Tristan Lorino

PROBABILITÉS ET STATISTIQUE

Février 2005

« JE NE PEUX PAS me tromper au sujet de $12 \times 12 = 144$. Et on ne peut pas opposer la sûreté de la *mathématique* au relatif manque de sûreté de propositions empiriques. En effet la proposition mathématique a été obtenue par une série d'actions qui ne se différencient d'aucune façon du reste des actions de la vie et qui sont tout aussi sujettes à l'oubli, l'inadvertence et l'illusion. »

WITTGENSTEIN, *De la certitude*.

« DIEU est une hypothèse dont je n'ai pas eu besoin. »

LAPLACE.

Sommaire

Première partie

VARIABLES ALÉATOIRES ET LOIS DE PROBABILITÉ

1

Rappels d'intégration – Variables aléatoires

1.1 Probabilité

Définition 1.1 — *Une* ***expérience aléatoire*** *se décrit mathématiquement par la donnée d'un ensemble qui représente les résultats possibles de l'expérience. On le note* Ω*. On note* ω *un résultat possible (ou épreuve, issue, réalisation, éventualité, événement élémentaire).*

Définition 1.2 — *Un* ***événement aléatoire*** A *sera toujours représenté par l'ensemble des résultats* ω *de l'expérience qui le réalise.*

$$A \;=\; \{\omega \mid A \text{ est réalisé si } \omega \text{ est le résultat de l'expérience}\}\ .$$

A est réalisé si le résultat de l'expérience ω appartient à A.

Définition 1.3 — *Une famille* $\mathcal{C}$ *de sous-ensembles de* Ω *est une* ***algèbre*** *sur* Ω *si*

1. $\Omega \in \mathcal{C}$ *;*
2. $\mathcal{C}$ *est stable par intersections finies ;*
3. $\mathcal{C}$ *est stable par complémentarité.*

On définit une algèbre d'événements $\mathcal{A}$.

Définition 1.4 — *Une* ***probabilité*** $\mathbb{P}$ *sur* $(\Omega, \mathcal{A})$ *— où* $\mathcal{A}$ *est une algèbre sur* Ω *— est une application additive de* $\mathcal{A}$ *dans [0,1] telle que* $\mathbb{P}(\Omega)=1$*.*

$$\mathbb{P}(A \cup B) \;=\; \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\ .$$

Définition 1.5 — *Pour définir une probabilité sur* $(\Omega, \mathcal{A})$*, il suffit de se donner une famille de nombres* $p(\omega) \geqslant 0$ *telle que*

$$\sum_{\omega\in\Omega} p\big(\{\omega\}\big) \;=\; 1 \;.$$

On pose

$$\mathbb{P}(A) \;=\; \sum_{\omega\in A} p\big(\{\omega\}\big) \;.$$

1.2 Échantillonnage

Soit S une population de taille N : $S = \{s_1, s_2, \ldots, s_N\}$.

Définition 1.6 — *On appelle* ***échantillon*** *de taille* r *une suite ordonnée* $(s_{i_1}, \ldots, s_{i_r})$ *de* r *éléments de* S*.*

Proposition 1.1 — *Le cardinal de l'ensemble des échantillons de taille* r ***avec répétition (replacement)*** *vaut*

$$Card\ \Omega_N^r = N^r \;.$$

Proposition 1.2 — *Le cardinal de l'ensemble des échantillons de taille* r ***sans répétition*** *vaut*

$$\begin{aligned} Card\ \Omega_N^r &= N \times (N-1) \times \cdots \times (N-r+1) \\ &= \frac{N!}{(N-r)!} \;. \end{aligned}$$

Définition 1.7 — *On appelle* ***sous-population*** *de taille* r *de* S *tout ensemble de* r *éléments distincts choisis dans* S*.*

Le cardinal de l'ensemble des sous-populations vaut

$$\begin{aligned} \text{Card}\ \Omega_N^r &= \frac{N \times (N-1) \times \cdots \times (N-r+1)}{r!} \\ &= C_N^r \;. \end{aligned}$$

1.3 Variable aléatoire – Loi de probabilité

Définition 1.8 — *On appelle* ***espace de probabilité*** *le triplet* $(\Omega, \mathcal{A}, \mathbb{P})$ *où :*

- Ω *est l'ensemble des réalisations ;*
- $\mathcal{A}$ *est une tribu sur* Ω *;*
- $\mathbb{P}$ *est une probabilité sur* $\mathcal{A}$*, i.e. une fonction de* $\mathcal{A}$ *dans [0,1]* σ*-additive telle que* $\mathbb{P}(\Omega)$*=1.*

Définition 1.9 — *Une* ***variable aléatoire discrète*** X *(i.e. dont l'ensemble des valeurs est dénombrable) est une application de* $(\Omega, \mathcal{A})$ *dans* E *dénombrable, telle que* $\forall x \in E$*,*

$$\begin{aligned} \{X = x\} &= \{\omega \in \Omega \mid X(\omega) = x\} \\ &= X^{-1}\big(\{x\}\big) \in \mathcal{A} \ . \end{aligned}$$

Définition 1.10 — *La famille de nombres* $\mathbb{P}_X(x) = \mathbb{P}(X = x)$ *est appelée* ***loi de probabilité*** *de* X*.*

Définition 1.11 — *On appelle* ***probabilité conditionnelle*** *de* B *sachant* A

$$\mathbb{P}(B \mid A) \ = \ \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \ .$$

Proposition 1.3 — $\forall A_1, \ldots, A_n \in \mathcal{A}$ *tels que* $\mathbb{P}(A_1 \cap \cdots \cap A_n) > 0$*,*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) \ = \ \mathbb{P}(A_1) \times \mathbb{P}(A_2 \mid A_1) \times \mathbb{P}(A_3 \mid A_1 \cap A_2) \times \cdots \times \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}) \ .$$

Théorème 1.1 (Bayes) — *Nous avons :*

a)

$$\mathbb{P}(B \mid A) \ = \ \frac{\mathbb{P}(B) \cdot \mathbb{P}(A \mid B)}{\mathbb{P}(B) \cdot \mathbb{P}(A \mid B) \ + \ \mathbb{P}(\bar{B}) \cdot \mathbb{P}(A \mid \bar{B})} \ ;$$

b) soit $(E_1, \ldots, E_n)$ *une partition de* Ω*, pour laquelle* $\mathbb{P}(E_i) > 0$*. Soit* $A \in \mathcal{A}$ *tel que* $\mathbb{P}(A) > 0$ *; alors* $\forall i$*,*

$$\mathbb{P}(E_i \mid A) \ = \ \frac{\mathbb{P}(E_i) \cdot \mathbb{P}(A \mid E_i)}{\sum_k \mathbb{P}(E_k) \cdot \mathbb{P}(A \mid E_k)} \ .$$

Définition 1.12 — *A et B sont deux événements **indépendants** si*

$$\mathbb{P}(A \cap B) \;=\; \mathbb{P}(A) \cdot \mathbb{P}(B) \;,$$

i.e. si $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ si $\mathbb{P}(A) \neq 0$. Ceci signifie que la réalisation de A ne donne aucune information sur la réalisation possible de B.

Lemme 1.1 — *Si A et B sont indépendants, alors A et B^c, A^c et B, A^c et B^c le sont également.*

Définition 1.13 — *Une suite finie $A_1, \ldots, A_n$ d'événements est dite indépendante si quelle que soit la suite extraite $i_1, \ldots, i_k$,*

$$\mathbb{P}\Big(\bigcap_{l=1}^{k} A_{i_l}\Big) \;=\; \prod_{l=1}^{k} \mathbb{P}(A_{i_l}) \;.$$

Proposition 1.4 — *$A_1, \ldots, A_n$ forment une suite d'événements indépendants si $\forall D_i \in \{A_i, A_i^c, \emptyset, \Omega\}$,*

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} D_i\Big) \;=\; \prod_{i=1}^{n} \mathbb{P}(D_i) \;.$$

1.4 Propriétés élémentaires des probabilités

Rappels — $\mathbb{P} : (\Omega, \mathcal{A}) \longrightarrow [0, 1]$ est une probabilité si $\mathbb{P}$ est σ-additive, c.-à-d. si $\mathbb{P}$ est additive et stable par limite croissante, c.-à-d. si $\mathbb{P}$ est additive et $A_n \searrow \emptyset \;\Rightarrow\; \mathbb{P}(A_n) \searrow 0$, $\forall A_n$, c.-à-d. encore si $\mathbb{P}$ est additive et $\forall A_n \searrow A$, $\mathbb{P}(A_n) \searrow \mathbb{P}(A)$.

Définition 1.14 — *Soit $(A_n)_n$ une suite infinie d'événements. Alors*

$$\begin{aligned} \overline{\lim_n}\, A_n &= \bigcap_n \bigcup_{k \geqslant n} A_k \\ &= \Big\{ \sum_n 1_{A_n} = \infty \Big\} \;. \end{aligned}$$

Définition 1.15 — *Soit $(A_n)_n$ une suite infinie d'événements. Alors*

$$\begin{aligned} \underline{\lim_n}\, A_n &= \bigcup_n \bigcap_{k \geqslant n} A_k \\ &= \{\textit{tous les } A_n \textit{ sont réalisés sauf un nombre fini}\} \;. \end{aligned}$$

Proposition 1.5 — *Nous avons :*

$$\begin{aligned} \big(\overline{\lim}\, A_n\big)^c &= \underline{\lim}\, A_n^c \ ; \\ \big(\underline{\lim}\, A_n\big)^c &= \overline{\lim}\, A_n^c \ . \end{aligned}$$

Proposition 1.6 — *Nous avons :*

$$\begin{aligned} 1_{\overline{\lim}\, A_n} &= \overline{\lim}\, 1_{A_n} \ ; \\ 1_{\underline{\lim}\, A_n} &= \underline{\lim}\, 1_{A_n} \ . \end{aligned}$$

Définition 1.16 — *Nous avons :*

$$\begin{aligned} A_n \longrightarrow A \quad &\Leftrightarrow \quad 1_{A_n} \longrightarrow 1_A \\ &\Leftrightarrow \quad \overline{\lim}\, 1_{A_n} \ = \ \underline{\lim}\, 1_{A_n} \\ &\Leftrightarrow \quad \overline{\lim}\, A_n \ = \ \underline{\lim}\, A_n \ . \end{aligned}$$

Proposition 1.7 — *Nous avons :*

$$\overline{\lim}\, A_n \ = \ \underline{\lim}\, A_n \quad \Rightarrow \quad A_n \longrightarrow A \ = \ \overline{\lim}\, A_n \ (= \underline{\lim}\, A_n) \ .$$

Proposition 1.8 — *Soit $(A_n)_n$ une suite d'événements.*

$$\mathbb{P}\big(\underline{\lim}\, A_n\big) \ \leqslant \ \underline{\lim}\, \mathbb{P}(A_n) \ \leqslant \ \overline{\lim}\, \mathbb{P}(A_n) \ \leqslant \ \mathbb{P}\big(\overline{\lim}\, A_n\big) \ .$$

Proposition 1.9 — *On dit qu'une suite infinie $(A_n)_n$ d'événements sont indépendants si toute sous-famille finie est formée d'événements indépendants.*

Lemme 1.2 (Borel-Cantelli) — *Soit une suite infinie $(A_n)_n$ d'événements.*
a)

$$\sum_n \mathbb{P}(A_n) < \infty \quad \Rightarrow \quad \mathbb{P}\big(\underline{\lim} A_n\big) = 0 \ .$$

b)

$$\left.\begin{array}{r} \text{les } (A_n)_n \text{ forment une suite indépendante} \\ \sum_n \mathbb{P}(A_n) = \infty \end{array}\right\} \quad \Rightarrow \quad \mathbb{P}\big(\overline{\lim}\, A_n\big) = 1 \ .$$

1.5 Variables aléatoires

Proposition 1.10 — *X est une v.a. discrète si X est mesurable de $(\Omega,\ \mathcal{A})$ dans $\big(E,\ \mathcal{P}(E)\big)$.*

Rappel — Soit $\mathcal{C}$ une classe d'ensembles et $\mathcal{B} = \sigma(\mathcal{C})$ la tribu engendrée par $\mathcal{C}$. Alors $\sigma\big(X^{-1}(\mathcal{C})\big) = X^{-1}\big(\sigma(\mathcal{C})\big)$.

$$\begin{aligned} \forall\ x \in E,\ \ \{X = x\} \in \mathcal{A} \ &\Leftrightarrow\ X^{-1}(\mathcal{C}) \subset \mathcal{A} \ \text{ où } \ \mathcal{C} = \big\{\{x\},\ x \in E\big\} \\ &\Leftrightarrow\ X^{-1}\big(\sigma(\mathcal{C})\big) \subset \mathcal{A}\ . \end{aligned}$$

Or $\sigma(\mathcal{C}) = \mathcal{P}(E)$.

Proposition 1.11 — *$X : (\Omega,\ \mathcal{A}) \longrightarrow \mathbb{R}$ est une v.a. réelle si X est mesurable de $(\Omega,\ \mathcal{A})$ dans $\big(\mathbb{R},\ \mathcal{B}(\mathbb{R})\big)$.*

Proposition 1.12 — *Nous avons :*

$$\begin{aligned} X \ \textit{ v.a. réelle} \quad &\Leftrightarrow \quad X \textit{ fonction borélienne de } (\Omega,\ \mathcal{A}) \textit{ dans } \mathbb{R} \\ &\Leftrightarrow \quad \forall \mathcal{O} \textit{ ouvert de } \mathbb{R},\ \{X \in \mathcal{O}\} = X^{-1}(\mathcal{O}) \in \mathcal{A}\ . \end{aligned}$$

Proposition 1.13 — *Nous avons :*

$$\begin{aligned} X \ \textit{ v.a. vectorielle} \quad &\Leftrightarrow \quad X : (\Omega,\ \mathcal{A}) \longrightarrow \big(\mathbb{R}^d,\ \mathcal{B}(\mathbb{R}^d)\big) \\ &\Leftrightarrow \quad \forall\ i \in \{1, \ldots, d\},\ \ X_i \ \textit{ v.a. réelle}\ . \end{aligned}$$

Proposition 1.14 — *L'espace des v.a. est stable par addition, multiplication, passage au sup et à l'inf. Ainsi, si une suite $(X_n)_n$ est une suite de v.a., alors $\underline{\lim} X_n$ et $\overline{\lim} X_n$ sont des v.a.*

Proposition 1.15 — *Soit $(X_n)_n$ une suite de v.a.*

$$\begin{aligned} \big\{\underline{\lim} X_n = \overline{\lim} X_n\big\} \ &= \ \big\{\omega \mid \underline{\lim} X_n(\omega) = \overline{\lim} X_n(\omega)\big\} \\ &= \ \textit{domaine de convergence de } \ (X_n)_n \end{aligned}$$

et ce domaine appartient à $\mathcal{A}$.

De plus, si $X_n \longrightarrow X$, alors X est une v.a.

Définition 1.17 — *Une propriété des points ω est dite **vraie presque sûrement** (noté **p.s.**) si l'ensemble où elle est fausse est contenu dans un événement A tel que $\mathbb{P}(A) = 0$.*

Remarque — Cette notion est l'équivalent du « presque partout » de la théorie de l'intégration.

Définition 1.18 — *Une v.a. X_n converge presque sûrement vers X si* $\mathbb{P}\big(\lim_n X_n = X\big) = 1$.

Proposition 1.16 — *Si $(X_n)_n$ est une suite de v.a. réelles telles que $\exists X$ v.a. satisfaisant*

$$\forall \epsilon > 0, \quad \sum_n \mathbb{P}\big(|X_n - X| > \epsilon\big) < \infty \ ,$$

alors $X_n \xrightarrow{p.s.} X$.

Définition 1.19 — *Une v.a. réelle est dite* ***étagée*** *si elle ne prend qu'un nombre fini de valeurs. X s'écrit sous la forme*

$$X \ = \ \sum_{i=1}^{n} a_i 1_{A_i} \ ,$$

où $A_i = \{X = a_i\}$.

Proposition 1.17 — *Toute v.a. positive est limite croissante de v.a. étagées.*

Proposition 1.18 — *Toute v.a. réelle est différence de deux v.a. positives : $X = X^+ - X^-$, où $X^+ = \sup(X, 0)$ et $X^- = \sup(-X, 0)$.*

1.6 Moment — Espérance

Définition 1.20 — *Soit $(\Omega, \mathcal{A}, \mathbb{P})$. Soit X une v.a. réelle. On dit que X admet un* ***moment d'ordre 1*** *si X est une fonction intégrable par rapport à $\mathbb{P}$:*

$$\int |X| \, \mathrm{d}\mathbb{P} \ < \ \infty \ .$$

X admet un **moment d'ordre q** si

$$\int |X|^q \, \mathrm{d}\mathbb{P} \ < \ \infty \ .$$

Définition 1.21 — *On appelle* ***espérance mathématique*** *(**moyenne**) de la v.a. X et on la note $\mathbb{E}(X)$ la quantité*

$$\int X \, d\mathbb{P} \ .$$

Théorème 1.2 (Convergence monotone) — *Soit $(X_n)_n$ une suite de v.a. positives tendant en croîssant vers X. Alors $\mathbb{E}(X_n)$ tend en croîssant vers $\mathbb{E}(X)$.*

Remarque — Dans le théorème précédant, $\mathbb{E}(X_n)$ et $\mathbb{E}(X)$ peuvent être infinies.

Lemme 1.3 (Fatou) — *Soit $(X_n)_n$ une suite de v.a. positives. Alors*

$$\mathbb{E}\big(\underline{\lim} X_n\big) \ \leqslant \ \overline{\lim} \ \mathbb{E}(X_n) \ .$$

Théorème 1.3 (Lebesgue — Convergence dominée) — *Soit $(X_n)_n$ telle que :*

$X_n \xrightarrow{p.s.} X$ *;*

$\forall n, |X_n| \leqslant Y$*, Y ayant un moment d'ordre 1.*

Alors $\mathbb{E}(X_n) \longrightarrow \mathbb{E}(X)$.

Proposition 1.19 — *Nous avons :*

$$\mathbb{E}(aX + bY) = a \ \mathbb{E}(X) + b \ \mathbb{E}(Y) \ .$$

Proposition 1.20 —

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \geqslant a) \ \mathrm{d}a \ .$$

1.7 Variance — Covariance — Corrélation — Écart-type

Définition 1.22 — *La* ***variance*** *est définie par*

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)^2\Big] \\ &= \mathbb{E}(X^2) - \big[\mathbb{E}(X)\big]^2 \ . \end{aligned}$$

Définition 1.23 — *La **covariance** entre deux v.a. X et Y est définie par*

$$\begin{aligned}\mathbb{C}ov(X,Y) &= \mathbb{E}\Big[\big(X-\mathbb{E}\,(X)\big)\big(Y-\mathbb{E}(Y)\big)\Big] \\ &= \mathbb{E}(XY)-\mathbb{E}(X)\cdot\mathbb{E}(Y)\ .\end{aligned}$$

Proposition 1.21 — *$(X,Y)\mapsto \mathbb{C}ov(X,Y)$ est bilinéaire.*

Interprétation — Si $\mathbb{C}\mathrm{ov}\,(X,Y)>0$, X et Y sont liées positivement, *i.e.* elles ont tendance à évoluer dans le même sens.

Propriété 1.1 — *Nous avons*

$$\mathbb{V}\,(aX+bY)=a^2\ \mathbb{V}\,(X)+b^2\ \mathbb{V}\,(Y)+2ab\ \mathbb{C}ov(X,Y)\ .$$

Propriété 1.2 — *Si X et Y sont indépendantes, alors $\mathbb{C}ov(X,Y)=0$.*
(La réciproque est fausse.)

Définition 1.24 — *Le **coefficient de corrélation** entre X et Y est défini par*

$$\rho(X,Y)=\frac{\mathbb{C}ov(X,Y)}{\sigma_X\cdot\sigma_Y}\ .$$

Propriété 1.3 — *Nous avons*

$$\big|\rho(X,Y)\big|\ \leqslant\ 1\ .$$

Propriété 1.4 — *Si $\big|\rho(X,Y)\big|=1$, la liaison est dite complète et linéaire entre X et Y : $Y=aX+b$.*
Si $\rho(X,Y)=1$, alors $a>0$.
Si $\rho(X,Y)=-1$, alors $a<0$.

Définition 1.25 — *L'**écart-type** est défini par*

$$\sigma(X)=\sqrt{\mathbb{V}\,(X)}\ .$$

Théorème 1.4 (Inégalité de Chebichev) — *Nous avons :*

$$\mathbb{P}\Big(\big|E-\mathbb{E}\,(X)\big|\ \geqslant\epsilon\Big)\ \leqslant\ \frac{\mathbb{V}\,(X)}{\epsilon^2}\ .$$

Définition 1.26 — *Soit X une v.a. de moyenne μ et de variance σ^2. On appelle* ***coefficient de variation*** *la quantité*

$$100 \times \frac{\mu}{\sigma} \ .$$

Exprimée en pourcentage, elle permet de comparer deux séries de moyennes différentes.

2

Lois de probabilité

2.1 Lois de variables aléatoires

Définition 2.1 — *On appelle* ***tribu engendrée par une v.a.*** $X : \Omega \longrightarrow \mathbb{R}^d$ *la classe* $\sigma(X)$ *définie par*

$$\begin{aligned} \sigma(X) &= X^{-1}\big(\mathcal{B}(\mathbb{R}^d)\big) \\ &= \big\{\{X \in A\},\, A \in \mathbb{R}^d\big\} \ . \end{aligned}$$

Proposition 2.1 — *Soit* $Y : \big(\Omega, \sigma(X)\big) \longrightarrow \big(\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'})\big)$ *mesurable. Alors il existe une fonction* f *mesurable de* $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$ *dans* $\big(\mathbb{R}^{d'}, \mathcal{B}(\mathbb{R}^{d'})\big)$ *telle que* $Y = f(X)$.

Définition 2.2 — *On appelle* ***loi de probabilité*** *de la v.a.* X *définie sur* $(\Omega, \mathcal{A}, \mathbb{P})$ *et à valeurs dans un espace* E *la probabilité* $\mathbb{P}_X$ *image de* $\mathbb{P}$ *par* X.
On dira que $\mathbb{P}_X$ *est la loi (ou* ***distribution****) de* X*, ou encore que* X *suit la loi* $\mathbb{P}_X$.

Formulation — Nous avons :

$$\begin{aligned} \mathbb{E}\,\big(f(X)\big) &= \int f(X)\,\mathrm{d}\mathbb{P} \\ &= \int_\Omega f(X(\omega))\,\mathrm{d}\mathbb{P}(\omega) \\ &= \int_{\mathbb{R}^d} f(x)\,\mathrm{d}\mathbb{P}_X(x) \ . \end{aligned}$$

puisque, pour $f = 1_A$,

$$\begin{aligned}\mathbb{E}\left(1_A(X)\right) &= \int 1_A(X)\,\mathrm{d}\mathbb{P} \\ &= \int_\Omega 1_{X^{-1}(A)}\,\mathrm{d}\mathbb{P} \\ &= \mathbb{P}_X(A) \\ &= \int 1_A(x)\,\mathrm{d}\mathbb{P}_X(x)\ . \end{aligned} \tag{2.1}$$

Ceci est vrai pour 1_A, donc pour les fonctions étagées, et par suite pour les fonctions positives (d'après le théorème de convergence monotone).

Proposition 2.2 — *Si l'un des deux membres de (2.1) a un sens, alors l'autre en a aussi un, et il y a égalité.*

Proposition 2.3 — *Si deux probabilités sont égales sur une classe $\mathcal{C}$ stable par intersections finies, alors elles sont égales sur la tribu engendrée par $\mathcal{C}$.*

Théorème 2.1 (Théorème de la classe monotone) — *Soit $\mathcal{C} \subset \mathcal{P}(\Omega)$ stable par intersections finies et contenant Ω.*

La plus petite classe monotone (i.e. stable par différence finie et limite croissante) contenant $\mathcal{C}$ est la tribu engendrée par $\mathcal{C}$.

Définition 2.3 — *Soit X et X' deux v.a. On dit qu'elles sont **équidistantes** si $\mathbb{P}_X = \mathbb{P}_{X'}$ et on note $X \overset{(d)}{=} X'$.*

Définition 2.4 — *La loi de X est dite **symétrique** si $\mathbb{P}_X = \mathbb{P}_{-X}$.*

Définition 2.5 — *La **fonction de répartition** de la v.a. réelle X est définie par $\forall t \in]-\infty, +\infty[$,*

$$\begin{aligned} F_X(t) &= \mathbb{P}(X \leqslant t) \\ &= \mathbb{P}_X\big(]-\infty, t]\big)\ . \end{aligned}$$

Proposition 2.4 — *L'application qui à $\mathbb{P}_X$ associe F_X est injective.*

Proposition 2.5 — *Toute fonction de répartition $F_X :]-\infty, +\infty[\longrightarrow [0,1]$ satisfait :*

(i) F_X croissante ;

(ii) F_X continue à droite ;

(iii) $F_X(t) \to 0 \quad (t \to -\infty)$;
(iv) $F_X(t) \to 1 \quad (t \to +\infty)$.

