé les deux médicaments est "quel est le meilleur des deux ?" (au lieu de noter les médicaments sur une échelle de 1 à 10 par exemple).*

Remarque 3.3. *Concrètement, il faut bien vérifier que la valeur 0 n'est pas dans l'échantillon.*

En Td et en examen, on ne se demandera pas si la condition "les X_i ont une médiane commune" est réaliste, on supposera simplement que cette condition est vérifiée.

Avec R

Comme il s'agit du test du signe sur les variables X_i, on utilise exactement la même procédure que dans le cas d'un seul échantillon. Il suffit donc de calculer l'échantillon des différences puis de faire le test sur cet échantillon à l'aide de la fonction `binom.test`.

(Remarque 26 enlevée)

Remarque 3.4. *On pourrait utiliser ce test pour tester que les échantillons sont de même loi. Supposons que l'on nous donne deux échantillons indépendants $U_1^n = (U_1,\dots,U_n)$ et $V_1^n = (V_1,\dots,V_n)$. Supposons que les U_i sont iid de loi F continue, et les V_i sont iid de loi G continue. Nous voulons donc tester l'égalité des lois :*

$$H_0 : F = G \quad \textit{contre} \quad H_1 : F \neq G.$$

Pour cela nous voulons utiliser le test du signe. Alors nous devons vérifier si, sous l'hypothèse $F = G$, l'hypothèse nulle associée au test du signe est vérifiée, c'est-à-dire si $m = 0$ en notant m la médiane de $V - U$. Comme $V - U$ est diffuse, cela revient à montrer que

$$\mathbf{P}(U \leq V) = \mathbf{P}(V \leq U)$$

Or on a, sous l'hypothèse $F = G$,

$$(U, V) \sim (V, U)$$

car U et V sont alors interchangeables. Donc

$$U - V \sim V - U$$

Donc on a bien $\mathbf{P}(V \leq U) = \mathbf{P}(U \leq V)$.

Remarquons d'une part que l'on ne pourra détecter qu'un changement de paramètre de position, contrairement au test de Kolmogorov-Smirnov qui est plus général. D'autre part, on peut avoir $\mathbf{P}(U \leq V) = \frac{1}{2}$ sans que U et V aient la même loi. Autrement dit, l'égalité des lois ne se traduit pas vraiment par la propriété $\mathbf{P}(U \leq V) = \frac{1}{2}$ (qui est la propriété réellement testée par le test du signe). L'égalité des lois est une propriété beaucoup plus forte et générale que le fait que la médiane des différences est égale à 0.

Il suffit par exemple que U et V soient symétriques, diffuses et indépendantes pour que la médiane de $V - U$ soit égale à 0. Par exemple, le test du signe ne sera pas capable de détecter la différence de loi entre un échantillon de loi normale standard et un échantillon de loi de Cauchy.

En effet si U et V sont symétriques, en plus d'être indépendantes et diffuses, on a

$$(U, V) \sim (-U, -V)$$

Donc

$$U - V \sim -U - (-V) = V - U$$

Ainsi

$$\mathbf{P}(U - V \geq 0) = \mathbf{P}(V - U \geq 0)$$

En combinant cette inégalité avec le fait que $\mathbf{P}(U - V = 0$ on obtient

$$\mathbf{P}(U \geq V) = \mathbf{P}(V \geq U) = 1/2.$$

Ainsi, si on voit le test du signe comme le test d'égalité des lois

$$H_0 : U \sim V$$

alors une p-valeur observée grande n'implique pas que H_0 est vraie. Par exemple, il n'est pas rare d'avoir une p-valeur grande si on fait le test du signe sur un échantillon de loi normale et l'autre de loi de Cauchy, alors que l'on est bien sous $H_1 : F \neq G$. Le test du signe, vu comme un test d'homogénéité, est donc un exemple de test particulièrement peu puissant : "il ne voit" pas certaines alternatives.

3.4 Statistiques d'ordre et de rang

Définition 3.5. *Soient $X_1, \ldots, X_n$ n v.a. réelles. La statistique d'ordre $(X_{(1)}, \ldots, X_{(n)})$ est définie par*

$$\{X_{(1)}, \ldots, X_{(n)}\} = \{X_1, \ldots, X_n\}$$

et

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

On pose

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

Il existe une permutation aléatoire $\hat{\sigma} \in \mathfrak{S}_n$ telle que

$$(X_{(1)}, \ldots, X_{(n)}) = (X_{\hat{\sigma}(1)}, \ldots, X_{\hat{\sigma}(n)}).$$

Evidemment, comme on peut avoir $X_i = X_j$ pour $i \neq j$, il n'y a pas toujours unicité de cette permutation.

On définit le vecteur des rangs R_X comme la permutation inverse de $\hat{\sigma}$. Evidemment, de la même manière que $\hat{\sigma}$, ce vecteur de rang n'est pas unique s'il existe $i \neq j$ tels que $X_i = X_j$.

En fait comme son nom l'indique, le vecteur de rangs donne le rang de chaque variable dans l'échantillon. Exemple :

$x = (4, 2, 1, 1, 2, 0, 1)$

x	4	2	1	1	2	0	1
R_x	7	5	2	3	6	1	4

En théorie, si les X_i sont iid et de loi continue alors, presque sûrement, il n'y a pas d'ex-aequo (cf TD1). En pratique, comme on l'a déjà signalé, à cause de la limitation de la précision des mesures et des arrondis, il peut y avoir des ex-aequo dans un échantillon issu d'une loi continue. Il faut être attentif à ce problème car, dans les logiciels, les procédures censées fonctionner sur des données de loi continue ne sont pas toujours prévues pour parer à l'éventualité d'un ex-aequo, et même si elles le sont, le résultat n'est pas toujours fiable (cf plus loin).

3.5 Test des rangs signés de Wilcoxon

3.5.1 Sur un échantillon

On va à nouveau faire un test sur la médiane d'un échantillon.

On considère des variables $(X_1, \ldots, X_n)$ diffuses et indépendantes, mais pas forcément de même loi. La proposition suivante montre que le vecteur de rangs de X est alors unique presque sûrement.

Proposition 3.6. *Si les variables aléatoires $X_1, \ldots X_n$ sont indépendantes et diffuses alors*

$$\mathbf{P}(\exists i \neq j : |X_i| = |X_j|) = 0$$

Démonstration. Pour tout $i \neq j$ on a

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

car les variables sont indépendantes et diffuses. Ainsi

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

□

On suppose disposer d'observations $(X_1, \ldots, X_n)$ qui vérifient les conditions suivantes :

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

Notez que nous avons besoin d'une hypothèse supplémentaire par rapport au test du signe : nous devons supposer que les X_i sont symétriques en loi par rapport à leur médiane commune m. Pour simplifier, nous ne nous intéresserons pas à la vérification de la condition de symétrie, ni à la condition de médiane commune, dans les exercices du TD ou à l'examen. Nous nous contenterons de supposer que ces conditions sont bien vérifiées, sans autre précision.

L'hypothèse H_0 est la suivante

$$H_0 : m = 0$$

On va utiliser à nouveau le signe des X_i, mais on suppose en plus qu'on dispose de la valeur des $|X_i|$. On compte le nombre de $X_i > 0$ mais on leur attribue un poids d'autant plus grand que $|X_i|$ est élevé. Si on dispose des valeurs des $|X_i|$, le test suivant est préférable au test du signe étudié précédemment, car il utilise plus d'information tout en ayant des conditions d'application presqu'aussi larges.

On considère la statistique d'ordre associée aux $\{|X_i|\}_{1 \leq i \leq n}$. On a donc

$$|X|_{(1)} < |X|_{(2)} < \ldots < |X|_{(n)}.$$

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

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

Exemple 3.7.

X_i	*-0,15*	*-0,42*	*0,22*	*0,6*	*-0,1*
$\lvert X_i \rvert$	*0,15*	*0,42*	*0,22*	*0,6*	*0,1*
$R_{\lvert X \rvert}(i)$	*2*	*4*	*3*	*5*	*1*

Remarque 3.8. *On a* $0 \leq W_n^+ \leq \frac{n(n+1)}{2}$. *Le cas* $W_n^+ = 0$ *correspond à tous les* $X_i < 0$, *le cas* $W_n^+ = \frac{n(n+1)}{2}$ *correspond au cas où tous les* $X_i > 0$.

Si on pose en plus $W_n^- = \sum_{i=1}^n R_{|X|}(i)\mathbb{1}_{\{X_i<0\}}$ *et si* $\mathbf{P}(X_i = 0) = 0$ *(ce qui est le cas si les variables sont diffuses), alors*

$$W_n^+ + W_n^- = \frac{n(n+1)}{2}$$

Expliquons rapidement l'idée derrière ce test. Supposons pour fixer les idées que

$$H_1 : m > 0.$$

L'idée est que, sous H_1, il y a plus de X_i positifs que de X_i négatifs. Jusque là c'est même idée que pour le test du signe. Mais en plus, du fait de la symétrie, les X_i positifs ont tendance à être plus grands en valeur absolue que les X_i négatifs, c'est là qu'on utilise une information supplémentaire par rapport au test du signe. Donc sous cette alternative, W_n^+ sera "grand".
Evidemment si l'alternative est $H_1 : m < 0$, alors W_n^+ sera au contraire "petit".

Exemple 3.9. *Prenons un exemple concret. On simule un premier échantillon* X_1^n *de taille* $n = 30$ *de loi de* $T(4)$ *localisé en* $m = 0$, *c'est-à-dire dont la densité est* $f(x) = \frac{1}{\pi(1+x^2)}$ *(en noir). On simule ensuite un échantillon de même taille de loi de Cauchy localisé en* $m = 1$, *c'est-à-dire dont la densité est* $f(x) = \frac{1}{\pi(1+(x-1)^2)}$. *Ces deux échantillons sont représentés dans la figure 3.9. On remarque que*

- *Le premier échantillon a quasiment autant de valeurs positives que de valeurs négatives. De plus, grâce à la symétrie de la loi pa rapport à 0, les valeurs absolues des* x_i *qui sont positifs n'ont pas tendance à être plus grandes que les valeurs absolues des* x_i *qui sont négaltifs, et vice versa.*
- *Le deuxième échantillon a plus de valeurs positives que de valeurs négatives. Mais en plus si on range par ordre croissant les valeurs absolues des* x_i, *ce sont les* x_i *positives qui ont les rangs les plus élevés.*

Exemple 3.10. *Prenons un autre exemple pour illustrer la nécessité de la condition de symétrie. On simule un échantillon de loi de densité* $f(x) = \frac{1}{2}\exp(-x)\mathbb{1}_{x>0} + \frac{1}{2}\exp(-3x)\mathbb{1}_{x<0}$. *C'est une loi de médiane 0 mais non symétrique. L'échantillon est représenté sur la figure 3.5.1. On voit qu'il y a à peu près autant de valeurs positives que négatives, mais que les* x_i *positifs ont tendance à prendre des valeurs absolues plus grandes.*

Théorème 3.11. *Les conditions 1,2,3 et 4 sont supposées vérifiées. On a,* ***sous*** H_0 ***: m=0****,*

1. W_n^+ *et* W_n^- *ont même distribution.*
2. $\mathbf{E}[W_n^+] = \frac{n(n+1)}{4}$
3. W_n^+ *et* W_n^- *sont libres en loi de* X.
4. $\mathbf{Var}(W_n^+) = \frac{n(n+1)(2n+1)}{24}$.

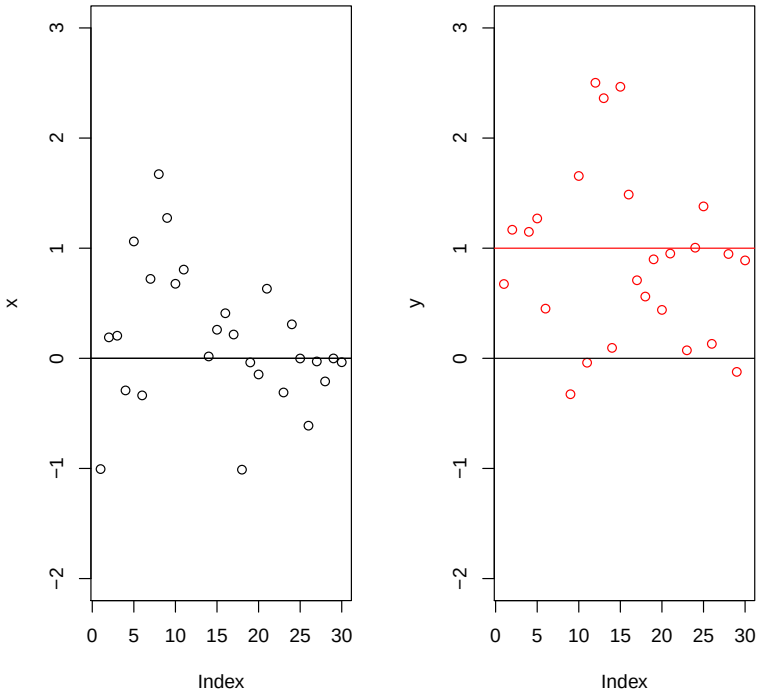


FIGURE 3.1 – échantillon de loi symétrique avec $m = 0$ (à gauche) et $m = 10$ à droite

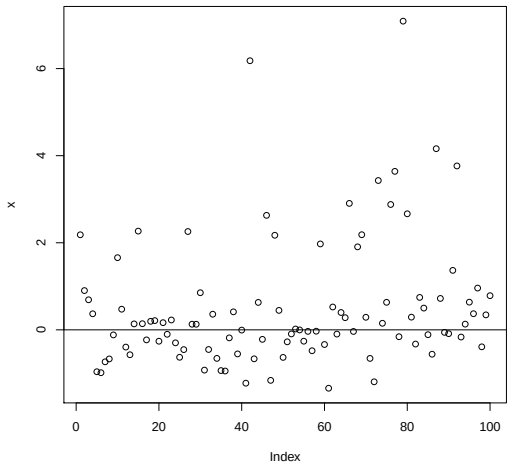


FIGURE 3.2 – échantillon de loi non symétrique avec médiane égale à 0

5. *Asymptotiquement, on a*

$$\frac{W_n^+ - \mathbf{E}[W_n^+]}{\sqrt{\mathbf{Var}(W_n^+)}} \xrightarrow{loi} \mathcal{N}(0,1)$$

On admet la preuve. Cependant pour les étudiants intéressés, voici une partie de la preuve.

Démonstration. On se place sous H_0.

1. On a

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

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

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

La loi des X_i est symétrique par rapport à 0, donc

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

Donc, pour toute fonction f (déterministe), on a

$$f(X_1^n) \sim f(-X_1^n).$$

Or W_n^+ est une fonction du vecteur X_1^n. Donc on a

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

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

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

Donc W_n^+ et W_n^- sont de même loi.

2. De la même manière que pour l'item 1, la symétrie de la loi de X_1^n implique que, pour tout $1 \leq j \leq n$, $X_{\sigma_{|X|}(j)} \sim -X_{\sigma_{|X|}(j)}$ et donc

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

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

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

et donc

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

Montrons donc que $X_{\sigma_{|X|}(j)}$ est diffuse. On a pour tout $x \in \mathbb{R}$

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

car les X_i sont diffuses.

3. Le point clé est que la symétrie de la loi des X_i par rapport à 0 implique que les vecteurs $(|X_1|, \ldots, |X_n|)$ et $(\mathbb{1}_{X_1>0}, \ldots, \mathbb{1}_{X_n>0})$ sont indépendants. En effet cette propriété, combinée au fait que les $X_{\sigma_{|X|}(i)}$ sont diffuses, implique que

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

où

$$Y_1, \ldots, Y_n \overset{iid}{\sim} Be(1/2).$$

En effet, soit $(\epsilon_1, \ldots, \epsilon_n) \in \{0,1\}^n$, on a

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

On a utilisé

- ligne 2 : les probabilités totales.
- ligne 3 : l'indépendance entre $(|X_1|, \ldots, |X_n|)$ et $(\mathbb{1}_{X_1>0}, \ldots, \mathbb{1}_{X_n>0})$ entraine l'indépendance entre $(|X_1|, \ldots, |X_n|)$ et $(\mathbb{1}_{X_{s(1)}>0}, \ldots, \mathbb{1}_{X_{s(n)}>0})$ car s est fixe ! (et $\sigma_{|X|}$ est une fonction de $|X|$).
- ligne 4 : s est fixe et les variables $X_1, \ldots, X_n$ sont indépendantes donc les variables $X_{s(1)}, \ldots, X_{s(n)}$ sont indépendantes. Donc

$$\begin{aligned}
&\mathbf{P}\Big(\mathbb{1}_{X_{s(1)}>0}, \ldots, \mathbb{1}_{X_{s(n)}>0}) = (\epsilon_1, \ldots, \epsilon_n)\Big) \\
&= \mathbf{P}(\mathbb{1}_{X_{s(1)}>0} = \epsilon_1) \ldots \mathbf{P}(\mathbb{1}_{X_{s(n)}>0} = \epsilon_n)
\end{aligned}$$

De plus $\mathbf{P}(X_{\sigma_{|X|}(i)} > 0) = 1/2$ d'après l'item 2.

Ceci prouve l'item 3 : en effet, $W_n^+ \sim \sum_{j=1}^n jY_j$. Ceci permet aussi de trouver la valeur de la variance : en effet

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

□

Remarque 3.12. *Sous H_0, la statistique W_n^+ a une distribution symétrique par rapport à sa moyenne $\frac{n(n+1)}{4}$.*

En effet, sous H_0, $W_n^+ \sim W_n^-$ et comme $W_n^+ + W_n^- = \frac{n(n+1)}{2}$ on a

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

c'est-à-dire

$$W_n^+ \sim 2b - W_n^+$$

avec

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

En conséquence, le test, pour l'alternative $H_1 : m \neq 0$, est

$$\phi(X_1, \ldots, X_n) = \mathbb{1}_{\{|W_n^+ - \frac{n(n+1)}{4}|>q\}}$$

pour une certaine valeur q à choisir, fonction du niveau souhaité. En raisonnant comme dans l'exercice 3 du TD1,on peut montrer que le test au niveau α est

$$\phi_\alpha(X_1, \ldots, X_n) = \mathbb{1}_{\{|W_n^+ - \frac{n(n+1)}{4}|>q_{1-\frac{\alpha}{2}} - \frac{n(n+1)}{4}\}}$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de W_n^+ sous H_0 (il s'agit de la loi de $\sum_{j=1}^n jY_j$ avec $Y_1, \ldots, Y_n \overset{iid}{\sim} Be(1/2)$ d'après la preuve).

Remarque 3.13. *Test exact ou asymptotique ?*

On utilise la loi exacte de W_n^+ sous H_0 quand $n \leq 20$.

Pour $n > 20$, on utilise un test asymptotique, conséquence de la convergence en loi.

Avec R

Pour tester $H_0 : m = 0$ contre $H_1 : m \neq 0$, si l'échantillon se trouve dans un vecteur `x`, on peut utiliser `wilcox.test(x,alternative="two.sided")`. Pour $H_1 : m > 0$, on met `alternative="greater"` et pour $H_1 : m < 0$, on met `alternative="less"`.

3.5.2 Echantillons appariées

On suppose disposer de deux échantillons appariés de taille n : $(U_1, \ldots, U_n)$ et $(V_1, \ldots, V_n)$. On veut savoir si l'un des échantillons a "tendance à prendre des valeurs plus grandes que l'autre" (penser à l'exemple des traitements médicamenteux). C'est la même problématique que pour le test du signe sur deux échantillons. On modélise le problème de la même manière.

Comme pour le test du signe, on pose

$$X_i = V_i - U_i$$

et nous utilisons le test de Wilcoxon des rangs signés sur l'échantillon X_1^n.

Nous supposons donc que

— Les X_i sont indépendants entre eux (mais pas forcément de même loi).

— Les X_i sont diffuses.

— Les X_i ont une médiane commune m.

— Les X_i sont de loi symétrique par rapport à m.

L'hypothèse nulle est

$$H_0 : m = 0$$

Si on pense que soit on est sous H_0, soit les U_i prennent des valeurs plus petites que les V_i , alors on pose comme hypothèse alternative

$$H_1 : m > 0$$

Remarque 3.14. *Comme on l'a déjà remarqué pour le cas du test sur la médiane d'un seul échantillon, le test des rangs signés est plus puissant que le test du signe. Donc il est conseillé d'utiliser le test des rangs signés plutôt que le test du signe si on peut le faire, c'est-à-dire si on a accès aux valeurs de l'échantillon des X_i et pas seulement à leur signe et si la loi des X_i est symétrique.*

Avec R

On peut utiliser la fonction `wilcox.test`. Si nos échantillons sont dans des vecteurs `x` et `y`, et si $H_1 : m \neq 0$, on écrit

`wilcox.test(x,y,paired=T, alternative="two.sided")`.

ou son équivalent `wilcox.test(x-y, alternative="two.sided")`.