Définition 2.6 — *Soit f une fonction de $\mathbb{R}$ dans $[0,1]$ vérifiant (i), (ii), (iii) et (iv). Alors il existe une v.a. X et donc une probabilité $\mathbb{P}_X$ qui admettent f comme fonction de répartition.*

Définition 2.7 — *Les points de discontinuité de F_X sont les points **chargés** par $\mathbb{P}_X$.*

Définition 2.8 — *Si $\mathbb{P}_X$ ne charge aucun point, on dit que la loi de X est **diffuse**.*

Proposition 2.6 — *Si la loi de X est diffuse, alors F_X est continue.*

Lemme 2.1 — *L'ensemble des points chargés par une probabilité est au plus dénombrable.*

Proposition 2.7 — *Toute probabilité $\mathbb{P}_X$ s'écrit comme somme d'une mesure chargeant un ensemble dénombrable de points (i.e. mesure discrète) et d'une mesure diffuse.*

Définition 2.9 — *On appelle **densité** de probabilité sur $\mathbb{R}^d$ toute fonction borélienne positive d'intégrale par rapport à la mesure de Lebesgue égale à 1.*

On appelle probabilité de densité f la probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ définie par : $\forall B \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}\mathbb{P}(B) &= \int_B f(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} 1_B f(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} 1_B(x_1,\ldots,x_d)\, f(x_1,\ldots,x_d)\,\mathrm{d}(x_1,\ldots,x_d)\ .\end{aligned}$$

Proposition 2.8 — *Si la probabilité $\mathbb{P}$ sur $\mathbb{R}^d$ a pour densités f et g, alors $f = g$ pp.*

Proposition 2.9 — *Toute probabilité définie par une densité par rapport à la mesure de Lebesgue est diffuse.*

Définition 2.10 — *On dit qu'une v.a. X à valeurs dans $\mathbb{R}^d$ a pour densité p si la loi de probabilité $\mathbb{P}_X$ a pour densité p sur $\mathbb{R}^d$.*

Définition 2.11 — *Soient X une v.a. à valeurs dans $\mathbb{R}^d$ de densité p, et f une fonction borélienne positive de $\mathbb{R}^d$ dans $\mathbb{R}$. Alors*

$$\begin{aligned}\mathbb{E}\left[f(X)\right] &= \int_{\mathbb{R}^d} f(x)\,\mathrm{d}\mathbb{P}_X(x) \\ &= \int_{\mathbb{R}^d} f(x)\,p(x)\,\mathrm{d}x\ .\end{aligned}$$

Proposition 2.10 — *Si X a une densité paire, sa loi est symétrique (i.e. $\mathbb{P}_X = \mathbb{P}_{-X}$).*

Proposition 2.11 — *Si la loi de X est symétrique, $\forall f$ impaire telle que $f(X)$ soit intégrable, $\mathbb{E}\left[f(X)\right] = 0$.*

Proposition 2.12 — *Soit $(X,\,Y)$ un couple de vecteurs aléatoires de dimensions respectives d et d', et ayant une densité $p(x,\,y)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^{d+d'}$. Alors X et Y ont pour densités respectives*

$$p_X(x) = \int_{\mathbb{R}^{d'}} p(x,\,y)\,\mathrm{d}y$$

et

$$p_Y(y) = \int_{\mathbb{R}^d} p(x,\,y)\,\mathrm{d}x\ .$$

Définition 2.12 — *$p(x,\,y)$ est appelée **loi conjointe** de X et Y. Quant à p_X et p_Y, elles sont appelées **lois marginales** de (respectivement) X et Y.*

Proposition 2.13 (Formule de changement de variables) — *Soit X un vecteur aléatoire à valeurs dans un ouvert U de $\mathbb{R}^k$. Soit $\phi = (\phi_1, \ldots, \phi_k)$ un difféomorphisme[1] de U dans un ouvert V de $\mathbb{R}^k$. Soit J_ϕ son jacobien :*

$$J_\phi = \begin{vmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_k} \\ \vdots & \vdots & \vdots \\ \frac{\partial \phi_k}{\partial x_1} & \cdots & \frac{\partial \phi_k}{\partial x_k} \end{vmatrix}\ .$$

Soit $Y = \phi(X)$. On suppose que X a une densité f_X par rapport à la mesure de Lebesgue et Y une densité f_Y. Alors

$$f_{\phi(X)}(x) = \frac{1}{\left|J_\phi\big(\phi^{-1}(x)\big)\right|}\ f_X\big(\phi^{-1}(x)\big)\ .$$

1. Fonction bijective de classe C^1.

Théorème 2.2 — *Soient X et Y deux v.a. indépendantes de densités f et g par rapport à la mesure de Lebesgue. Alors XY a pour densité*

$$x \longmapsto \int f\left(\frac{x}{y}\right) \cdot g(y) \cdot \frac{1}{|y|} \, \mathrm{d}y \ .$$

2.2 Lois discrètes usuelles

2.2.1 Loi de Bernouilli

On la note $\mathcal{B}(p)$. C'est la représentation de l'alternative oui/non :

$$X = \begin{cases} 1 & \text{si oui (avec la probabilité } p), \\ 0 & \text{si non (avec la probabilité } 1-p). \end{cases}$$

La densité est :

$$\mathbb{P}(X = x) \ = \ p^x (1-p)^{1-x}$$

et les premiers moments donnent :

$$\begin{aligned} \mathbb{E}(X) &= p \ , \\ \mathbb{V}(X) &= p(1-p) \ . \end{aligned}$$

2.2.2 Loi binomiale

On la note $\mathcal{B}(n, p)$. C'est la répétition (n fois et de façon indépendante) de l'alternative précédente.

La densité est

$$\mathbb{P}(X = x) \ = \ C_n^x \, p^x (1-p)^{n-x} \ .$$

et les premiers moments donnent

$$\begin{aligned} \mathbb{E}(X) &= np \ , \\ \mathbb{V}(X) &= np(1-p) \ . \end{aligned}$$

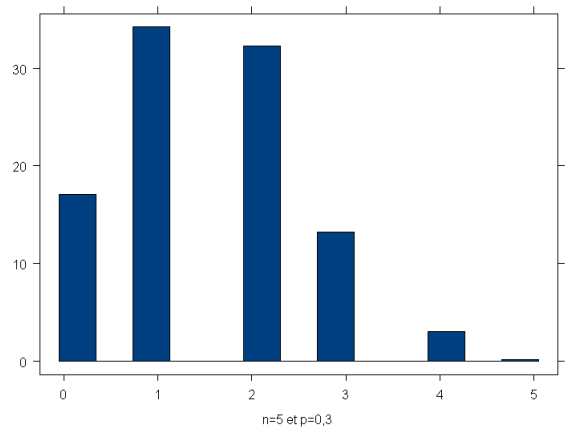


FIGURE 2.1 — Loi binomiale.

Proposition 2.14 — *Quand $n \to \infty$ et $p \to 0$ de manière à ce que $np \to \lambda$, $\mathcal{B}(n,p) \to \mathcal{P}(\lambda)$ — qui est une loi de Poisson (voir section suivante).*

2.2.3 Loi de Poisson

On la note $\mathcal{P}(\lambda)$, avec $\lambda \in \mathbb{R}^+$.

La densité est

$$\mathbb{P}(X = n) \;=\; \frac{e^{-\lambda}\lambda^n}{n!} \;.$$

et les premiers moments donnent

$$\begin{aligned} \mathbb{E}(X) &= \lambda \,, \\ \mathbb{V}(X) &= \lambda \,. \end{aligned}$$

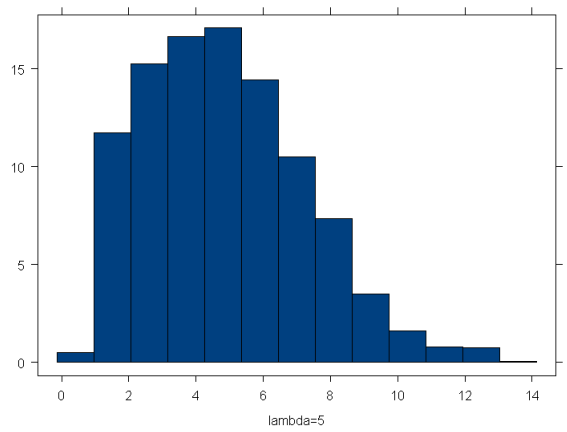


FIGURE 2.2 — Loi de Poisson.

2.2.4 Loi géométrique (ou de Pascal)

On la note $\mathcal{G}(p)$. Elle représente le nombre d'expériences nécessaires pour avoir le premier succès — sachant que la probabilité de succès est p.

La densité est

$$\mathbb{P}(X = x) \;=\; p(1-p)^{x-1} \;.$$

et les premiers moments donnent

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{p} \,, \\ \mathbb{V}(X) &= \frac{1-p}{p^2} \,. \end{aligned}$$

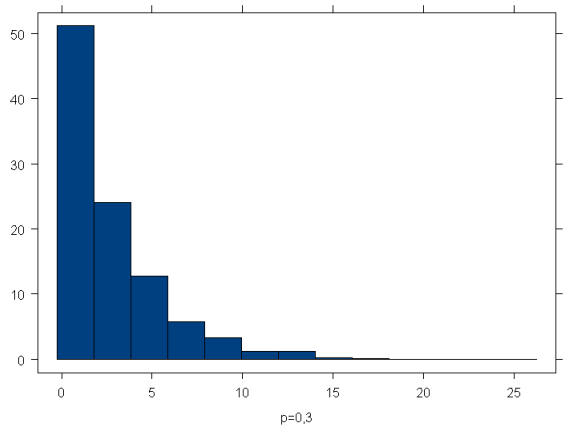


FIGURE 2.3 — Loi géométrique.

Proposition 2.15 — *Si X suit une loi géométrique, alors*

$$\mathbb{P}(X \geqslant m+n) = \mathbb{P}(X \geqslant m) \times \mathbb{P}(X \geqslant n) \ ,$$

$\forall n, m \in \mathbb{N}$.

2.2.5 Loi géométrique généralisée

On la note $\mathcal{R}(n, p)$. Elle représente le nombre d'expériences nécessaires pour obtenir n succès.

La densité est

$$\mathbb{P}(X = x) \ = \ C_{x-1}^{n-1} \, p^n (1-p)^{x-n} \ .$$

et les premiers moments donnent

$$\begin{aligned} \mathbb{E}(X) \ &= \ \frac{n}{p} \ , \\ \mathbb{V}(X) \ &= \ \frac{n(1-p)}{p^2} \ . \end{aligned}$$

2.2.6 Loi hypergéométrique

On la note $\mathcal{H}(N, n, p)$.

Soient N boules dans une urne réparties comme suit : $N \times p$ boules rouges (p est la proportion de boules rouges) et $N \times (1-p)$ boules blanches. On tire n boules et on s'intéresse au nombre de boules rouges (soit X) sur ces n.

Si on tire les boules les unes après les autres avec remise immédiate, $X \rightsquigarrow \mathcal{B}(n, p)$.

Si le tirage est global, ou si l'on tire les n boules les unes après les autres sans remise, alors il s'agit d'un tirage **exhaustif** et d'une loi hypergéométrique.

La densité est

$$\mathbb{P}(X = x) \ = \ \frac{C_{Np}^{x} \cdot C_{N(1-p)}^{n-x}}{C_N^n} \ .$$

et les premiers moments donnent

$$\begin{aligned} \mathbb{E}(X) \ &= \ np \ , \\ \mathbb{V}(X) \ &= \ np(1-p)\frac{N-n}{N-1} \ . \end{aligned}$$

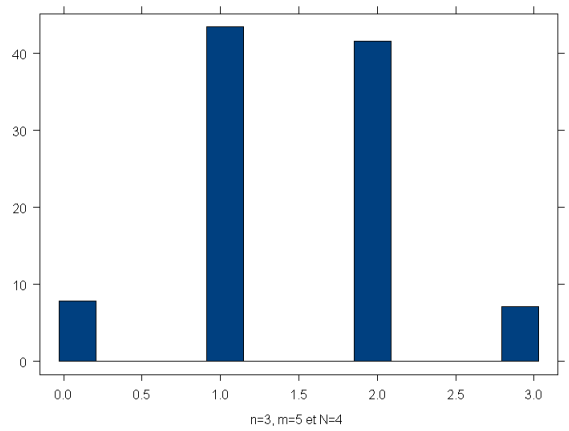


FIGURE 2.4 — Loi hypergéométrique.

Définition 2.13 — $\dfrac{N-n}{N-1}$ *est le* ***coefficient d'exhaustivité****.*

2.2.7 Loi binomiale négative

Cette loi admet deux paramètres : r et p ($0 \leqslant p \leqslant 1$), et elle représente la probabilité d'obtenir $r-1$ succès et x échecs sur $x+r-1$ tentatives.

La densité est

$$\mathbb{P}(X=x) \;=\; \binom{x+r-1}{r-1} p^r (1-p)^x$$

et les deux premiers moments donnent

$$\begin{aligned} \mathbb{E}(X) &= \frac{r(1-p)}{p}\,, \\ \mathbb{V}(X) &= \frac{r(1-p)}{p^2}\,. \end{aligned}$$

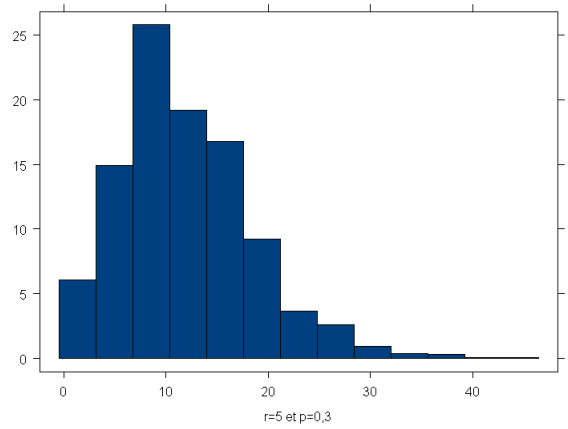


FIGURE 2.5 — Loi binomiale négative.

2.2.8 Loi discrète

Les paramètres sont $\{x_1, \ldots, x_n\}$, $p_1, \ldots, p_n$ ($\forall i$, $0 \leqslant p_i \leqslant 1$). Cette loi représente la probabilité d'obtenir l'une quelconque des valeurs $\{x_1, \ldots, x_n\}$, sachant que chacune de ces valeurs a respectivement la probabilité $p_1, \ldots, p_n$ d'être tirée au sort.

Nous avons :

$$\begin{aligned} \mathbb{P}(X=x_i) &= p_i\,, \\ \mathbb{E}(X) &= \sum_i x_i p_i\,, \\ \mathbb{V}(X) &= \sum_i x_i^2 p_i - (\sum_i x_i p_i)^2 \quad . \end{aligned}$$

2.2.9 Loi multinomiale

Dans une population de N individus, on distingue r types distincts ; soient $N_1, N_2, \ldots, N_r$ les nombres respectifs d'individus de type $1, 2, \ldots, r$.

On fait un sondage portant sur n individus : soit X_i la réponse du i^{e} individu. Soit enfin

$$Z_j = \sum_{i=1}^{n} \mathbb{1}\!\mathbb{1}_{\{X_i=j\}}$$

le nombre de réponses de type j.

Si on effectue le sondage en prélevant globalement un groupe de n individus, il s'agit d'un sondage « sans remise » : la loi de $(Z_1, \dots, Z_n)$ est alors hypergéométrique.

On effectue ici un sondage avec remise. Pour $j = 1, \dots, r$, posons $p_j = N_j/N$. On suppose qu'après chaque tirage, l'individu interrogé est remis dans la population, et que les tirages successifs sont indépenants. Les v.a. $(X_1, \dots, X_n)$ sont indépendantes et $\mathbb{P}(X_i = j) = p_j$.

Soit $E = \big\{(i_1, \dots, i_r) \in \mathbb{N}^r \ ; \ i_1 + \cdots + i_r = n\big\}$. Pour $(i_1, \dots, i_r) \in E$, on a :

$$\mathbb{P}(Z_1 = i_1, Z_2 = i_2, \dots, Z_r = i_r) \quad = \quad \frac{n!}{i_1! \dots i_r!} \, p_1^{i_1} \ \dots \ p_r^{i_r} \ .$$

Nous avons

$$\begin{aligned} \mathbb{E}\,(Z_i) &= Np_i \ , \\ \mathbb{V}\,(Z_i) &= Np_i(1-p_i) \end{aligned}$$

et

$$\mathbb{C}\mathrm{ov}\,(Z_i, Z_j) \ = \ -Np_ip_j \ .$$

2.3 Lois continues usuelles

2.3.1 Loi uniforme

On la note $\mathcal{U}\big([a, b]\big)$.

Nous avons :

$$f(x) \quad = \quad \begin{cases} 0 \text{ si } x < a \text{ ou } x > b \\ \\ \frac{1}{b-a} \text{ si } a < x < b \\ \\ \text{pas définie si } x = a \text{ ou } b \end{cases}$$

$$F(x) \quad = \quad \begin{cases} 0 \text{ si } x < a \\ \\ \frac{x-a}{b-a} \text{ si } a \leqslant x \leqslant b \\ \\ 1 \text{ si } x \geqslant b \end{cases}$$

$$\mathbb{E}\,(X) \quad = \quad \frac{a+b}{2} \ ,$$

$$\mathbb{V}\,(X) \quad = \quad \frac{(b-a)^2}{12} \ .$$

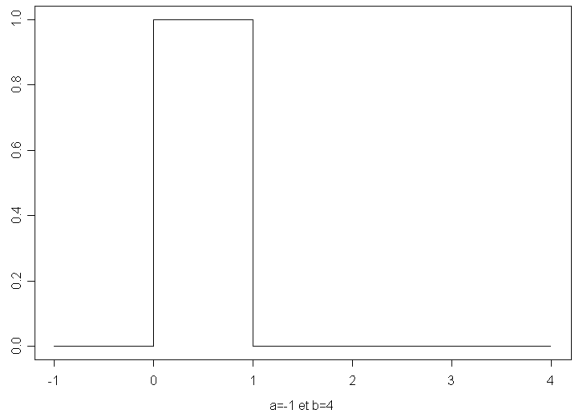


FIGURE 2.6 — Loi uniforme.

2.3.2 Loi exponentielle

On la note $\mathcal{E}(\lambda)$, avec $\lambda > 0$.

Nous avons :

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}^+}(x) \ , \\ F(x) &= 1 - e^{-\lambda x} \mathbf{1}_{\mathbb{R}^+}(x) \ , \\ \mathbb{E}(X) &= \frac{1}{\lambda} \ , \\ \mathbb{V}(X) &= \frac{1}{\lambda^2} \ . \end{aligned}$$

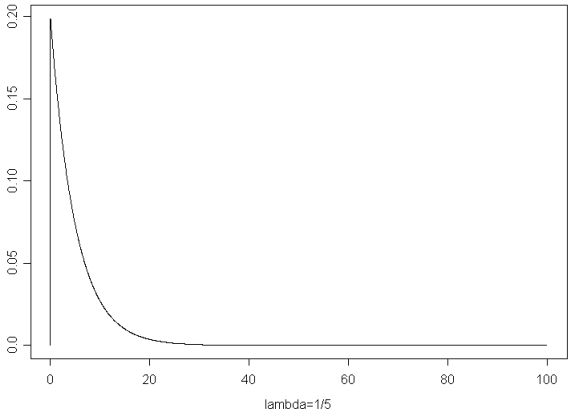


FIGURE 2.7 — Loi exponentielle.

Proposition 2.16 — *Si X suit une loi exponentielle, X a des moments de tout ordre et*

$$\mathbb{E}(X^p) = \frac{p!}{\lambda^p} \ .$$

Proposition 2.17 — *Si X suit une loi exponentielle,*

$$\mathbb{P}(X > a + b \mid X > b) = \mathbb{P}(X > a) \ .$$

2.3.3 Loi normale

On la note $\mathcal{N}(\mu, \sigma^2)$. Nous avons [a] :

$$\begin{aligned} f(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} , \\ \mathbb{E}(X) &= \mu , \\ \mathbb{V}(X) &= \sigma^2 . \end{aligned}$$

a. La fonction de répartition n'est pas définie.

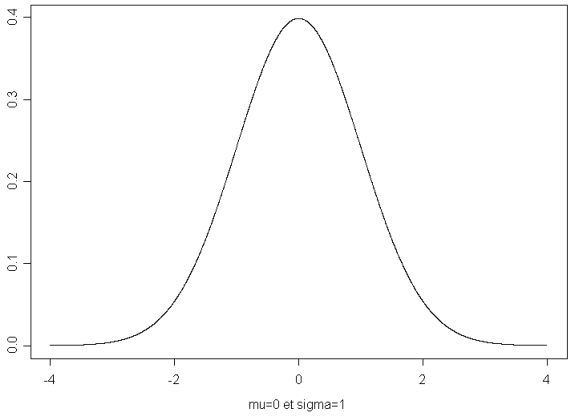


FIGURE 2.8 — Loi normale.

Proposition 2.18 — *Si X suit une loi normale, X a des moments de tout ordre.*

Une loi normale $\mathcal{N}(\mu,\sigma^2)$ vérifie :
68 % de la distribution est dans l'intervalle $[\mu - \sigma\ ,\ \mu + \sigma]$
95 % de la distribution est dans l'intervalle $[\mu - 2\sigma\ ,\ \mu + 2\sigma]$
99,8 % de la distribution est dans l'intervalle $[\mu - 3\sigma\ ,\ \mu + 3\sigma]$.

2.3.4 Loi de Cauchy

X suit une loi de Cauchy de paramètre 1 si X a pour densité

$$\frac{1}{\pi} \cdot \frac{1}{1+x^2} .$$

Elle est portée par $\mathbb{R}$ et est symétrique.
Elle n'a pas de moment d'ordre 1.

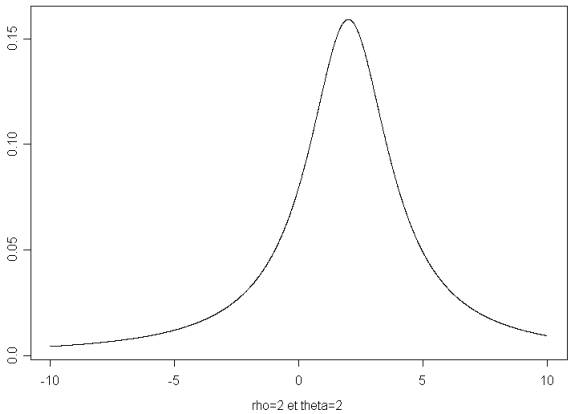


FIGURE 2.9 — Loi de Cauchy.