(Mêmes changements possibles d'alternative que précédemment.)

On peut aussi avoir des données correspondants à deux colonnes d'un dataframe. Un exemple : on veut savoir si les salaires des hommes d'une entreprise sont du même ordre que les salaires des femmes ou bien plus élevés. On suppose que l'on a apparié les données (par exemple on a rassemblés les salaires selon l'âge de la personne). On suppose que ces salaires apparaissent dans un dataframe nommé `salaires` avec pour colonnes `femmes` et `hommes`, on peut alors utiliser

```
wilcox.test(data=salaires, hommes~femmes, paired=T, alternative="greater")
```

Remarque 3.15. *Il faut alors faire attention au problème des ex aequo ("ties" en anglais) quand on utilise la procédure* `wilcox.test`*. On peut quand même faire le test, mais il n'est jamais exact, c'est-à-dire qu'il repose automatiquement sur une approximation gaussienne.*

Comme l'approximation gaussienne n'est valable que pour des grands échantillons, on ne peut pas trop se fier au résultat de `wilcox.test` *quand il y a des ex aequo et quand la taille d'échantillon est trop petite (pas de problème en revanche avec les éventuels ex aequo si la taille est suffisamment grande).*

Dans le cas d'ex aequos, on reçoit le message suivant `warning message : cannot compute exact p-values with ties`*.*

Ce qu'on entend par ex aequo ici, c'est un ex aequo dans l'échantillon x_1^n des différences ou un ex aequo dans l'échantillon des valeurs absolues des différences $(|x_1|, \ldots, |x_n|)$.

Remarque 3.16. *Pour tester l'hypothèse de symétrie des X_i quand il s'agit d'un échantillon i.i.d, on peut par exemple commencer par représenter les données (histogramme par exemple ou densité cf chap4), (ou utiliser un des nombreux tests de symétrie : par exemple le test* `symmetry.test` *du package* `lawstat` *(attention cette fonction est un peu lente). En cas d'asymétrie sur l'échantillon des différences il semble préférable d'utiliser le test du signe.*

Certains praticiens utilisent une transformation des X_i pour rendre l'échantillon symétrique (mais le test fait sur l'échantillon transformé n'est alors pas un test sur la médiane des X_i)

3.6 Wilcoxon de la somme des rangs/Mann-Whitney

3.6.1 Résultats préliminaires sur le vecteur des rangs

On commence par quelques résultats liés au vecteur des rang dans le cas de données i.i.d.

Théorème 3.17. *Soient $X_1, \ldots, X_n$ n v.a. i.i.d. de loi continue et de statistique d'ordre X^* et de vecteur des rangs R_X. Alors X^* et R_X sont indépendants et de plus R_X est distribué uniformément sur $\mathfrak{S}_n$.*

Démonstration. La loi est continue donc presque sûrement il n'y a pas d'ex-aequo. R_X est clairement à valeurs dans $\mathfrak{S}_n$. Comme R_X est la permutation inverse de σ, il suffit en fait de montrer que

1. σ suit une loi uniforme sur $\mathfrak{S}_n$.
2. σ et X^* sont indépendants.

Puisque les X_i sont indépendantes et de même loi, elles sont interchangeables donc

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

Par exemple, pour $n = 3$, on a

$$\begin{aligned}\mathbf{P}(X_1 < X_2 < X_3) = \mathbf{P}(X_1 < X_3 < X_2) = \mathbf{P}(X_2 < X_1 < X_3) = \mathbf{P}(X_2 < X_3 < X_1)\\ = \mathbf{P}(X_3 < X_2 < X_1) = \mathbf{P}(X_3 < X_1 < X_2) = 1/6\end{aligned}$$

On montre maintenant que σ et X^* sont indépendantes. On veut montrer que, pour tout borélien B de $\mathbb{R}^d$ et toute permutation s de $\mathfrak{S}_n$, on a

$$\mathbf{P}(X^* \in B \cap \sigma = s) = \mathbf{P}(X^* \in B)\mathbf{P}(\sigma = s).$$

Et comme $\mathbf{P}(\sigma = s) = \frac{1}{n!}$, cela revient à montrer que, pour toute permutation $s \in \mathfrak{S}_n$

$$\mathbf{P}(X^* \in B) = n!\mathbf{P}(X^* \in B \cap \sigma = s)$$

Comme les X_i sont indépendantes et de même loi, elles sont interchangeables et donc, pour tout s et tout B,

$$\begin{aligned}\mathbf{P}\Big(X^* \in B \cap \sigma = s\Big) &= \mathbf{P}\Big(X_{s(1)} < \ldots < X_{(s(n)}, (X_{s(1)}, \ldots, X_{s(n)}) \in B\Big) \\ &= \mathbf{P}\Big(X_1 < \ldots < X_n, (X_1, \ldots, X_n) \in B)\end{aligned}$$

D'autre part, le théorème des probabilités totales permet d'écrire :

$$\begin{aligned}\mathbf{P}(X^* \in B) &= \sum_{s \in \mathfrak{S}_n} \mathbf{P}(X^* \in B \cap \sigma = s) \\ &= \sum_{s \in \mathfrak{S}_n} \mathbf{P}\Big(X_1 < \ldots < X_n, (X_1, \ldots, X_n) \in B) \\ &= n!\mathbf{P}\Big(X_1 < \ldots < X_n, (X_1, \ldots, X_n) \in B) \\ &= n!\mathbf{P}\Big(X^* \in B \cap \sigma = s\Big)\end{aligned}$$

où s est une permutation quelconque. □

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

Remarque 3.18. *Pour tout s fixé (non aléatoire) dans $\mathfrak{S}_n$ on a*

$$(X_{s(1)}, \ldots, X_{s(n)}) \sim (X_1, \ldots, X_n)$$

Mais ça n'est pas vrai si la permutation est aléatoire (à moins qu'elle ne soit indépendante des X_i). Par exemple, on a évidemment

$$X^* = (X_{\sigma(1)}, \ldots, X_{\sigma(n)}) \nsim (X_1, \ldots, X_n)$$

Proposition 3.19. *Soient $X_1, \ldots, X_n$ n v.a. i.i.d. de loi continue de vecteur des rangs $R_X = (R_1, \ldots, R_n)$. Et, pour tout entier s tel que $1 \leq s \leq n$, et pour toute suite d'entiers distincts $(r_1, \ldots, r_s)$ dans $\{1, \ldots, n\}$, on a*

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

En particulier, pour tout $i \in \{1, \ldots, n\}$, R_i suit une loi uniforme sur $\{1, \ldots, n\}$.

Démonstration. Pour simplifier, considérons d'abord trois cas simples.

— $s = n$: $\mathbf{P}\Big((R_1, \dots, R_n) = (r_1, \dots, r_n)\Big) = \mathbf{P}(R = r)$ où $r = (r_1, \dots, r_n) \in \mathfrak{S}_n$.
Donc, d'après le théorème précédent, on a

$$\mathbf{P}\Big((R_1, \dots, R_n) = (r_1, \dots, r_n)\Big) = \frac{1}{n!},$$

ce qui est bien le résultat annoncé.

— $s = 1 : \mathbf{P}(R_1 = s) = \mathbf{P}($ « X_1 est le s-ème plus petit élément de l'échantillon » $) = \frac{1}{n}$, toujours du fait que les X_i sont interchangeables.

— $s = 2$: alors

$$\mathbf{P}\Big((R_1, R_2) = (s_1, s_2)\Big) = \mathbf{P}(R_1 = s_1)\mathbf{P}(R_2 = s_2 \mid R_1 = s_1) = \frac{1}{n}\frac{1}{n-1}$$

Le cas général se traite de la même manière que le cas $s = 2$.

□

3.6.2 Test de Mann-Whitney

Nous allons maintenant décrire le **test de Wilcoxon de la somme des rangs**, encore appelé **test de Mann-Whitney.**

On se donne deux échantillons U_1^n et V_1^p tels que

1. $U_1, \dots, U_n \overset{iid}{\sim} U$ et $V_1, \dots, V_p \overset{iid}{\sim} V$
2. U_1^n et V_1^p sont **indépendants**
3. U et V sont diffuses.

A noter que les échantillons ne sont pas forcément de même taille.

On note F la fonction de répartition des U_i et G celle des V_j. On veut tester

$$H_0 : F = G$$

Ce n'est donc pas un test sur la médiane ou la moyenne contrairement aux tests précédents (signe et Wilcoxon des rangs signés). Cependant l'alternative **n'est pas** $F \neq G$. D'ailleurs si vous faites le test avec la commande R associée, il sera écrit `alternative hypothesis: true location shift is not equal to 0`.

On suppose en fait que U et V ont la même loi, à un paramètre de position près, c'est-à-dire

— soit U et V ont la même loi (H_0)

— soit U a tendance à prendre des valeurs plus grandes que V, ou le contraire (H_1).

Autrement dit

$$H_0 : F = G \text{ contre } H_1 : \exists \theta \neq 0 \text{ tel que } F(\cdot) = G(\cdot - \theta)$$

Exemple 3.20. *veut tester un nouveau médicament par rapport à un ancien médicament. On donne le premier à un groupe de n personnes, et le deuxième à un groupe de p personnes, ces deux groupes étant cette fois-ci indépendants. On veut voir si le nouveau médicament est plus efficace que l'ancien.*

Remarque 3.21. *On peut voir le test de Student comme un test d'égalité en loi, quand on suppose que les données sont gaussiennes et ne peuvent différer (éventuellement) que par leur moyenne. En ce sens le test de Mann-Whitney peut être vu comme vu comme une version non-paramétrique (et plus généralement robuste) du test de Student sur deux échantillons indépendants.*

On met les deux échantillons ensemble pour former un seul échantillon global de taille $n+p$: $(U_1, \ldots, U_n, V_1, \ldots, V_p)$. On classe ensuite les variables $\{U_i, V_j\}$ par leur rang global dans cet échantillon global : cela donne un vecteur de rangs que l'on note $R_{U,V}$. On note $R_1, \ldots, R_n$ les rangs associés aux variables U_i et $S_1, \ldots, S_p$ les rangs associés aux variables V_j.

Exemple 3.22. *soient* $U_1 = 3.5 \quad U_2 = 4.7 \quad U_3 = 1.2 \quad V_1 = 0.7 \quad V_2 = 3.9$. *Alors on a :* $V_1 < U_3 < U_1 < V_2 < U_2$.

$$R_1 = 3, \quad R_2 = 5, \quad R_3 = 2, \quad S_1 = 1, \quad S_2 = 4$$

On pose

$$\Sigma_1 = R_1 + R_2 + \ldots + R_n, \quad \Sigma_2 = S_1 + S_2 + \ldots + S_p$$

Principe : pour simplifier, prenons d'abord le cas simple où les deux échantillons sont de même taille. Alors, sous H_0, on s'attend à ce que Σ_1 et Σ_2 soit à peu près égaux. Pour fixer les idées, imaginons que l'alternative corresponde au fait que les U_i ont tendance à prendre des valeurs supérieures aux V_i. Alors, sous H_1, les rangs R_i des U_i dans l'échantillon global seront dans l'ensemble supérieurs aux rangs S_j des V_j dans l'échantillon global. Donc sous H_1, Σ_1 sera "grand" (c'est-à-dire "anormalement grand" par rapport à ce qui se passe sous H_0).

Maintenant, même si les échantillons ne sont pas de même taille, sous H_1, Σ_1 aura tendance à être anormalement grand par rapport à ce qui se passe sous H_0.

Plus généralement, si on pense que U et V n'ont pas la même loi et que l'une des deux variables a tendance à prendre des valeurs supérieures à l'autre mais on n'a pas d'intuition sur laquelle des deux, alors s'attend à ce que Σ_1 soit "anormalement grand" ou "anormalement petit" (toujours par rapport à ce qui se passe sous H_0).

Maintenant la question est : qu'est-ce qu'une valeur "normale" sous H_0 ? Les résultats suivants répondent à cette question.

Proposition 3.23. *On a*

$$\frac{n(n+1)}{2} \leq \Sigma_1 \leq np + \frac{n(n+1)}{2}$$

$$\frac{p(p+1)}{2} \leq \Sigma_2 \leq np + \frac{p(p+1)}{2}$$

Sous $H_0 : F = G$, et sous les conditions 1, 2 et 3 ci-dessus, on a, pour tout i et tout j,

$$\mathbf{E}(R_i) = \mathbf{E}(S_j) = \frac{n+p+1}{2}$$

$$\mathbf{Var}(R_i) = \mathbf{Var}(S_j) = \frac{(n+p)^2 - 1}{12}$$

$$\mathbf{E}(\Sigma_1) = \frac{n(n+p+1)}{2}, \quad \mathbf{E}(\Sigma_2) = \frac{p(n+p+1)}{2}$$

$$\mathbf{Var}(\Sigma_1) = \mathbf{Var}(\Sigma_2) = \frac{np(n+p+1)}{12}$$

Démonstration.

$$\Sigma_1 \geq 1 + 2 + \ldots + n = \frac{n(n+1)}{2}$$

et

$$\Sigma_2 \geq 1 + 2 + \ldots + p = \frac{p(p+1)}{2}$$

Comme $\Sigma_1 + \Sigma_2 = \sum_{i=1}^{n+p} i = \frac{(n+p)(n+p+1)}{2}$, on a

$$\Sigma_1 \leq \frac{(n+p)(n+p+1)}{2} - \frac{p(p+1)}{2} = np + \frac{n(n+1)}{2}$$

De même

$$\Sigma_2 \leq np + \frac{p(p+1)}{2}$$

On se place désormais sous H_0.

Alors toutes les variables $U_1, \ldots, U_n, V_1, \ldots, V_p$ sont i.i.d. Donc on a un échantillon global i.i.d. de taille $N = n + p$. D'après le théorème 7 (cas $s = 1$), pour tout i, la v.a. R_i, qui est donc une composante du vecteur de rang de l'échantillon global $R_{U,V}$, suit une loi uniforme sur $\{1, \ldots, N\}$. De même pour chaque S_j. Donc l'espérance et la variance de chacune de ces variables est simplement l'espérance et la variance d'une variable de loi uniforme sur $\{1, \ldots, N\}$. Donc on a, pour tout $i = 1, \ldots, n$, et pour tout $j = 1, \ldots, p$,

$$\mathbf{E}(R_i) = \mathbf{E}(S_j) = \frac{1}{N}\sum_{i=1}^{N} i = \frac{N+1}{2} = \frac{n+p+1}{2}$$

et

$$\begin{aligned}\mathbf{Var}(R_i) = \mathbf{Var}(S_j) &= \frac{1}{N}\sum_{i=1}^{N} i^2 - \Big(\frac{N+1}{2}\Big)^2 \\ &= \frac{(N+1)(2N+1)}{6} - \Big(\frac{N+1}{2}\Big)^2 \\ &= \frac{N^2-1}{12}\end{aligned}$$

Donc on a

$$\mathbf{E}\Sigma_1 = \mathbf{E}(R_1 + \ldots R_n) = n\mathbf{E}R_1 = \frac{n(n+p+1)}{2}$$

et

$$\mathbf{E}\Sigma_2 = \mathbf{E}(S_1 + \dots S_p) = p\mathbf{E}S_1 = \frac{p(n+p+1)}{2}$$

Il reste le calcul des variances de Σ_1 et Σ_2. Attention les variables R_i et S_j sont de même loi mais pas indépendantes ! On a

$$\mathbf{Var}\Sigma_1 = \mathbf{Var}(R_1 + \dots + R_n) = \sum_{i=1}^{n} \mathbf{Var}(R_i) + \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{Cov}(R_i, R_j)$$

On a déjà calculé les variances, il faut donc calculer maintenant les covariances. Soit donc $i \neq j$,

$$\begin{aligned}\mathbf{Cov}(R_i, R_j) &= \mathbf{E}\Big[(R_i - \mathbf{E}R_i)(R_j - \mathbf{E}R_j)\Big] \\ &= \sum_{1 \leq k,l \leq N, k \neq l} (k - \frac{N+1}{2})(l - \frac{N+1}{2})\mathbf{P}(R_i = k, R_j = l)\end{aligned}$$

Or, (R_i, R_j) a la même loi que (R_1, R_2) et, d'après le théorème 7, on a, pour $k \neq l$,

$$\mathbf{P}(R_1 = k, R_2 = l) = \frac{1}{N(N-1)}$$

Donc

$$\mathbf{Cov}(R_i, R_j) = \mathbf{Cov}(R_1, R_2) = \frac{1}{N(N-1)} \sum_{k \neq l} (k - \frac{N+1}{2})(l - \frac{N+1}{2})$$

Or on a

$$\begin{aligned}\sum_{k \neq l} (k - \frac{N+1}{2})(l - \frac{N+1}{2}) &= \sum_{1 \leq k,l \leq N} (k - \frac{N+1}{2})(l - \frac{N+1}{2}) - \sum_{k=1}^{N} (k - \frac{N+1}{2})^2 \\ &= \Big[\sum_{k=1}^{N} (k - \frac{N+1}{2})\Big]^2 - \sum_{k=1}^{N} (k - \frac{N+1}{2})^2\end{aligned}$$

De plus

$$\sum_{k=1}^{N} (k - \frac{N+1}{2}) = \sum_{k=1}^{N} k - \frac{N(N+1)}{2} = 0$$

Donc

$$\mathbf{Cov}(R_i, R_j) = -\frac{1}{N(N-1)} \sum_{k=1}^{N} (k - \frac{N+1}{2})^2 = -\frac{1}{N-1} \mathbf{Var} R_1$$

Et finalement

$$\begin{aligned}
\mathbf{Var}(\Sigma_1) &= \sum_{i=1}^{n} \mathbf{Var}(R_i) + \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{Cov}(R_i, R_j) \\
&= n\mathbf{Var}(R_1) + n(n-1)\mathbf{Cov}(R_1, R_2) \\
&= n\mathbf{Var}(R_1) - \frac{n(n-1)}{N-1}\mathbf{Var}(R_1) \\
&= \frac{n(N-n)}{N-1}\frac{N^2-1}{12} \\
&= \frac{n(N-n)(N+1)}{12} \\
&= \frac{np(n+p+1)}{12}
\end{aligned}$$

Le calcul de $\mathbf{Var}(\Sigma_2)$ se déduit de celui de $\mathbf{Var}(\Sigma_1)$ en échangeant les rôles de n et p. □

Au vu de la proposition, on considère naturellement les statistiques suivantes :

$$M_U = \Sigma_1 - \frac{n(n+1)}{2} \in \{0, 1, \ldots, np\}$$

$$M_V = \Sigma_2 - \frac{p(p+1)}{2} \in \{0, 1, \ldots, np\}$$

Proposition 3.24. *On suppose les conditions 1,2 et 3 du début de la sous-section vérifiées. Alors*

1. $M_U + M_V = np \quad p.s.$
2. *Sous H_0 : $F = G$, la loi de M_U est symétrique par rapport à $\frac{np}{2}$*
3. *Sous H_0 : $F = G$, $M_U \sim M_V$.*
4. *M_V est égal au nombre de paires (U_i, V_j), parmi toutes les paires possibles, telles que $U_i < V_j$.*

Démonstration.