2.3.5 Loi gamma

On la note $\Gamma(\rho, \theta)$, avec $\rho, \theta > 0$.
Nous avons :

$$f(x) = \frac{\theta^\rho}{\Gamma(\rho)} e^{-\theta x} x^{\rho-1} \mathbf{1}_{\mathbb{R}^+}(x) \ .$$

avec

$$\begin{aligned} \Gamma(\rho) &= \int_0^{+\infty} e^{-x} x^{\rho-1} \, \mathrm{d}x \\ &= (\rho-1)! \quad \text{si} \quad p \in \mathbb{N}^* \ . \end{aligned}$$

$$\begin{aligned} \mathbb{E}(X) &= \frac{\rho}{\theta} \ , \\ \mathbb{V}(X) &= \frac{\rho}{\theta^2} \ . \end{aligned}$$

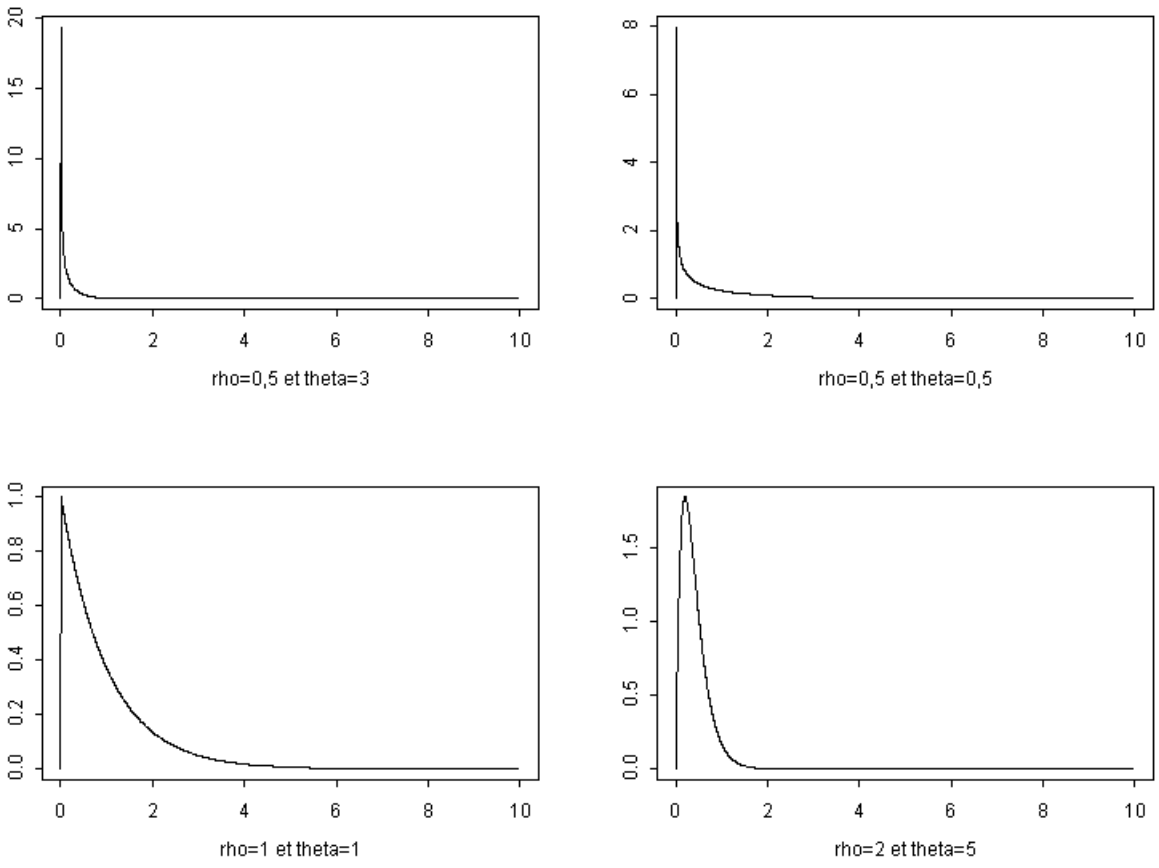


FIGURE 2.10 — Loi gamma.

Nota — Si $p = 1$, on retrouve la loi exponentielle.

Rappel — Nous avons :

$$\Gamma(n+1) \;=\; n\,!$$

et

$$\Gamma(x+1) \;=\; x\,\Gamma(x).$$

2.3.6 Loi Bêta

On la note $\beta(\rho, \theta)$, avec $\rho, \theta > 0$.
Nous avons :

$$\begin{aligned} f(x) &= \frac{\Gamma(\rho+\theta)}{\Gamma(\rho)\cdot\Gamma(\theta)}\; x^{\rho-1}\,(1-x)^{\theta-1}\;\mathbf{1}_{]0,1[}(x). \\ \mathbb{E}\,(X) &= \frac{\rho}{\rho+\theta}, \\ \mathbb{V}\,(X) &= \frac{\rho\theta}{(\rho+\theta)^2\,(\rho+\theta+1)}\;. \end{aligned}$$

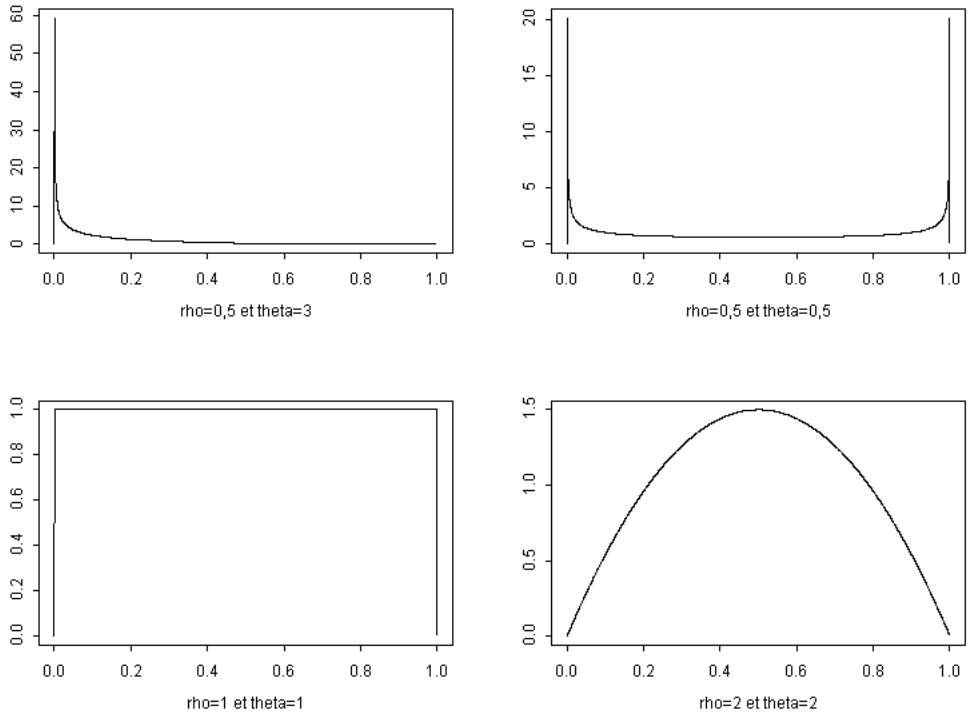


FIGURE 2.11 — Loi Bêta.

2.3.7 Loi logistique

Elle admet deux paramètres a et b.
Nous avons :

$$\begin{aligned}\mathbb{P}(X \leqslant x) &= \frac{\frac{1}{b}\exp(x-a)}{|b|\Big[1+\frac{1}{b}\exp(x-a)\Big]^2}\,,\\ \mathbb{E}(X) &= a,\\ \mathbb{V}(X) &= \frac{1}{3}(\pi b)^2\end{aligned}$$

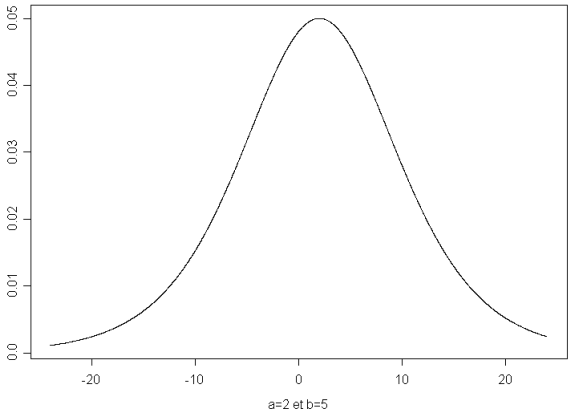


FIGURE 2.12 — Loi logistique.

Cette loi est symétrique et unimodale, et elle présente une queue.

2.3.8 Loi log-normale

Elle admet deux paramètres μ et σ^2 (μ réel quelconque, σ^2 réel positif).
Nous avons :

$$\begin{aligned}\mathbb{P}(X \leqslant x) &= \frac{1}{\sigma x\sqrt{2\pi}}\,\exp\left[\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right],\\ \mathbb{E}(X) &= \exp\Big[(\mu+\sigma^2)/2\Big],\\ \mathbb{V}(X) &= \exp(\sigma^2+2\mu)\big[\exp(\sigma^2)-1\big]\,.\end{aligned}$$

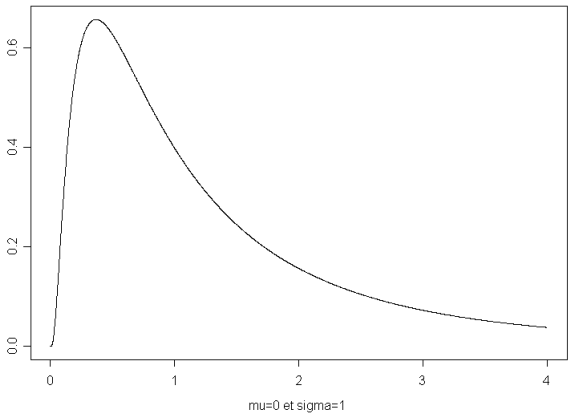


FIGURE 2.13 — Loi log-normale.

Proposition 2.19 — *X suit une loi log-normale si $\log(X)$ suit une loi normale.*

2.3.9 Loi du chi-deux

Elle admet un paramètre r (positif) et sert essentiellement à la réalisation de tests statistiques.
Nous avons :

$$\begin{aligned} \mathbb{P}(X \leqslant x) &= \frac{x^{\frac{r}{2}-1}\ \exp(-x/2)}{\Gamma(r/2)\ 2^{r/2}}, \\ \mathbb{E}(X) &= r, \\ \mathbb{V}(X) &= 2r\ . \end{aligned}$$

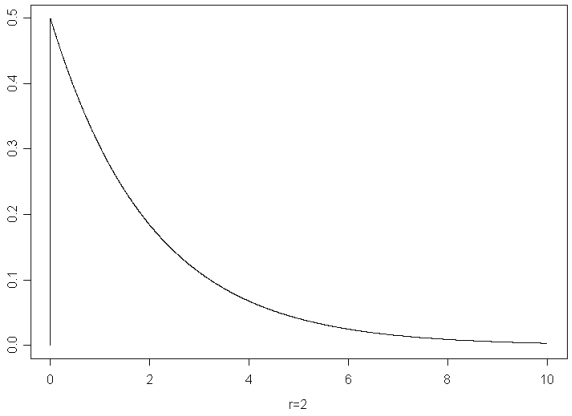


FIGURE 2.14 — Loi du chi-deux.

Remarque — Il s'agit d'une loi gamma avec $\rho = r/2$ et $\theta = 1/2$.

2.3.10 Loi normale tronquée

Elle admet quatre paramètres : a, b, μ et σ^2 (a, b réels quelconques, $\mu \in [a,b]$ réel et σ^2 réel positif) et consiste en une loi normale restreinte à l'intervalle $[a,b]$.
Nous avons :

$$\begin{aligned} \mathbb{E}(X) &= \mu, \\ \mathbb{V}(X) &= \sigma^2\ . \end{aligned}$$

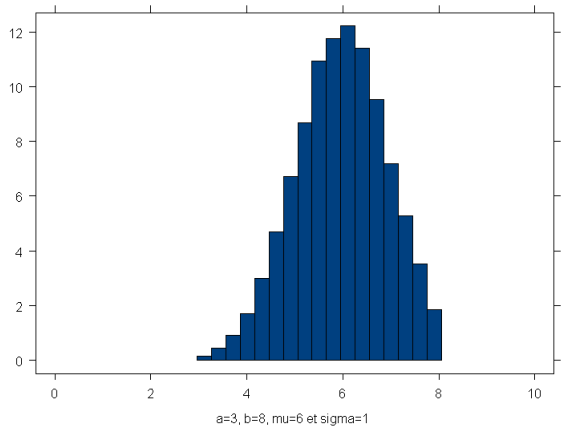


FIGURE 2.15 — Loi normale tronquée.

2.3.11 Loi de Weibull

Elle admet deux paramètres α,β. Cette loi est très utilisée pour caractériser la fiabilité des matériels. Elle est reliée à la loi exponentielle par la relation suivante : X suit une loi de Weibull de paramètre β si X^β suit une loi exponentielle. β est le paramètre de forme :

- le cas où $\beta > 1$ correspond à un matériel qui se dégrade avec le temps (usure) ;
- le cas où $\beta < 1$ correspond à un matériel qui se bonifie avec le temps ;
- le cas où $\beta = 1$ (la loi est alors une loi exponentielle) correspond à un matériel sans usure (pannes purement accidentelles).

Nous avons :

$$\begin{aligned}
\mathbb{P}(X \leqslant x) &= \alpha\beta^{-\alpha}x^{\alpha-1}\exp\Big[-\big(\frac{x}{\beta}\big)^{\alpha}\Big], \\
\mathbb{E}(X) &= \beta\,\Gamma\big(1+\frac{1}{\alpha}\big), \\
\mathbb{V}(X) &= \beta^2\Big[\Gamma\big(1+\frac{2}{\alpha}\big)-\Gamma^2\big(1+\frac{1}{\alpha}\big)\Big].
\end{aligned}$$

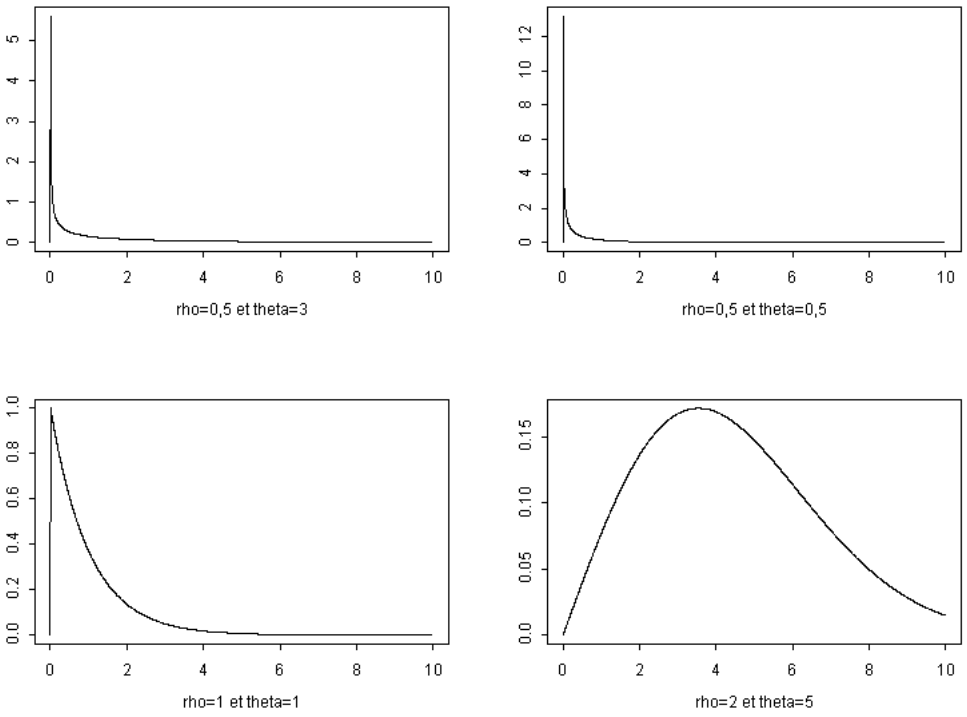


FIGURE 2.16 — Loi de Weibull.

2.3.12 Loi triangulaire

Les paramètres sont a, b (réels) et $c \in [a,b]$ (réel).
Nous avons :

$$\mathbb{P}(X \leqslant x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} \text{ si } a \leqslant x \leqslant c \\ \frac{2(b-x)}{(b-a)(b-c)} \text{ si } c \leqslant x \leqslant b \\ 0 \text{ sinon,} \end{cases}$$

$$\mathbb{E}(X) = \frac{1}{3}(a+b+c).$$

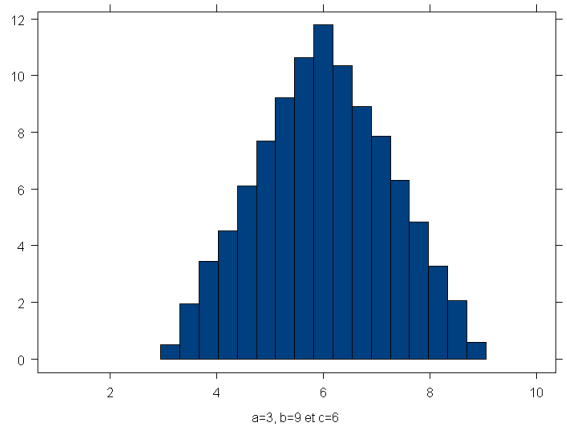


FIGURE 2.17 — Loi triangulaire.

Il s'agit d'une loi flexible portant sur un certain intervalle, et dont le mode est connu.

2.3.13 Loi de la valeur extrême

Les distributions de la valeur extrême sont les distributions limites du minimum ou du maximum d'un très grand ensemble d'observations aléatoires issues d'une même loi. Notons M_n la statistique d'ordre extrême $X^{(n)}$ relative à une distribution de n v.a. X_i suivant une même loi.

Si la loi commune est une loi uniforme sur [0,1], alors

$$\mathbb{P}(M_n < x) = \begin{cases} 0 \text{ si } x < 0 \\ x^n \text{ si } 0 \leqslant x \leqslant 0 \\ 1 \text{ si } x > 1 \end{cases}$$

et dans ce cas,

$$\begin{aligned} \mathbb{E}(M_n) &= \frac{n}{n+1}\ , \\ \mathbb{V}(M_n) &= \frac{n}{(n+1)^2(n+2)}\ . \end{aligned}$$

Si la loi commune est une loi normale centrée réduite, *i.e.*

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt \\ &= \frac{1}{2} + \Phi(x)\ , \end{aligned}$$

où $\Phi(x)$ est la fonction de distribution normale, alors

$$\begin{aligned} \mathbb{P}(M_n < x) &= \big[F(x)\big]^n \\ &= \frac{n}{\sqrt{2\pi}} \int_{-\infty}^{x} \big[F(t)\big]^{n-1}\, e^{-\frac{t^2}{2}}\, \mathrm{d}t \end{aligned}$$

et dans ce cas,

$$\begin{aligned}
\mathbb{E}(M_1) &= 0 \,, \\
\mathbb{V}(M_1) &= 1 \,, \\
\mathbb{E}(M_2) &= \frac{1}{\sqrt{\pi}} \,, \\
\mathbb{V}(M_2) &= 1 - \frac{1}{\pi} \,, \\
\mathbb{E}(M_3) &= \frac{3}{2\sqrt{\pi}} \,, \\
\mathbb{V}(M_3) &= \frac{4\pi - 9 + 2\sqrt{3}}{4\pi} \,, \\
&\vdots
\end{aligned}$$

Un théorème analogue à celui de la limite centrale établit que la distribution asymptotique normalisée de M_n satisfait l'une des trois distributions de probabilités suivantes :

1. Loi de **Gumbel** :
$$F(y) = \exp(-e^{-y}) \,.$$
2. Loi de **Fréchet** :
$$F(y) = \begin{cases} 0 \text{ si } y \leqslant 0 \,, \\ \exp(-y^{-a}) \text{ si } y > 0 \,. \end{cases}$$
3. Loi de **Weibull** :
$$F(y) = \begin{cases} \exp\big[-(-y)^{a})\big] \text{ si } y \leqslant 0 \,, \\ 1 \text{ si } y > 0 \,. \end{cases}$$

Dans le contexte des modèles de fiabilité, les distributions de la valeur extrême pour le minimum sont fréquemment utilisées. Ainsi, si un système consiste en n composantes identiques placées en série, et si le système tombe en panne lorsque la première de ces composantes défaille, alors le temps auquel le système tombe en panne est le minimum des n temps aléatoires de survenue d'une panne. La théorie de la valeur extrême dit que, indépendamment du choix du modèle des composantes, le modèle du système va approcher une distribution de Weibull à mesure que n devient très grand. Le même raisonnement peut être appliqué à chacune des composantes du système, si nous supposons que la survenue d'une panne d'une composante a lieu lorsque la première défaillance est due à un mécanisme agissant parmi de nombreux mécanismes similaires.

2.3.14 Loi de Fisher-Tippett (ou log-Weibull)

Cette distribution est aussi appelée **distribution de la valeur extrême**. Ses paramètres sont a et b.

Nous avons :

$$\begin{aligned}
f(x) &= \frac{e^{(a-x)/b - e^{(a-x)/b}}}{b} \,, \\
F(x) &= e^{-e^{(a-x)/b}} \,.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(X) &= a + b\gamma \,, \\
\mathbb{V}(X) &= \frac{1}{6}\pi^2 b^2 \,,
\end{aligned}$$

où γ est la constante d'Euler-Mascheroni.

2.3.15 Loi de Fisher

Ses paramètres sont n_1 et n_2.
Nous avons :

$$\begin{aligned} f(X) &= \frac{n_1\left(\frac{n_1 X}{n_2}\right)^{n_1/2-1}\left(1+\frac{n_1 X}{n_2}\right)^{-(n_1+n_2)/2}}{n_2\, B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}, \\ \mathbb{E}(X) &= \frac{n_2}{n_2-1}\,. \end{aligned}$$

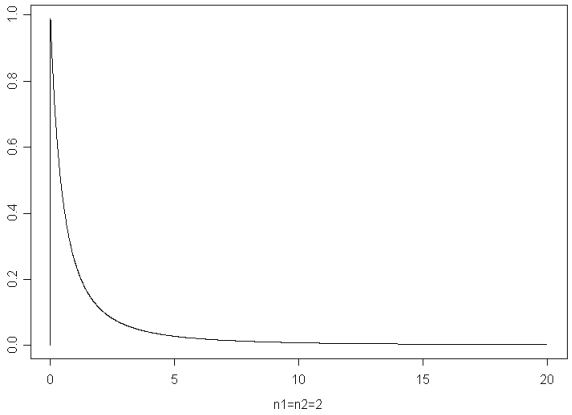


FIGURE 2.18 — Loi de Fisher.

2.3.16 Loi de Gumbel

Il s'agit d'un cas particulier de la loi de Fisher-Tippett pour $a = 0$ et $b = 1$.

2.3.17 Loi de Pareto

Ses paramètres sont a et b.
Nous avons :

$$\begin{aligned} f(X) &= \frac{ab^a}{x^{a+1}}\,, \\ F(x) &= 1-\left(\frac{b}{x}\right)^a\,, \\ \mathbb{E}(X) &= \frac{ab}{a-1}\,, \\ \mathbb{V}(X) &= \frac{ab^2}{(a-1)^2(a-2)}\,. \end{aligned}$$

2.3.18 Loi de Laplace

Ses paramètres sont μ et b.

Cette loi de probabilité est aussi appelée **distribution exponentielle double**. Il s'agit de la distribution de la différence entre deux variables indépendantes de même loi exponentielle.

Nous avons :

$$\begin{aligned} f(X) &= \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} , \\ F(x) &= \frac{1}{2}\Big[1 + \operatorname{sgn}(x-\mu)\Big(1 - e^{-\frac{|x-\mu|}{b}}\Big)\Big] , \end{aligned}$$

$$\begin{aligned} \mathbb{E}(X) &= \mu , \\ \mathbb{V}(X) &= 2b^2 . \end{aligned}$$

Deuxième partie

INDÉPENDANCE

3

Généralités

3.1 Présentation

Définition 3.1 — $(\Omega, \mathcal{A}, \mathbb{P})$. *Soient* $\mathcal{B}_1, \ldots, \mathcal{B}_n$ *des sous-tribus de* $\mathcal{A}$.

$$\begin{aligned} \mathcal{B}_1, \ldots, \mathcal{B}_n \text{ indépendantes} \iff & \forall B_i \in \mathcal{B}_i, \\ & \mathbb{P}\Big(\bigcap_{i=1}^{n} B_i\Big) = \prod_{i=1}^{n} \mathbb{P}(B_i) \ . \end{aligned}$$

Définition 3.2 — $X_i : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (\mathbb{R}^{d_i}, \mathcal{B}(\mathbb{R}^{d_i}))$.

$$\begin{aligned} X_1, \ldots, X_n \text{ indépendantes} &\iff \mathcal{B}_{X_1}, \ldots, \mathcal{B}_{X_n} \text{ indépendantes} \\ &\iff \forall A_i \in \mathcal{B}(\mathbb{R}^{d_i}), \\ &\quad \mathbb{P}\Big(\bigcap_{i=1}^{n} \{X_i \in A_i\}\Big) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) \ . \end{aligned}$$

Proposition 3.1 — *Soient* $\mathcal{C}_i$, $1 \leq i \leqslant n$, *des classes d'ensembles de* $\mathcal{A}$, *stables par intersections finies et contenant* Ω. *Si* $\forall i \in \{1, \ldots, n\}$, $\forall C_i \in \mathcal{C}_i$,

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} C_i\Big) = \prod_{i=1}^{n} \mathbb{P}(C_i) \ ,$$

alors les tribus $\sigma(\mathcal{C}_i)$, $1 \leq i \leqslant n$, *sont indépendantes.*

Proposition 3.2 — *Soient $X_i : \Omega \longrightarrow E$ avec E dénombrable.*

$$X_1, \ldots, X_n \text{ indépendantes} \iff \forall x_i, \ldots, x_n \in E, \quad \mathbb{P}\Big(\bigcap_{i=1}^{n} \{X_i = x_i\}\Big) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \ .$$

Remarque — X_1 indépendante de X_2 et X_1 indépendante de X_3 n'entraîne pas que X_1 soit indépendante de (X_2, X_3).