1. $\Sigma_1 + \Sigma_2$ est égal à la somme des rangs de toutes les N variables. Donc $\Sigma_1 + \Sigma_2 = \sum_{i=1}^{N} i = \frac{N(N+1)}{2}$. Donc $M_U + M_V = \Sigma_1 - \frac{n(n+1)}{2} + \Sigma_2 - \frac{p(p+1)}{2} = \frac{(n+p)(n+p+1)}{2} - \frac{n(n+1)}{2} - \frac{p(p+1)}{2} = np$.
2. On se place sous H_0. On introduit $(S'_1, \ldots, S'_p)$ les rangs des $V_1, \ldots, V_p$ dans l'échantillon global lorsque les variables sont ordonnées de façon décroissante. On montre exactement de la même manière que dans la proposition 3.19 que, pour toute suite d'entiers distincts $(r_1, \ldots, r_p)$ dans $\{1, \ldots, N\}$, on a

$$\mathbf{P}\Big((S'_1, \ldots, S'_p) = (r_1, \ldots, r_p)\Big) = \frac{1}{N(N-1)\ldots(N-p+1)}.$$

Donc $(S'_1, \ldots, S'_p) \sim (S_1, \ldots, S_p)$. Donc

$$\Sigma_2 \sim \Sigma'_2 \tag{3.1}$$

où $\Sigma_2' = S_1' + \ldots + S_p'$. Or, pour tout $j \in [p]$, $S_j' = N + 1 - S_j$. Donc

$$\Sigma_2' = N+1-S_1+\cdots+N+1-S_j+\ldots+N+1-S_p = (N+1)p - \Sigma_2 \quad (3.2)$$

Ainsi, en combinant (3.1) et (3.2), on obtient

$$\Sigma_2 \sim (N+1)p - \Sigma_2$$

Ceci implique que $M_V + \frac{p(p+1)}{2} \sim (n+p+1)p - (M_V + \frac{p(p+1)}{2})$. Autrement dit on a

$$M_V \sim (n+p+1)p - p(p+1) - M_V = np - M_V,$$

ce qui se traduit par : M_V est symétrique par rapport à $\frac{np}{2}$.

3. On se place sous H_0. En combinant l'item 2 et l'item 1 on a

$$M_V \sim np - M_V = np - (np - M_U) = M_U.$$

4. La démonstration de l'item 4 est admise. Cependant, on donne la preuve ici pour les étudiants intéressés. On se place sous H_0. Sans perte de généralité, on suppose que σ est égale à l'identité, autrement dit $v_1 < \ldots < v_p$. On va compter, pour tout $j \in [p]$, le nombre d'éléments du premier échantillon $u_1, \ldots, u_n$ qui sont plus petits que v_j. On rappelle que, pour tout $j \in [p]$, s_j est le rang de v_j dans l'échantillon global $u_1, \ldots, u_n, v_1, \ldots, v_p$. Commençons par $j = 1$: il y a $s_1 - 1$ valeurs plus petites que v_1 dans l'échantillon global. Ces valeurs ne peuvent être que des valeurs du premier échantillon car v_1 est la plus petite valeur de l'échantillon $v_1, \ldots, v_p$. Donc il y a $s_1 - 1$ couples (u_i, v_1) tels que $u_i < v_j$. Passons au cas $j = 2$. Il y a s_2 valeurs de l'échantillon global qui sont plus petites que v_2, et comme il y a une seule valeur (c'est v_1) du second échantillon qui est plus petite que v_2, il y a $s_2 - 2$ couples (u_i, v_2) tels que $u_i < v_2$. De manière générale, pour tout $j \in [p]$ fixé, il y a $s_j - j$ couples (u_i, v_j) tels que $u_i < v_j$. Donc le nombre total de couples (u_i, v_j) tels que $u_i < v_j$ est égal à

$$s_1 - 1 + \ldots + s_j - j + \ldots + s_p - p = \sum_{j=1}^{p} s_j - \sum_{j=}^{p} j = \Sigma_2 - \frac{p(p+1)}{2} = M_V.$$

□

Théorème 3.25. *On suppose les conditions 1, 2 et 3 vérifiées.*

Les lois de M_U et M_V sont libres sous H_0 : $F = G$ (i.e. elles ne dépendent pas de F, fonction de répartition des U_i et des V_j). Elles ne dépendent que de n et p. Asymptotiquement, sous H_0, quand n et p tendent vers $+\infty$,

$$\frac{M_U - \mathbf{E}(M_U)}{\sqrt{\mathbf{Var}(M_U)}} \xrightarrow{loi} N(0,1)$$

(et la même chose pour M_V puisque $M_U \sim M_V$)

$$\mathbf{E}(M_U) = \frac{np}{2} \quad \mathbf{Var}(M_U) = \frac{np(n+p+1)}{12}.$$

Démonstration. On admet la convergence en loi.

$M_U = \Sigma_1 - \frac{n(n+1)}{2} = R_1 + \ldots + R_n - \frac{n(n+1)}{2}$ est une fonction du vecteur $(R_1, \ldots, R_n)$. On connait la loi de ce vecteur sous H_0, cette loi est donnée par le théorème 7 : pour toute suite d'entiers $(r_1, \ldots, r_n)$ à valeur dans $[N]$, on a

$$\mathbf{P}_{F=G}\Big((R_1, \ldots, R_n) = (r_1, \ldots, r_n)\Big) = \frac{1}{N(N-1)\ldots(N-n+1)}$$

On voit donc qu'elle ne dépend pas de F et ne dépend que de n et p.

L'espérance et la variance de M_U se déduisent l de l'espérance et de la variance de Σ_1, qu'on a obtenues dans la proposition 3.23. □

Remarque 3.26. *(Test exact ou asymptotique) Pour les valeurs de n et p plus petites que 10, la loi de ω_X est tabulée. Pour les grandes valeurs, on utilise l'approximation gaussienne.*

Remarque 3.27. *(Correction de continuité)*
Supposons que la statistique de test T_n prenne des valeurs discrètes, disons entières, mais n étant grand, la loi de T_n peut être approchée par une loi gaussienne, qui est une loi continue. Alors $\mathbf{P}(T_n \geq p) = \mathbf{P}(T_n \geq p-u)$, pour tout $u \in [0,1[$ et pour tout $p \in \mathbb{N}$. La correction du continu consiste à remplacer la valeur p dans l'approximation gaussienne par $p-0,5$: plus précisément, si on a $a_n(T_n - t_n) \overset{loi}{\to} \mathcal{N}(0,1)$, on approche comme suit :

$$\mathbf{P}(T_n \geq p) = \mathbf{P}(a_n(T_n - t_n) \geq a_n(p - t_n)) \approx 1 - \Phi(a_n(p - 0.5 - t_n))$$

Avec R

C'est exactement la même formulation que le test de Wilcoxon des rangs signés, sauf qu'on met `paired=F`. C'est en fait `False` par défaut.

Dans l'exemple lié aux salaires, en supposant cette fois que les échantillons de salaires d'hommes et de femmes sont i.i.d. et indépendants entre eux, on peut utiliser

```
wilcox.test(data=salaires,hommes~femmes,alternative="greater")
```

Quelques détails de plus : l'argument `exact` indique si on veut le test exact ou l'approximation gaussienne. Cet argument est par défaut à `true` si l'un des échantillons a une taille supérieure à 50 et à `false` dans le cas contraire. L'argument `correct` indique si on veut la correction de continuité quand on utilise l'approximation gaussienne. Il est par défaut à `TRUE`.

Remarque 3.28. *En plus d'être adaptés à un plus grand nombre de lois, les tests basés sur les rangs sont plus robustes à la présence d'observations aberrantes, ou "outliers", dans l'échantillon (penser à la différence médiane/moyenne) .*

Remarque 3.29. *Certains auteurs préconisent, avant l'utilisation éventuelle de Mann-Whitney, de tester si les deux échantillons ont le même paramètre d'échelle (même variance par exemple). En effet, si on considère Mann-Whitney comme un test d'égalité en loi supposé détecter une différence de position, alors il ne semble pas judicieux d'utiliser Mann-Whitney si les échelles diffèrent (ni d'ailleurs si la forme général de l'histogramme est très différente). Si on fait ce test dans cette optique-là, alors il parait judicieux de vérifier cette condition sur un graphique par exemple (de toute façon il faut toujours représenter les données avant toute chose). Il existe aussi des tests d'échelle (par exemple le test de Levene, qui a des propriétés de robustesse). Citation de Zimmerman (2004) : "for a wide variety of non-normal distributions, especially skewed distributions, the Type I error probabilities of both the t test and the Wilcoxon-Mann-Whitney test are substantially inflated by heterogeneous variances, even when sample sizes are equal."*

Cependant, certains praticiens utilisent le test de Mann-Whitney comme un test pour savoir en gros si l'une des deux populations (U ou V) a tendance à prendre des valeurs plus grandes que l'autre. Il n'est alors pas vu comme un test d'égalité en loi. A ce moment-là, on n'a pas besoin de vérifier si les lois semblent les mêmes à un paramètre de position près (et donc pas besoin de vérifier que l'échelle est la même).

Remarque 3.30. *Une question naturelle : quel type de test (paramétrique/ non paramétrique) choisir ?*

Souvent, si le modèle paramétrique est correct, les tests paramétriques sont plus puissants que les tests non paramétriques. Cependant, ils sont aussi plus contraignants, car il faut vérifier les conditions d'application qui sont plus nombreuses dans ce cas. On choisira généralement un test non paramétrique lorsque

- *les conditions d'application du test paramétrique ne sont pas vérifiées*
- *ou il est impossible de vérifier ces conditions.*

"On préconise aussi parfois l'utilisation de tests non paramétriques dans le cas de petits échantillons, mais le fait d'avoir de petits échantillons ne justifie pas à lui seul l'utilisation de tests non paramétriques : si les échantillons sont petits, mais que ce type de données a été suffisamment étudié pour que l'on puisse supposer la normalité de la distribution, pas de problème pour utiliser des tests paramétriques. Ce type de conseils est en général donné par prudence, parce que le petit nombre de données ne permet pas de vérifier, à partir de l'échantillon, la normalité de la distribution. Dans le doute, on peut donc choisir un test non paramétrique. Les tests non paramétriques sont certes un peu moins puissants que les tests paramétriques, mais leur efficacité relative reste bonne" (citation de C. Chabanet, cf biblio).

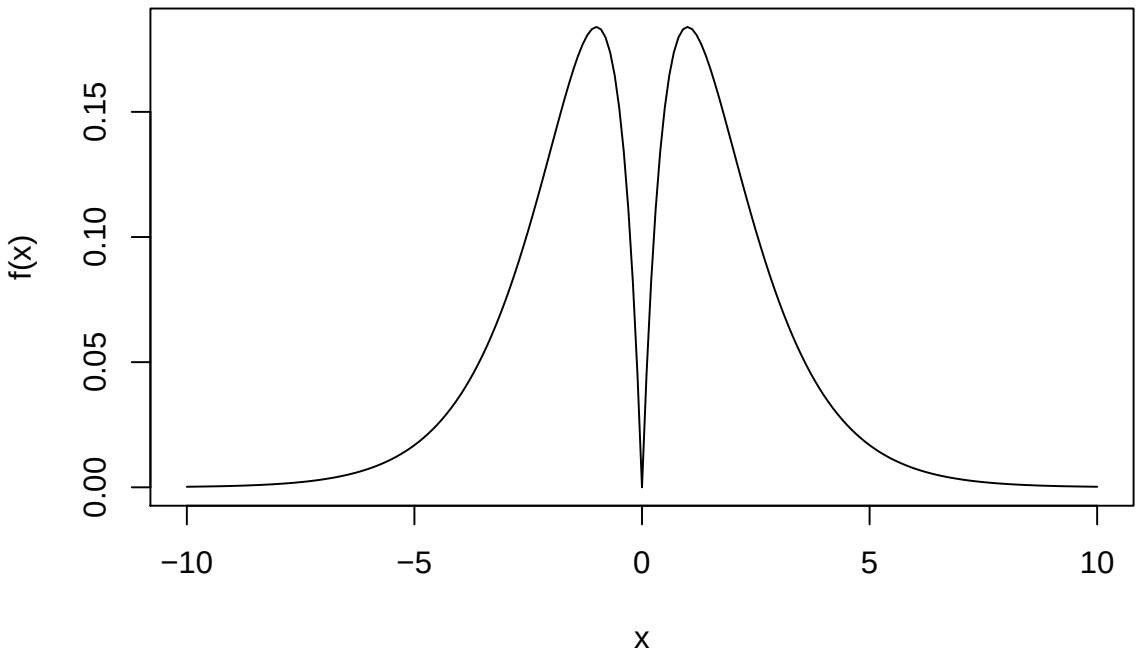
Remarque 3.31. *De la même manière que le test de Student pour comparer les moyennes de deux échantillons se généralise à plus de deux échantillons par l'analyse la variance (cf cours de MLG, cas de la régression sur une variable qualitative), "l'équivalent" du test de Wilcoxon de la somme des rangs pour plus de deux échantillons existe et s'appelle le test de Kruskal-Wallis.*

A nouveau, pour Kruskal-Walllis, les données sont remplacées par leur rang dans l'échantillon global mais cette fois on calcule les sommes de carrés intra-groupe. L'idée est que sous l'hypothèse nulle (la loi ne dépend pas du groupe) le problème se réduit à nouveau à un problème combinatoire (il y a une uniformité sous-jacente).

Remarque 3.32. *Comparons maintenant les tests de Kolmogorov-Smirnov (noté KS) et le test de Mann-Whitney (noté MW). Le test KS est sensible à tout changement dans les deux distributions. Des différences substantielles dans la forme, l'étendue ou la médiane vont amener à une petite p-valeur. En revanche, le test MW test est seulement sensible à un changement de position (cf plus loin pour une illustration).*

Illustration : On regarde ci-dessous les performances respectives des tests de Kolmogorov-Smirnov et de Mann-Whitney sur un cas particulier. Plus précisément, on utilise deux échantillons qui ne sont pas de même loi : l'un est de loi normale, l'autre est un mélange de deux lois gamma dont on représente la densité ci-dessous. On utilise KS et MW pour tester l'égalité des lois sur ces deux échantillons. On regarde la p-valeur de chaque test.

```
f=function(x){0.5*(dgamma(x,shape=2,rate=1)+dgamma(-x,2,1))}#densité de probabilité,
#mélange d'une loi gamma et de sa symétrisée
x=seq(-10,10,by=0.1)
plot(x,f(x),type="l")
```




```
#simulation de 1000 expériences de test  et calcul des p-valeurs
KS=rep(0,1000)
MW=rep(0,1000)
for (i in 1:1000){
  z=rnorm(100)#simulation de N(0,1)
  # simulation d'un échantillon y de loi de densité f
  #simulation d'un échantillon t de loi gamma(2,1)
  t=c(rgamma(100,shape=2,rate=1))
  #simulation de Rademacher
  rad=2*rbinom(100,size=1,0.5)-1
  y=rad*t
  KS[i]=ks.test(y,z)$p.value #p-valeur du test de kolmogorov Smirnov
  MW[i]=wilcox.test(y,z)$p.value# pvaleur du test de Mann-Whitney
}
#moyennes des p-valeur de chaque test sur les 1000 simulations

mean(KS)
```

```
## [1] 0.002863565
```

```
mean(MW)
```

```
## [1] 0.4830098
```

Chapitre 4

Estimation de densités par estimateurs à noyau

4.1 Quelques rappels d'analyse utiles pour les chapitres 4 et 5

Définition de la différentiabilité : soit $\ell : \mathbb{R}^m \to \mathbb{R}^p$. L'application ℓ est différentiable en u s'il existe une application linéaire $D\ell(u) : \mathbb{R}^m \to \mathbb{R}^p$ (qu'on peut donc représenter par une matrice élément de $M_{pm}(\mathbb{R})$) telle que :

$$\forall \epsilon > 0, \exists \delta > 0 : \|x - u\| \leq \delta \longrightarrow \|\ell(x) - \ell(u) - D\ell(u)(x - u)\| \leq \epsilon \|x - u\|$$

Formule de Taylor-Lagrange : soit $f : I \to \mathbb{R}$ où I est un intervalle de $\mathbb{R}$. On suppose que f est n fois dérivable sur I. Alors pour tout x et y de l'intérieur de I, il existe $\eta \in]0, 1[$ tel que

$$f(y) = \sum_{k=0}^{n-1} f^{(k)}(x) \frac{(y-x)^k}{k!} + f^{(n)}(x + \eta(y - x)) \frac{(y-x)^n}{n!}$$

Formule de Taylor avec reste intégral : on suppose cette fois que $f \in \mathcal{C}^n(I)$ (n fois continument dérivable) alors, pour tout couple (x, y) de l'interieur de I,

$$f(y) = \sum_{k=0}^{n-1} f^{(k)}(x) \frac{(y-x)^k}{k!} + \int_x^y \frac{(y-t)^{n-1}}{(n-1)!} f^{(n)}(t) dt$$

4.2 Introduction

Dans tout le chapitre, l'objectif sera d'estimer une densité f. Pour cela, on s'appuiera sur un n-échantillon iid $X = (X_1, \ldots, X_n)$ où chacune des variables X_i admet la densité f (par rapport à la mesure de Lebesgue).

Mesure de la qualité d'un estimateur :

1. Définition d'une distance sur l'espace des fonctions :

— Distance L_p : $d(f,g) = \|f-g\|_p = \left(\int |f(x)-g(x)|^p dx \right)^{\frac{1}{p}}$
Cas usuel $p=2$ ou $p=1$.
— distance L_∞ : $d(f,g) = \|f-g\|_\infty = \sup_{x\in\mathbb{R}} |f(x)-g(x)|$.
— Distance ponctuelle en x_0 : $d(f,g) = |f(x_0)-g(x_0)|$

2. Définition d'une fonction de perte $\omega : \mathbb{R} \to \mathbb{R}^+$ telle que ω est convexe et $\omega(0) = 0$.
Exemple : $\omega(x) = x^2$.
3. Définition du risque d'un estimateur $\hat{f}_n$:

$$R(\hat{f}_n, f) = \mathbf{E}[\omega(d(\hat{f}_n, f))]$$

où $\mathbf{E}$ désigne l'espérance sous la loi des X_i.

Attention en non-paramétrique, on estime donc des fonctions et non plus des vecteurs (dimension infinie contre dimension finie en gros). Il y a deux "variables" : la variable x et le vecteur aléatoire $X = (X_1, \ldots, X_n)$. On a donc : $\hat{f}_n = \hat{f}_n(x, X)$. On a donc à la fois, pour chaque valeur de X, une fonction en x (ou plus généralement un élément de L_p) et, pour chaque valeur fixée de x, une variable aléatoire réelle.

Exemples usuels :
— $d(f,g) = |f(x_0)-g(x_0)|, \quad \omega(x) = x^2$:

$$R(\hat{f}_n, f) = \mathbf{E}[|\hat{f}_n(x_0) - f(x_0)|^2]$$

— $d(f,g) = \|f-g\|_2, \quad \omega(x) = x^2$

$$R(\hat{f}_n, f) = \mathbf{E}[\|\hat{f}_n - f\|_2^2]$$

On cherche à déterminer $\hat{f}_n$ tel que $R(\hat{f}_n, f)$ soit minimal. Comme expliqué dans l'introduction, on ne suppose pas que la fonction de densité f appartient à une famille paramétrique. On va faire une hypothèse moins précise : f appartient à une classe fonctionnelle qu'on note $\mathcal{F}$. On peut alors définir un risque, qu'on appelle risque minimax de $\hat{f}_n$ sur la classe $\mathcal{F}$, par

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

On va donc chercher un estimateur $\hat{f}_n$ tel que le risque $R(\hat{f}_n, \mathcal{F})$ tende vers zéro le plus vite possible quand n tend vers l'infini.

Définition 4.1. *soit $(r_n)_n$ une suite et une constante C telles que*

$$\forall n \quad R(\hat{f}_n, \mathcal{F}) \leq C r_n$$

On dit que la suite d'estimateurs $(\hat{f}_n)_n$ atteint la vitesse (ou le taux) r_n sur la classe $\mathcal{F}$ (pour la distance d et la perte ω.)

Nous verrons que la vitesse sera d'autant plus grande que la classe $\mathcal{F}$ sera une classe de régularité élevée.

Exemple de classes de fonctions : $\mathcal{C}^k$, la classe de Holder (cf définition ci-dessous), boule dans un espace de Sobolev (cf cours d'analyse fonctionnelle).

Définition 4.2. *Si $\beta \in \mathbb{R}$ on note $\lfloor \beta \rfloor$ l'entier naturel qui soit le plus grand entier strictement inférieur à β.*

ex : si $\beta = 3,5$ alors $\lfloor \beta \rfloor = 3$ et si $\beta = 4$ alors $\lfloor \beta \rfloor = 3$.

Définition 4.3. *Pour tout $\beta > 0$ et tout $L > 0$, on définit la classe de Holder de régularité β et de rayon L par*

$$\Sigma(\beta, L) = \{g : \mathbb{R} \to \mathbb{R} \text{ } t.q. \text{ } g \text{ } est \text{ } \lfloor \beta \rfloor \text{ } fois \text{ } dérivable \text{ } et \\ \forall (x,y) \in \mathbb{R}^2 \quad |g^{(\lfloor \beta \rfloor)}(y) - g^{(\lfloor \beta \rfloor)}(x)| \leq L|x-y|^{\beta - \lfloor \beta \rfloor}\}$$

Quand on intersecte $\Sigma(\beta, L)$ avec l'ensemble des densités, on note $\Sigma_d(\beta, L)$ cette intersection.

Remarque 4.4.
— *Si $\beta = 1$ on obtient l'ensemble des fonctions lipschitziennes .*
— *Si $\beta > 1$ alors $f' \in \Sigma(\beta - 1, L)$.*

Proposition 4.5. *(admise) Soit $\beta > 0$ et $L > 0$, il existe une constante $M(\beta, L)$ telle que*

$$\sup_{f \in \Sigma_d(\beta, L)} \|f\|_\infty = \sup_{x \in \mathbb{R}} \sup_{f \in \Sigma_d(\beta, L)} f(x) \leq M(\beta, L)$$

4.3 Estimation non paramétrique de la densité

L'approche classique pour estimer une densité est de supposer un modèle paramétrique : par exemple, en dimension 1, on représente les données par un histogramme, et si la courbe est en cloche avec des queues légères, on conclut qu'il y a de fortes chances que le modèle suive une loi gaussienne. Il n'y a alors plus qu'à estimer la moyenne et la variance (μ, σ^2), c'est-à-dire un paramètre de dimension 2. On peut aussi se trouver dans un cas où on a des connaissances a priori sur les données, nous amenant à poser encore une loi paramétrique (ex typique : nombre de voitures passant par un carrefour par jour, représenté en général par une loi de poisson).

Il y a plusieurs problèmes possibles avec cette approche : en dimension supérieure à 2 il sera difficile de représenter les données et d'intuiter une loi connue, parfois on n'a pas de connaissances a priori sur le sujet etc.

De plus, si on se trompe de modèle, on arrivera à une interprétation erronée des données.

Un modèle non paramétrique est moins rigide, et fait moins de suppositions a priori sur les données.

Evidemment, comme pour le cas des tests, si on a des connaissances a priori fiables sur les données nous indiquant un modèle paramétrique, il faut utiliser le modèle paramétrique. Autrement dit, si le modèle paramétrique choisi est correct, ou plus précisément suffisamment proche de la réalité, alors le modèle paramétrique sera en général meilleur qu'un modèle non paramétrique.

4.3.1 Un estimateur simple de la densité : l'histogramme

Supposons pour simplifier qu'on soit en dimension 1 et que les variables de l'échantillon soient à valeurs dans $[0,1]$ donc $f : [0,1] \to \mathbb{R}^+$.

On se donne un découpage de $[0,1]$ en un certain nombre de classes $]a_1, a_2], \ldots,]a_p, a_{p+1}]$. Pour simplifier encore, on suppose que les classes sont de même longueur $a_{i+1} - a_i = a_i - a_{i-1}$. Cette longueur est notée h. Estimer f par la méthode de l'histogramme consiste simplement à estimer f par une fonction constante sur chaque classe, cette constante étant liée à la proportion de X_i tombant dans cette classe. Plus exactement on pose, pour $t \in]a_j, a_{j+1}]$,

$$\hat{f}_n(t) = \frac{1}{nh}\mathbf{Card}\{i : X_i \in]a_j, a_{j+1}]\}$$

Pour voir très exactement d'où vient cette formule : on a, si f est égale à une constante c_j constante sur $]a_j, a_{j+1}]$,

$$F(a_{j+1}) - F(a_j) = \int_{a_j}^{a_{j+1}} f(t)dt = c_j h$$

Ensuite on approche la probabilité $F(a_{j+1}) - F(a_j)$, qui correspond à la probabilité que $X \in]a_j, a_{j+1}]$, par la proportion de X_j se trouvant dans $]a_j, a_{j+1}]$. On a alors

$$c_j = \frac{F(a_{j+1}) - F(a_j)}{h} \approx \frac{1}{nh}\mathbf{Card}\{i : X_i \in]a_j, a_{j+1}]\}$$

La performance de cet estimateur dépend fortement du nombre de classes.

Code R et illustration graphique du choix du nombre de classes.

On va illustrer l'importance de bien choisir le nombre de classes par un exemple faisant intervenir une densité bimodale. On va pour cela simuler un mélange de deux lois gaussiennes : la densité simulée est

$$f(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\Big(\exp(-\frac{(x-2)^2}{2}) + \exp(-\frac{(x-6)^2}{2})\Big)$$

On devrait donc, si l'approximation par l'histogramme est bien faite, se retrouver avec deux "cloches" qui se chevauchent un petit peu (écart-type=1) et qui sont centrées en 2 et 6 respectivement.

Simulation d'un échantillon de taille `n=500` de loi de densité f :

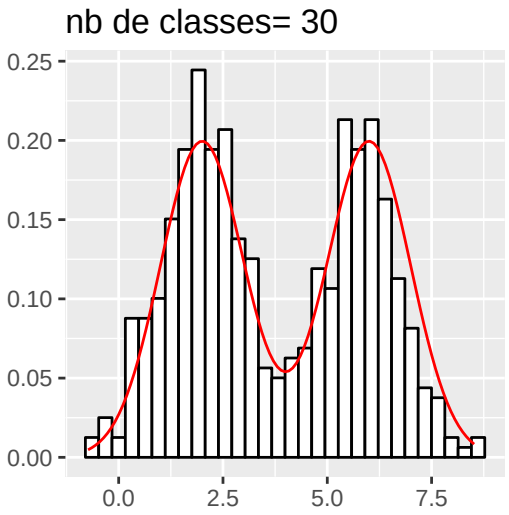
```
f=function(x){0.5*dnorm(x,mean=2)+0.5*dnorm(x,mean=6)}

sim=function(n){
X=rnorm(n,2,1)
Y=rnorm(n,6,1)
ber=rbinom(n=n,size=1,prob=0.5)
return(ber*X+(1-ber)*Y)}
Z=sim(500)
```

On estime la densité par un histogramme (on utilise ici la bibliothèque ggplot2) et on rajoute la vraie densité f en rouge :

```
library(ggplot2)
p<-ggplot(data.frame(x=Z),aes(x))+labs(x="",y="")
p1<-p+ geom_histogram(aes(y=..density..),color="black",fill="white")+
  stat_function(fun=f,col='red')+
  labs(title="nb de classes= 30")
p1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

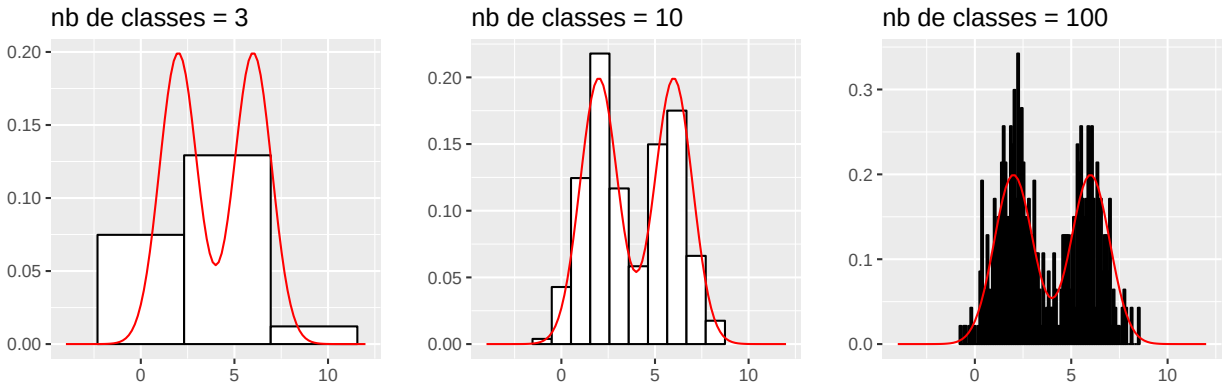


La fonction histogram dans ggplot calcule un histogramme avec 30 classes par défaut (ce qu'il signale d'ailleurs). Ce n'est donc pas la valeur optimale en général. Essayons avec d'autres valeurs du nombre de classes (=bins).

```
p1<-p+
  geom_histogram(aes(y=..density..),bins=3, color="black",fill="white")+
  stat_function(fun=f,col='red',xlim=c(-4,12))+
  labs(title="nb de classes = 3")
p2<-p+
  geom_histogram(aes(y=..density..),bins=10, color="black",fill="white")+
  stat_function(fun=f,col='red',xlim=c(-4,12))+
  labs(title="nb de classes = 10")
p3<-ggplot(data.frame(x=Z),aes(x))+
  geom_histogram(aes(y=..density..),bins=100, color="black",fill="white")+
  stat_function(fun=f,col='red',xlim=c(-4,12))+
  labs(title="nb de classes = 100",x="",y="")



library(gridExtra)#pour faire apparaitre les trois figures en même temps
grid.arrange(p1,p2,p3,nrow=1)
```



On peut aussi indiquer le pas h (`binwidth`) plutôt que le nombre de classes (bins).

On constate donc que, avec une fenêtre h trop petite, c'est-à-dire avec un trop grand nombre de classes, on fait apparaitre trop de variations souvent insignifiantes (variance trop grande). Au contraire avec une fenêtre h trop grande, on a une approche trop grossière (biais trop grand) et une distribution peu discriminante : en particulier ici on ne voit même plus qu'il s'agit d'une distribution bimodale . On voit qu'il faut trouver un compromis entre le biais (au carré) et la variance, compromis qu'on va illustrer plus en détail plus loin, par le calcul.

Il existe d'ailleurs dans R des estimations de la taille optimale du pas h, cf l'aide en ligne ou la page wikipedia sur l'histogramme. L'estimateur par histogramme étant présenté ici essentiellement à titre illustratif, nous ne donnons pas plus de détails sur le sujet. Des détails plus précis seront donnés pour l'estimateur qui nous intéresse vraiment : l'estimateur à noyau.

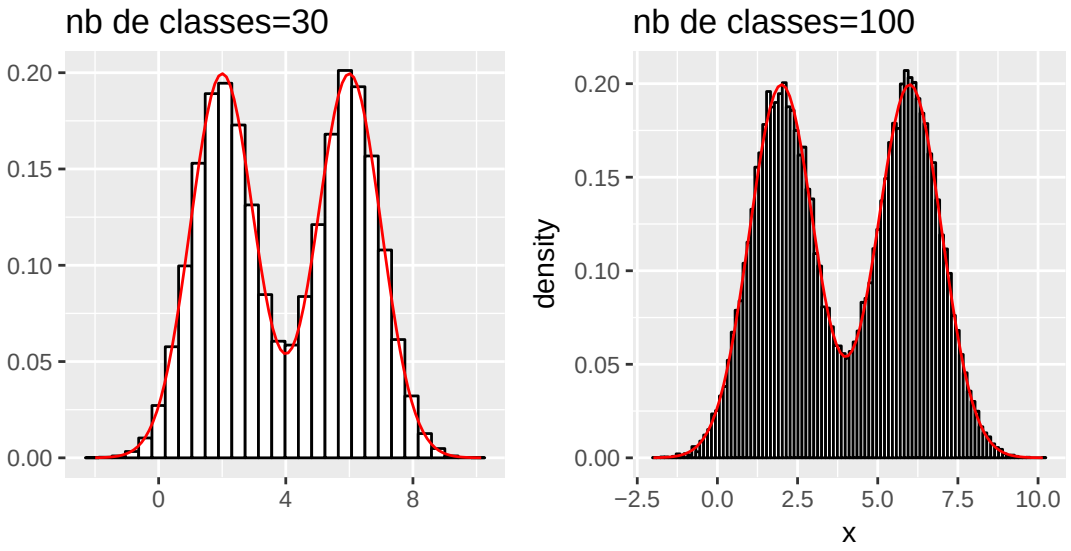
Evidemment le nombre optimal de classes dépend de n. Illustrons ceci en changeant la taille de l'échantillon : on passe de 500 à 50000.

```
Z=sim(50000)
p<-ggplot(data.frame(x=Z),aes(x))+labs(x="",y="")

p1<-p+
  geom_histogram(aes(y=..density..),color="black",fill="white")+
  stat_function(fun=f,col='red')+
labs(title="nb de classes=30")
p2<-ggplot(data.frame(x=Z),aes(x))+
```

```
  geom_histogram(aes(y=..density..),bins=100,color="black",fill="white")+
  stat_function(fun=f,col='red')+
labs(title="nb de classes=100")
grid.arrange(p1,p2,nrow=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



On voit donc qu'avec un nombre de classes égal à 100, on a, conrairement à précédemment, un très bon choix. La taille optimale du nombre de classes est croissante avec n, autrement dit, le pas h optimal décroit avedc n, ce que l'on va illustrer plus tard avec l'estimateur à noyau de fenêtre h.

Remarquez que l'on fait deux approximation successives : une première approximation quand on approche la densité par une fonction constante par morceaux, et ensuite une deuxième approximation quand on approche chaque constante à l'aide des données.

4.3.2 Estimateurs à noyaux

Un inconvénient de l'estimateur par histogramme précédent est que la fonction de densité résultante $\hat{f}_n$ n'est pas régulière : il s'agit d'une fonction constante par morceau, qui a donc des sauts aux extrémités de chaque classe. En général, la densité à estimer est plus lisse, au moins continue.

L'estimation par noyau a pour but de répondre à cet écueil.

Principe : Si f est continue en x (ce qui va être le cas pour les classes de fonctions qu'on va considérer) alors

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

L'idée est donc d'utiliser l'approximation suivante, pour h petit,

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

Pour estimer la densité f on peut donc passer par un estimateur $\hat{F}_n$ de la cdf F. Voyons ce qui se passe si on choisit comme estimateur la fonction de répartition empirique F_n. (On rappelle que $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \leq n}$) On choisit un $h > 0$ petit pour que l'approximation ci-dessus soit valable, et on pose

$$\tilde{f}_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{n}\sum_{i=1}^n \frac{1}{2h} \mathbb{1}_{X_i \in]x-h, x+h]}$$

Si on pose $K_0(x) = \frac{1}{2}\mathbb{1}_{]-1,1](u)}$ alors on a

$$\tilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h} K_0\left(\frac{X_i - x}{h}\right)$$

K_0 est appelé le noyau de Rosenblatt. Cet estimateur a le même inconvénient d'irrégularité que l'estimateur par histogramme.

On a donc l'idée d'utiliser des noyaux plus réguliers.

Définition 4.6. *Soit $K : \mathbb{R} \to \mathbb{R}$ intégrable et tel que*

$$\int K(y)dy = 1$$

alors K est appelé noyau (kernel).

Exemples :

— Noyau triangulaire : $K(u) = (1 - |u|)\mathbb{1}_{[-1,1]}(u)$

— Noyau d'Epanechnikov : $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}_{[-1,1]}(u)$

— Noyau Biweight : $K(u) = \frac{15}{16}(1 - u^2)^2\mathbb{1}_{[-1,1]}(u)$

— Noyau Gaussien : $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$

On définit alors un estimateur à noyau dès qu'on se donne un noyau K et une fenêtre $h > 0$.

Définition 4.7. *Etant donné K un noyau et $h > 0$, on pose*

$$\forall x \in \mathbb{R}, \quad \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K(\frac{X_i - x}{h})$$

Remarque 4.8.
- *La plupart des noyaux sont symétriques, positifs et sont décroissants sur $\mathbb{R}^+$ comme le noyau Gaussien : plus y est proche de 0, plus $K(y)$ est grand. Donc, pour un $x \in \mathbb{R}$ donné, plus une observation X_i est proche de x, plus $K(\frac{X_i - x}{h})$ est grand. Donc $\hat{f}_n(x)$ est d'autant plus grand que x est proche de beaucoup d'observations X_i (somme de beaucoup de grandes valeurs $K(\frac{X_i - x}{h})$).*
- *L'estimateur est somme de fonctions $K(\frac{X_i - x}{h})$ qui sont continues si K est continu. Donc $\hat{f}_n$ est continu si K est continu.*
- *$\int \hat{f}_n(x)dx = 1$, donc, si $K(x) \geq 0 \quad \forall x \in \mathbb{R}$, alors $\hat{f}_n$ est une densité.*
- *Le paramètre $h > 0$ est appelé fenêtre (bandwidth). C'est un paramètre de lissage : plus h est grand, plus l'estimateur est régulier. Comme dans le cas de l'estimateur à histogramme, le choix de h est délicat, la fenêtre h optimale devant réaliser un équilibre biais/variance (cf section suivante).*
- *Dans la pratique, le choix du noyau est peu influent, contrairement au choix de la fenêtre !*

Illustration graphique et code R

On va utiliser le même exemple de distribution bimodale que précédemment. L'estimation par noyaux peut se faire avec différentes méthodes. On peut utiliser la fonction `density` du package `stat`. Cette procédure n'estime que des densités à une seule variable. Pour des fonctions multivariées, on peut utiliser par exemple la fonction `kde` du package `ks` (de 1 à 6 variables).

Par défaut le noyau utilisé est le noyau gaussien, il est possible de changer de noyau avec l'option `kernel`.

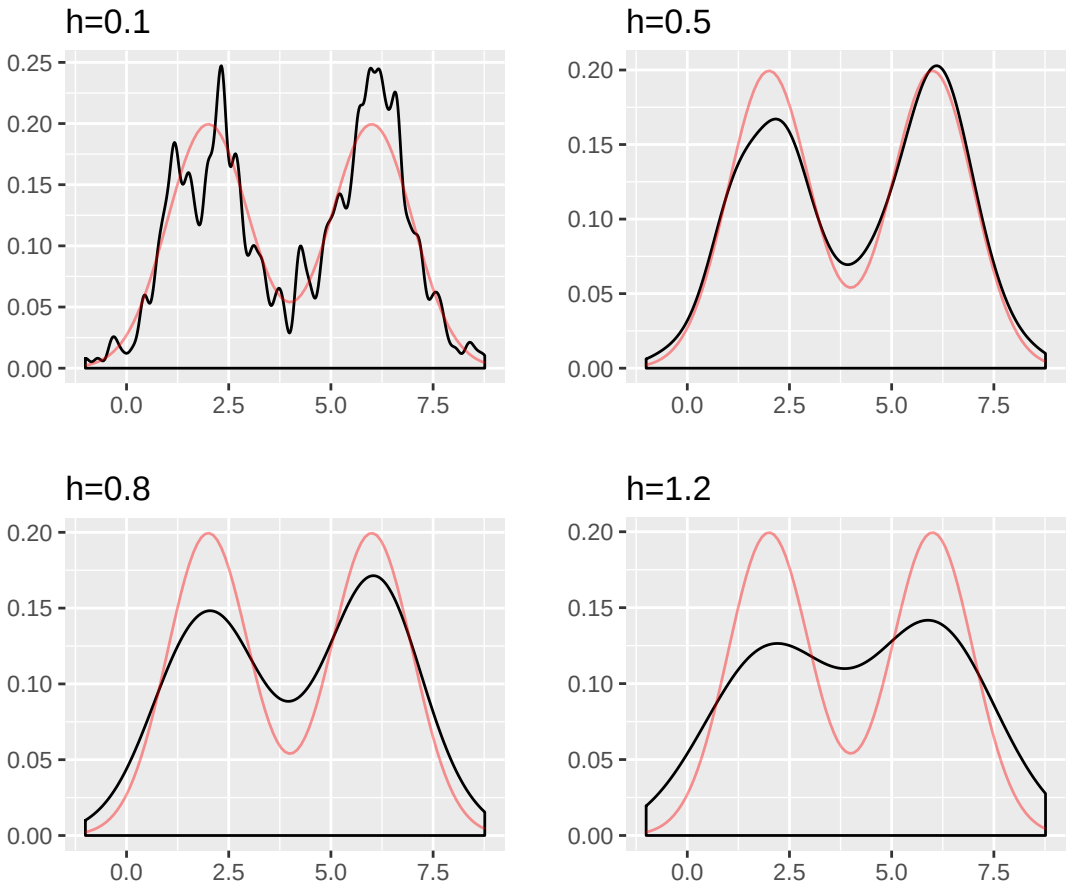
On va en fait utiliser la version de ggplot pour représenter l'estimateur à noyau. La fonction qui permet de dessiner l'estimateur à noyau est

`geom_density`

Le paramètre représentant le fenêtre h s'appelle `bw` (comme bandwidth).

On illustre l'influence du choix de la fenêtre. On tire les mêmes conclusions que pour l'histogramme.