Proposition 3.3 (Indépendance par paquets) — *Si $(X_i)_{i=1,\ldots,n}$ est une suite de v.a. indépendantes d_i-dimensionnelles et si $n_0 < n_1 < \ldots < n_k = n$ est une suite d'entiers, alors les vecteurs aléatoires $(Y_j)_{j=1,\ldots,k}$ où $Y_j = \{X_{n_{j-1}+1}, \ldots, X_{n_j}\}$ sont indépendantes.*

Proposition 3.4 — *Si $X_1, \ldots, X_n$ sont indépendantes, alors toute sous-famille extraite est formée de v.a. indépendantes.*

Proposition 3.5 — *Si $X_1, \ldots, X_n$ sont indépendantes, avec $X_i : \Omega \to \mathbb{R}^{d_i}$, alors $\forall f_i$ boréliennes avec $f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$, $f_1(X_1), \ldots, f_n(X_n)$ sont indépendantes.*

Rappel d'intégration — Soit $(E_1, \mathcal{A}_1, \mu_1)$ et $(E_2, \mathcal{A}_2, \mu_2)$ deux espaces de mesure, avec μ_1 et μ_2 positives et σ-finies. Alors il existe une unique mesure μ sur $(E_1 \times E_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ telle que $\forall A_1 \in \mathcal{A}_1$, $\forall A_2 \in \mathcal{A}_2$,

$$\mu(A_1 \times A_2) \ = \ \mu_1(A_1) \cdot \mu_2(A_2) \ .$$

μ est appelée la **mesure-produit** et est notée $\mu = \mu_1 \otimes \mu_2$.

Proposition 3.6 — *Une suite de v.a. X_i d_i-dimensionnelles est indépendante ssi la loi du vecteur $X = (X_1, \ldots, X_n)$ d-dimensionnel avec $d = \sum_{i=1}^{n}$ est le produit des lois des v.a. X_i, i.e.*

$$\mathbb{P}_{(X_1, \ldots, X_n)} = \mathbb{P}_{X_1} \otimes \ldots \otimes \mathbb{P}_{X_n} \ .$$

Proposition 3.7 — *$(X_1, \ldots, X_n)$ v.a. indépendantes $\iff$ $\forall p \in \{2, \ldots, n\}$, X_p est indépendante de $(X_1, \ldots, X_{p-1})$.*

Lemme 3.1 — *Si X_i a pour densité p_{X_i}, alors $\mathbb{P}_{X_1} \otimes \ldots \otimes \mathbb{P}_{X_n}$ a pour densité*

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \ \cdots \ p_{X_n}(x_n) \ .$$

Proposition 3.8 — *Si les v.a. X_i sont indépendantes et ont pour densité p_i sur $\mathbb{R}^{d_i}$, le vecteur aléatoire $X = (X_1, \ldots, X_n)$ a une densité p sur $\mathbb{R}^d$ ($d = \sum_{i=1}^n d_i$) définie par*

$$p(x_1, \ldots, x_n) = p_1(x_1) \cdot p_2(x_2) \ \cdots \ p_n(x_n) \ .$$

Proposition 3.9 (Réciproque) — *On suppose que $X = (X_1, \ldots, X_n)$ admet sur $\mathbb{R}^d$ une densité $p(x_1, \ldots, x_n)$ qui s'écrive sous la forme $p(x_1, \ldots, x_n) = f_1(x_1) \times \ldots \times f_n(x_n)$, avec $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$, $f_i \geqslant 0$ et f_i borélienne.*

Alors les v.a. X_i sont indépendantes et ont pour densité

$$p_{X_i}(x_i) = \frac{f_i(x_i)}{\int_{\mathbb{R}^{d_i}} f_i(x_i) \, \mathrm{d}x_i} \ .$$

Corollaire 3.1 — *On suppose que $X = (X_1, \ldots, X_n)$ a pour densité p. Alors*

$$X_1, \ldots, X_n \text{ indépendantes} \iff \begin{array}{l} \forall i \in \{1, \ldots, n\}, \exists f_i \geqslant 0 \text{ t.q.} \\ p(x_1, \ldots, x_n) = f_1(x_1) \cdot f_2(x_2) \ \cdots \ f_n(x_n) \quad pp \ . \end{array}$$

Proposition 3.10 — *Étant donnée une suite finie de probabilités μ_i sur $\mathbb{R}^{d_i}$, il existe une suite de v.a. X_i indépendantes telle que $\mathbb{P}_{X_i} = \mu_i$.*

Proposition 3.11 — *Soient $(X_i)_i$ une suite de v.a. indépendantes à valeurs dans $\mathbb{R}^{d_i}$, et $f_i : \mathbb{R}^{d_i} \to \mathbb{R}$ boréliennes. On suppose $f_i \geqslant 0$, $\forall i \in \{1, \ldots, n\}$.*

Alors

$$\mathbb{E}\left(\prod_{i=1}^n f_i(x_i)\right) \ = \ \prod_{i=1}^n \mathbb{E}\left(f_i(X_i)\right) \ .$$

Proposition 3.12 — *Soient $X_1, \ldots, X_n$ indépendantes. On suppose que pour tout i, $f_i(X_i)$ a un moment d'ordre 1.*

Alors $\prod_{i=1}^n f_i(X_i)$ a un moment d'ordre 1 et

$$\mathbb{E}\left(\prod_{i=1}^n f_i(x_i)\right) \ = \ \prod_{i=1}^n \mathbb{E}\left(f_i(X_i)\right) \ .$$

Définition 3.3 — *Soit $(\Omega, \mathcal{A}, \mathbb{P})$. Une famille $(\mathcal{B}_i)_{i \in I}$ de sous-tribus de $\mathcal{A}$ est dite indépendante si toute famille finie extraite de la famille $(\mathcal{B}_i)_{i \in I}$ est indépendante.*

Une famille $(X_i)_{i \in I}$ de v.a. est dite indépendante si la famille $(\mathcal{B}_{X_i})_{i \in I}$ est indépendante.

Proposition 3.13 — *Soit $(X_n)_{n \in \mathbb{N}^*}$ une suite de v.a. Les propositions suivantes sont équivalentes :*

(i) la suite $(X_n)_{n \in \mathbb{N}^*}$ est une suite de v.a. indépendantes ;

(ii) $\forall n \in \mathbb{N}^*$, *la suite* $X_1, \ldots, X_n$ *est indépendante ;*
(iii) $\forall n \in \mathbb{N}^*$, X_{n+1} *est indépendante du vecteur* $(X_1, \ldots, X_n)$.

Proposition 3.14 (Indépendance par paquets) — *Si* $(\mathcal{B}_i)_{i\in I}$ *est une famille indépendante de tribus, et si* $\{I_j,\ j \in J\}$ *est une partition de* I, *la famille de tribus* $\big(\sigma(\cup_{i\in I_j}\mathcal{B}_i)\big)_{j\in J}$ *est formée de tribus indépendantes.*

Théorème 3.1 — *Si* $(\mu_n)_{n\in\mathbb{N}^*}$ *est une suite de probabilités sur* $\mathbb{R}^{d_n}$, *il existe un espace* $(\Omega, \mathcal{A}, \mathbb{P})$ *sur lequel on peut définir une suite* $(X_n)_{n\in\mathbb{N}^*}$ *de v.a. indépendantes et telles que* $\forall n \in \mathbb{N}^*$, $\mathbb{P}_{X_n} = \mu_n$.

Définition 3.4 — *Soit* $(X_n)_n$ *une suite de v.a. On considère, pour* $p \in \mathbb{N}$, *la tribu*

$$\begin{aligned}\mathcal{A}_p &= \sigma(X_n,\, n \geqslant p)\\ &= \sigma\Big(\bigcup_{n\geqslant p} \mathcal{B}_{X_n}\Big)\ .\end{aligned}$$

On pose

$$\mathcal{B}_\infty = \bigcap_{p\in\mathbb{N}} \mathcal{A}_p\ .$$

C'est la ***tribu asymptotique****.*

Proposition 3.15 (Loi du tout ou rien) — *Si* $(X_n)_n$ *est une suite de v.a. indépendantes, la tribu asymptotique associée à la famille* $(X_n)_n$ *est p.s. grossière, i.e.*

$$\forall B \in \mathcal{B}_\infty, \quad \mathbb{P}(B) = 0 \ \ ou\ \ 1\ .$$

Conséquence — Toute v.a. mesurable par rapport à la tribu asymptotique est p.s. constante.

3.2 Loi des grands nombres

Théorème 3.2 (Loi forte des grands nombres) — *Soit* $(X_n)_n$ *une suite de v.a. indépendantes, de même loi et ayant un moment d'ordre 1. Alors*

$$\frac{1}{n}\,(X_1 + \ldots + X_n) \ \longrightarrow\ \mathbb{E}\,(X_1) \quad p.s.$$

Théorème 3.3 — *Soit* $(X_n)_n$ *une suite de v.a. indépendantes et de même loi. Alors*

$$\frac{1}{n}\,(X_1 + \ldots + X_n) \ \xrightarrow{p.s.}\ \text{constante finie} \quad\Longleftrightarrow\quad \mathbb{E}\,\big(|X_1|\big) < \infty\ .$$

3.3 Fonctions caractéristiques

Définition 3.5 — *Soit X une v.a.r. On appelle* ***fonction caractéristique*** *de X l'application*

$$\begin{array}{rccl} \phi_X : \mathbb{R} & \longrightarrow & \mathbb{C} & \\ t & \longmapsto & \phi_X(t) & = \mathbb{E}(e^{itX}) \\ & & & = \displaystyle\int_{\mathbb{R}} e^{itx}\, \mathrm{d}\mathbb{P}_X \ . \end{array}$$

Remarques — Elles sont au nombre de trois :

- elle existe toujours car $|e^{itx}| = 1$;
- c'est la transformée de Fourier sur $\mathbb{R}$ de $\mathbb{P}_X$;
- si X et Y sont indépendantes, alors $\phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t)$.

Définition 3.6 — *Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$. On appelle* ***fonction caractéristique*** *de X l'application*

$$\begin{array}{rccl} \phi_X : \mathbb{R}^d & \longrightarrow & \mathbb{C} & \\ t & \longmapsto & \mathbb{E}(e^{i<t,X>}) & = \mathbb{E}\big(e^{i\sum_{j=1}^d t_j x_j}\big) \ . \end{array}$$

Proposition 3.16 — *Nous avons $|\phi X(t)| \leqslant 1,\ \forall t \in \mathbb{R}^d,$ et $\phi_X(0) = 1$. Par ailleurs, la fonction $t \mapsto \phi_X(t)$ est uniformément continue.*

Propriété 3.1 — *Si X est un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, alors $\phi_X(-t) = \overline{\phi_X(t)}$.*

Proposition 3.17 — *Soit X une v.a. à valeurs dans $\mathbb{R}^d$. X a une loi symétrique ssi ϕ_X est réelle et paire, ce qui équivaut à ϕ_X réelle.*

3.3.1 Dans le cas gaussien

Proposition 3.18 — *Si $X \rightsquigarrow \mathcal{N}(m, \sigma^2)$, alors*

$$\phi_X(t) = e^{itm - t^2\sigma^2/2}, \quad \forall t \in \mathbb{R} \ .$$

Proposition 3.19 — *L'extension à $\mathbb{R}^d$ du résultat précédent se formule ainsi :*

$$\phi_X(t) = e^{i<t,m> - \frac{1}{2}\sigma^2\|t\|^2}, \quad \forall t \in \mathbb{R}^d \ .$$

Proposition 3.20 — *Si $X \rightsquigarrow \mathcal{N}(0,1)$, alors $\forall k$*

$$\begin{aligned} \mathbb{E}(Z^{2k+1}) &= 0 \ , \\ \mathbb{E}(Z^{2k}) &= \frac{2k!}{2^k \cdot k!} \ . \end{aligned}$$

Proposition 3.21 (Formule d'inversion de Fourier) — *Soit X une v.a.r. de fonction caractéristique ϕ. On suppose que $\phi \in L^1(\mathbb{R}, dt)$. Alors X admet une densité f donnée par*

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \ \phi(t) \ \mathrm{d}t \ .$$

Proposition 3.22 — *Soit $\phi : \mathbb{R} \mapsto \mathbb{C}$ continue avec $\phi(0) = 1$. On suppose que $\phi \in L^1(\mathbb{R}, \mathrm{d}t)$. On pose*

$$g(x) \ = \ \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \ \phi(t) \ \mathrm{d}t \ .$$

Si g est réelle, positive et dans $L^1(\mathbb{R}, \mathrm{d}x)$, alors

$$\phi(t) \ = \ \int_{\mathbb{R}} e^{itx} \ g(x) \ \mathrm{d}x$$

et ϕ est la fonction caractéristique de la v.a. de loi de densité g.

Proposition 3.23 — *Soit X une var admettant un moment d'ordre $r \in \mathbb{N}^*$. Alors la fonction caractéristique ϕ_X de X est de classe $\mathcal{C}^r$ et on a*

$$\frac{\partial^r \phi_X(t)}{\partial t^r} \ = \ \mathbb{E}\left[(iX)^r \ e^{itX}\right] \ .$$

Par conséquent,

$$\mathbb{E}\left[(iX)^r\right] \ = \ \frac{\partial^r \phi_X}{\partial t^r}(0) \ .$$

Lemme 3.2 — *Si X admet un moment d'ordre $r \in \mathbb{N}^*$, alors X admet un moment d'ordre p, $\forall p < r$, $p \in \mathbb{N}^*$.*

3.4 Formule de Taylor pour les fonctions caractéristiques

Proposition 3.24 — *Soit X une var ayant un moment d'ordre 2. Alors ϕ_X est de classe $\mathcal{C}^2$ et au voisinage de 0,*

$$\begin{aligned} \phi_X(t) &= 1 + it \ \mathbb{E}(X) - \frac{t^2}{2} \ \mathbb{E}(X^2) + o(t^2) \ , \\ \ln \phi_X(t) &= it \ \mathbb{E}(X) - \frac{t^2}{2} \ \mathbb{V}ar(X) + o(t^2) \ . \end{aligned}$$

3.5 Indépendance

Proposition 3.25 — *Une suite $(X_i)_i$ de v.a. d_i-dimensionnelles est indépendante ssi la fonction caractéristique du vecteur $X = (X_1, \ldots, X_n)$ de dimension $d = \sum_{i=1}^n d_i$ est le produit des fonctions caractéristiques des X_i, i.e. $\forall t_i \in \mathbb{R}^{d_i}$, $\forall i$,*

$$\phi_{(X_1,\ldots,\, X_n)}(t_1, \ldots, t_n) \;=\; \phi_{X_1}(t_1) \;\times\; \phi_{X_2}(t_2) \;\times\; \cdots \;\times\; \phi_{X_n}(t_n) \;.$$

Lemme 3.3 —

$$\phi_{\mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}}(t_1, \ldots, t_n) \;=\; \prod_{i=1}^n \phi_{X_i}(t_i) \;.$$

Proposition 3.26 — *Si X et Y sont indépendantes,*

$$\phi_{X+Y}(u) = \phi_X(u) \cdot \phi_Y(u) \;.$$

Proposition 3.27 — *Si X et Y sont indépendantes,*

$$\mathbb{P}_{X+Y} \;=\; \mathbb{P}_X * \mathbb{P}_Y \;.$$

Rappel — Le **produit de convolution** $\mu * \nu$ est l'image de $\mu \otimes \nu$ par l'application $(x, y) \mapsto x + y$.

Proposition 3.28 — *Quelle que soit f borélienne,*

$$\begin{aligned} \int f(u)\, \mathrm{d}\mathbb{P}_{X+Y}(u) &= \mathbb{E}\left[f(X+Y)\right] \\ &= \int\int f(x+y)\, \mathrm{d}\mathbb{P}_X(x)\, \mathrm{d}\mathbb{P}_Y(y) \;. \end{aligned}$$

Proposition 3.29 — *Si X a pour densité p, alors $X + Y$ a pour densité*

$$\int p(x, y)\; \mathrm{d}\mathbb{P}_Y(y) \;.$$

Proposition 3.30 — *Si X a pour densité p et si Y a pour densité q, alors $X + Y$ a pour densité*

$$\begin{aligned} \int p(x-y)\, q(y)\, \mathrm{d}y &= \int p(y)\, q(x-y)\, \mathrm{d}y \\ &= p * q \;. \end{aligned}$$

Proposition 3.31 — *Si X et Y sont à valeurs dans $\mathbb{Z}$,*

$$\mathbb{P}(X + Y = n) \;=\; \sum_{p \in \mathbb{Z}} \mathbb{P}(X = p) \times \mathbb{P}(Y = n - p) \;.$$

3.6 Caractéristiques $\mathfrak{L}^2$

3.6.1 Moments

Proposition 3.32 — *Si $p < q$,*

$$\mathbb{E}\big(\|X\|^p\big)^{\frac{1}{p}} \leqslant \mathbb{E}\big(\|X\|^q\big)^{\frac{1}{q}} ,$$

i.e.

$$\|X\|_p \leqslant \|X\|_q .$$

Proposition 3.33 (Inégalité de Jensen) — *Soit X intégrable et f convexe. Alors*

$$f\big(\mathbb{E}(X)\big) \leqslant \mathbb{E}\big(f(X)\big) .$$

Proposition 3.34 (Inégalité de Minkowski) — *Nous avons :*

$$\|X + Y\|_p \leqslant \|X\|_p + \|Y\|_p .$$

Proposition 3.35 (Inégalité de Hölder) — *Soient $r > p > q$ avec $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$:*

$$\|XY\|_r \leqslant \|X\|_p \cdot \|Y\|_q .$$

Définition 3.7 — *$X = (X_1, \ldots, X_n)^t$ v.a. d-dimensionnelle a un moment d'ordre 1 si $\forall i$, X_i a un moment d'ordre 1.*

Proposition 3.36 — *Soit X une v.a. d-dimensionnelle, T une matrice $d' \times d$ et $Y = T \cdot X$. Si X a un moment d'ordre 1, alors $T \cdot X$ aussi et*

$$\mathbb{E}(T \cdot X) = T \cdot \mathbb{E}(X) .$$

En particulier, si T est un vecteur colonne d-dimensionnel $(d \times 1)$ noté a,

$$\begin{aligned} \mathbb{E}\left(< a, X >\right) &= \mathbb{E}\left(a^t X\right) \\ &= a^t \, \mathbb{E}\left(X\right) \\ &= < a, \mathbb{E}\left(X\right) > . \end{aligned}$$

Définition 3.8 — *Une v.a. d-dimensionnelle X a un moment d'ordre 2 si $\forall i$, X_i a un moment d'ordre 2. Par ailleurs,*

$$X \text{ a un moment d'ordre 2} \iff \sum_{i=1}^{d} X_i^2 \text{ a un moment d'ordre 1.}$$

Définition 3.9 — *Si X est de carré intégrable, la matrice des moments d'ordre 2 est*

$$M_X \;=\; \mathbb{E}(XX^t)\ .$$

La matrice de covariance est

$$K_X = \mathbb{E}\left(\big[X-\mathbb{E}(X)\big]\big[X-\mathbb{E}(X)\big]^t\right).$$

Proposition 3.37 — *$\mathbb{E}(XX^t)$ et K_X sont des matrices symétriques.*

Définition 3.10 — *L'espérance et la matrice de covariance sont les caractéristiques $\mathfrak{L}^2$ de X, vecteur aléatoire ayant un moment d'ordre 2.*

Proposition 3.38 —

$$K_X = \mathbb{E}(XX^t) - \mathbb{E}(X)\cdot\mathbb{E}(X^t)\ .$$

Proposition 3.39 — *Soient X une v.a. d-dimensionnelle ayant un moment d'ordre 2, A une matrice $d'\times d$ et $Y=AX$. Alors Y a un moment d'ordre 2 et (à un coefficient constant près) :*

$$\begin{aligned} M_Y &= \mathbb{E}\,(YY^t) \\ &= A\,M_X\,A^t\ , \end{aligned}$$

$$K_Y \;=\; A\,K_X\,A^t\ .$$

Proposition 3.40 — *Les matrices M_X et K_X sont symétriques de type positif. La matrice M_X est dite **définie positive** ssi il n'existe pas de relation linéaire entre les coordonnées de X (au sens p.s.).*

*La matrice de covariance K_X est définie positive ssi il n'existe pas de relation affine entre les coordonnées de X (au sens p.s.) — ce qui équivaut à K_X **inversible**).*

Proposition 3.41 — *Si K_X n'est pas inversible, la loi de X n'a pas de densité.*

Proposition 3.42 — *Soit X v.a. d-dimensionnelle. Si les v.a. X_i sont indépendantes, alors K_X est une matrice diagonale.*

Proposition 3.43 — *K matrice de covariance $\iff$ K symétrique de type positif.*

Proposition 3.44 — *Soient $Y_1,\dots,Y_n$ n vecteurs aléatoires d-dimensionnels indépendants. Alors*

$$K_{Y_1+\cdots+Y_n} = K_{Y_1}+\cdots+K_{Y_n}\ .$$

3.6.2 Vecteurs gaussiens

Proposition 3.45 — *Soient $X_1, \ldots, X_n$ n v.a. gaussiennes indépendantes. Alors*

$$\sum_{i=1}^{n} a_i X_i \quad \rightsquigarrow \quad \mathcal{N}\Big(\sum_{i=1}^{n} a_i m_i \ , \ \sum_{i=1}^{n} a_i^2 \sigma^2 \Big) \ .$$

Définition 3.11 — *Un vecteur $X = (X_1, \ldots, X_d)^t$ d-dimensionnel est dit* ***vecteur gaussien*** *si $\forall a \in \mathbb{R}^d$,*

$$< a, X >= \sum_{i=1}^{d} a_i X_i$$

est une v.a. gaussienne réelle.

Proposition 3.46 — *Soient $X_1, \ldots, X_d$ des v.a. réelles gaussiennes indépendantes. Alors le vecteur $X = (X_1, \ldots, X_d)^t$ est gaussien.*

Proposition 3.47 —

$$X \text{ vecteur gaussien de dimension } d \iff \left\{ \begin{array}{l} \forall d', \forall A \text{ application linéaire de } \mathbb{R}^d \text{ dans } \mathbb{R}^{d'}, \\ AX \text{ est un vecteur gaussien .} \end{array} \right.$$

Proposition 3.48 — *La fonction caractéristique du vecteur gaussien d-dimensionnel X est, pour $t \in \mathbb{R}^d$,*

$$\begin{aligned} \phi_X(t) &= \mathbb{E}\ (e^{i<t,X>}) \\ &= \mathbb{E}\ (e^{i\sum_{i=1}^{d} t_j X_j}) \\ &= \phi_{<t,\ X>}(1) \\ &= \exp\Big[i\mathbb{E}(< t,\ X >) - \frac{1}{2}\mathbb{V}ar(< t,\ X >) \Big] \\ &= \exp\Big(i < t, \mathbb{E}(X) > - \frac{1}{2}\ t^t\ K_X\ t \Big) \ . \end{aligned}$$

Proposition 3.49 — *Soit (X, Y) un couple gaussien (i.e. toute combinaison linéaire de X et de Y est gaussienne). Alors*

$$X \text{ et } Y \text{ indépendantes} \iff \mathbb{C}ov(X, Y) = 0 \ .$$

Proposition 3.50 — *Soit $X = (X_1, \ldots, X_d)$ un vecteur gaussien. Alors*

$$\begin{aligned} X_1, \ldots, X_d \text{ v.a. réelles indépendantes} &\iff \forall i \neq j,\ \mathbb{C}ov(X_i, X_j) = 0 \\ &\iff K_X \text{ diagonale .} \end{aligned}$$

Proposition 3.51 — *Soit $Y = (Y_1, \ldots, Y_d)$ où Y_j est k_j-dimensionnelle et gaussienne. Alors*

$$Y_1, \ldots, Y_d \text{ indépendantes} \iff K_Y \text{ diagonale par blocs .}$$

i.e.

$$K_Y = \begin{pmatrix} K_{Y_1} & 0 & \ldots & \ldots & 0 \\ 0 & K_{Y_2} & 0 & \ldots & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & K_{Y_d} \end{pmatrix} .$$

Théorème 3.4 — *Soit m un vecteur de $\mathbb{R}^d$ et K une matrice $d \times d$ de type positif. Alors il existe un vecteur gaussien d-dimensionnel de moyenne m et de matrice de covariance $K_X = K$.*

Théorème 3.5 — *Si K est inversible, alors la loi de X a pour densité*

$$\frac{1}{(\sqrt{2\pi})^d \sqrt{det\ K}} \exp\left[- \frac{(x-m)^t K^{-1} (x-m)}{2} \right] .$$

Sinon, la loi de X étant portée par un hyperplan (i.e. $\exists \alpha_1, \ldots, \alpha_d, b$ réels t.q. $\sum_{i=1}^d \alpha_i X_i = b$ p.s.), cette loi n'a pas de densité.

4

Conditionnement

4.1 Espérance conditionnelle

Soient $Y \in \mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ et $X_1, \ldots, X_d$ des v.a. On cherche une fonction f mesurable telle que $\left\|Y - f(X_1, \ldots, X_d)\right\|_2$ soit minimale.

$f(X_1, \ldots, X_d)$ sera alors la meilleure approximation dans $\mathfrak{L}^2$ de Y par une fonction de $(X_1, \ldots, X_d)$.

On va être amené à projeter Y sur l'espace

$$\begin{aligned} M &= \left\{Z = f(X_1, \ldots, X_d) \text{ p.s. , } f \text{ borélienne de } \mathbb{R}^d \text{ dans } \mathbb{R} \text{ , } Z \in \mathfrak{L}^2\right\} \\ &= \left\{Z \in \mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P}) \text{ , } Z \text{ admet un représentant } \sigma(X_1, \ldots, X_d)\text{–mesurable}\right\} . \end{aligned}$$

U $\sigma(X_1, \ldots, X_d)$–mesurable $\iff$ $\exists \phi$ borélienne t.q. $U = \phi(X_1, \ldots, X_d)$.

Lemme 4.1 — *M est un sous-espace vectoriel fermé de $\mathfrak{L}^2$.*

Définition 4.1 — *La **meilleure approximation** de $Y \in \mathfrak{L}^2$ **au sens des moindres carrés** par une fonction de $X_1, \ldots, X_d$ est la projection orthogonale de Y sur M, soit $\hat{Y}$. $\hat{Y}$ existe et est unique.*

Définition 4.2 — *$(\Omega, \mathcal{A}, \mathbb{P})$, $\mathcal{B} \subset \mathcal{A}$. On note*

$$\mathfrak{L}^2(\mathcal{B}) = \left\{Z \in \mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P}) \text{ , } Z \text{ admet un représentant } \mathcal{B}\text{–mesurable}\right\} .$$

$\mathfrak{L}^2(\mathcal{B})$ est un sev fermé de $\mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P})$.

Définition 4.3 — *Soit $Y \in \mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P})$. On appelle* ***espérance conditionnelle*** *de Y sachant $\mathcal{B}$ et on note $\mathbb{E}(Y \mid \mathcal{B})$ la classe d'équivalence de la projection de Y sur $\mathfrak{L}^2(\mathcal{B})$.*

Propriété 4.1 — *$Y \mapsto \mathbb{E}(Y \mid \mathcal{B})$, de $\mathfrak{L}^2$ dans $\mathfrak{L}^2$, est une contraction (car une projection) :*

$$\big\|\mathbb{E}(Y \mid \mathcal{B})\|_2 \leqslant \big\|Y\|_2 \ .$$

Propriété 4.2 — *Si Y est $\mathcal{B}$–mesurable, $\mathbb{E}(Y \mid \mathcal{B}) = Y$ p.s.*

Propriété 4.3 — *$\mathcal{B} = \{\emptyset, \Omega\}$. $\mathfrak{L}^2(\mathcal{B}) = \{$ v.a. p.s. constantes $\}$. Alors*

$$\mathbb{E}(Y \mid \mathcal{B}) = \mathbb{E}(Y) \text{ p.s.}$$

Proposition 4.1 — *Soit $Y \in \mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P})$. L'espérance conditionnelle $\mathbb{E}(Y \mid \mathcal{B})$ est caractérisée par l'une des deux propriétés suivantes, qui sont équivalentes :*

(i) c'est l'unique élément Z de $\mathfrak{L}^2(\mathcal{B})$ vérifiant

$$\forall B \in \mathcal{B}, \quad \int_B Z \, \mathrm{d}\mathbb{P} = \int_B Y \, \mathrm{d}\mathbb{P} \ ;$$

(ii) c'est l'unique élément Z de $\mathfrak{L}^2(\mathcal{B})$ vérifiant

$$\forall U \in \mathfrak{L}^2(\mathcal{B}), \ \ \mathbb{E}\ (ZU) = \mathbb{E}\ (YU) \ .$$

Proposition 4.2 — *Si Y est indépendante de $\mathcal{B}$ (i.e. $\sigma(Y)$ indépendante de $\mathcal{B}$), alors*

$$\mathbb{E}(Y \mid \mathcal{B}) = \mathbb{E}(Y) \ \text{ p.s.}$$

Proposition 4.3 — *Si $(X_1, \ldots, X_n, Y)$ est un vecteur gaussien, l'espérance conditionnelle $\mathbb{E}\left[Y \mid (X_1, \ldots, X_n)\right]$ est égale à $\hat{Y}$, meilleure approximation affine dans $\mathfrak{L}^2$ de Y par les v.a. X_i, $i = 1, \ldots, n$, i.e.*

$$\hat{Y} = \sum_{i=1}^{n} a_i X_i + b \ .$$

Ce qui équivaut à dire que la meilleure approximation dans $\mathfrak{L}^2$ de Y par une fonction de $(X_1, \ldots, X_n)$ est une fonction affine de $(X_1, \ldots, X_n)$.

Proposition 4.4 — *Si U est $\mathcal{B}$-mesurable bornée et si Y est bornée, alors*

$$\mathbb{E}(UY \mid \mathcal{B}) = U\ \mathbb{E}(Y \mid \mathcal{B}) \text{ p.s.}$$

Proposition 4.5 — *Les propositions suivantes sont équivalentes :*
1) Y *et* $\mathcal{B}$ *sont indépendantes ;*
2) $\forall f$ *borélienne,* $\mathbb{E}\big(f(Y) \mid \mathcal{B}\big) = \mathbb{E}\big[f(Y)\big]$ *p.s. ;*
3) $\forall t \in \mathbb{R}$, $\mathbb{E}(e^{itY} \mid \mathcal{B}) = \mathbb{E}(e^{itY})$ *p.s.*

Proposition 4.6 — *Soient* $\mathcal{B}_1 \subset \mathcal{B}_2$, $Y \in \mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P})$. *Alors* $\mathfrak{L}^2(\mathcal{B}_1) \subset \mathfrak{L}^2(\mathcal{B}_2)$ *et*

$$\mathbb{E}\left[\, \mathbb{E}\big(Y \mid \mathcal{B}_2\big) \;\middle|\; \mathcal{B}_1 \right] = \mathbb{E}(Y \mid \mathcal{B}_1) \quad p.s.$$

Proposition 4.7 — *Soient* $Y \in \mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, $T : \Omega \to \mathbb{N}$ *et* $N_1 = \big\{n \in \mathbb{N},\, \mathbb{P}(T = n) > 0\big\}$. *Alors*

$$\mathbb{E}(Y \mid T) = h(T) \quad p.s.$$

et si $n \in N_1$,

$$h(n) = \frac{\mathbb{E}(Y \mathbf{1}_{\{T=n\}})}{\mathbb{P}(T = n)} \ .$$

Interprétation
Si $Y = \mathbf{1}_A$,

$$\mathbb{E}(\mathbf{1}_A \mid T) = h_1(T)$$

où

$$\begin{aligned} h_1(n) &= \frac{\mathbb{E}(\mathbf{1}_A \cdot \mathbf{1}_{\{T=n\}})}{\mathbb{P}(T = n)} \\ &= \frac{\mathbb{P}(A \cap T = n)}{\mathbb{P}(T = n)} \\ &= \mathbb{P}(A \mid T = n) \ . \end{aligned}$$

Cette dernière probabilité conditionnelle a un sens puisque $\mathbb{P}(T = n) > 0$.

Proposition 4.8 —

$$\mathbb{E}(Y \mid T)\Big|_{T=n} = \int Y \, \mathrm{d}\mathbb{P}(. \mid T = n) \ .$$

Donc si T est discrète, l'espérance conditionnelle sachant T, calculée en une valeur n telle que $\mathbb{P}(T = n) > 0$, est l'espérance de Y par rapport à la probabilité conditionnelle $\mathbb{P}(. \mid T = n)$.

Proposition 4.9 — $(X,\, Y)$ *a pour densité* $p(x,\, y)$. *Soit* $D = \big\{x \mid p(x) > 0\big\}$. *On suppose* $Y \in \mathfrak{L}^2$. *Alors*

$$\mathbb{E}(Y \mid X) = h(X)$$

et $\forall x \in D$, *on a*

$$h(x) = \int_{\mathbb{R}} y \, \frac{p(x,\, y)}{p(x)} \, \mathrm{d}y \ .$$

Proposition 4.10 — *Soit x fixé tel que $p(x) > 0$. L'application $y \mapsto p(x, y)/p(x)$ définit une densité de probabilité que l'on notera $p(y \mid x)$, appelée* ***densité conditionnelle*** *de Y sachant $\{X = x\}$. L'espérance conditionnelle $\mathbb{E}(Y \mid X)$ calculée pour $X = x$ est l'espérance de Y par rapport à la loi conditionnelle de Y sachant $\{X = x\}$, de densité $p(y \mid x)$.*

Interprétation — Nous avons :

$$\mathbb{E}(Y \mid X)\Big|_{X=x} \;=\; \int y\, p(y|x)\, \mathrm{d}y \;.$$

4.2 Extension au cas où $Y \notin \mathfrak{L}^2$

Proposition 4.11 — *$Y \mapsto \mathbb{E}(Y \mid \mathcal{B})$ est une application croissante de $\mathfrak{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, i.e. pour $Y, Z \in \mathfrak{L}^2$,*

$$Y < Z \text{ p.s.} \quad \Longrightarrow \quad \mathbb{E}(Y \mid \mathcal{B}) \;\leqslant\; \mathbb{E}(Z \mid \mathcal{B}) \text{ p.s.}$$

Définition 4.4 — *Y est dite* ***quasi-intégrable*** *(noté* ***q.i.****) si une des v.a. Y^+ ou Y^- est intégrable. On peut alors définir l'espérance de Y (éventuellement infinie) par*

$$\mathbb{E}(Y) \;=\; \mathbb{E}(Y^+) - \mathbb{E}(Y^-) \;.$$

Les v.a. q.i. contiennent les v.a. positives et les v.a. appartenant à $\mathfrak{L}^1$.

Théorème 4.1 — *$\forall Y$ q.i., $\exists!$ v.a. $\mathcal{B}$–mesurable Z notée $\mathbb{E}(Y \mid \mathcal{B})$ telle que $\forall B \in \mathcal{B}$,*

$$\int_B Y \,\mathrm{d}\mathbb{P} \;=\; \int_B Z \,\mathrm{d}\mathbb{P} \;.$$

La v.a. Z est aussi q.i.

Proposition 4.12 — *Si $Y \geqslant 0$, $\mathbb{E}(Y \mid \mathcal{B}) \geqslant 0$.*

Proposition 4.13 — *Si $Y \in \mathfrak{L}^1$, $\mathbb{E}(Y \mid \mathcal{B}) \in \mathfrak{L}^1$ et*

$$\big\|\mathbb{E}(Y \mid \mathcal{B})\big\|_{\mathfrak{L}^1} \;\leqslant\; \|Y\|_{\mathfrak{L}^1} \;.$$

Proposition 4.14 — *Si Y est q.i., alors $\mathbb{E}(Y \mid \mathcal{B})$ est q.i. et*

$$\mathbb{E}(Y) \;=\; \mathbb{E}(Y \mid \mathcal{B}) \;.$$

4.3 Propriétés

L'espérance conditionnelle vérifie les mêmes propriétés que l'espérance ordinaire, données ci-dessous.

Propriété 4.4 — *Si X et Y sont q.i., alors $aX + bY$ est q.i. et*

$$\mathbb{E}(aX + bY \mid \mathcal{B}) \;=\; a\,\mathbb{E}(X \mid \mathcal{B}) + b\,\mathbb{E}(Y \mid \mathcal{B}) \;.$$

Propriété 4.5 — *Si X et Y sont q.i. et si $X \geqslant Y$, alors*

$$\mathbb{E}(X \mid \mathcal{B}) \;\geqslant\; \mathbb{E}(Y \mid \mathcal{B}) \text{ p.s.}$$

Propriété 4.6 — *Soit $X_n \geqslant 0$, $(X_n)_n$ croissante. Alors*

$$\lim \nearrow \mathbb{E}(X_n \mid \mathcal{B}) \;=\; \mathbb{E}\big(\lim \nearrow X_n \mid \mathcal{B}\big) \;.$$

Propriété 4.7 (Fatou) — *Soit $X_n \geqslant 0$. Alors*

$$\mathbb{E}\big(\underline{\lim}\, X_n \mid \mathcal{B}\big) \;\leqslant\; \underline{\lim}\, \mathbb{E}(X_n \mid \mathcal{B}) \;.$$

Propriété 4.8 — *Soit $X_n \to X$ p.s. et $\forall n$, $|X_n| \leqslant Y$ avec Y intégrable. Alors*

$$\mathbb{E}(X_n \mid \mathcal{B}) \;\longrightarrow\; \mathbb{E}(X \mid \mathcal{B}) \;.$$

Propriété 4.9 (Inégalité de Jensen) — *Soient ϕ convexe positive et X q.i. Alors*

$$\mathbb{E}\big(\phi(X) \mid \mathcal{B}\big) \;\geqslant\; \phi\big[\mathbb{E}(X \mid \mathcal{B})\big] \text{ p.s.}$$

4.4 Lois conditionnelles et probabilités de transition

4.4.1 Lois conditionnelles

Remarques — Elles sont au nombre de deux :

1) connaître la loi de Y revient à connaître la loi de $\mathbb{E}\big(f(Y)\big)$, $\forall f$ borélienne bornée — et réciproquement ;
2) connaître la loi conditionnelle de Y sachant X revient à connaître la loi de $\mathbb{E}\big(f(Y) \mid X\big)$, $\forall f$ borélienne bornée — et réciproquement.

Définition 4.5 — *La loi conditionnelle de Y sachant $\{X = n\}$ est définie par $\forall B \in \mathcal{B}(\mathbb{R})$,*

$$N(n, B) = \mathbb{P}\left(Y \in B \mid \{X = n\}\right)$$

pour $n \in N_1 = \{n \mid \mathbb{P}(X = n) > 0\}$.

$\forall n \in N_1$, $N(n, .)$ est une probabilité sur $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$.

Proposition 4.15 — *Soit $\forall f$ borélienne bornée, $f : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ avec*

$$Nf(n) = \int_{\mathbb{R}} f(n, y)\, N(n, \mathrm{d}y).$$

On a les deux égalités suivantes :

$$\mathbb{E}\big(f(X, Y) \mid X\big) \;=\; Nf(X)$$

et

$$\mathbb{E}\big(f(X, Y)\big) = \sum_n \mathbb{P}_X\big(\{n\}\big) \int f(n,y)\, N(n, \mathrm{d}y)\ .$$

Définition 4.6 — *Soit un couple de v.a. (X, Y) de densité $p(x, y)$. Soit $p(x) = \int p(x, y)\, \mathrm{d}y$ la densité de X. À tout x tel que $p(x) > 0$, on associe la probabilité $N(x, dy)$ définie par sa densité*

$$p(y \mid x) = \frac{p(x, y)}{p(x)}\ .$$

$y \mapsto p(y \mid x)$ est positive et d'intégrale 1 : c'est la densité conditionnelle de Y sachant $\{X = x\}$, et $N(x, \mathrm{d}y)$ est la loi conditionnelle de Y sachant $\{X = x\}$.

Proposition 4.16 — *On a les deux égalités suivantes :*

$$\mathbb{E}\big(f(X, Y) \mid X\big) = Nf(X)$$

et

$$\mathbb{E}\big(f(X, Y)\big) = \mathbb{E}\big(Nf(X)\big)\ .$$

Proposition 4.17 — *Si X et Y sont indépendantes, la loi conditionnelle de Y sachant $\{X = x\}$ ne dépend pas de x : c'est alors la loi de Y.*

4.4.2 Probabilités de transition

Définition 4.7 — *On appelle **probabilité de transition** de $\mathbb{R}^{d'}$ dans $\mathbb{R}$ une famille $\{N(x, .),\, x \in \mathbb{R}^{d'}\}$ de probabilités sur $\mathbb{R}$ telle que, $\forall A \in \mathcal{B}(\mathbb{R}^d)$,*

$$\begin{array}{ccc} \mathbb{R}^{d'} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & N(x, A) \end{array}$$

est borélienne.

Conséquences — Elles sont au nombre de deux :

a) soit $A \in \mathcal{B}(\mathbb{R}^d)$ fixé : $x \mapsto N(x, A)$ est borélienne ;

b) soit $x \in \mathbb{R}^{d'}$ fixé : $A \mapsto N(x, A)$ est une probabilité.

Théorème 4.2 — *$\forall (X, Y)$ couple de v.a. à valeurs dans $\mathbb{R}^d \times \mathbb{R}^{d'}$, il existe une probabilité de transition $N : \mathbb{R}^{d'} \to \mathbb{R}^d$ telle que $\forall f \geqslant 0$,*

$$\begin{aligned} \mathbb{E}\big[f(X, Y)\big] &= \int_{\mathbb{R}^d} \mathrm{d}\mathbb{P}_X(x) \int_{\mathbb{R}^{d'}} f(x, y)\, N(x, \mathrm{d}y) \\ &= \mathbb{E}\big[Nf(X)\big] \ , \end{aligned}$$

si l'on pose $Nf(x) = \int f(x, y)\, N(x, \mathrm{d}y)$.

De plus,

$$\mathbb{E}\big(f(X, Y) \mid X\big) = Nf(X) \ p.s.$$

$N(x, dy)$ est la loi conditionnelle de Y sachant $\{X = x\}$.

Proposition 4.18 — *Si X et Y sont indépendantes, alors $N(x, \mathrm{d}y) = \mathbb{P}_Y$: la loi conditionnelle de Y sachant $\{X = x\}$ est la loi de Y. De plus,*

$$\begin{aligned} \mathbb{E}\big[f(X, Y) \mid X\big] &= \mathbb{E}\big[f(X, Y) \mid \{X = x\}\big] \\ &= Nf(X) \\ &= \mathbb{E}\big[f(x, Y)\big] \ . \end{aligned}$$

Proposition 4.19 — *Si le couple (X, Y) est à valeurs dans $\mathbb{R}^{d+d'}$, si la loi de X a pour densité p et si la loi conditionnelle de Y sachant $\{X = x\}$ a pour densité $p(y \mid x)$, alors le couple (X, Y) a pour densité*

$$p(x, y) = p(x) \cdot p(y \mid x) \ .$$

Corollaire 4.1 — *Si X_1 a pour densité $p(x_1)$, si X_2 sachant $\{X_1 = x_1\}$ a pour densité $p(x_2 \mid x_1)$, si X_3 sachant $\{(X_1, X_2) = (x_1, x_2)\}$ a pour densité $p(x_3 \mid x_1, x_2)$, si ..., et si X_n sachant $\{(X_1, \ldots, X_{n-1}) = (x_1, \ldots, x_{n-1})\}$ a pour densité $p(x_2 \mid x_1, \ldots, x_{n-1})$, alors $(X_1, \ldots, X_n)$ a pour densité*

$$p(x_1) \times p(x_2 \mid x_1) \times p(x_3 \mid x_1, x_2) \times \cdots \times p(x_2 \mid x_1, \ldots, x_{n-1}) \ .$$

Proposition 4.20 — *Si (X, Y) est un vecteur gaussien, alors :*

1) l'espérance conditionnelle $\mathbb{E}(Y \mid X)$ est la meilleurs approximation affine $aX + b$ de Y par X au sens des moindres carrés ;

2) la loi conditionnelle de Y sachant $\{X = x\}$ est une loi gaussienne d'espérance $ax + b$ et de variance σ^2.

Proposition 4.21 — *Si* $(X_1, \ldots, X_n, Y)$ *est un vecteur gaussien,* $X_i : \Omega \to \mathbb{R}$, $Y : \Omega \to \mathbb{R}$, *alors la loi conditionnelle de* Y *sachant* $\big\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n)\big\}$ *est une loi normale d'espérance* $\sum a_i x_i + b$ *et de variance* $\sigma^2 = \mathbb{E}\big[(Y - \hat{Y})^2\big]$, *sachant que* $\hat{Y} = \sum a_i x_i + b$ *est la meilleure approximation affine de* Y *par* $(X_1, \ldots, X_n)$.

5

Convergences

5.1 Introduction

5.1.1 Différents types de convergence

$$\begin{aligned} X_n \to X \text{ p.s.} &\Leftrightarrow \underline{\lim}\, X_n = \overline{\lim}\, X_n = X \text{ p.s.} \\ &\Leftrightarrow \text{pour presque tout } \omega \in \Omega,\ X_n(\omega) \to X(\omega)\ . \end{aligned}$$

$$X_n \to X \text{ dans } \mathfrak{L}^1 \quad \Leftrightarrow \quad \|X_n - X\|_1 = \mathbb{E}\left(|X_n - X|\right) \to 0 \quad (n \to \infty)\ .$$

$$X_n \to X \text{ dans } \mathfrak{L}^p \quad \Leftrightarrow \quad \|X_n - X\|_p = \mathbb{E}\left(|X_n - X|^p\right) \to 0 \quad (n \to \infty)\ .$$

Proposition 5.1 — *Si $p < q$, $\|.\|_p < \|.\|_q$, c'est-à-dire que $\mathfrak{L}^q \subset \mathfrak{L}^p$. Par conséquent, **la convergence dans $\mathfrak{L}^q$ entraîne la convergence dans $\mathfrak{L}^p$**.*

Définition 5.1 (Convergence en probabilité) — *X_n converge vers X en probabilité ssi $\forall \epsilon > 0$,*

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad (n \to \infty)\ .$$

On note $X_n \xrightarrow{\mathbb{P}} X$ cette convergence.

Proposition 5.2 — *La convergence p.s. entraîne la convergence en proba.*

Proposition 5.3 — *La convergence dans $\mathfrak{L}^p$ entraîne la convergence en proba.*

Proposition 5.4 — *Pour tout $\epsilon > 0$,*

$$\sum_n \mathbb{P}\left(|X_n - X| > \epsilon\right) < \infty \quad \Longrightarrow \quad X_n \to X \quad p.s.$$

Proposition 5.5 — *Nous avons l'implication suivante :*

$$\left.\begin{array}{c} \exists(\epsilon_n)_n \text{ t.q. } \epsilon \to 0 \text{ et t.q.} \\ \sum_n \mathbb{P}\left(|X_n - X| > \epsilon_n\right) < \infty \end{array}\right\} \quad \Longrightarrow \quad X_n \to X \quad p.s.$$

Proposition 5.6 —

$$\begin{array}{rcl} (X_n)_n \text{ converge en proba} & \Leftrightarrow & (X_n)_n \text{ est une suite de Cauchy en proba} \\ & \Leftrightarrow & \forall\epsilon > 0, \forall\delta > 0, \exists N, \forall n,m > N, \\ & & \displaystyle\sum_n \mathbb{P}\left(|X_n - X_m| > \delta\right) < \epsilon \ . \end{array}$$

Proposition 5.7 — *De toute suite convergente en proba, on peut extraire une sous-suite qui converge p.s.*

Conséquence — De toute suite convergente dans $\mathfrak{L}^p$, on peut extraire une sous-suite qui converge p.s.

Proposition 5.8 — *Si $X_n \xrightarrow{\mathbb{P}} X$ et $X_n \xrightarrow{\mathbb{P}} Y$, alors $X_n = Y$ p.s.*

Proposition 5.9 — *Si $X_n \xrightarrow{\mathbb{P}} X$ et $Y_n \xrightarrow{\mathbb{P}} Y$, alors*

$$X_n \cdot Y_n \xrightarrow{\mathbb{P}} X \cdot Y$$

et

$$X_n + Y_n \xrightarrow{\mathbb{P}} X + Y.$$

5.1.2 Loi faible des grands nombres

Théorème 5.1 — *Soit $(X_n)_n$ une suite de v.a. ayant un moment d'ordre 2, deux à deux non corrélées, ayant même espérance m et même variance σ^2. Alors*

$$\frac{1}{n}\big(X_1 + \cdots + X_n\big) \xrightarrow{\mathbb{P}} m \ .$$

TABLE 5.1 — Les différents types de loi des grands nombres.