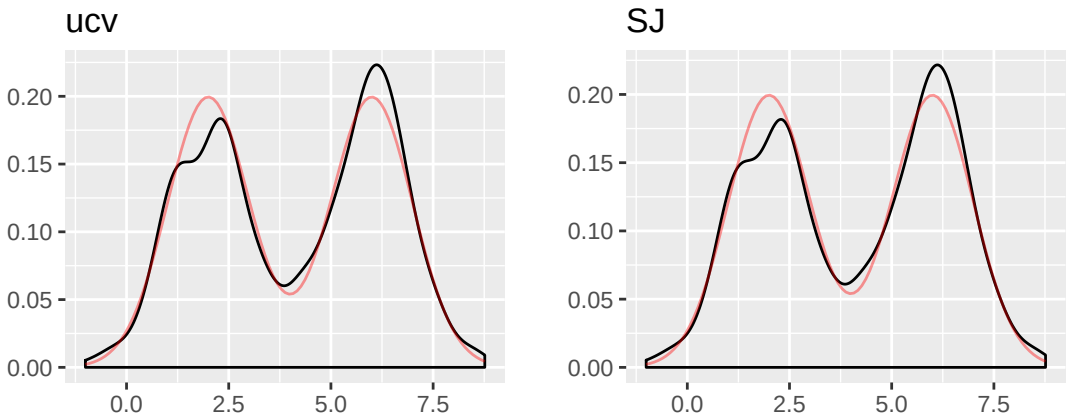
```
p<-ggplot(data.frame(x=Z),aes(x))+labs(x="",y="")
p1<-p+geom_density(bw=0.1)+stat_function(fun=f,col='red',alpha=0.4)+ggtitle("h=0.1")
p2<-p+geom_density(bw=0.5)+stat_function(fun=f,col='red',alpha=0.4)+ggtitle("h=0.5")
p3<-p+geom_density(bw=0.8)+stat_function(fun=f,col='red',alpha=0.4)+ggtitle("h=0.8")
p4<-p+geom_density(bw=1.2)+stat_function(fun=f,col='red',alpha=0.4)+ggtitle("h=1.2")
grid.arrange(p1,p2,p3,p4,nrow=2,ncol=2)
```



Pour finir, on illustre le choix de deux fenêtres calculées à partir des données. L'une est la méthode de Sheather et Jones (SJ) et l'autre est basée sur la validation croisée, qui sera vue en fin de chapitre (ucv=unbiased

cross-validation). Pour d'autres méthodes, consultez la documentation liée à `bw`.

```
p<-ggplot(data.frame(x=Z),aes(x))+labs(x="",y="")
p5<-p+geom_density(bw="ucv")+stat_function(fun=f,col='red',alpha=0.4)+ggtitle("ucv")
p6<-p+geom_density(bw="SJ")+stat_function(fun=f,col='red',alpha=0.4)+ggtitle("SJ")
grid.arrange(p5,p6,ncol=2)
```



Il existe une version en dimension 2 de cette fonction dans `ggplot2` qui s'appelle

`geom_density_2d`

.

4.4 Risque quadratique ponctuel des estimateurs à noyau sur la classe des espaces de Holder

Dans cette section, on s'intéresse au risque quadratique ponctuel de $\hat{f}_n$, i.e. étant donné $x_0 \in \mathbb{R}$

$$R(\hat{f}_n, f) = \mathbf{E}\Big[|\hat{f}_n(x_0) - f(x_0)|^2\Big]$$

Rappelons la décomposition "biais au carré+ variance" du risque quadratique :

$$\mathbf{E}\Big[|\hat{f}_n(x_0) - f(x_0)|^2\Big] = \Big(\mathbf{E}[\hat{f}_n(x_0)] - f(x_0)]\Big)^2 + \mathbf{Var}(\hat{f}_n(x_0))$$

Définition 4.9. *Soit $\ell \in \mathbb{N}^*$. On dit que le noyau K est d'ordre ℓ si $\forall j \in \{1, \dots, \ell\}$, $u \to u^j K(u)$ est intégrable et $\int u^j K(u)du = 0$.*

Proposition 4.10. *Si $f \in \Sigma(\beta, L)$ avec $\beta > 0$ et $L > 0$ et si K est un noyau d'ordre $\ell = \lfloor \beta \rfloor$ tel que $\int |u|^\beta |K(u)|du < \infty$ alors pour tout $x_0 \in \mathbb{R}$, et pour tout $h > 0$ le biais peut être borné comme suit :*

$$|\mathbf{E}[\hat{f}_n(x_0)] - f(x_0)| \leq \frac{h^\beta L}{\ell!} \int |u|^\beta |K(u)|du$$

Démonstration. On a

$$\begin{aligned}
\mathbf{E}[\hat{f}_n(x_0)] &= \mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^n \frac{1}{h} K(\frac{X_i - x_0}{h})\Big] \\
&\mathbf{E}\Big[\frac{1}{h} K(\frac{X_1 - x_0}{h})\Big] \\
&= \frac{1}{h}\int K(\frac{u - x_0}{h}) f(u)du \\
&= \int K(v) f(x_0 + hv)dv
\end{aligned}$$

De plus

$$f(x_0) = f(x_0) \times 1 = f(x_0) \int K(v)dv.$$

Donc

$$\mathbf{E}\Big[\hat{f}_n(x_0)\Big] - f(x_0) = \int K(v)\Big[f(x_0 + hv) - f(x_0)\Big]dv$$

Comme $f \in \Sigma(\beta, L)$, f admet $\lfloor \beta \rfloor$ dérivées et par un développement de Taylor-Lagrange (cf rappel chapitre 1) on a, pour tout $x \in \mathbb{R}$,

$$f(x) = \sum_{k=0}^{\ell-1} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) + \frac{(x - x_0)^\ell}{\ell!} f^{(\ell)}(x_0 + \xi(x - x_0))$$

avec $\xi \in]0, 1[$. Autrement dit on a, avec $x = x_0 + hv$,

$$f(x_0 + hv) - f(x_0) = \sum_{k=1}^{\ell-1} \frac{(hv)^k}{k!} f^{(k)}(x_0) + f^{(\ell)}(x_0 + hv\xi)\frac{(hv)^\ell}{\ell!}$$

pour un certain $\xi \in]0,1[$.Donc

$$\int K(v)\Big[f(x_0+hv)-f(x_0)\Big]dv = \int K(v)\Big[\sum_{k=1}^{\ell-1}\frac{(hv)^k}{k!}f^{(k)}(x_0)+f^{(\ell)}(x_0+hv\xi)\frac{(hv)^\ell}{\ell!}\Big]dv$$
$$= \frac{h^\ell}{\ell!}\int K(v)v^\ell f^{(\ell)}(x_0+hv\xi)dv$$

Comme K est d'ordre ℓ, on a aussi $\int K(v)v^\ell f^{(\ell)}(x_0)dv = 0$. Donc on a

$$\int K(v)\Big[f(x_0+hv)-f(x_0)\Big]dv = \frac{h^\ell}{\ell!}\int K(v)v^\ell\Big[f^{(\ell)}(x_0+hv\xi)-f^{(\ell)}(x_0)\Big]dv$$

Or, comme $f \in \Sigma(\beta, L)$, on a $|f^{(\ell)}(x_0+hv\xi)-f^{(\ell)}(x_0)| \leq L|hv|^{\beta-\ell}$. Et finalement

$$\left|\int K(v)\Big[f(x_0+hv)-f(x_0)\Big]dv\right| \leq \frac{|h|^\ell}{\ell!}\int |K(v)||v|^\ell L|hv|^{\beta-\ell}dv$$

ce qui signifie que

$$\Big|\mathbf{E}\Big[\hat{f}_n(x_0)\Big]-f(x_0)\Big| \leq \frac{L|h|^\beta}{\ell!}\int |K(v)||v|^\beta dv$$

□

Le biais au carré tend donc vers zéro à la vitesse $h^{2\beta}$. Plus la fonction f est régulière, plus le biais tend vite vers zéro quand h tend vers zéro (à condition bien sûr que l'ordre du noyau soit suffisamment grand).

Proposition 4.11. *Si f est bornée et si K est de carré intégrable alors*

$$\mathbf{Var}(\hat{f}_n(x_0)) \leq \frac{\|f\|_\infty \|K\|_2^2}{nh}$$

En particulier, si $f \in \Sigma(\beta, L)$ alors

$$\mathbf{Var}(\hat{f}_n(x_0)) \leq \frac{M(\beta,L)\|K\|_2^2}{nh}$$

Démonstration.

$$\begin{aligned}
\mathbf{Var}(\hat{f}_n(x_0) &= \mathbf{Var}\Big(\frac{1}{nh}\sum_{i=1}^n K(\frac{X_i-x_0}{h})\Big)\\
&= \sum_{i=1}^n \mathbf{Var}\Big(\frac{1}{nh}K(\frac{X_i-x_0}{h})\Big)\\
&= \sum_{i=1}^n \frac{1}{n^2h^2}\mathbf{Var}\Big(K(\frac{X_i-x_0}{h})\Big)\\
&= \frac{1}{nh^2}\mathbf{Var}\Big(K(\frac{X_1-x_0}{h})\Big)\\
&\leq \frac{1}{nh^2}\mathbf{E}\Big(K^2(\frac{X_1-x_0}{h})\Big)\\
&= \frac{1}{nh^2}\int K^2\Big(\frac{u-x_0}{h}\Big)f(u)du\\
&= \frac{1}{nh}\int K^2(v)f(x_0+vh)dv
\end{aligned}$$

Et enfin, on utilise la proposition 10 : il existe une constante positive $M(\beta, L)$ tel que $\|f\|_\infty \leq M(\beta, L)$. Ceci implique que

$$\mathbf{Var}(\hat{f}_n(x_0) \leq \frac{1}{nh} M(\beta, L) \int K^2(v)dv$$

□

Pour que la variance tende vers zéro, il faut que nh tende vers l'infini. En particulier, à n fixé, la variance est une fonction décroissante de h au contraire du biais qui est une fonction croissante de h. Il y a donc une valeur optimale de h qui doit réaliser l'équilibre entre le biais au carré et la variance. On peut à présent donner un contrôle du risque quadratique.

Théorème 4.12. *Soit $\beta > 0$ et $L > 0$ et K un noyau de carré intégrable et d'ordre $\lfloor \beta \rfloor$ tel que $\int |u|^\beta |K(u)|du < \infty$. Alors, en choisissant une fenêtre de la forme $h = cn^{-\frac{1}{2\beta+1}}$ avec une constante $c > 0$, on obtient*

$$\forall x_0 \in \mathbb{R}, \quad R(\hat{f}_n(x_0), \Sigma_d(\beta, L) := \sup_{f \in \Sigma_d(\beta,L)} \mathbf{E}[|\hat{f}_n(x_0) - f(x_0)|^2] \leq Cn^{-\frac{2\beta}{2\beta+1}}$$

où C est une constante dépendant de L, β, c et K.

Démonstration. On a

$$R\Big(\hat{f}_n(x_0), f(x_0)\Big) = \text{ Biais au carré + Variance}$$

Le terme de biais a été traité dans la proposition 11 et le terme de variance a été traité dans la proposition 12. On trouve

$$R\Big(\hat{f}_n(x_0), f(x_0)\Big) \leq \Big(\frac{h^\beta L}{\ell!} \int |u|^\beta |K(u)|du\Big)^2 + \frac{M(\beta, L)\|K\|_2^2}{nh}$$

On cherche ensuite la fenêtre h qui optimise cette quantité. Comme on ne soucie pas vraiment des constantes exactes quand on cherche la vitesse d'un estimateur, on utilise la notation $c_1 = \Big(\frac{L}{\ell!} \int |u|^\beta |K(u)|du\Big)^2$ et $c_2 = M(\beta, L)\|K\|_2^2$. On doit alors minimiser en h la quantité

$$c_1 h^{2\beta} + \frac{c_2}{nh}$$

On a une quantité croissante et une quantité décroissante en h. Encore une fois, comme on ne se soucie pas des constantes, donc on cherche seulement la fenêtre h qui nous donne l'ordre minimal du risque. Quand h est trop grand, le biais est trop grand, et quand h est trop petit, c'est la variance qui est trop grande. On cherche donc la fenêtre h qui réalise un équilibre entre le biais au carré et la variance :

$$h^{2\beta} \approx \frac{1}{nh}$$

où le signe $\approx$ signifie ici "de l'ordre de". Cela donne

$$h \approx n^{-\frac{1}{2\beta+1}}$$

Autrement dit, pour une fenêtre h de l'ordre de $n^{-\frac{1}{2\beta+1}}$, le biais au carré et la variance sont de même ordre. Plus exactement, si on choisit la fenêtre $h_* = cn^{-\frac{1}{2\beta+1}}$, avec c une constante positive, on a

$$\text{Biais au carré} \approx h_*^{2\beta} \approx \text{ variance } \approx \frac{1}{nh_*}$$

De plus on a alors

$$h_*^{2\beta} \approx n^{\frac{-2\beta}{2\beta+1}}$$

Autrement dit, il existe une certaine constante C telle que, pour cette fenêtre h_*, on a

$$R\Big(\hat{f}_n(x_0), \Sigma_d(\beta, L)\Big) \leq Cn^{\frac{-2\beta}{2\beta+1}}$$

Cette fenêtre est donc optimale à une constante près (si on change c, on change C mais ça ne change pas le taux qui est $n^{\frac{-2\beta}{2\beta+1}}$). □

Remarque 4.13. — *l'estimateur dépend de β à travers la fenêtre h. Or, sans connaissance a priori sur la régularité de la fonction f, on ne peut donc pas utiliser cet estimateur. On essaie alors de trouver un choix de fenêtre ne dépendant que des données et qui soit aussi performant (ou presque aussi performant si ce n'est pas possible d'être aussi performant) que l'estimateur utilisant cette fenêtre optimale. A ce sujet, on introduit plus loin un choix de fenêtre ne dépendant que des données et qui est basé sur ce qu'on appelle la validation croisée (ou "cross validation").*

— *Plus β est grand, plus la vitesse est grande. A la limite $\beta \to \infty$ on obtient une vitesse paramétrique.*

— *On peut généraliser le concept des estimateurs à noyaux pour une densité à plusieurs variables. Mais attention, en grande dimension, le problème du "fléau de dimension" ("curse of dimensionality") se pose souvent. En fait, l'estimateur à noyau en dimension d donne une vitesse de $n^{-\frac{2\beta}{2\beta+d}}$ (on retrouve bien le résultat du théorème avec $d = 1$). Donc cette vitesse se dégrade très vite avec la dimension. On évite donc en général d'utiliser un estimateur à noyau en dimension supérieure à 4 ou 5.*

4.5 Construction de noyaux d'ordre ℓ

La section 4.4 est de lecture facultative.

On va montrer que pour tout $\ell \in \mathbb{N}^*$ des noyaux d'ordre ℓ existent bien.

Soit $(\phi_m)_{m\in\mathbb{N}}$ la base orthonormée des polynômes de Legendre dans $L_2([-1,1])$ définie par

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

Cette base est obtenue par orthonormalisation de Gram-Schmidt de la base $(x \to x^k)_{k\geq 0}$. Elle a les propriétés suivantes :

— $\int_{-1}^{1} \phi_m(u)\phi_k(u)du = \mathbb{1}_{m=k}$

— ϕ_m est un polynôme de degré m.

— ϕ_{2m} est pair et ϕ_{2m+1} est impair $\forall m \geq 0$.

Proposition 4.14. *Soit $K_\ell : u \to \sum_{m=0}^{\ell} \phi_m(0)\phi_m(u)\mathbb{1}_{|u|\leq 1}$. Alors K_ℓ est un noyau d'ordre ℓ.*

Démonstration. $\forall j \in \mathbb{N}$, $u \mapsto u^j K(u)$ est intégrable sur $\mathbb{R}$. De plus $\forall j \in \mathbb{N}$, $\exists (a_q)_{q\geq 0}$ telle que $\forall u \in [-1, 1]$,

$$u^j = \sum_{q\geq 0} a_q \phi_q(u) = \sum_{q=0}^{j} a_q \phi_q(u)$$

Donc

$$\begin{aligned}
\int u^j K(u)du &= \int_{-1}^{1} \sum_{q=0}^{j} a_q \phi_q(u) K(u)du \\
&= \sum_{q=0}^{j} a_q \int_{-1}^{1} \phi_q(u) \sum_{m=0}^{\ell} \phi_m(0)\phi_m(u)du \\
&= \sum_{q=0}^{j} \sum_{m=0}^{\ell} a_q \phi_m(0) \int_{-1}^{1} \phi_q(u)\phi_m(u)du \\
&= \sum_{q=0}^{j} a_q \phi_q(0) \\
&= \begin{cases} 0 & \text{si } j \geq 1 \\ 1 & \text{si } j = 0 \end{cases}
\end{aligned}$$

□

Remarque 4.15. *Comme ϕ_{2k+1} est impaire, on a $\phi_{2k+1}(0) = 0$ et donc $K_{2k} = K_{2k+1}$. Et donc l'ordre maximal de K_ℓ est impair.*

4.6 Choix de la fenêtre h par validation croisée

Le choix de la fenêtre dans la section précédente est critiquable : comme on l'a mentionné, il dépend de la régularité qui est en général inconnue. On peut donc essayer d'estimer cette fenêtre idéale par un estimateur $\hat{h}$. De façon à souligner la dépendance à la fenêtre h, on va noter $\hat{f}_{n,h}$ l'estimateur associé à un choix de fenêtre h. L'estimateur final sera $\hat{f}_{n,\hat{h}}$, une fois le choix de $\hat{h}$ fait.

On cherche à minimiser en h le risque quadratique pour la distance L_2 :

$$R(\hat{f}_{n,h}, f) = \mathbf{E}[\|\hat{f}_{n,h} - f\|_2^2]$$

Or la fonction f étant inconnue, ce risque n'est pas calculable à partir des données. On cherche donc à estimer ce risque en utilisant uniquement les données. Remarquons tout de suite que minimiser en h la quantité $R(\hat{f}_{n,h}, f)$ est équivalent à

minimiser en h la quantité $R(\hat{f}_{n,h}, f) - \|f\|_2^2$. On va en fait remplacer la minimisation de la quantité inconnue $R(\hat{f}_{n,h}, f) - \|f\|_2^2$ par la minimisation d'un estimateur $\hat{R}(h)$ de cette quantité. Plus précisément on va chercher un estimateur sans biais de $R(\hat{f}_{n,h}, f) - \|f\|_2^2$.

Pour simplifier on suppose dans le théorème suivant que K est positif (on aurait pu aussi supposer que f et K sont tels que $\int |K(\frac{u-v}{h})| f(u)f(v)dudv$ est finie). De cette manière toutes les quantités que l'on manipulera seront positives (car K et f sont positives) et on pourra appliquer Fubini. On suppose aussi que $R(\hat{f}_{n,h}, f) < \infty$ et $f \in L_2$.

Théorème 4.16. *Si on pose*

$$\hat{R}(h) = \|\hat{f}_{n,h}\|_2^2 - \frac{2}{n(n-1)} \sum_{i=1} \sum_{j=1, j\neq i}^{n} \frac{1}{h} K\Big(\frac{X_i - X_j}{h}\Big)$$

alors $\hat{R}(h)$ est un estimateur sans biais de $R(\hat{f}_{n,h}, f) - \|f\|_2^2$.

Démonstration. On veut montrer que

$$\mathbf{E}\hat{R}(h) = R(\hat{f}_{n,h}, f) - \|f\|_2^2$$

Or

$$\begin{aligned} R(\hat{f}_{n,h}, f) - \|f\|_2^2 &= \mathbf{E}\Big(\|\hat{f}_{n,h}\|_2^2 - 2\int \hat{f}_{n,h}(x) f(x) dx\Big) \\ &= \mathbf{E}\|\hat{f}_{n,h}\|_2^2 - 2\int \mathbf{E}\hat{f}_{n,h}(x) f(x) dx \end{aligned}$$

(on a appliqué Fubini pour la seconde égalité)

Il suffit donc de montrer que

$$\int \mathbf{E}\hat{f}_{n,h}(x) f(x) dx = \frac{1}{n(n-1)} \mathbf{E}\Big[\sum_{i=1} \sum_{j=1, j\neq i}^{n} \frac{1}{h} K\Big(\frac{X_i - X_j}{h}\Big)\Big]$$

Le côté gauche donne, d'après le calcul fait dans la proposition 11,

$$\int \Big[\mathbf{E}\hat{f}_{n,h}(x)\Big] f(x) dx = \int \Big[\int \frac{1}{h} K(\frac{u-x}{h}) f(u) du\Big] f(x) dx$$

Le côté droit donne

$$\begin{aligned} \frac{1}{n(n-1)} \mathbf{E}\Big[\sum_{i=1} \sum_{j=1, j\neq i}^{n} \frac{1}{h} K\Big(\frac{X_i - X_j}{h}\Big)\Big] &= \mathbf{E}\Big[\frac{1}{h} K\Big(\frac{X_1 - X_2}{h}\Big)\Big] \\ &= \int\int \frac{1}{h} K\Big(\frac{u-v}{h}\Big) f(u) f(v) du dv \end{aligned}$$

On applique Fubini.

□

On définit alors

$$\hat{h} = \arg\min_{h \in H} \hat{R}(h)$$

si ce minimum est atteint. On cherche une fenêtre parmi une grille finie de valeurs, grille qu'on a notée H dans la formule ci-dessus.

L'estimateur $\hat{f}_{n,\hat{h}}$ a de bonnes propriétés pratiques et des propriétés de consistance.

La validation croisée est une méthode très générale dont on reparlera plus en détail dans le prochain chapitre. L'idée d'utiliser un estimateur sans biais du risque est aussi une idée assez générale (cf critère Cp).

Chapitre 5

Régression non paramétrique

5.1 Introduction

Dans ce chapitre, on cherche à expliquer les valeurs que peut prendre une variable Y à partir des valeurs que peut prendre une variable X.

Exemples :

— Y est le taux d'insuline dans le sang, qu'on explique (ou prédit) à l'aide de $X=$ (IMC, pression du sang, concentration de molécules).

— Y est le niveau de diplôme obtenu , qu'on explique à l'aide de $X =$ (âge, sexe, revenu des parents, métier des parents).

On suppose que la variable Y est intégrable $\mathbf{E}|Y| < \infty$ et on note r la fonction de régression de Y sur X :

$$r(x) = \mathbf{E}(Y|X = x)$$

L'objectif est d'estimer la fonction r pour expliquer et prédire Y à partir de X. Pour cela on dispose des réalisations de n couples de variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. On va supposer que les (X_i, Y_i) sont indépendants.

Vocabulaire

— les Y_i sont les variables à expliquer ou les variables réponses ou variables de sortie.

— les X_i constituent le design, les variables explicatives, les covariables, ou variables d'entrée.

Modélisation

Le design pourra être aléatoire ou déterministe. Dans ce dernier cas, on notera plutôt x_i à la place de X_i.

Le fait que $r(x) = \mathbf{E}(Y|X = x)$ se réécrit

$$Y = r(X) + \epsilon \quad \text{avec} \quad \mathbf{E}(\epsilon|X) = 0$$

On aura donc pour l'échantillon

$$Y_i = r(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \quad \mathbf{E}(\epsilon_i|X_i) = 0$$

En particulier on a donc $\mathbf{E}(\epsilon) = 0$.

Les ϵ_i sont appelées erreurs et jouent le rôle de bruit. Dans la suite, on va faire une hypothèse très forte :

$$\mathbf{Var}(\epsilon_i) = \sigma^2 < \infty \quad \text{variance finie et indépendante de } i$$

On va comme dans le chapitre précédent estimer une fonction. Précédemment une densité, ici une fonction de régression. Des méthodes similaires vont s'appliquer.

5.2 EMC non paramétrique

5.2.1 Modèle linéaire : rappels

Le modèle linéaire consiste à supposer que r s'écrit, si $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$,

$$r(x) = \beta_0 + \beta_1 x_1 + \ldots, \beta_p x_p$$

On a donc, pour tout $i = 1, \ldots, n$,

$$\begin{aligned} r(X_i) &= \beta_0 + \beta_1 X_{i1} + \ldots, \beta_p X_{ip} \\ &= X_i^T \boldsymbol{\beta} \end{aligned}$$

On note $\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$ et $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$

Dans ce cas, l'estimation de r revient à l'estimation du vecteur $\boldsymbol{\beta}$. C'est un problème paramétrique. Quand on ne sait rien sur la loi des observations, on utilise les moindres carrés ordinaires :

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

Si $\boldsymbol{X}$ est injective (i.e. de plein de rang en colonnes) alors $\boldsymbol{X}^T\boldsymbol{X}$ est inversible et $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$ et $\hat{Y} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y = AY$ où $A = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Et finalement, l'estimateur de la fonction de régression est,

$$\hat{r}(x) = (1, x^T)\hat{\beta}$$

pour $x \in \mathbb{R}^p$.

Un exemple : la hauteur des eucalyptus

Lorsqu'un forestier essaie de quantifier le volume de bois fourni par un arbre, il est nécessaire de connaitre sa hauteur. Or il est parfois impossible d'effectuer une telle mesure. Une mesure plus simple est la mesure de la circonférence de l'arbre à une hauteur fixée du sol. Le forestier souhaite trouver une formule, si celle-ci existe, permettant de déduire la hauteur de l'arbre à partir de sa circonférence. Pour cela

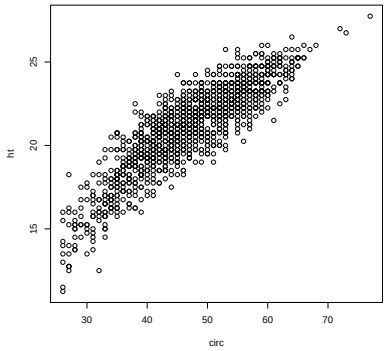


FIGURE 5.1 – Représentation hauteur versus circonférence pour les 1429 eucalyptus mesurés

il dispose d'un ensemble de $n = 1429$ couples de mesures circonférence-hauteur effectuées sur n arbres.

Pour commencer, comme il n'y a qu'une seule variable, on représente les données.

Cela nous permet de savoir qu'une régression simple semble indiquée, les points étant disposés grossièrement le long d'une droite.

Si les données se trouvent dans un data.frame appelé `euca` et si les noms des variables sont `ht` et `circ` alors on peut utiliser

```
reg=lm(ht~circ,data=euca)
```

On peut ensuite représenter le nuage de points avec la droite de régression, ainsi que l'intervalle de confiance sur un ensemble de valeurs de prévisions (à 95%) .

```
> plot(ht~circ,data=euca)
> circ=euca[,'circ']
> grille<-seq(min(circ),max(circ),length=100)
> grilledataframe<-data.frame(circ=grille)
> ICpred<-predict(reg,new=grilledataframe,interval="pred",level=0.95)
> matlines(grille,ICpred,lty=c(1,2,2),col=c('red','blue','blue'))
```

Nous constatons que les observations sont globalement bien ajustées par le modèle, sauf peut-être pour les faibles valeurs de circonférences, qui semblent en majorité situées en dessous de la droite. Ceci suggère d'utiliser plutôt le modèle de régression suivant

$$\texttt{ht} = a_1 + a_2\texttt{circ} + a_3\sqrt{\texttt{circ}} + \epsilon$$

On peut donc utiliser un modèle linéaire avec une transformation de la variable d'origine. On peut d'ailleurs vérifier qu'en introduisant la variable `sqrt(circ)`, on a bien un meilleur modèle :

```
> reg1=lm(ht~circ,data=euca)
> reg2=lm(ht~circ+I(sqrt(circ)),data=euca)
> anova(reg1,reg2)
```

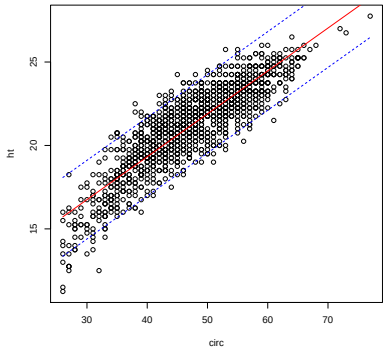


FIGURE 5.2 – Droite de régression et intervalles de confiance sur la prévision

```
Analysis of Variance Table

Model 1: ht ~ circ
Model 2: ht ~ circ + I(sqrt(circ))
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   1427 2052.1
2   1426 1840.7  1    211.43 163.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

De manière générale, on peut utiliser le modèle linéaire avec n'importe quelle transformation de la variable d'origine (ou des variables d'origine si on est en dimension supérieure à 1).

5.2.2 EMC non paramétrique

Revenons sur notre problème général : on cherche à expliquer une variable Y par une variable explicative X. On suppose pour simplifier que X est de dimension 1.

On peut penser que la fonction est bien approchée par un polynôme :

$$r(x) \approx \theta_1 + \theta_2 x + \ldots + \theta_3 x^M$$

mais on ne sait pas quel degré M choisir. Si on choisit le degré 2 par exemple, on a alors l'estimateur

$$\hat{r}(x) = \hat{\theta}_1 + \hat{\theta}_2 x + \hat{\theta}_3 x^2$$

où $\hat{\theta}$ est l'estimateur des moindres carrés de θ dans le modèle

$$Y_i = \theta_1 + \theta_2 z_{i1} + \theta_3 z_{i2} + \epsilon_i, \quad \ldots i = 1, \ldots, n$$

avec $z_{i1} = X_i$ et $z_{i2} = X_i^2$ pour tout $i = 1, \ldots, n$. . On trouve donc cet estimateur $\hat{\theta}$ par la commande

```
\color{red}
lm(y~x+x^2))$coefficients\color{black}
```

qu'on peut aussi écrire sous la forme