Loi forte	**Loi faible**
Moment d'ordre 1	Moment d'ordre 2
Indépendance	Non corrélation 2 à 2
Même loi	Même espérance et même variance

Théorème 5.2 — *Soit $(X_n)_n$ une suite de v.a. ayant un moment d'ordre 2, deux à deux non corrélées, ayant même espérance m et même variance σ^2. Alors*

$$\frac{1}{n}\big(X_1 + \cdots + X_n\big) \xrightarrow{p.s.} m \ .$$

5.2 Convergence en loi

5.2.1 Introduction

Remarque — Comment peut-on connaître la loi d'une v.a. ? En connaissant :

1. $\forall A$ borélien, $\mathbb{P}(X_n \in A)$;
1'. $\forall f$ bornée, $\mathbb{E}\big[f(X_n)\big]$;
2. $\forall t$, $\mathbb{E}\big[\exp(itX_n)\big] = \phi_{X_n}$;
3. $\forall f$ continue bornée (C_b), $\mathbb{E}\big[f(X_n)\big]$;
3'. $\forall f$ continue à support compact (C_K), $\mathbb{E}\big[f(X_n)\big]$;
3". $\forall f$ continue et $\lim\limits_{x\to\infty} f(x) = 0$, $\mathbb{E}\big[f(X_n)\big]$.

Rappel — Les fonctions de $C_K(\mathbb{R}^d)$ sont denses dans $\mathfrak{L}^1(\mathbb{R}^d, \mathrm{d}\mathbb{P}_{X_n})$ pour la norme 1.

Définition 5.2 (Convergence en loi) — *Soit $(X_n)_n$ une suite de v.a. On dit que X_n **converge en loi** (noté $\mathcal{L}$) vers X si $\forall f$ continue bornée,*

$$\mathbb{E}\big[f(X_n)\big] \longrightarrow \mathbb{E}\big[f(X)\big] \ .$$

Conséquence — Nous avons :

$$X_n \xrightarrow{\text{p.s.}} X \quad \Longrightarrow \quad X_n \xrightarrow{\mathcal{L}} X \ .$$

Définition 5.3 — *Une suite de mesures $\{m_n\}_n$ positives et bornées sur $\mathcal{B}(\mathbb{R}^d)$* ***converge étroitement*** *vers une mesure m bornée si*

$$\forall f \in C_b(\mathbb{R}^d), \quad \int f \, \mathrm{d}m_n \longrightarrow \int f \, \mathrm{d}m \ .$$

Conséquence — Nous avons :

$$X_n \xrightarrow{\mathcal{L}} X \quad \Longrightarrow \quad \mathbb{P}_{X_n} \xrightarrow{\text{étroit.}} \mathbb{P}_X \ .$$

Définition 5.4 (Convergence faible) — *Une suite de mesures $\{m_n\}_n$ positives et bornées sur $\mathcal{B}(\mathbb{R}^d)$* ***converge faiblement*** *vers une mesure m bornée si*

$$\forall f \in C_0(\mathbb{R}^d), \quad \int f \, \mathrm{d}m_n \longrightarrow \int f \, \mathrm{d}m \ .$$

Conséquence — La convergence étroite entraîne la convergence faible.

Proposition 5.10 — *Soient $\{m_n\}_n$ une suite de mesures positives et bornées, et m une mesure positive et bornée. Alors*

$$m_n \xrightarrow{\textit{étroit.}} m \quad \Longleftrightarrow \quad \begin{cases} m_n \xrightarrow{\textit{faibl.}} m, \\ m_n(\mathbb{R}^d) \longrightarrow m(\mathbb{R}^d). \end{cases}$$

Corollaire 5.1 — *Soient $(\mathbb{P}_n)_n$ une suite de proba. et $\mathbb{P}$ une proba. Alors*

$$\mathbb{P}_n \xrightarrow{\textit{étroit.}} \mathbb{P} \quad \Longleftrightarrow \quad \mathbb{P}_n \xrightarrow{\textit{faibl.}} \mathbb{P} \ .$$

Corollaire 5.2 — *Il y a équivalence entre :*

(i) $X_n \xrightarrow{\mathcal{L}} X$;

(ii) $\mathbb{P}_{X_n} \xrightarrow{\textit{étroit.}} \mathbb{P}_X$;

(iii) $\mathbb{P}_{X_n} \xrightarrow{\textit{faibl.}} \mathbb{P}_X$;

(iv) $\forall f \in C_b, \mathbb{E}\big[f(X_n)\big] \to \mathbb{E}\left[f(X)\right]$;

(iv) $\forall f \in C_0, \mathbb{E}\big[f(X_n)\big] \to \mathbb{E}\left[f(X)\right]$.

Proposition 5.11 — $X_n \xrightarrow{\mathbb{P}} X \quad \Longrightarrow \quad X_n \xrightarrow{\mathcal{L}} X.$

Proposition 5.12 — $X_n \xrightarrow{\mathcal{L}} a = \textit{cste} \quad \Longrightarrow \quad X_n \xrightarrow{\mathbb{P}} a.$

Proposition 5.13 — *Pour qu'une suite de probas $\mathbb{P}_n$ converge faiblement vers $\mathbb{P}$, il suffit que*

$$\int f \, \mathrm{d}\mathbb{P}_n \quad \longrightarrow \quad \int f \, \mathrm{d}\mathbb{P}$$

pour tout f appartenant à un ensemble total dans C_0.

Rappel — Un ensemble A est dit **total** dans C_0 si l'espace vectoriel engendré par A est dense dans C_0 muni de la norme sup.

Corollaire 5.3 — $X_n \xrightarrow{\mathcal{L}} X \quad \Longleftrightarrow \quad \mathbb{E}\big[f(X_n)\big] \longrightarrow \mathbb{E}\big[f(X)\big] \quad \forall f \in C_K(\mathbb{R}^d).$

Théorème 5.3 — $X_n \xrightarrow{\mathcal{L}} X \quad \Longleftrightarrow \quad \forall t \in \mathbb{R}^d, \ \phi_{X_n}(t) \longrightarrow \phi_X(t).$

Théorème 5.4 (Lévy) — $X_n \xrightarrow{\mathcal{L}} X \quad \Longleftrightarrow \quad \phi_{X_n} \longrightarrow \phi_X$ *continue en* 0.

Interprétation — Si $\phi_{X_n} \longrightarrow \phi_X$ continue en 0, alors ϕ est une fonction caractéristique, *i.e.* $\exists \mathbb{P}_X$ t.q.

$$\begin{cases} \phi = \phi_X \\ X_n \xrightarrow{\mathcal{L}} X \end{cases}$$

Proposition 5.14 (Slutsky) — *Nous avons :*

$$\left.\begin{array}{c} X_n \xrightarrow{\mathcal{L}} X \\ A_n \xrightarrow{\mathbb{P}} 0 \end{array}\right\} \quad \Longrightarrow \quad \begin{cases} A_n \cdot X_n \xrightarrow{\mathbb{P}} 0 \\ A_n + X_n \xrightarrow{\mathcal{L}} X. \end{cases}$$

$$\left.\begin{array}{c} X_n \xrightarrow{\mathcal{L}} X \\ A_n \xrightarrow{\mathbb{P}} a \\ B_n \xrightarrow{\mathbb{P}} b \\ a, b \text{ constantes} \end{array}\right\} \quad \Longrightarrow \quad A_n \cdot X_n + B_n \xrightarrow{\mathcal{L}} a \cdot X + b \ .$$

$$\left.\begin{array}{c} a_n \longrightarrow +\infty \\ a_n(X_n - b) \xrightarrow{\mathcal{L}} X \\ f \text{ différentiable au point } b \end{array}\right\} \quad \Longrightarrow \quad a_n\big[f(X_n) - f(b)\big] \xrightarrow{\mathcal{L}} f'(b) \cdot X \ .$$

5.2.2 Cas gaussien

Proposition 5.15 — *Soit $(X_n)_n$ une suite de v.a. gaussiennes telles que $X_n \xrightarrow{\mathfrak{L}^2} X$. Alors X est une v.a. gaussienne.*

Remarque — Si $X_n \xrightarrow{\mathfrak{L}^2} X$, alors

$$\mathbb{V}\mathrm{ar}(X_n) \longrightarrow \mathbb{V}\mathrm{ar}(X)$$

et

$$\mathbb{E}(X_n) \longrightarrow \mathbb{E}(X)$$

Notons que la réciproque est fausse.

Proposition 5.16 — *Soit $(X_n)_n$ une suite de v.a. gaussiennes telles que $X_n \xrightarrow{\mathcal{L}} X$. Alors X est une v.a. gaussienne.*

Théorème 5.5 (Limite centrale) — *Soit $(X_n)_n$ une suite de v.a. indépendantes, de même loi et dans $\mathfrak{L}^2$. On note respectivement m et σ^2 l'espérance et la variance de cette loi. Alors*

$$\sqrt{n}\left[\frac{1}{n}\left(X_1 + \cdots + X_n\right) - m\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Interprétation — La loi des grands nombres dit que $(X_1 + \cdots + X_n)/n$ tend p.s. vers m. Le théorème de la limite centrale signifie que la vitesse de convergence de $(X_1 + \cdots + X_n)/n$ vers m est de l'ordre de $1/\sqrt{n}$.

Théorème 5.6 (Théorème de la limite centrale vectorielle) — *Soit $(X_n)_n$ une suite de v.a. à valeurs dans $\mathbb{R}^d$, i.i.d. et dans $\mathfrak{L}^2$. On note respectivement m et K l'espérance et la matrice de covariance de cette loi. Alors*

$$\sqrt{n}\left[\frac{1}{n}\left(X_1 + \cdots + X_n\right) - m\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, K)\ .$$

Corollaire 5.4 — *Il y a équivalence entre :*

(i) $X_n \xrightarrow{\mathcal{L}} X$;

(ii) $\forall F$ *fermé de* $\mathbb{R}^d$, $\overline{\lim}\, \mathbb{P}_{X_n}(F) \leqslant \mathbb{P}_X(F)$;

(iii) $\forall O$ *ouvert de* $\mathbb{R}^d$, $\underline{\lim}\, \mathbb{P}_{X_n}(O) \geqslant \mathbb{P}_X(O)$;

(iv) $\forall A$ *borélien de* $\mathbb{R}^d$ *tel que* $\mathbb{P}_X(\delta A) = \mathbb{P}_X(\bar{A} - \mathring{A}) = 0$, $\mathbb{P}_{X_n}(A) \longrightarrow \mathbb{P}_X(A)$.

Corollaire 5.5 — *Si $\mathbb{P}_X$ a une densité, alors $\mathbb{P}_{X_n}\big(]a, b[\big) \longrightarrow \mathbb{P}_X\big(]a, b[\big)$.*

Proposition 5.17 — *Soit $(X_n)_n$ une suite de v.a. réelles. Alors*

$$X_n \xrightarrow{\mathcal{L}} X \iff F_{X_n}(t) \longrightarrow F_X(t) \text{ en tout point de discontinuité de } F_X\ .$$

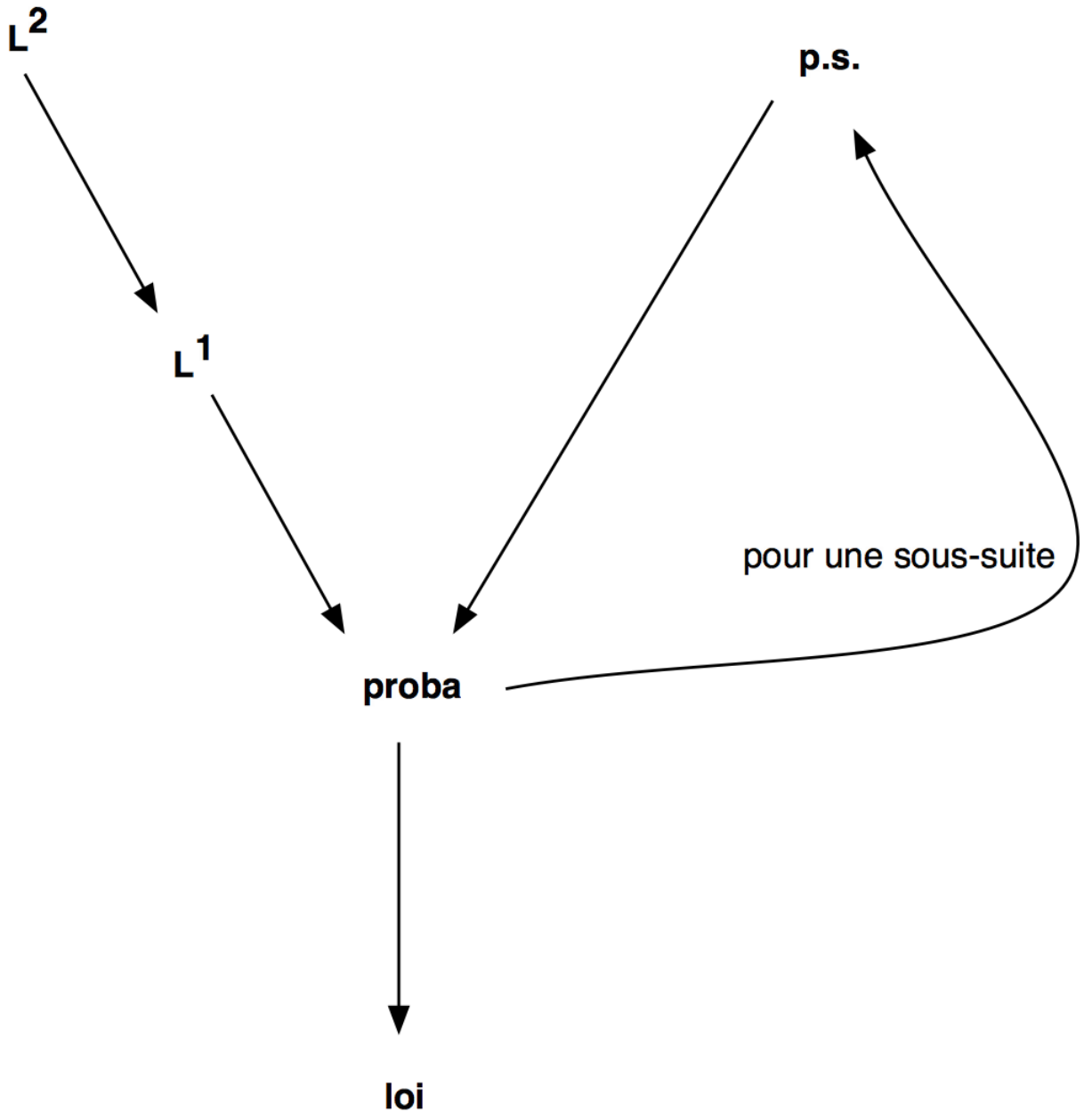


FIGURE 5.1 — Schématisation des différents types de convergence

Troisième partie

TEST

6

Introduction

Un test statistique est appelé à dégager un résultat significatif au milieu d'un ensemble de données expérimentales aléatoires. La méthodologie des tests consiste à répondre à l'aide de résultats expérimentaux à une question concernant les paramètres [1] de la loi de probabilité des variables aléatoires. Quatre conditions préalables au calcul d'un test doivent être réunies :

- la question doit être posée de telle sorte qu'il n'y ait que deux réponses possibles : oui et non ;
- on doit avoir des données chiffrées résultant d'un échantillon ou d'une expérimentation ;
- ces données doivent pouvoir être considérées comme la réalisation de variables aléatoires dont la forme de la loi de probabilité est connue ;
- la question doit concerner un ou plusieurs paramètres de cette loi.

Une fois posée cette dernière, la réponse du test est :

- soit l'acceptation de l'hypothèse, ce qui signifie que les donneés ne sont pas en contradiction avec l'hypothèse ;
- soit le rejet de cette hypothèse, ce qui signifie qu'il est très peu probable d'obtenir les résultats que l'on a trouvés si l'hypothèse est vraie, ou encore que les données sont en contradiction avec elle.

En un sens, le test d'hypothèse est une généralisation probabiliste du raisonnement par l'absurde, mais alors que ce dernier met en contradiction logique deux affirmations formelles, le premier oppose une affirmation formelle (l'hypothèse) avec des résultats du monde réel (les résultats de l'expérience).

De plus, le premier ne donne pas une certitude logique (l'hypothèse est fausse), mais seulement une forte présomption mesurée par une probabilité.

Enfin les deux formes du raisonnement ont en commun qu'elles ne peuvent que prouver (ou donner une présomption de preuve de) la fausseté de l'hypothèse et non sa vérité : ce n'est que parce qu'une expérience ne conduit pas au rejet de l'hypothèse que cette dernière est vraie : on peut imaginer d'autres expériences qui pourraient peut-être la rejeter.

1. Nous nous plaçons dans le cas paramétrique...

7

Théorie de Neyman-Pearson

7.1 Hypothèses simples

7.1.1 Introduction

Soit un modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$ tel que $\Theta = \Theta_0 \cup \Theta_1$ et $\Theta_0 \cap \Theta_1 = \emptyset$. On veut répondre à la question : « θ appartient-il à Θ_0 » ?

Définition 7.1 — *On appelle* ***hypothèse nulle*** $H_0 = \{\theta \in \Theta_0\}$.

Définition 7.2 — *On appelle* ***hypothèse alternative***[1] $H_1 = \{\theta \in \Theta_1\}$.

Définition 7.3 — *On appelle* ***test*** *une statistique* $\phi : \Omega \to \{0, 1\}$ *mesurable telle que*

$$\begin{cases} \text{si } \phi(\omega) = 0, & \text{on décide } H_0, \\ \text{si } \phi(\omega) = 1, & \text{on décide } H_1. \end{cases}$$

Définition 7.4 — *On appelle* ***région de rejet*** *(de* H_0*) l'ensemble* $\{\omega \mid \phi(\omega) = 1\}$.

Définition 7.5 — *On appelle* ***région d'acceptation*** *(de* H_0*) l'ensemble* $\{\omega \mid \phi(\omega) = 0\}$.

Définition 7.6 — *On appelle* ***hypothèse de base*** *l'hypothèse dont le rejet à tort a les conséquences les plus graves. C'est habituellement* H_0.

1. Parfois appelée *contre-hypothèse*.

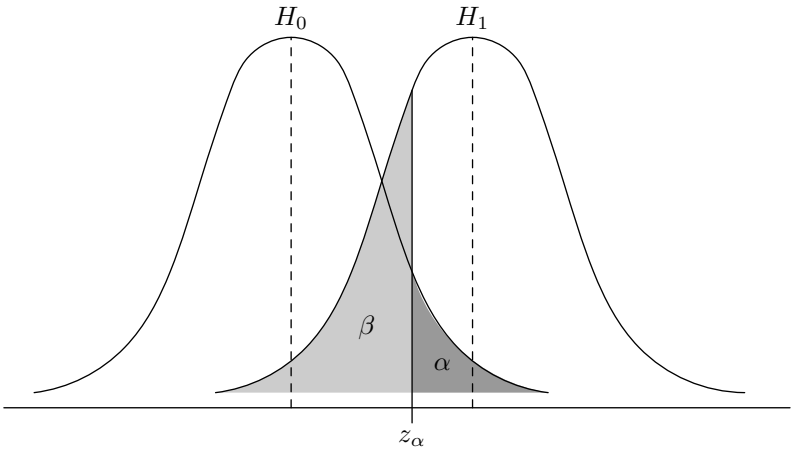


FIGURE 7.1 — Risques de première et de seconde espèce.

Définition 7.7 — *Un test d'hypothèse est dit* ***conservatif*** *car il conserve H_0 sauf si les données conduisent à la rejeter — H_0 est l'hypothèse privilégiée, i.e. celle que l'on garde si le résultat de l'expérience n'est pas clair.*

Définition 7.8 — *On appelle* ***risque de première espèce du test*** *ϕ la probabilité de rejeter à tort l'hypothèse de base, soit $\mathbb{P}_\theta(\phi = 0)$, $\theta \in \Theta_1$.*

Définition 7.9 — *On appelle* ***risque de seconde espèce du test*** *ϕ la probabilité de rejeter à tort l'hypothèse alternative, soit $\mathbb{P}_\theta(\phi = 1)$, $\theta \in \Theta_0$.*

Définition 7.10 — *On dit que le test est* ***exactement de niveau*** *α, $\alpha \in [0, 1]$, ssi*

$$\forall \theta \in \Theta_0, \quad \mathbb{P}_\theta(\phi = 1) \leqslant \alpha \ .$$

Définition 7.11 — *Un test ϕ de niveau α pour tester θ_0 contre θ_1, c.-à-d. tel que $\mathbb{E}_\theta(\phi) \leqslant \alpha$, $\forall \theta \in \Theta_0$, est dit* ***sans biais*** *si*

$$\mathbb{E}_\theta(\phi) \geqslant \alpha, \quad \forall \theta \in \Theta_1 \ .$$

7.1.2 Test randomisé

Définition 7.12 — *On appelle* ***test randomisé*** *une fonction ϕ mesurable de $(\mathfrak{X}, \mathcal{A})$ dans $[0, 1]$ telle que $\phi(\omega) = \gamma$ avec :*

- *si $\gamma = 0$: on décide H_0 ;*

- *si* $\gamma = 1$ *: on décide* H_1 *;*
- *si* $0 < \gamma < 1$ *: on effectue un tirage au sort auxiliaire, indépendant de l'expérience, à valeurs dans* $\{0, 1\}$ *:*

$$\begin{cases} \mathbb{P}\big(\{1\}\big) = \gamma, \\ \mathbb{P}\big(\{0\}\big) = 1 - \gamma. \end{cases}$$

Proposition 7.1 — *Il existe toujours un test randomisé de niveau exactement* α*.*

Remarque — Par opposition, il n'existe pas toujours de test (non randomisé) de niveau exactement α.

Définition 7.13 — *Un test* ϕ *randomisé est de niveau (exactement)* α *ssi*

$$\begin{cases} \forall \theta \in \Theta_0, \quad \mathbb{E}_\theta(\phi) \leqslant \alpha, \\ \exists \theta \in \Theta_0 \quad t.q. \quad \mathbb{E}_\theta(\phi) = \alpha. \end{cases}$$

Proposition 7.2 — *Le risque de première espèce vaut* $\mathbb{E}_\theta(\phi)$, $\theta \in \Theta_0$.

Proposition 7.3 — *Le risque de seconde espèce vaut* $1 - \mathbb{E}_\theta(\phi)$, $\theta \in \Theta_1$.

7.1.3 Puissance

Définition 7.14 — *On appelle* ***puissance*** *du test* ϕ *la quantité* $\mathbb{E}_\theta(\phi)$, $\theta \in \Theta_1$.

Définition 7.15 — *Un test* ϕ *est dit* ***uniformément le plus puissant (UPP)*** *de niveau* α *si*

$$\begin{cases} \phi \text{ est de niveau } \alpha, \\ \forall \phi' \text{ test de niveau } \alpha, \ \mathbb{E}_\theta(\phi) \geqslant \mathbb{E}_\theta(\phi'), \ \theta \in \Theta_1. \end{cases}$$

Théorème 7.1 — *Soit* $0 < \alpha < 1$.
1) Il existe $k \in \mathbb{R}^+$ *et* $\gamma \in]0,1[$ *tels que le test défini par*

$$\phi(x) = \begin{cases} 1 \text{ si } p_1(x) > k p_0(x), \\ 0 \text{ si } p_1(x) < k p_0(x), \\ \gamma \text{ si } p_1(x) = k p_0(x), \end{cases}$$

soit exactement de niveau α*.*
2) Soit $\phi^\star$ *un test de niveau* α*. Alors* ϕ *est plus puissant que* $\phi^\star$*, i.e.*

$$\mathbb{E}_{\theta_1}(\phi) \geqslant \mathbb{E}_{\theta_1}(\phi^\star) \ .$$

3) Si ϕ' *est un test de niveau tel que*

$$\mathbb{E}_{\theta_1}(\phi') \geqslant \mathbb{E}_{\theta_1}(\phi) \qquad \big(\Rightarrow \ \mathbb{E}_{\theta_1}(\phi') = \mathbb{E}_{\theta_1}(\phi) \big)$$

alors ϕ' vérifie

$$\phi'(x) \;=\; \begin{cases} 1 \text{ si } p_1(x) > kp_0(x), \\ 0 \text{ si } p_1(x) < kp_0(x). \end{cases}$$

Remarques — Ainsi :

1) ϕ est une réponse à la question : « parmi les tests de niveau α, existe-t-il un test UPP ? » ;
2) ϕ est « à peu près » la seule réponse, *i.e.* $\phi' = \phi$ sur $\{p_1 \neq k \cdot p_0\}$.