```
lm(y~poly(x,2))$coefficients
```

De manière générale, on se donne un ensemble de fonctions $\varphi_1, \varphi_2, \ldots$ et on suppose que r est bien approchée par une combinaison linéaire d'éléments de cet ensemble : $\exists M$ tel que

$$r \approx \theta_1\varphi_1 + \ldots + \theta_M\varphi_M$$

On peut choisir une autre base que les polynômes, par exemple la base de Fourier, une base d'ondelettes etc. On peut même choisir un ensemble de fonctions qui n'est pas une base. Si on choisit d'utiliser les M premières fonctions du dictionnaire alors on calcule l'EMC $\hat{\theta}$ en utilisant la matrice $\boldsymbol{X}$ telle que $X_{ij} = \varphi_j(X_i)$ (du moment que la matrice X est bien de plein rang). On obtient alors directement l'estimateur $\hat{r}$:

$$\forall x, \quad \hat{r}(x) = \hat{\theta}_1\varphi_1(x) + \ldots + \hat{\theta}_M\varphi_M(x)$$

La question qui se pose alors est celle du nombre d'éléments du dictionnaire (par exemple si on choisit les polynôme, quel degré ?). Plus on choisir M grand, meilleure est l'approximation de départ $r \approx \theta_1\varphi_1 + \ldots + \theta_M\varphi_M$. Cependant, on sait (cf cours de modèle linéaire et/ou cours de grande dimension), que plus on choisit M grand, plus la variance augmente. Le biais et la variance se comportent de façon contraire vis-à-dis de M. Le paramètre M joue en fait le même rôle que la fenêtre h dans le chapitre précédent. Il s'agit donc de trouver un équilibre entre l'erreur d'approximation et la variance.

Il y a diverses méthodes pour choisir M. On y revient en fin de chapitre.

L'EMC non-paramétrique est une méthode globale : on fait la même approximation sur tout l'espace de départ.

Dans la suite, nous utiliserons une autre méthode que l'EMC non-paramétrique. L'estimateur que l'on présente dans la suite est appelé l'estimateur par polynôme local et comme son nom l'indique, c'est au contraire une méthode locale.

5.3 Estimateur de Nadaraya-Watson

On suppose que les (X_i, Y_i) admettent une densité $f : \mathbb{R}^2 \to \mathbb{R}$ et on suppose que pour tout $x > 0$, $f_X(x) = \int f(x, y)dy > 0$ (f_X est la densité de X). On peut alors écrire

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

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

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

On peut utiliser les estimateurs à noyau du chapitre précédent :

$$\hat{f}_{n,X}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h})$$

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

Proposition 5.1. *Si K est un noyau d'ordre 1 alors $\forall x \in \mathbb{R}$*

$$\hat{r}_n(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i K(\frac{X_i - x}{h})}{\sum_{i=1}^n K(\frac{X_i - x}{h})} \text{ si } \sum_{i=1}^n K(\frac{X_i - x}{h}) \neq 0 \\ 0 \text{ sinon} \end{cases}$$

Démonstration. $\hat{f}_{n,X}(x) = 0$ est équivalent à $\sum_{i=1}^n K(\frac{X_i - x}{h}) = 0$.

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

$$\begin{aligned}
\hat{r}_n(x) &= \int \frac{y\hat{f}_n(x,y)}{\hat{f}_{n,X}(x)} dy \\
&= \frac{1}{\hat{f}_{n,X}(x)} \int y \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{X_i - x}{h})K(\frac{Y_i - y}{h})dy \\
&= \frac{nh}{\sum_{i=1}^n K(\frac{X_i - x}{h})} \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{X_i - x}{h}) \int yK(\frac{Y_i - y}{h})dy \\
&= \frac{1}{\sum_{i=1}^n K(\frac{X_i - x}{h})} \sum_{i=1}^{n} K(\frac{X_i - x}{h}) \frac{1}{h} \int yK(\frac{Y_i - y}{h})dy \\
&= \frac{1}{\sum_{i=1}^n K(\frac{X_i - x}{h})} \sum_{i=1}^{n} K(\frac{X_i - x}{h}) Y_i
\end{aligned}$$

Pour la dernière ligne, on a utilisé le fait que

$$\frac{1}{h}\int yK(\frac{Y_i - y}{h})dy = \frac{1}{h}\int (Y_i - uh)K(u)hdu = Y_i \int K(u)du - h\int uK(u)du = Y_i$$

□

Exemple 5.2. *Prenons le noyau triangulaire $K(u) = \frac{1}{2}\mathbb{1}_{|u|\leq 1}$.*

Alors $\hat{r}_n(x)$ est la moyenne des Y_i tels que $X_i \in [x-h, x+h]$. Pour n fixé, les deux cas extrêmes pour la fenêtre sont :

- *$h \to \infty$. Quand h est suffisamment grand, tous les X_i se trouvent dans l'intervalle $[x-h, x+h]$. Alors $\hat{r}_n$ est la moyenne des Y_i, c'est donc une fonction constante de x. L'erreur d'approximation est alors trop grande.*
- *$h \to 0$. Soit x distinct de tous les X_i, si h est assez petit, très exactement $h < \min_{1\leq i\leq n}\{|X_i - x|\}$, on a $\hat{r}_n(x) = 0$. Et si $x = X_j$ pour un certain $j = 1, \ldots, n$, on a $r_n(X_j) = Y_j$ dès que $h < \min_{1\leq i\leq n}\{|X_i - X_j|\}$. L'estimateur $\hat{r}_n$ est donc très oscillant : il reproduit les données Y_i aux points X_i et il s'annule partout ailleurs. L'erreur stochastique est trop grande.*

La fenêtre optimale équilibrant biais (au carré) et variance se trouve entre ces deux extrêmes.

Remarque 5.3. *Si K est continu, positif et à support sur $\mathbb{R}$ (par ex le noyau gaussien) alors $\hat{r}_n(x)$ est continu.*

Remarque 5.4. *On peut écrire*

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

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

Remarquons aussi que, si $\sum_{i=1}^{n} K(\frac{X_i - x}{h}) = 0$, i.e. si x se trouve dans une zone où il n'y a pas de X_i, alors $\hat{r}(x) = 0$. Et sinon , comme $\sum_{i=1}^{n} \omega_{n,i}(x) = 1$, alors Y_i est une moyenne pondérée des Y_i qui correspondent aux points X_i proches de x.

Dans la pratique, comme K est en général symétrique et décroissant sur $\mathbb{R}^+$, le poids associé à Y_i dans cette moyenne pondérée est d'autant plus grand que X_i est proche de x. Les Y_i associés à des points X_i qui sont loin de x n'ont pas ou peu d'impact sur l'estimation de $r(x)$. C'est en cela que la méthode est locale, au contraire de l'EMC non paramétrique.

Remarque 5.5. *Il se peut que la densité f_X soit connue. Dans ce cas, il est préférable d'utiliser*

$$\tilde{r}_n(x) = \begin{cases} \int \frac{y \hat{f}_n(x,y)}{f_X(x)} dy \text{ si } f_X(x) \neq 0 \\ 0 \text{ si } f_X(x) = 0 \end{cases}$$

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

$$\tilde{r}_n(x) = \begin{cases} \frac{1}{nh f_X(x)} \sum_{i=1}^{n} Y_i K(\frac{X_i - x}{h}) \text{ si } f_X(x) \neq 0 \\ 0 \text{ si } f_X(x) = 0 \end{cases}$$

Proposition 5.6. *On suppose f_X connue. On s'intéresse à l'estimation de $r(x)$ pour x fixé. Soit K un noyau d'ordre 1. On suppose de plus que*

- *$f_X(x) > 0$.*
- *Il existe $\epsilon > 0$ tel que les fonctions f_X et r sont continument dérivables sur $[x - \epsilon, x + \epsilon]$*
- *Pour tout y, si $|u| \leq \epsilon$*

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

  où

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

- *K est un noyau à support dans $[-1, 1]$ et de carré intégrable*

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

$$\mathbf{E}[(\tilde{r}_n(x) - r(x))^2] \leq C(x)(h^2 + \frac{1}{nh})$$

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

$$\mathbf{E}[(\tilde{r}_n(x) - r(x))^2] \leq C'(x)n^{-2/3}$$

Démonstration. On utilise la décomposition biais/variance :

$$\mathbf{E}[(\tilde{r}_n(x) - r(x))^2] = \text{ Biais}^2 + \text{ variance}$$

— Biais

On va prouver dans le calcul de la variance que, sous les hypothèses de l'énoncé, $\mathbf{Var}[Y_1 K(\frac{X_1 - x}{h})] < \infty$. Ceci implique que $\mathbf{E}\Big[|Y_1 K(\frac{X_1 - x}{h})|\Big] < \infty$. On va pouvoir utiliser le théorème du transfert pour calculer cette intégrale ainsi que le théorème de Fubini si besoin.

On a

$$\begin{aligned}
\mathbf{E}[\tilde{r}_n(x)] &= \mathbf{E}\Big(\frac{1}{nhf_X(x)} \sum_{i=1}^{n} Y_i K(\frac{X_i - x}{h})\Big) \\
&= \frac{1}{hf_X(x)} \mathbf{E}[Y_1 K(\frac{X_1 - x}{h})] \\
&= \frac{1}{hf_X(x)} \int\int yK(\frac{t-x}{h}) f(t,y) dt dy \\
&= \frac{1}{f_X(x)} \int\int yK(v) f(x + vh, y) dv dy
\end{aligned}$$

De plus

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

Donc

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

Donc on a aussi

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

Donc

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

On a utilisé le fait que K est à support dans $[-1,1]$ dans la dernière égalité. On applique l'inégalité des accroissements finis à r et f_X car elles sont continument dérivables au voisinage de x. Il existe une constante $C(x)$ telle que, pour tout $|u| \leq \epsilon$,

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

On peut donc appliquer ces inégalités avec $u = vh$ pour $|v| \leq 1$ et $|h| \leq \epsilon$, ce qui donne

$$
\begin{aligned}
&|\mathbf{E}[\tilde{r}_n(x)] - r(x)| \\
&\leq \frac{1}{f_X(x)}\Big[\int_{-1}^{1} |K(v)|\Big|f_X(x+vh) - f_X(x)\Big||r(x+vh)|dv\Big] + \int_{-1}^{1} |K(v)|\Big|r(x+vh) - r(x)\Big|dv\Big] \\
&\leq \frac{C(x)}{f_X(x)}\Big[\int_{-1}^{1} |K(v)|\Big|hv\Big||r(x+vh)|dv\Big] + C(x)\int_{-1}^{1} |K(v)|\Big|hv\Big|\Big|dv
\end{aligned}
$$

De plus r étant continue sur $[x-\epsilon, x+\epsilon]$, il existe une constante $c(x)$ telle que $|r(x+hv)| \leq c(x)$ pour tout $|h| \leq \epsilon$ et tout $|v| \leq 1$. Donc on a

$$|\mathbf{E}[\tilde{r}_n(x)] - r(x)| \leq C_1(x)h$$

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

— Variance

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

Comme $|h| \leq \epsilon$, on a $|hv| \leq \epsilon$ pour tout $v \in [-1,1]$. Donc , d'après la troisième hypothèse de l'énoncé,

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

et donc

$$f(x+hv,y) \leq f(x,y) + M(x,y)\epsilon \tag{5.1}$$

Ainsi

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

Finalement la variance vérifie, si $|h| \leq \epsilon$,

$$\mathbf{Var}(\tilde{r}_n(x)) \leq \frac{C_2(x)}{nh}$$

où $C_2(x) = \frac{\int K^2(v)}{f_X^2(x)}\Big(\epsilon\int y^2M(x,y)dy + \int y^2 f(x,y)dvdy\Big)$.Cette quantité est finie d'après les hypothèses de l'énoncé (3ème et 4ème).

— Calcul du risque quadratique

$$\mathbf{E}[(\tilde{r}_n(x) - r(x))^2] \leq C_1^2(x)h^2 + \frac{C_2(x)}{nh}$$

On équilibre les deux termes

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

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

$$\mathbf{E}[(\tilde{r}_n(x) - r(x))^2] \leq C_3(x)n^{-2/3}$$

□

L'estimateur de Nadaraya-Watson est un cas particulier des estimateurs par polynomes locaux.

5.4 Estimateur par polynomes locaux

Proposition 5.7. *Si $\hat{r}_n$ est l'estimateur de Nadaraya-Watson associé à un noyau $K \geq 0$ alors $\hat{r}_n$ est solution de*

$$\hat{r}_n(x) = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)(Y_i - \theta)^2$$

$\hat{r}_n(x)$ est donc un estimateur des moindres carrés pondéré si $\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big) \neq 0$

Démonstration.

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

où

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

τ est un polynôme du second degré en θ. Recherche d'un point critique :

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

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

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

L'estimateur par polynômes locaux est une généralisation de l'estimateur de Nadaraya-Watson associée à sa caractérisation par la proposition précédente. Il faut garder à l'esprit ici que l'idée est de regarder les choses localement, et donc que x est fixé. On aura donc calculé pour ce x fixé un estimateur de $r(x)$ mais si on veut $\hat{r}(y)$ il faut faire un autre calcul.

L'idée associée à l'estimateur par polynômes locaux est de reprendre le problème de minimisation de la proposition précédente mais au lieu d'utiliser une constante θ, on utilise un polynôme.

Plus précisément, si r est régulière alors, autour de x, r est proche du polynôme associé à son développement de Taylor-Lagrange en x : pour u proche de x on a

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

avec

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

Evidemment $P_{\ell,x}$ est tout aussi inconnu que $r(x)$ (ses coefficients dépendent de la quantité que l'on cherche à estimer $r(x)$ mais aussi des dérivées $r'(x), \ldots, r^{(\ell)}(x)$).

On va en fait essayer d'estimer ce polynôme $P_{\ell,x}$. Si on écrit

$$P_{\ell,x}(u) = \mu_0 + \mu_1(u-x) + \ldots + \mu_\ell(u-x)^\ell,$$

on cherche donc à estimer les coefficients $\mu_0, \ldots, \mu_\ell$ de ce polynôme par des estimateurs $\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_\ell$.

Remarquez que si l'on arrive à estimer les coefficients de ce polynôme, qui est le polynôme de Taylor-Lagrange de r en x de degré ℓ, alors, comme $\mu_0 = r(x)$, l'estimateur $\hat{\mu}_0$ sera donc l'estimateur $\hat{r}(x)$ recherché.

En particulier, on a

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

donc on est tenté de chercher un polynôme $\hat{P}$ qui soit tel que

$$\hat{P}(X_i) \text{ est proche de } r(X_i) \text{ pour les } X_i \text{ proches de } x.$$

Comme on n'a pas accès à $r(X_i)$ mais à sa donnée bruitée Y_i, on cherche en fait $\hat{P}$ tel que

$$\hat{P}(X_i) \text{ est proche de } Y_i \text{ pour les } X_i \text{ proches de } x.$$

Autrement dit

$$(\hat{P}(X_i) - Y_i)^2 \text{ petit pour les } X_i \text{ proches de } x.$$

Des poids $K(\frac{X_i - x}{h})$ sont ajoutés pour prendre en compte cette notion de proximité. On pose alors

Définition 5.8. *Si K est un noyau positif, $h > 0$ une fenêtre et $\ell \geq 0$ un entier, on définit $\forall x \in \mathbb{R}$,*

$$\hat{\theta}(x) = \arg \min_{\theta=(\theta_0,\ldots,\theta_\ell)\in\mathbb{R}^{\ell+1}} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big)\Big[Y_i - \sum_{k=0}^{\ell} \frac{\theta_k}{k!}\Big(\frac{X_i - x}{h}\Big)^k\Big]^2$$

On pose $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_\ell)$. L'estimateur par polynôme local d'ordre ℓ est alors défini par

$$\hat{r}_n^\ell(x) = \hat{\theta}_0$$

Remarque 5.9. *Si $\ell = 0$ alors $\hat{r}_n^\ell(x)$ est égal à l'estimateur de Nadaraya-Watson.*

Définition 5.10. *Un estimateur $\hat{r}$ de la fonction de régression r est linéaire s'il s'écrit*

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

où les $\omega_i(x)$ ne dépendent pas des Y_i.

On peut aussi écrire $\hat{r}(x) = \omega(x)^T \boldsymbol{Y}$ où $\boldsymbol{Y}$ est le vecteur $(Y_1, \dots, Y_n)^T$ et $\omega(x) = (\omega_1(x), \dots, \omega_n(x))^T$.

On a vu que l'estimateur de Nadaraya-Watson est linéaire.

Attention : ne pas confondre le fait que l'estimateur soit linéaire, ce qui sous entend linéaire en Y, et le fait que la fonction de régression soit linéaire, ce qui signifie que $r(x)$ est linéaire en x (et on cherche alors un estimateur linéaire en x). L'estimateur associé aux MCO $\hat{r}(x) = \hat{\beta}^T x$ est linéaire en x et c'est également un estimateur linéaire : $\hat{r}(x) = x^T\hat{\beta} = x^T(X^TX)^{-1}X^TY = \omega(x)^TY$ où $\omega(x) = [x^T(X^TX)^{-1}X^T]^T$ est un vecteur qui ne dépend pas de Y.

Introduisons, pour la proposition suivante, quelques notations : pour tout $i = 1, \dots, n$ et tout $u \in \mathbb{R}$,

$$Z_i = \frac{X_i - x}{h}, \quad V_\ell(u) = \begin{pmatrix} 1 \\ u \\ \vdots \\ \frac{u^\ell}{\ell!} \end{pmatrix}$$

Et on pose

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

Proposition 5.11. *Si la matrice $B_{n,x}$ est définie positive alors l'estimateur par polynômes locaux $\hat{r}_n^\ell(x)$ est un estimateur linéaire.*

Démonstration. On a

$$\hat{r}_{n,\ell}(x) = \hat{\theta}_0(x) = e_1^T \hat{\theta}(x)$$

avec

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

$$\hat{\theta}(x) = \arg\min_{\theta \in \mathbb{R}^{\ell+1}} \tau(\theta)$$

où

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

On a

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

avec $a = \sum_{i=1}^n K(Z_i)Y_i^2$ et $b = \sum_{i=1}^n K(Z_i)Y_i V_\ell(Z_i)$

Rappels :

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

Recherche de point critique :

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

Donc

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

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

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

Ce point critique correspond bien à un minimum global car la fonction est convexe. En effet

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

On a donc

$$\begin{aligned}
\hat{r}_{n,\ell}(x) &= e_1^T B_{n,x}^{-1} b \\
&= e_1^T B_{n,x}^{-1} \Big[\sum_{i=1}^{n} K(Z_i) Y_i V_\ell(Z_i) \Big] \\
&= \sum_{i=1}^{n} \omega_i(x) Y_i
\end{aligned}$$

avec

$$\omega_i(x) = K(Z_i) e_1^T B_{n,x}^{-1} V_\ell(Z_i)$$

$\omega_i(x)$ ne dépend que de x, K, ℓ, h, et des X_i et pas des Y_i. Donc $\hat{r}_{n,\ell}$ est bien un estimateur linéaire. □

Remarque 5.12. *On a*

$$\sum_{i=1}^{n} \omega_i(x) = 1$$

pour la preuve : cf TD 5 exercice 2.

Remarque 5.13. *Comme pour l'estimation de densités par noyaux, en pratique le choix du noyau n'est pas très important. Quant au degré de polynôme, on choisit souvent 1 ou 2. Le choix de la fenêtre est en revanche crucial.*

5.5 Choix des paramètres de régularisation

5.5.1 Risque empirique, surajustement

On va supposer dans la suite pour simplifier que les X_i sont aléatoires. On suppose de plus que les X_i, Y_i sont iid. On suppose toujours que $\mathbf{E}\epsilon_i^2 = \sigma^2$.

On note maintenant r_h l'estimateur utilisant la fenêtre h. Si on enlève une partie de l'échantillon $(X_i, Y_i)_{i\in I}$ avec I une partie de $\{1,\dots,n\}$ on notera $\hat{r}_h^{-I}$ l'estimateur calculé à partir de l'échantillon auquel on a ôté $(X_i, Y_i)_{i\in I}$.

Remarquez que la fonction de régression r est telle que

$$r = \arg\min_{f\in L_2(\mathbf{P}_X)} \mathbf{E}\Big[(Y - f(X))^2\Big].$$

On veut trouver la fenêtre h qui minimise le risque

$$R(h) = \mathbf{E}\Big[(\hat{r}_h - r)^2(X)\Big] = \mathbf{E}\Big[\|\hat{r}_h - r\|^2_{L_2(\mathbf{P}^X)}\Big].$$

On ne peut pas minimiser ce risque puisque r est inconnu. Une première idée est de remplacer $r(X_i)$ par son observation bruitée Y_i et d'oublier l'espérance, c'est-à-dire de minimiser

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

NB : cette quantité est connue sous le nom de "erreur d'apprentissage" (training error).

C'est en général une très mauvaise idée d'utiliser ce risque comme substitut du vrai risque pour la sélection de modèle ! En effet les mêmes données sont utilisées à la fois pour estimer r et estimer le risque. Il y a un manque d'indépendance.

Prenons l'exemple de l'EMC non paramétrique. Imaginons qu'on cherche à ajuster un polynôme. On se pose donc la question du degré M. Pour chaque M on calcule $\hat{\beta}^M$ l'EMC associé au design $\boldsymbol{X} = (X_{ij})_{1\le j\le M, 1\le i\le n}$ avec $X_{ij} = x_i^{j-1}$. Si M est assez grand et si les points du design sont distincts alors le risque empirique est égal à 0. On a obtenu un polynôme qui passe par tous les points (X_i, Y_i) ("on recopie les données"). Mais la variance de cet estimateur risque fort d'être trop grande.

L'erreur d'apprentissage est trop optimiste. On aura en général $\mathbf{E}[\hat{R}_n(h)] < R(h)$. Utiliser cette erreur pousse au sur-ajustement (overfitting) : l'estimateur associé sera trop adapté aux données particulières qu'on a et ne se généralisera pas bien à de nouvelles données.

Remarque 5.14. — *Si $Y_1,\dots,Y_n \overset{iid}{\sim} Y$ alors pour estimer $\mathbf{E}(Y)$ on utilise souvent son équivalent empirique $\frac{1}{n}\sum_{i=1}^n Y_i$.*

— *Si g est une fonction **fixe** (i.e. ne dépendant pas des données) alors $Y_i - g(X_i) \overset{iid}{\sim} Y - g(X)$. Et il est alors naturel d'utiliser $\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))^2$ pour estimer $\mathbf{E}(Y - g(X))^2$. En effet si g est fixe,*

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(g(X_i) - Y_i)^2\Big] = \|g - r\|^2_{L_2(\mathbf{P}^X)} + \sigma^2,$$

et

$$\mathbf{Var}\Big[\frac{1}{n}\sum_{i=1}^{n}(g(X_i) - Y_i)^2\Big] = \frac{1}{n}\mathbf{Var}(g(X) - Y)^2.$$

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

5.5.2 Validation croisée

La technique de validation croisée est très générale et s'applique à de nombreuses procédures d'estimation. Ici on va l'appliquer pour le choix de la fenêtre h de l'estimateur par polynômes locaux, mais elle aurait pu être utilisée pour le choix d'un autre paramètre d'ajustement (le degré du polynôme si on ajuste un polynôme par les moindres carrés par exemple).

On se donne une grille de valeurs H de fenêtres, parmi lesquelles on veut choisir une fenêtre optimale $\hat{h}$ en se basant sur les données uniquement.

Le principe général est de diviser l'échantillon en un ensemble d'apprentissage (training set) et un ensemble de validation (validation set). On fabrique des estimateurs à partir de l'ensemble d'apprentissage et ensuite l'ensemble de validation est utilisé pour estimer leur risque de prédiction. Les schémas les plus populaires sont les suivants :

— Hold-out CV : on divise l'échantillon en deux parties I_1 et I_2 (I_1 et I_2 sont donc deux ensembles disjoints de $\{1, \ldots, n\}$). On calcule les estimateurs $(\hat{r}_h^{I_1})_{h\in H}$ à partir de $(X_i, Y_i)_{i\in I_1}$. Puis on calcule les estimateurs des risques associés

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

où on a noté $n_2 = \text{Card}(I_2)$.

— V-fold CV : les données sont divisées en V ensembles disjoints $I_1, \ldots, I_V$. Chacun des V sous-ensembles est utilisé à tour de rôle comme ensemble de validation, le reste étant donc utilisé pour l'apprentissage : on calcule, pour chaque $j \in \{1, \ldots, V\}$, l'ensemble des estimateurs $(\hat{r}_h^{-I_j})_{h\in H}$ fabriqués avec $(X_i, Y_i)_{i\notin I_j}$. Ensuite le risque de prédiction pour une fenêtre h est estimé par

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

où on a noté $n_j = \text{Card}(I_j)$.

Dans la pratique on choisit souvent $V = 5$ ou $V = 10$.

— Leave-one out : cas particulier du V-fold CV avec $V = n$.

— Leave-q-out : tout sous-ensemble de cardinal q de l'échantillon est utilisé comme ensemble de validation et le reste comme ensemble d'apprentissage.

On choisit

$$\hat{h} = \arg\min_{h\in H} \hat{R}(h)$$

Et l'estimateur final est

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

où $\hat{r}_{n,h}$ est l'estimateur par polynômes locaux calculé avec la fenêtre h et en utilisant tout l'échantillon.

Le V-fold est la méthode la plus populaire.

Les méthodes ci-dessus sont présentées par ordre d'intensité de calculs, le leave-q out ou le leave-one out étant les plus intensives en calculs.

Explicitons un peu plus le cas particulier du "leave-one out". Pour chaque valeur h de la grille de valeurs H et pour chaque $i \in \{1, \dots, n\}$, on construit un estimateur $\hat{r}_h^{(-i)}$ en utilisant toutes les observations sauf la $i$ème. La ième observation est ensuite utilisée pour mesurer la performance de $\hat{r}_h^{(-i)}$ par $(Y_i - \hat{r}_h^{(-i)}(X_i))^2$. On pose donc

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

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

Dans la suite on explicite les calculs pour voir le problème de dépendance lié au risque empirique.

On note $X_1^n = (X_1, \dots, X_n)$ et $Y_1^n = (Y_1, \dots, Y_n)$.

On cherche h tel que $\mathbf{E}\Big[(Y - \hat{r}_h(X))^2\Big]$ soit minimal. Remarquez que l'on pourrait comparer aussi des estimateurs de nature différente. On fait donc disparaitre la dépendance à h dans la notation.

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

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

En effet on a, si l'estimateur est symétrique en ses variables (ce qui semble raisonnable et est le cas des estimateurs par polynômes locaux)

$$\mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}(X_i))^2\Big] = \mathbf{E}\Big[(Y_1 - \hat{r}(X_1))^2\Big]$$

On indique la dépendance de $\hat{r}$ à (X^n, Y^n) en écrivant $\hat{r}(x) = g(X_1, \dots, X_n, Y_1, \dots, Y_n, x)$. On rappelle qu'on a noté f la densité du couple (X, Y). On a alors

$$\begin{aligned}\mathbf{E}\Big[(Y_1 - \hat{r}(X_1))^2\Big] &= \mathbf{E}\Big[(Y_1 - g(X_1, \dots, X_n, Y_1, \dots, Y_n, X_1))^2\Big] \\ &= \int (y_1 - g(x_1, \dots, x_n, y_1, \dots, y_n, x_1))^2 f(x_1, y_1) \dots f(x_n, y_n) dx_1 dy_1 \dots dx_n dy_n\end{aligned}$$

Tandis que

$$\begin{aligned}&\mathbf{E}\Big[(Y - \hat{r}(X))^2\Big] = \mathbf{E}\Big[(Y - g(X_1, \dots, X_n, Y_1, \dots, Y_n, X))^2\Big] \\ &= \int (y - g(x_1, \dots, x_n, y_1, \dots, y_n, x))^2 f(x_1, y_1) \dots f(x_n, y_n) f(x, y) dx_1 dy_1 \dots dx_n dy_n dx dy\end{aligned}$$

Le risque empirique est un mauvais estimateur du "vrai" risque $\mathbf{E}\Big[(Y - \hat{r}(X))^2\Big]$.

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

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

On a finalement juste utilisé le fait que

$$Y_{n+1}-\hat{r}(X_{n+1}) = Y_{n+1}-g(X_1,\dots,X_n,Y_1,\dots,Y_n,X_{n+1}) \sim Y-g(X_1,\dots,X_n,Y_1,\dots,Y_n,X) = Y-\hat{r}(X)$$

D'où l'idée de séparer l'échantillon en deux si on a suffisamment de données : si on a $n+p$ données, on sépare l'échantillon en prenant $(X_1,Y_1),\dots,(X_n,Y_n)$ pour estimer $\hat{r}$ puis $(X_{n+1},Y_{n+1}),\dots,(X_{n+p},Y_{n+p})$ pour valider l'estimateur (ou estimer le risque de cet estimateur ou faire un choix de paramètre d'ajustement comme le choix de la fenêtre h pour un estimateur par polynômes locaux). On a alors un bon estimateur du risque $\mathbf{E}\Big[(Y-\hat{r}(X))^2\Big]$ en posant

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

En effet on a, en conditionnant sur $(X_1,\dots,X_n)$,

$$Y_{n+1}-\hat{r}(X_{n+1}),\dots,Y_{n+p}-\hat{r}(X_{n+p}) \overset{iid}{\sim} Y-\hat{r}(X)$$

C'est l'idée du Hold-out.

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

$$\mathbf{E}\Big[(Y_i-\hat{r}_{n-1}^{(-i)}(X_i))^2\Big] = \mathbf{E}\Big[(Y-\hat{r}_{n-1}(X))^2\Big]$$

Si on note $\hat{r}_{n-1}$ l'estimateur fabriqué avec seulement $n-1$ données.

Donc la moyenne empirique $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{r}_{n-1}^{(-i)}(X_i))^2$ semble un bon estimateur (en particulier sans biais) de $\mathbf{E}\Big[(Y-\hat{r}_{n-1}(X))^2\Big]$ qui est le "vrai" risque de l'estimateur $\hat{r}_{n-1}$ fabriqué à partir de $n-1$ données (on s'attend à ce que $\mathbf{E}\Big[(Y-\hat{r}_{n-1}(X))^2\Big]$ soir proche de $\mathbf{E}\Big[(Y-\hat{r}_n(X))^2\Big]$ où $\hat{r}_n$ est l'estimateur de départ, fabriqué avec n données).

On admet la proposition suivante, qui relie les poids associés à l'estimateur $\hat{r}_h^{(-i)}$ à ceux associés à l'estimateur $\hat{r}_h$.

Proposition 5.15. *Si* $\hat{r}_h(x)=\sum_{i=1}^{n}\omega_{i,h}(x)Y_i$ *et, pour* $1\le i\le n$, $\hat{r}_h^{(-i)}=\sum_{j\neq i}\tilde{\omega}_{j,h}(x)Y_j$ *alors, pour tout* $j\neq i$

$$\tilde{\omega}_{j,h}(X_i)=\frac{\omega_{j,h}(X_i)}{1-\omega_{i,h}(X_i)}$$

Remarque 5.16. *Cette proposition est également vérifiée pour d'autres estimateurs linéaires (par exemples les splines).*

Pour calculer $(\hat{r}_h^{(-i)})_{1\le i\le n}$ dans le cas des polynômes locaux, on n'a donc pas besoin de faire de calculs supplémentaires. Grâce à la proposition précédente on a facilement le résultat suivant.

Proposition 5.17. *Si* $\hat{r}_h(x) = \sum_{i=1}^n \omega_{i,h}(x) Y_i$ *alors*

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{Y_i - \hat{r}_h(X_i)}{1 - \omega_{i,h}(X_i)} \Big)^2$$

Démonstration. On a

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

avec

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

□

Il existe une alternative qui consiste à remplacer les $\omega_{i,h}(x_i)$ par leur moyenne. Cette alternative s'appelle la validation croisée généralisée. : on pose $\Omega = \sum_{i=1}^n \omega_{i,h}(x_i)$ puis

$$GCV(h) = \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{Y_i - \hat{r}_h(x_i)}{1 - \Omega/n} \Big)^2 = \frac{1}{(1 - \frac{\Omega}{n})^2} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \hat{r}_h(x_i) \Big)^2$$

On minimise ensuite GCV par rapport à h.

Remarquons que si $\Omega \ll n$ alors $(1 - \frac{\Omega}{n})^{-2} \approx 1 + 2\frac{\Omega}{n}$ et donc

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

Code R et exemples

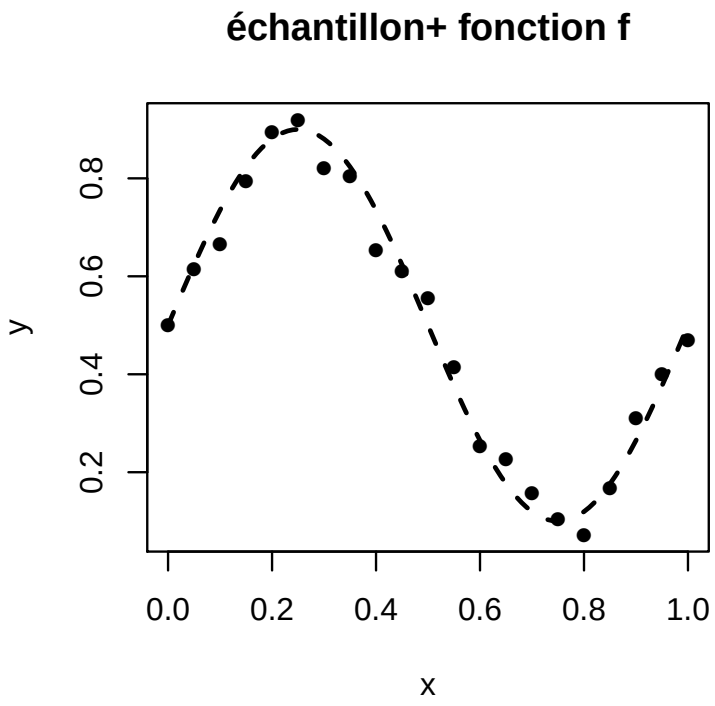
On illustre la méthode des polynômes locaux avec une simulation. La fonction utilisée s'appelle `locpoly` et appartient au package `Kernsmooth`. On peut aussi obtenir une estimation de la fenêtre idéale par la fonction `dpill`. On va représenter les résultats associés à diverses fenêtres (une fenêtre sur-lissant, une sous-lissant, et la fenêtre calculée par la fonction `dpill` associée à un noyau gaussien). Un noyau gaussien est utilisé et cette fonction ne permet que l'estimation d'une fonction à une seule variable. Possibilité d'estimer une dérivée avec l'argument `drv` (mis à zéro par défaut) ou bien une densité. Le degré du polynôme correspond à l'argument `degree` (par défaut à 1).

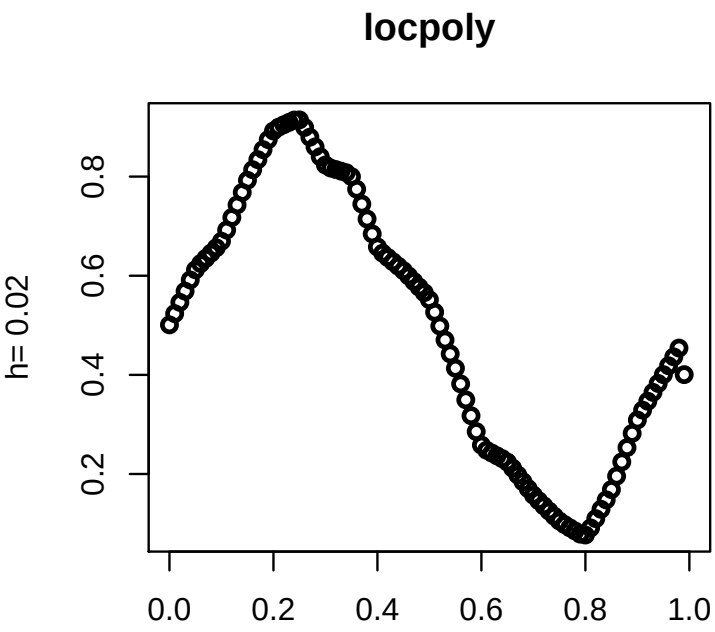
Simulation d'un échantillon associé à une fonction r :

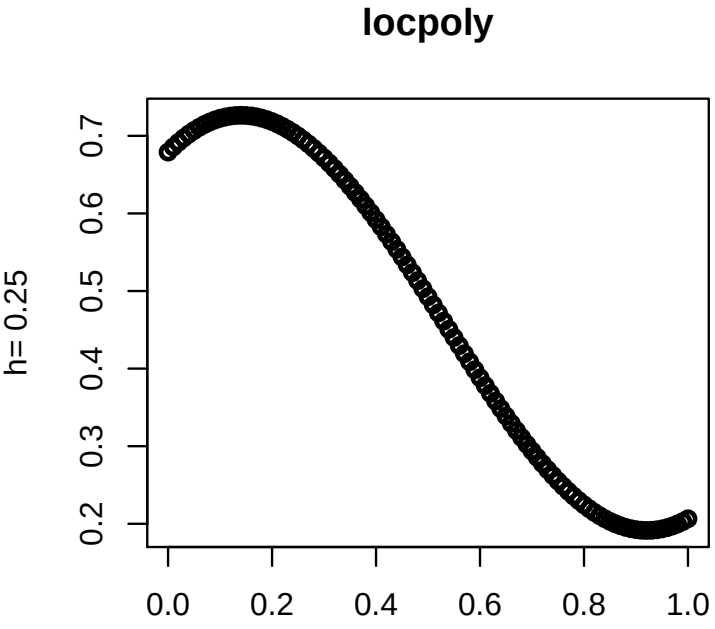
```
>x <- seq(0,1,0.05)
>r <- function(x){0.5 + 0.4*sin(2*pi*x)}
>set.seed(10)
>y <- r(x) + rnorm(n=length(x), sd=0.05)
>par(mfrow=c(2,2))
>plot(x, y, pch=16,main="échantillon+ fonction r")
>xtemp <- seq(0,1,0.01)
>lines(xtemp, r(xtemp), lty=2, lwd=2)
```

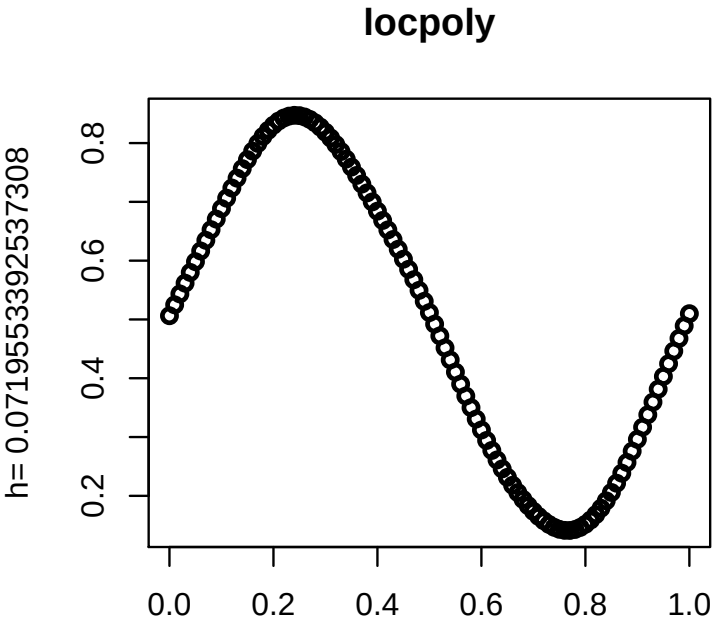
Prediction avec la fonction locpoly : on ne peut pas définir une grille de prédiction quelconque avec cette fonction, seulement une grille de points espacés uniformément