Théorème 7.2 — *Soit ϕ le test de niveau α de la forme*

$$\phi(x) \;=\; \begin{cases} 1 \text{ si } p_1(x) > kp_0(x), \\ 0 \text{ si } p_1(x) < kp_0(x), \\ \gamma \text{ si } p_1(x) = k \cdot p_0(x). \end{cases}$$

Alors

$$\left.\begin{array}{r} \phi'' \text{ t.q. } \mathbb{E}_{\theta_0}(\phi'') \leqslant \alpha \\ \text{et t.q. } \mathbb{E}_{\theta_1}(\phi'') \geqslant \mathbb{E}_{\theta_1}(\phi) \end{array}\right\} \qquad \Longrightarrow \qquad \phi'' = \phi \quad \text{sur} \quad \frac{p_{\theta_1}}{p_{\theta_0}} \neq k \ .$$

Remarque — Le choix de la valeur de $\phi^\star$ optimal sur $\{p_1 = k \cdot p_0\}$ n'est pas nécessairement déterminé.

Définition 7.16 — *Tout test $\phi^\star$ qui coïncide avec ϕ sur $\{p_1 = k \cdot p_0\}$ et qui vérifie $\mathbb{E}_{\theta_0}(\phi^\star) = \alpha$ est dit **optimal**. Un tel test s'appelle **test de Neyman–Pearson**.*

Proposition 7.4 — *Un test de Neyman–Pearson est nécessairement sans biais.*

Définition 7.17 — *Un test de Neyman–Pearson est nécessairement strictement sans biais, à condition que le modèle soit identifiable.*

7.2 Hypothèses multiples

7.2.1 Tests unilatères (*one-tailed tests*)

Il s'agit de tester $\Theta_0 = \{\theta \leqslant \theta_0\}$ (respectivement $\Theta_0 = \{\theta \geqslant \theta_0\}$) contre $\Theta_1 = \{\theta > \theta_0\}$ (respectivement $\Theta_1 = \{\theta < \theta_0\}$).

Définition 7.18 — *Soit $(\mathbb{P}_\theta)$ un modèle dominé. La famille $\mathbb{P}_\theta$ est dite **à rapport de vraisemblance monotone (RVM)** s'il existe une fonction $T : (\mathfrak{X}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ mesurable telle que $\forall \theta < \theta'$,*

$$\Phi_{\theta,\theta'}\big(T(x)\big) \;=\; \frac{p_{\theta'}}{p_\theta}(x)$$

soit une fonction croissante de $T(x)$.

Théorème 7.3 — *Soit une famille RMV. Le test*

$$\phi \;=\; \begin{cases} 1 \text{ si } T > \tau, \\ 0 \text{ si } T < \tau, \\ \gamma \text{ si } T = \tau, \end{cases}$$

est UPP pour tester $\Theta_0 = \{\theta \leqslant \theta_0\}$ *contre* $\Theta_1 = \{\theta > \theta_0\}$ *dès que* $\mathbb{E}_{\theta_0}(\phi) = \alpha$.

De même, le test

$$\phi \;=\; \begin{cases} 0 \text{ si } T > \tau, \\ 1 \text{ si } T < \tau, \\ \gamma \text{ si } T = \tau, \end{cases}$$

est UPP pour tester $\Theta_0 = \{\theta \geqslant \theta_0\}$ *contre* $\Theta_1 = \{\theta < \theta_0\}$ *dès que* $\mathbb{E}_{\theta_0}(\phi) = \alpha$.

7.2.2 Tests bilatères (*two-tailed tests*)

Soient $\theta_1 < \theta_2$. Il s'agit de tester l'une des trois hypothèses suivantes :

- $\theta \in [\theta_1, \theta_2]$ contre $\theta \notin [\theta_1, \theta_2]$;
- $\theta \notin [\theta_1, \theta_2]$ contre $\theta \in [\theta_1, \theta_2]$;
- $\theta = \theta_0$ contre $\theta \neq \theta_0$.

Théorème 7.4 — *Il n'existe pas de test UPP de* $\theta \in [\theta_1, \theta_2]$ *contre* $\theta \notin [\theta_1, \theta_2]$, $\forall \theta_1 \leqslant \theta_2$.

Théorème 7.5 — *Si la famille est exponentielle, soit de la forme*

$$\exp\big[C(\theta) \cdot T(x) - \psi(\theta)\big]$$

avec C *croisante, alors le test*

$$\phi \;=\; \begin{cases} 0 \text{ si } T \notin [\tau_1, \tau_2], \\ 1 \text{ si } T \in]\tau_1, \tau_2[, \\ \gamma_1 \text{ si } T = \tau_1, \\ \gamma_2 \text{ si } T = \tau_2, \end{cases}$$

est UPP, parmi les tests de niveau α*, à condition que* $\mathbb{E}_{\theta_1}(\phi) = \mathbb{E}_{\theta_2}(\phi) = \alpha$.

De même, le test

$$\phi \;=\; \begin{cases} 0 \text{ si } T \in]\tau_1, \tau_2[, \\ 1 \text{ si } T \notin [\tau_1, \tau_2], \\ \gamma_1 \text{ si } T = \tau_1, \\ \gamma_2 \text{ si } T = \tau_2, \end{cases}$$

est UPP, parmi les tests de niveau α*, à condition que* $\mathbb{E}_{\theta_1}(\phi) = \mathbb{E}_{\theta_2}(\phi) = \alpha$.

Degré de signification	Probabilité critique	Notation
Test significatif	$0{,}01 < P_c(t) \leqslant 0{,}05$	*
Test très significatif	$0{,}001 < P_c(t) \leqslant 0{,}01$	**
Test hautement significatif	$P_c(t) \leqslant 0{,}001$	***

TABLE 7.1 — Correspondance entre degré de signification et probabilité critique.

7.3 Probabilité critique et règle de décision associée

7.3.1 Définition

Soit t une réalisation de la statistique de test T. La **probabilité critique (*p-value*)** mesure la probabilité d'obtenir t ou une valeur encore plus éloignée de θ_0 si H_0 est vraie. C'est une mesure de l'accord entre l'hypothèse testée et le résultat obtenu. Plus elle est proche de 0, plus forte est la contradiction entre H_0 et le résultat obtenu. La contradiction au sens logique du terme correspond à une valeur nulle de la probabilité critique (le résultat obtenu est impossible quand H_0 est vraie).

Nous distinguons les deux cas suivants :

1° cas d'un test unilatéral : la région de rejet est de la forme $\mathcal{R} = \{T \geqslant l\}$; on appelle probabilité critique et on note P_c, $P_c(t) = \mathbb{P}(T \geqslant t \mid \theta = \theta_0)$;

2° cas d'un test bilatéral : la région de rejet est de la forme $\mathcal{R} = \{|T| \geqslant l\}$; on appelle probabilité critique et on note P_c, $P_c(t) = \mathbb{P}(|T| \geqslant t \mid \theta = \theta_0)$.

On a dans les deux cas la propriété suivante, qui permet de procéder à la décision d'acceptation ou de rejet au vu de la probabilité critique :

$$P_c(t) < \alpha \;\Leftrightarrow\; t \in \mathcal{R}$$

où $\mathcal{R}$ est la région de rejet d'un test de niveau α.

7.3.2 Signification statistique et importance de la distance entre θ et H_0

On mesure le degré de « signification [1] statistique » d'un test par $P_c(t)$: l'usage courant veut que l'on utilise la correspondance donnée par le tableau 7.1.

Remarque — On ne doit pas confondre le degré de signification statistique avec l'importance de la distance entre θ et H_0. On peut ainsi avoir un test hautement significatif avec un écart faible entre θ et θ_0 si le test est très puissant — inversement, avoir un test non significatif avec une différence réelle importante si le test est peu puissant.

1. Ou encore *significativité*...

8

Fisher et Cramer-Rao

8.1 Introduction

Définition 8.1 — *La fonction f est dite* ***absolument continue*** *de dérivée f' s'il existe une fonction f' intégrable sur tout intervalle $[a, b]$ et telle que*

$$\int_a^b f'(x)\ \mathrm{d}\lambda(x) \ = \ f(b) - f(a)\ .$$

Remarque — f' n'est définie que λ–p.s. Elle est appelée **dérivée faible**.

Remarque — Si la dérivée « classique » existe et est continue, alors f est absolument continue. Mais que la dérivée « classique » existe p.s. et et soit intégrable n'implique pas que f soit absolument continue.

Proposition 8.1 — *Nous avons*

$$f \text{ absolument continue } \iff \left\{ \begin{array}{l} \forall \phi \in C_K^\infty, \\ \int \phi'(x)\ f(x)\ \mathrm{d}x \ = \ -\int \phi(x)\ f'(x)\ \mathrm{d}x\ . \end{array} \right.$$

Proposition 8.2 — *Soit ou bien $g \in C^1(\mathbb{R})$ et f absolument continue, ou bien g absolument continue et $f \in C^1(\mathbb{R})$. Alors $g \circ f$ est absolument continue et sa dérivée faible est $g'(f) \times f'$.*

Proposition 8.3 — *Soient f et g absolument continues. Alors $f \cdot g$ est absolument continue de dérivée faible $f' \cdot g + f \cdot g'$.*

8.2 Modèles réguliers

Définition 8.2 — *Soit Θ un ouvert de $\mathbb{R}$ et $(\mathfrak{X}, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$ un modèle dominé, avec*

$$p(\theta, x) = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x) \ .$$

*La **régulariré** est caractérisée par :*

1° $\theta \mapsto \sqrt{p(\theta, x)}$ *est absolument continue pour tout x, μ–p.s. ;*

2° $x \mapsto \frac{\partial}{\partial\theta}\sqrt{p(\theta, x)} \in \mathcal{L}^2(\mu)$ *;*

3° *la fonction*

$$\theta \mapsto \mathcal{I}(\theta) = 4\int \left(\frac{\partial}{\partial\theta}\sqrt{p(\theta, x)}\right)^2 \mathrm{d}\mu(x)$$

est localement bornée.

Définition 8.3 — *La quantité $\mathcal{I}(\theta)$ s'appelle **l'information de Fischer**.*

Proposition 8.4 — *Dans un modèle régulier, quelle que soit la statistique T telle que $\theta \mapsto \mathbb{E}_\theta(T^2)$ soit localement bornée, $\theta \mapsto \mathbb{E}_\theta(T)$ est absolument continue de dérivée*

$$2\int T(x) \cdot \left(\frac{\partial}{\partial\theta}\sqrt{p(\theta, x)}\right) \cdot \sqrt{p(\theta, x)}\ \mathrm{d}\mu(x) \ .$$

Corollaire 8.1 — *Si $T = 1$,*

$$\int \left(\frac{\partial}{\partial\theta}\sqrt{p(\theta, x)}\right) \cdot \sqrt{p(\theta, x)}\ \mathrm{d}\mu(x) = 0 \ .$$

8.3 Information de Fischer

8.3.1 Changement de paramètres

Soit h bijective : h et h^{-1} sont continûment différentiables. On pose

$$\begin{cases} \eta = h(\theta), \\ \theta = h^{-1}(\eta). \end{cases}$$

On opère le changement suivant : $\mathbb{P}_\theta \to \mathbb{Q}_\eta = \mathbb{P}_{h^{-1}(\eta)}$. Alors

$$\mathcal{I}(\eta) = \left.\frac{\mathcal{I}(\theta)}{h'^2(\theta)}\right|_{\theta = h^{-1}(\eta)} .$$

8.3.2 Échantillonnage

Soit $(\mathfrak{X},\, \mathcal{A},\, \mathbb{P}_\theta,\, \theta \in \Theta)$ un modèle régulier, et $\mathcal{I}(\theta)$ l'information de Fischer associée. Soit un n-échantillon de ce modèle : $(\mathfrak{X}^{\otimes n},\, \mathcal{A}^{\otimes n},\, \mathbb{P}_\theta^{\otimes n},\, \theta \in \Theta)$ est régulier. On note $\mathcal{I}_n(\theta)$ l'information de Fischer associée.

Théorème 8.1 —
$$\mathcal{I}_n(\theta) \;=\; n \cdot \mathcal{I}(\theta)\ .$$

Proposition 8.5 — *Sous l'hypothèse (plus forte) que $\theta \mapsto \log p(\theta,\, x)$ est absolument continue, alors*
$$\begin{aligned}\mathcal{I}(\theta) &= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p(\Theta,\, X)\right]^2 \\ &= \int p(\theta,\, x)\cdot\left(\frac{\partial}{\partial\theta}\log p(\theta,\, x)\right)^2 \mathrm{d}\mu(x)\ .\end{aligned}$$

Proposition 8.6 — *Sous l'hypothèse (plus forte encore) que $\theta \mapsto \frac{\partial}{\partial\theta}\log p(\theta,\, x)$ est absolument continue, et que $\frac{\partial^2}{\partial\theta^2}\log p(\theta,\, x)$ est localement bornée, alors*
$$\mathcal{I}(\theta) \;=\; \mathbb{E}_\theta\left[\frac{\partial^2}{\partial^2\theta}\log p(\Theta,\, X)\right]\ .$$

8.4 Calculs de l'information de Fischer dans des cas particuliers

8.4.1 Familles exponentielles

Modèle droit Soit le modèle $\dfrac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x) = \exp\big\{\theta\cdot T(x) - \phi(\theta)\big\}$. Alors
$$\begin{aligned}\mathcal{I}(\theta) &= \phi''(\theta) \\ &= \mathbb{V}\mathrm{ar}_\theta(T)\ .\end{aligned}$$

Modèle courbe Soit le modèle $\dfrac{d\mathbb{P}_\theta}{d\mu}(x) = \exp\Big\{c(\lambda)\cdot T(x) - \phi\big[c(\lambda)\big]\Big\}$. Alors
$$\mathcal{I}(\lambda) = \big[c'(\lambda)\big]^2\ \mathbb{V}\mathrm{ar}_\lambda(T)\ .$$

8.4.2 Modèle de translation

Soit μ la mesure de Lebesgue. Soit f une densité de probabilité telle que $\sqrt{f}$ soit absolument continue de dérivée g, avec $g \in \mathcal{L}^2(\mu)$. Soit $\mathbb{P}_\theta$ telle que

$$\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x) \;=\; f(x-\theta)\;,$$

i.e. on observe Y sous $\mathbb{P}_\theta$: $Y = \theta + \epsilon$, ϵ de densité de probabilité f.

Alors

$$\mathcal{I}(\theta) \;=\; \int g^2 \; d\mu\;.$$

8.5 Autres résultats

Proposition 8.7 — *Soit T une statistique de loi $\mathbb{P}_\theta^T$. On suppose que $\left\{\mathbb{P}_\theta^T, \theta \in \Theta\right\}$ est un modèle régulier. On note respectivement $\mathcal{I}(\theta)$ et $\mathcal{I}^T(\theta)$ les informations de Fischer sur le modèle global et sur le modèle $\left\{\mathbb{P}_\theta^T, \theta \in \Theta\right\}$. Alors*

$$\mathcal{I}^T(\theta) \;\leqslant\; \mathcal{I}(\theta)$$

et

$$\mathcal{I}^T(\theta) \;=\; \mathcal{I}(\theta) \quad\Leftrightarrow\quad T \text{ exhaustive}\;.$$

Définition 8.4 — *T est une statistique **libre** sur le modèle $\left\{\mathbb{P}_\theta^T, \theta \in \Theta\right\}$ ssi la loi de T sous $\mathbb{P}_\theta$ ne dépend pas de θ.*

Proposition 8.8 — *T libre* $\Leftrightarrow$ $\mathcal{I}^T(\theta) \;=\; 0$.

8.6 Inégalité de Cramer-Rao

Théorème 8.2 (Inégalité de Cramer-Rao) — *Si le modèle est régulier et si T est une statistique telle que*

- *$\theta \mapsto \mathbb{E}_\theta(T^2)$ est localement bornée,*
- *$\mathcal{I}(\theta) > 0$,*

et si $\phi(\theta) = \mathbb{E}_\theta(T)$, alors ϕ est absolument continue de dérivée ϕ' et

$$\mathbb{V}ar_\theta(T) \;\geqslant\; \frac{\phi'(\theta)^2}{\mathcal{I}(\theta)}\;.$$

Définition 8.5 — *Soit T un estimateur sans biais de $\phi(\theta)$. T est dit* ***efficace*** *si*

$$\mathbb{V}ar_\theta(T) \;=\; \frac{\phi'(\theta)^2}{\mathcal{I}(\theta)} \;.$$

Théorème 8.3 — *Soit Θ un ouvert non vide. Si $\theta \mapsto \sqrt{p(\theta,\,x)}$ est continûment différentiable, alors $p(\theta,\,x) \neq 0$ $\forall x$, $\forall \theta$. On suppose $\phi(\theta)$ non constante. Dans un modèle régulier, s'il existe un estimateur T d'une quantité $\phi(\theta)$ qui soit sans biais et efficace, alors :*

- *le modèle est exponentiel ;*
- *T est la statistique du modèle exponentiel, i.e.*

$$p(\theta,\,x) \;=\; \exp\big\{v(\theta)\cdot T(x) - h(\theta)\big\} \;.$$

Soit $(\mathfrak{X},\ \mathcal{A},\ \mathbb{P}_\theta,\ \theta\in\Theta)$ un modèle régulier. Soit un n-échantillon de ce modèle : $(\mathfrak{X}^{\otimes n},\, \mathcal{A}^{\otimes n},\, \mathbb{P}_\theta^{\otimes n},\, \theta\in\Theta)$ est régulier. Soit $q(\theta)$ la quantité à estimer.

Définition 8.6 — *$(T_n)_n$ est une* ***suite d'estimateurs convergents*** *si*

$$\forall\theta\in\Theta,\quad T_n \;\xrightarrow{\;\mathbb{P}_\theta^n\;}\; q(\theta) \;,$$

i.e.

$$\forall\theta\in\Theta,\,\forall\epsilon>0,\quad \mathbb{P}_\theta^n\big(|t_n - q(\theta)|>\epsilon\big) \;\longrightarrow\; 0 \quad (n\to\infty) \;.$$

Définition 8.7 — *Soit $(\zeta_n)_n$ une suite de modèles. Soit $(T_n)_n$ une suite d'estimateurs associée à $q(\theta)$. On dit que T_n* ***converge en loi le long de*** *ζ_n* ***à la vitesse de*** *$(a_n)_n$ ssi l'une des trois conditions suivantes (équivalentes) est satisfaite :*

(i) si $X_n = a_n\big[T_n - q(\theta)\big]$, il existe une v.a. de loi $\mathbb{P}_\theta^X$ telle que

$$\forall\theta\in\Theta,\quad (\mathbb{P}_\theta)^{X_n} \;\longrightarrow\; \mathbb{P}_\theta^X \;;$$

(ii) $\forall\theta\in\Theta$, $\forall f$ continue bornée,

$$\int f\big[a_n(T_n - q(\theta)\big]\;\mathrm{d}\mathbb{P}_\theta^{X_n} \;\longrightarrow\; \int f(X)\;\mathrm{d}\mathbb{P}_\theta^X \qquad (n\to\infty) \;;$$

(iii) $\forall a,b\notin A_\theta$, $\forall\theta\in\Theta$,

$$\mathbb{P}_\theta^n\Big(a_n\big[T_n - q(\theta)\big]\in[a,b[\Big) \;\longrightarrow\; \mathbb{P}_\theta\big(X\in[a,b[\big) \;.$$

Définition 8.8 — *$[\alpha_n,\beta_n]$ est* ***asymptotiquement un intervalle de confiance*** *au niveau α si*

$$\forall\theta\in\Theta,\quad \lim_{n\to\infty}\mathbb{P}_\theta^n\big(q(\theta)\in[\alpha_n,\beta_n]\big) \;\geqslant\; 1-\alpha \;.$$

Remarque — Dans la précédente définition, θ est fixé.

Proposition 8.9 — *Dans un modèle régulier, tout estimateur « intéressant » converge à la vitesse $a_n = \sqrt{n}$ et la loi limite est normale, i.e. sous $\mathbb{P}_\theta$,*

$$X \rightsquigarrow \mathcal{N}\big(\theta, v(\theta)\big)$$

où

$$v(\theta) \;=\; \frac{1}{\mathcal{I}(\theta)} \;.$$

Définition 8.9 — *On dit qu'un estimateur T_n de θ est* ***asymptotiquement efficace*** *ssi*

$$\sqrt{n}\,(T_n - \theta) \quad \xrightarrow[\mathbb{P}_\theta^n]{\mathcal{L}} \quad \mathcal{N}\left(0, \frac{1}{\mathcal{I}(\theta)}\right) \qquad \forall \theta \in \Theta \;.$$

Définition 8.10 — *On dit qu'un estimateur T_n de θ est* ***super efficace*** *si $\forall \theta \in \Theta$,*

$$\sqrt{n}\,(T_n - \theta) \quad \xrightarrow[\mathbb{P}_\theta^n]{\mathcal{L}} \quad \mathcal{N}\big(0, V(\theta)\big)$$

où

$$\forall \theta, \quad V(\theta) \leqslant \frac{1}{\mathcal{I}(\theta)} \quad \textit{et} \quad \exists \theta_0 \textit{ t.q. } V(\theta_0) < \frac{1}{\mathcal{I}(\theta_0)} \;.$$

Théorème 8.4 — *Soit le modèle exponentiel $p(\theta, x) = \exp\big\{C(\theta)\cdot T(x) - \phi(\theta)\big\}$, avec C de classe C^1, bijective et telle que $C'(\theta) \neq 0$, $\forall \theta \in \Theta$. Soit un n-échantillon et $\hat{\theta}_n$ l'EMV associé. Alors*

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \quad \xrightarrow[\mathbb{P}_\theta^n]{\mathcal{L}} \quad \mathcal{N}\left(0, \frac{1}{\mathcal{I}(\theta)}\right) \qquad \forall \theta \in \Theta \;.$$

8.6.1 Maximum de vraisemblance en modèle exponentiel

Théorème 8.5 — *Soit θ appartenant à un ouvert de $\mathbb{R}$ et $p(\theta, x) = \exp\big\{C(\theta)\cdot T(x) - \Phi(\theta)\big\}$. Si $\hat{\theta}_n$ (existe et) est l'EMV calculé sur un n-échantillon de $\mathbb{P}_\theta$,*

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \quad \xrightarrow[\mathbb{P}_\theta]{\mathcal{L}} \quad \mathcal{N}\left(0\;,\; \frac{1}{\mathcal{I}(\theta)}\right) \qquad \forall \theta \in \Theta \;.$$

Remarque — Nous avons :

$$\hat{\theta}_n = \phi^{-1}\big[\frac{1}{n}\sum_{i=1}^{n} T(X_i)\big]$$

où

$$\begin{aligned} \phi(\theta) &= \mathbb{E}_\theta(T) \\ &= \frac{\phi'\big[C(\theta)\big]}{C'(\theta)} \;, \end{aligned}$$

i.e.

$$\phi'(\theta) = \frac{\mathcal{I}(\theta)}{C'(\theta)} \ .$$

Remarques — Elles sont au nombre de trois :

1) dans les modèles réguliers, la vitesse de convergence est plus petite ou égale à $\sqrt{n}$;
2) dans les modèles exponentiels, l'EMV atteint la vitesse $\sqrt{n}$ et est le plus efficace asymptotiquement, *i.e.* sa variance asymptotique est égale à la borne de Cramer-Rao d'un 1-échantillon :

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \quad \xrightarrow[\mathbb{P}_\theta]{\mathcal{L}} \quad \mathcal{N}\left(0\ ,\ \frac{1}{\mathcal{I}(\theta)}\right) \ ;$$

3) dans les modèles réguliers, parmi les estimateurs qui possèdent une certaine stabilité en loi, la « meilleure » variance asymptotique est $1/\mathcal{I}(\theta)$.