```
library(KernSmooth)
>h=dpill(x,y)        # calcul d'une fenêtre "idéale"
>fenetres=c(0.02,0.25,h)
>for (i in fenetres) {
plot(locpoly(x, y, bandwidth=i,gridsize=101),ylab=paste("h=",i),xlab="",
lwd=2,main="locpoly")
}
```


échantillon+ fonction f
y
x
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
1.0


locpoly
h= 0.02
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
1.0


locpoly
h= 0.25
0.2
0.3
0.4
0.5
0.6
0.7
0.0
0.2
0.4
0.6
0.8
1.0


locpoly
h= 0.0719553392537308
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
1.0

Il existe aussi une fonction appelée `loess` du package `stats` qui permet aussi l'estimation par polynômes locaux, et ce jusqu'à la dimension 4 (de toute façon ce n'est pas très raisonnable d'aller plus loin en dimension).

Et enfin, il existe aussi le package `locfit` dont voici quelques paramètres : `deg` pour le degré du polynôme local (à 2 par défaut, on utilise rarement au-delà de 3) , `kern` pour le noyau (tricube par défaut) , `deriv` pour estimer une dérivée de la fonction de régression.

Le choix de la fenêtre est régi par le paramètre `alpha`. Si on met `alpha=c(0,h)` ça donne un estimateur avec une fenêtre égale à h.

Par exemple, si on veut le polynôme local de degré 1 associé à la régression d'une variable y sur deux variables explicatives x et z avec une fenêtre égale à 0.5, on utilise `locfit(y∼x+z,deg=1,alpha=c(0,0.5))`.

Si `resultat=locfit(..)` alors `fitted(resultat)` donne les $\hat{r}(X_i)$ et
`residuals(resultat)` donne les résidus $r(X_i) - \hat{r}(X_i)$.

On va illustrer l'utilisation de la fonction `gcvplot` associée au package locfit, fonction qui calcule la validation croisée généralisée pour une série de valeurs de alpha et fait le graphique correspondant (attention, en abscisse, ce ne sont pas les valeurs de `alpha`).

Pour cela on va utiliser les mêmes données simulées.

On utilise une grille de 30 valeurs pour la fenêtre :

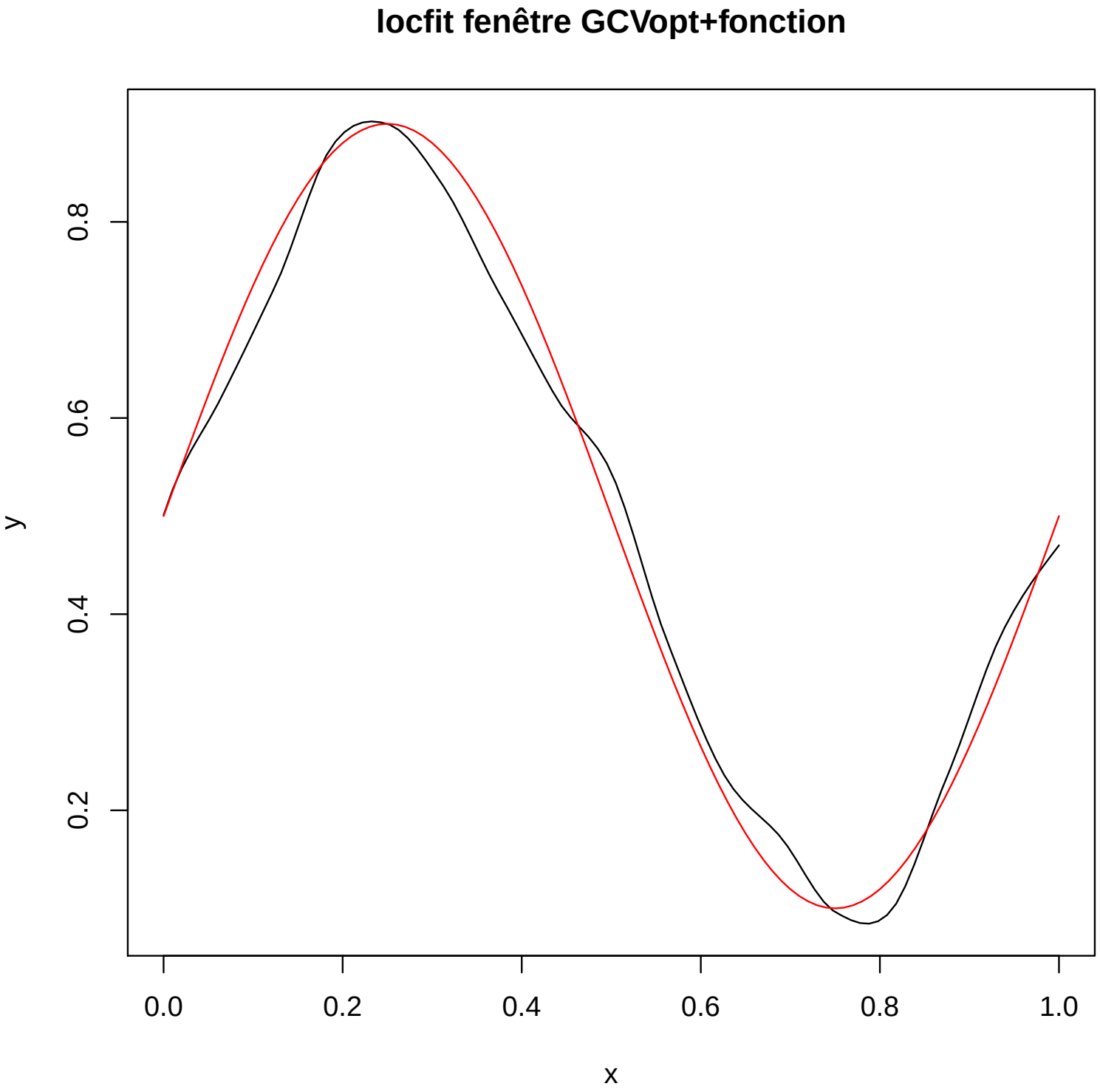
```
>alphamat= matrix(0,ncol=2,nrow=30)
>alphamat[,2]= seq(from=0.1,to=0.8,length=30)
>gcvs= gcvplot(y∼x ,alpha=alphamat,maxk=1000)
```

La fonction `gcvplot` est telle que `gcvplot$values` contient les valeurs de la validation croisée généralisée (GCV en anglais) et `gcvplot$alpha` contient les valeurs de alpha correspondantes. Donc `gcvs$values == min(gcvs$values)` donne la ligne i correspondant à la valeur minimale de la GCV, et avec `gcvs$alpha[i,2]` on obtient la valeur de la fenêtre correspondante. Il se peut que plusieurs valeurs donnent le minimum, auquel cas on prend souvent la plus grande fenêtre donnant le miminum :

```
>optband= max(gcvs$alpha[gcvs$values == min(gcvs$values),2])
```

On peut ensuite fabriquer l'estimateur correspondant à cette fenêtre :

```
>locfitopt= locfit(y∼x,alpha=c(0,optband),maxk=1000)
>plot(locfitopt,main="locfit fenêtre GCV opt+fonction")
>lines(xtemp,r(xtemp),col='red')
```


locfit fenêtre GCVopt+fonction
y
0.2
0.4
0.6
0.8
0.0
0.2
0.4
0.6
0.8
1.0
x

Il y a aussi la possibilité de spécifier une fenêtre différemment, qui n'est pas une fenêtre constante : pour chaque x où la fonction est évaluée, on utilise une fenêtre h_x telle que qu'il y ait une fraction donnée des X_i dans $[x - h_x, x + h_x]$ (ou dans la boule de centre x et de rayon h_x si on est en dimension > 1). Par exemple, si on met `alpha=0.5`, on utilise toujours la moitié des données dans l'intervalle $[x-h_x, x+h_x]$). Ce type de choix est censé être adapté au cas où le design n'est pas distribué assez uniformément et où on peut avoir peu de données à certains endroits.

5.6 Estimateurs par projection

Cette section n'est pas au programme et est donc de lecture facultative.

On se place à nouveau dans le cadre de la régression à effets fixes sur $[0, 1]$. On suppose à présent que la fonction de régression r vérifie $r \in L_2([0; 1])$. On considère $(\phi_j)_{j\geq 1}$ une base orthonormale de $L_2([0; 1])$. On peut écrire

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

au sens de la convergence dans $L_2([0; 1])$ et avec

$$\theta_j = \int_0^1 r(x)\phi_j(x)dx.$$

On a donc, quand N tend vers l'infini, et au sens de la convergence dans L_2,

$$\sum_{j=1}^{N} \theta_j \phi_j \to r.$$

Si on fixe un N grand, et si on arrive à estimer les coefficients θ_j par des estimateurs $\hat{\theta}_j$, il semble naturel d'estimer r par l'estimateur

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

Evidemment, on a le problème du choix de N, qui est équivalent au problème du choix de h pour les estimateurs à noyau. En effet N trop grand donnera une variance trop grande (overfitting) et N trop petit donnera un biais trop grand (underfitting).

Exemple 5.18. *Prenons le cas du dispositif fixe uniforme sur* $[0, 1]$. *Alors on observe*

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

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

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

Bien sûr on ne connait pas $r(i/n)$ *donc on le remplace par son observation bruitée* Y_i, *ce qui donne l'estimateur suivant pour* θ_j

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

et donc l'estimateur suivant pour la fonction de régression

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

On remarque qu'il s'agit d'un estimateur linéaire.

Le choix de la base s'apparente plus au choix du noyau. Les bases les plus fréquemment utilisées sont la base trigonométrique et les bases d'ondelettes.

Base Trigonométrique (de Fourier). Elle est donnée par

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

Base d'ondelettes Soit ψ une fonction suffisamment régulière, à support compact. On définit $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ pour tous $k, j \in \mathbb{Z}$. Alors, sous certaines hypothèses sur ψ, les fonctions $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ forment une base orthonormale de $L_2(\mathbb{R})$.

Chapitre 6

Bibliographie conseillée

— pour les chapitres 4 et 5 : le chapitre 1 [Tsy08] (existe aussi en français), [Gir14]

— pour les chapitres 2 et 3 : [HWC13]

— autres : [LD98, Was06, Dal08, CHJ+12, Loa99]

Bibliographie utilisée pour écrire le poly (ou pour les TDs)

— Notes de cours : introduction à la statistique non paramétrique, Catherine Mathias.

— http://astrostatistics.psu.edu/samsi06/tutorials/tut2larryl_all.pdf

— http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Nonparametriques.pdf

— Statistique inférentielle avancée, notes cours, Olivier Gaudouin http://www-ljk.imag.fr/membres/Olivier.Gaudoin/SIA.pdf

— C.Chabanet, Formation "initiation aux statistiques avec R" : https://informatique-mia.inra.fr/r4ciam/sites/ciam.inra.fr.r4ciam/files/Tutoriels/tp_R.pdf

— F-G Carpentier. Univ Brest, "tests de Kolmogorov-Smirnov et Lilliefors" : http://geai.univ-brest.fr/carpentier/

— "théorème de Dini et application au théorème de Glivenko-Cantelli" dans http://perso.eleves.ens-rennes.fr/people/adrien.fontaine/agregation3.html

— Christophe Chesneau. Sur l'adéquation à une loi de probabilité avec R. Licence. France. 2016. <cel-01387705>

Index

Bibliographie

[CHJ+12] Pierre-André Cornillon, François Husson, Nicolas Jégou, Eric Matzner-Lober, and Collectif. *Statistiques avec R*. PU Rennes, Rennes, 3e édition revue et augmentée edition, May 2012.

[Dal08] Peter Dalgaard. *Introductory Statistics with R*. Springer Science & Business Media, August 2008.

[Gir14] Christophe Giraud. *Introduction to High-Dimensional Statistics*. CRC Press, December 2014.

[HWC13] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, November 2013.

[LD98] Erich Leo Lehmann and H. J. M. D'Abrera. *Nonparametrics : Statistical Methods Based on Ranks*. Prentice Hall, 1998.

[Loa99] Clive Loader. *Local Regression and Likelihood*. Springer, New York, 1999 edition edition, July 1999.

[Tsy08] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York ; London, 1st edition. 2nd printing. 2008 edition edition, November 2008.

[Was06] Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, September 2006.