Théorème 8.6 — *Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi $\mathbb{P}_\theta$, avec $\theta \in \Theta$ ouvert de $\mathbb{R}$. On pose*

$$\begin{aligned} p(\theta, x) &= \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu} \ , \\ l(\theta, x) &= \log p(\theta, x) \ . \end{aligned}$$

On suppose que $\theta^\star$ est la vraie valeur du paramètre, et qu'il existe un voisinage $V(\theta^\star)$ inclus dans Θ et tel que :

1° $l^{(3)}(\theta, x)$ *existe* $\forall \theta \in V(\theta^\star)$, *$\mu$-p.s. en x ;*

2° $l(\theta, x)$ *est deux fois absolument continue et* $l^{(2)}(\theta, x) \in \mathcal{L}^1(\mathbb{P}_\theta)$ *;*

3° $\exists H(x) : \mathbb{R} \to \mathbb{R}^+$ *telle que* $H \in \mathcal{L}^1(\mathbb{P}_{\theta^\star})$ *et telle que*

$$\left|l^{(3)}(\theta, x)\right| \ \leqslant \ H(x) \quad \forall x \ \mu - p.s. \, ;$$

4° $\mathcal{I}(\theta^\star) > 0$ *;*

5° $\forall n \geqslant n_0$, $L_n(\theta) = \sum_{i=1}^n l(\theta, X_i)$ *(i.e. la log-vraisemblance) admet un maximum unique sur Θ.*

Alors

$$\hat{\theta}_n \ \longrightarrow \ \theta^\star \quad \textit{(en proba ou p.s.)}$$

et

$$\sqrt{n}\,(\hat{\theta}_n - \theta^\star) \ \xrightarrow{\mathcal{L}_{\mathbb{P}_{\theta^\star}}} \ \mathcal{N}\left(0\ ,\ \frac{1}{\mathcal{I}(\theta^\star)}\right) \ .$$

Remarque — Ces résultats demeurent vrais dans le cas vectoriel. En particulier, l'inégalité de Cramer-Rao pour un estimateur non biaisé T d'une quantité $\phi(\theta)$ devient

$$\mathbb{V}\mathrm{ar}_\theta(T) \ \geqslant \ \left[\bigtriangledown \phi(\theta)\right]^t \cdot \mathcal{I}^{-1}(\theta) \cdot \left[\bigtriangledown \phi(\theta)\right] \ .$$

Quatrième partie

STATISTIQUE GAUSSIENNE

9

Statistique gaussienne

9.1 Dans $\mathbb{R}$

Définition 9.1 — *Une variable aléatoire réelle (v.a.r.) Z est une* ***gaussienne standard*** *si et seulement si sa densité par rapport à la mesure de Lebesgue est*

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \ .$$

On note $\mathcal{N}(0,1)$ sa loi de probabilités.

Propriété 9.1 — *Les moments de Z sont :*

$$\begin{aligned} \mathbb{E}(Z) &= 0 \ , \\ \mathbb{E}(Z^{2k+1}) &= 0 \ , \\ \mathbb{E}(Z^{2k}) &= \frac{2 \cdot k!}{2^k \cdot k!} \ . \end{aligned}$$

Propriété 9.2 — *La fonction caractéristique de Z vaut*

$$\begin{aligned} \mathbb{E}(e^{-\phi z}) &= e^{\frac{\phi^2}{2}} \quad \forall \phi \in \mathbb{C} \ , \\ \mathbb{E}(e^{i\omega z}) &= e^{-\frac{\omega^2}{2}} \quad \forall \omega \in \mathbb{R} \ . \end{aligned}$$

Propriété 9.3 — *La fonction de répartition de Z est*

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, \mathrm{d}u \ .$$

Le tableau 9.1 fournit quelques valeurs de $\Phi(x) = 1 - F(x)$.

TABLE 9.1 — Valeurs de $\Phi(x) = 1 - F(x)$.

x	0,67	1,00	1,96	2,00	3,00
$\Phi(x)$	0,250	0,159	0,025	0,022	0,001

Nous pouvons obtenir l'encadrement de $\Phi(x)$ suivant :

$$\frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}} \cdot \frac{x^2}{1+x^2} \leqslant \Phi(x) \leqslant \left\{\frac{1}{2}e^{-\frac{x^2}{2}}\right\} \wedge \left\{\frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}}\right\},$$

$\forall x > 0$.

Si $x > 1$,

$$\frac{1}{2}\frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}} \leqslant \Phi(x) \leqslant \frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}}.$$

Quand $x \longrightarrow \infty$,

$$\Phi(x) \approx \frac{e^{-\frac{x^2}{2}}}{x\sqrt{2\pi}}.$$

Définition 9.2 — *Y est une gaussienne réelle ssi $Y = m + \sigma Z$ où $m \in \mathbb{R}$, $\sigma \geqslant 0$ et $Z \rightsquigarrow \mathcal{N}(0,1)$. On a :*

$$\begin{aligned} \mathbb{E}(Y) &= m, \\ \mathbb{V}ar(Y) &= \sigma^2, \\ \mathbb{E}(e^{i\omega Y}) &= e^{i\omega m - \frac{\omega^2\sigma^2}{2}} \end{aligned}$$

et

$$\mathbb{E}(e^{-pY}) = e^{-pm + \frac{p^2\sigma^2}{2}}.$$

Remarque — Une v.a.r. gaussienne est caractérisée par sa moyenne et sa variance.

9.2 Vecteurs gaussiens

Définition 9.3 — *$Y = (Y_1, \ldots, Y_n)^t$ est un vecteur gaussien ssi toute combinaison linéaire des Y_i est une gaussienne réelle.*

Proposition 9.1 — *Soit $Y = (Y_1, \ldots, Y_n)^t$ un vecteur aléatoire tel que $\mathbb{E}(Y) = m$ et $V = \mathbb{C}ov(Y)$ — matrice de variance-covariance. Alors Y est un vecteur gaussien ssi $\forall \omega \in \mathbb{R}^n$,*

$$\mathbb{E}(e^{i\omega^t Y}) = e^{i\omega^t m - \frac{1}{2}\omega^t V \omega}.$$

Proposition 9.2 — *Si $Y = (Y_1, \dots, Y_n)^t$ suit une loi $\mathcal{N}(m, V)$ et si $Z = AY + b$, alors Z suit une loi $\mathcal{N}(Am + b, AVA^t)$.*

Proposition 9.3 — *Soit $Z \rightsquigarrow \mathcal{N}(m,\, V)$, et $V = MDM^t$ la décomposition de V — cette décomposition existe puisque V est une matrice symétrique positive — avec M orthogonale et*

$$D = \begin{pmatrix} r_1^2 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & \vdots \\ \vdots & 0 & r_k^2 & 0 & \cdots & \vdots \\ \vdots & 0 & \cdots & 0 & \cdots & \vdots \\ \vdots & 0 & \cdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix} .$$

où $r_i \neq 0 \quad \forall i = 1, \dots, k$.

Alors il existe $X = (X_1, \dots, X_k)^t$ de loi $\mathcal{N}(0, I_k)$ tel que

$$Z = m + \sum_{i=1}^{k} r_i \, v_i \, X_i \ ,$$

où les v_i sont les k premiers vecteurs-colonnes de M.

Nota — Le théorème se réécrit matriciellement sous la forme $Z = m + BX$ avec

$$B = \begin{pmatrix} \overset{r_1 v_1}{\begin{pmatrix} \\ \vdots \\ \\ \end{pmatrix}} & \cdots & \overset{r_k v_k}{\begin{pmatrix} \\ \vdots \\ \\ \end{pmatrix}} \end{pmatrix} .$$

Cette matrice B, de dimensions $n \times k$, est injective (*i.e.* exactement de rang k).

9.3 Normes de vecteurs gaussiens

Définition 9.4 — *Soit $X \rightsquigarrow \mathcal{N}(0, I_n)$. Alors la norme de X est*

$$\|X\|^2 = \sum_{i=1}^{n} X_i^2 \ .$$

Proposition 9.4 — *La norme définie ci-dessus suit un $\chi^2(n)$.*

Proposition 9.5 — *Si $X \rightsquigarrow \mathcal{N}(0,1)$, alors X^2 suit une loi gamma $\Gamma\Big(\frac{1}{2}, \frac{1}{2}\Big)$.*

Rappel — Une loi $\Gamma(\rho, \lambda)$ a pour densité

$$\frac{\lambda^\rho}{\Gamma(\rho)} e^{-\lambda x}\, x^{\rho-1}\, \mathbf{1}_{\mathbb{R}^+}(x)\ .$$

La somme de deux lois gamma indépendantes vérifie

$$\Gamma(\rho_1, \lambda) + \Gamma(\rho_2, \lambda) \ = \ \Gamma(\rho_1 + \rho_2, \lambda)\ .$$

Par conséquent, si $X \rightsquigarrow \mathcal{N}(0,1)$, alors $\displaystyle\sum_{i=1}^{n} X_i^2 \rightsquigarrow \Gamma\Big(\frac{n}{2}, \frac{1}{2}\Big)$.

Nota — Si $X \rightsquigarrow \Gamma(\rho, \lambda)$, alors

$$\mathbb{E}(e^{-tX}) = \Big(\frac{\lambda}{\lambda + t}\Big)^\rho$$

pour tout $t > -\lambda$.

Remarque — Si $X \rightsquigarrow \mathcal{N}(0, I_n)$ et si P est une matrice $n \times n$ de projection (*i.e.* $P = P^t = P^2$), alors $\|PX\|^2 \rightsquigarrow \chi^2(p)$, où p est le rang de P.

Définition 9.5 — *On appelle **loi de Student** $t(k)$ la loi de la variable*

$$Z = \frac{X}{\sqrt{\frac{Y}{k}}}$$

si

$$\begin{cases} X \rightsquigarrow \mathcal{N}(0,1) \\ Y \rightsquigarrow \chi^2(k) \\ X \text{ et } Y \text{ sont indépendantes.} \end{cases}$$

Propriété 9.4 — *Pour tout x,*

$$t_n(x) \longrightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad (n \to \infty)\ .$$

Théorème 9.1 (Student) — *Si $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathcal{N}(m, \sigma^2)$, alors :*

1) $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \ \rightsquigarrow \mathcal{N}(m, \frac{\sigma^2}{n})$;
2) $R_n = \sum_{i=1}^n (X_i - \bar{X}_n)^2 \ \rightsquigarrow \sigma^2 \chi^2(n-1)$;
3) $\bar{X}_n$ *et* R_n *sont indépendants ;*
4) $T_n = \dfrac{\sqrt{n}(\bar{X}_n - m)}{S_n}$, *où* $S_n = \sqrt{\dfrac{R_n}{n-1}}$, *suit une loi de Student* $t(n-1)$.

10

Estimations et tests

10.1 Estimation de la moyenne

10.1.1 Cas où la variance est connue

Définition 10.1 — *L'estimateur de la moyenne μ de $(X_1, \ldots, X_n)$ est*

$$\hat{\mu} \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i \;.$$

D'après les résultats précédents, nous avons :

$$\bar{X_n} \rightsquigarrow \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big) \;,$$

c.-à-d.

$$\frac{\sqrt{n}(\bar{X_n} - \mu)}{\sigma} \rightsquigarrow \mathcal{N}(0,1) \;.$$

Par conséquent,

$$\mathbb{P}\bigg(\Big|\frac{\sqrt{n}(\bar{X_n} - \mu)}{\sigma}\Big| > 1{,}96\bigg) \;=\; 0{,}05$$

et donc

$$\mathbb{P}\bigg(\bar{X_n} - 1{,}96\,\frac{\sigma}{\sqrt{n}} \;\leqslant\; \mu \;\leqslant\; \bar{X_n} + 1{,}96\,\frac{\sigma}{\sqrt{n}}\bigg) \;=\; 1 - 0{,}05 \;.$$

Définition 10.2 — *On dit que* $\Big[\bar{X_n} - 1{,}96\,\frac{\sigma}{\sqrt{n}}\,, \bar{X_n} + 1{,}96\,\frac{\sigma}{\sqrt{n}}\Big]$ *est un* ***intervalle de confiance*** *pour μ de niveau d'erreur 5 %.*

10.1.2 Cas où la variance est inconnue

On se sert du théorème de Student : on sait que $\frac{\sqrt{n}(\bar{X}_n - m)}{S_n} \rightsquigarrow t(n-1)$. Par conséquent,

$$\mathbb{P}\left(\left| \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \right| > \tau_\alpha \right) = \alpha$$

et donc

$$\mathbb{P}\left(\bar{X}_n - \tau_\alpha \frac{S_n}{\sqrt{n}} \leqslant \mu \leqslant \bar{X}_n + \tau_\alpha \frac{S_n}{\sqrt{n}} \right) = 1 - \alpha \, .$$

Définition 10.3 — $\left[\bar{X}_n - \tau_\alpha \frac{S_n}{\sqrt{n}} \, , \bar{X}_n + \tau_\alpha \frac{S_n}{\sqrt{n}} \right]$ *est un intervalle de confiance pour* μ *de niveau d'erreur* α.

10.1.3 Test

À partir d'un échantillon dont on connait la moyenne empirique, on calcule un intervalle de confiance, et l'on regarde si la moyenne correcte (moyenne de référence) appartient à cet intervalle.

10.2 Estimation de la variance

La statistique de la variance σ^2 est

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

qui suit une loi $\dfrac{\sigma^2}{n-1} \, \chi^2(n-1)$.

Si $C_\alpha(n-1)$ est donné par $\mathbb{P}\big(\chi^2(n-1) > C_\alpha(n-1)\big) = \alpha$, alors

$$\mathbb{P}\left(\frac{n-1}{C_{\frac{\alpha}{2}}(n-1)} S_n^2 \leqslant \sigma^2 \leqslant \frac{n-1}{C_{1-\frac{\alpha}{2}}(n-1)} S_n^2 \right) = 1 - \alpha \, .$$

Définition 10.4 — *L'intervalle* $\left[\frac{n-1}{C_{\frac{\alpha}{2}}(n-1)} S_n^2 \, , \frac{n-1}{C_{1-\frac{\alpha}{2}}(n-1)} S_n^2 \right]$ *est un intervalle de confiance pour* σ^2 *de niveau d'erreur* α.

10.3 Comparaison des moyennes de deux populations

Soient $(X_1, \ldots, X_n)$ i.i.d. de loi $\mathcal{N}(\mu_1, \sigma_1^2)$, et $(Y_1, \ldots, Y_m)$ i.i.d. de loi $\mathcal{N}(\mu_2, \sigma_2^2)$. Les deux échantillons sont supposés indépendants.

10.3.1 Cas où les variances sont connues

D'après les résultats précédents, $\bar{X}_n - \bar{Y}_m$ suit une loi $\mathcal{N}\Big(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\Big)$.

La statistique de la différence des moyennes $\mu_1 - \mu_2$,

$$Z \;=\; \frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

et elle suit une loi normale centrée réduite.

Par conséquent,

$$\mathbb{P}\left(\bar{X}_n - \bar{Y}_m - 1{,}96\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \;\leqslant\; \mu_1 - \mu_2 \;\leqslant\; \bar{X}_n - \bar{Y}_m + 1{,}96\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right) \;=\; 0{,}95\ .$$

Définition 10.5 — *L'intervalle* $\left[\bar{X}_n - \bar{Y}_m - 1{,}96\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\,, \bar{X}_n - \bar{Y}_n + 1{,}96\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right]$ *est un intervalle de confiance pour* $\mu_1 - \mu_2$ *de niveau 5 %.*

10.3.2 Cas où les variances sont inconnues mais égales

Supposons que $\sigma_1^2 = \sigma_2^2 = \sigma^2$. La statistique de la différence des moyennes est

$$W \;=\; \frac{\bar{X}_n - \bar{Y}_m - \dfrac{\mu_1 - \mu_2}{\sqrt{\frac{1}{n} + \frac{1}{m}}}}{\dfrac{\sqrt{\sum (X_i - \bar{X}_n)^2 + \sum (Y_i - \bar{Y}_m)^2}}{n + m - 2}}$$

et elle suit une loi de Student $t(n + m - 2)$.

10.3.3 Test de l'hypothèse d'égalité des variances

On va construire une statistique pour le rapport des variances σ_1^2/σ_2^2, puis comparer cette statistique à 1.

Définition 10.6 — *On appelle* ***loi de Fischer-Snedecor*** $F(n_1,\ n_2)$ *la loi de la variable*

$$Z = \frac{X/n_1}{Y/n_2}\ ,$$

avec

$$\left\{\begin{array}{l} X \rightsquigarrow \chi^2(n_1) \\ Y \rightsquigarrow \chi^2(n_2) \\ X \text{ et } Y \text{ sont indépendantes.} \end{array}\right.$$

Remarque — $F(n_1, n_2) = 1/F(n_2, n_1)$.

La statistique du rapport des variances est

$$\frac{S_n^2}{S_m^2} ,$$

et elle suit une $F(n-1, n-2)$.

Par suite, si l'on note $f_x(n, m)$ la quantité telle que $\mathbb{P}\big(F(n-1, m-1) > f_x(n, m)\big) = x$, on obtient un intervalle de confiance pour le rapport des variances de niveau d'erreur α :

$$\left[\frac{S_1^2}{S_2^2} \cdot \frac{1}{f_{1-\frac{\alpha}{2}}(n, m)} \, , \, \frac{S_1^2}{S_2^2} \cdot \frac{1}{f_{\frac{\alpha}{2}}(n, m)}\right] .$$

11

Modèle linéaire

11.1 Présentation

Définition 11.1 — *On appelle* ***modèle linéaire gaussien unidimensionnel*** *une relation de la forme :*

$$Y = X\beta + \epsilon \ ,$$

où $Y = (Y_1, \ldots, Y_n)$ est un vecteur de n observations, X est une matrice $n \times p$ (observée elle aussi), β est un vecteur de p paramètres (inconnus) et ϵ est un vecteur aléatoire de dimension n, supposé suivre une loi $\mathcal{N}(0, \sigma^2 I_n)$.

Dans le cadre de la régression linéaire, on cherche à obtenir d'un modèle linéaire théorique

$$y_t \ = \ \beta x_t + \delta + \epsilon_t$$

un ajustement

$$\hat{y}_t \ = \ \hat{\beta} x_t + \hat{\delta} + \hat{\epsilon}_t$$

tel que $\hat{y}_t$ soit le plus « proche » possible de y_t. Ceci revient à chercher, parmi les droites d'équation $y = \beta x + \delta$, celle qui est telle que la somme des carrés des écarts $\hat{\epsilon}_t$ soit minimum.

Interprétation des paramètres — β est la pente de la droite : β_i représente la variation de la moyenne des Y_i lorsque X_i augmente d'une unité, *mutatis mutandis*. Quant à δ, il représente la valeur moyenne de Y lorsque $X_i = 0$.

Définition 11.2 — *La droite telle que la somme des $\hat{\epsilon}_t$ soit minimum est appelée* ***droite de régression*** *de y en x.*

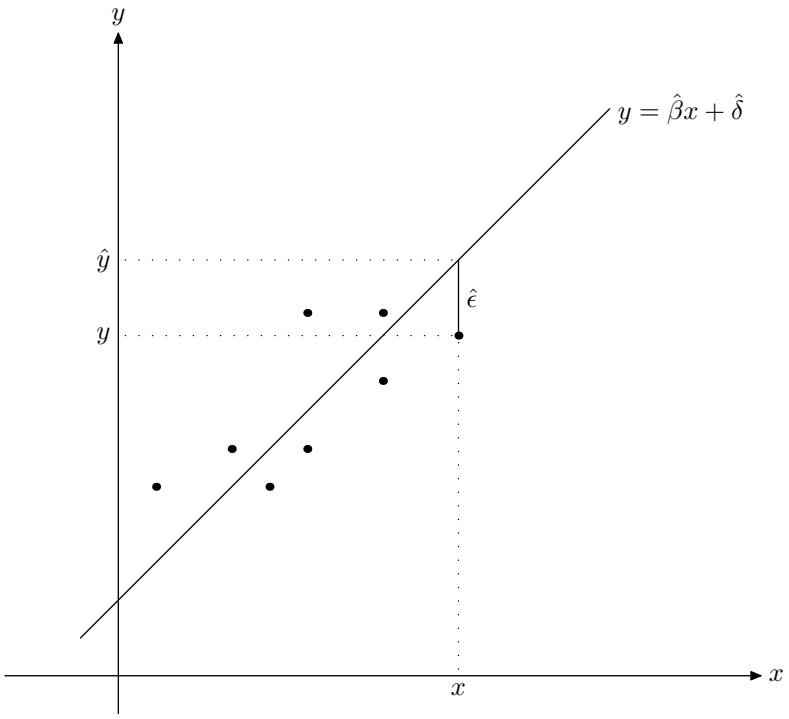


FIGURE 11.1 — Régression linéaire

11.2 Estimateur des moindres carrés

Définition 11.3 — *Y étant donné, soit la fonction γ définie comme suit :*

$$\begin{array}{rcl} \mathbb{R}^p & \to & \mathbb{R}^+ \\ \beta & \mapsto & \gamma(Y,\beta) = \|Y - X\beta\|^2 \ . \end{array}$$

On appelle ***estimateur des moindres carrés*** $\hat{\beta} = \arg\min_{\beta} \gamma(\beta, Y)$.

11.2.1 Interprétation géométrique

Soit $V \subset \mathbb{R}^n$ défini par

$$\begin{array}{rcl} V & = & \Im\mathrm{m}(X) \\ & = & X(\mathbb{R}^p) \\ & = & \{X\beta\,,\, \beta \in \mathbb{R}^p\} \ . \end{array}$$

$$\begin{array}{rcl} \hat{\beta} \text{ minimise } \|Y - X\beta\|^2 & \Leftrightarrow & X\hat{\beta} = \mathrm{proj}_V(Y) \\ & \Rightarrow & \hat{\beta} \text{ existe toujours, mais il n'est pas nécessairement unique.} \end{array}$$

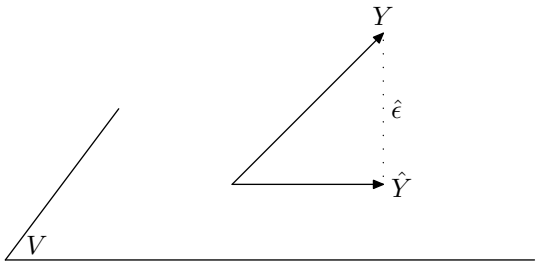


FIGURE 11.2 — Interprétation géométrique

Proposition 11.1 — *Nous avons :*
1) $\hat{\beta}$ *unique* $\Leftrightarrow$ X *injective ;*
2) pour $p \leqslant n$ *:* $X_{n\times p}$ *injective* $\Leftrightarrow$ X^tX *inversible.*

11.2.2 Expression algébrique de l'estimateur

Si X^tX est inversible,

$$\hat{\beta} \;=\; (X^tX)^{-1}X^tY \;.$$

Si X^tX n'est pas inversible, on définit la **pseudo-inverse** de X^tX : si X^tX est symétrique positive, alors $X^tX = M^tDM$ avec M orthogonale et

$$D = \begin{pmatrix} r_1^2 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & \vdots \\ \vdots & 0 & r_q^2 & 0 & \cdots & \vdots \\ \vdots & 0 & \cdots & 0 & \cdots & \vdots \\ \vdots & 0 & \cdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix} .$$

avec $r_1^2 > r_2^2 > \ \ldots\ > r_q^2 \neq 0$.

La pseudo-inverse $(X^tX)^{[-1]}$ est

$$(X^tX)^{[-1]} \;=\; M^t \cdot \begin{pmatrix} \frac{1}{r_1^2} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & \vdots \\ \vdots & 0 & \frac{1}{r_q^2} & 0 & \cdots & \vdots \\ \vdots & 0 & \cdots & 0 & \cdots & \vdots \\ \vdots & 0 & \cdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix} \cdot M \;.$$

On vérifie facilement que

$$\hat{\beta} \;=\; (X^tX)^{[-1]}X^tY \;.$$

11.2.3 Expression algébrique de l'opérateur de projection sur V

Pour tout Z,

$$\mathrm{proj}_V(Z) = X(X^tX)^{-1}X^tZ \ .$$

Définition 11.4 — *On appelle* **vecteur des résidus**

$$\hat{\epsilon} \ = \ Y - X\hat{\beta} \ .$$

11.3 Théorèmes de Cochran

Lemme 11.1 — *Soient $U_n \rightsquigarrow \mathcal{N}(\zeta,\, I_n)$ et A_i une matrice $n_i \times n$. Une condition nécessaire et suffisante pour que A_iU soit indépendante de A_jU est que $A_j^tA_i = 0$.*

Lemme 11.2 — *Soit $U_n \rightsquigarrow \mathcal{N}(\zeta,\, I_n)$. Alors $\|U\|^2$ a une loi qui ne dépend que de n, et si l'on pose $\|\zeta\| = \lambda^2$, cette loi est un* **chi-deux décentré** *$\chi'^2(n,\, \lambda^2)$.*

Lemme 11.3 — *Soient $P_1, \ldots, P_k \in \mathfrak{L}(\mathbb{R}^n,\mathbb{R}^n)$ telles que :*

- *$P_i = P_i^t$;*
- *$I_n = \sum_{i=1}^k P_i$.*

Alors on a les équivalences entre (i), (ii) et (iii) :

(i) $\sum_{i=1}^k \mathrm{rg}(P_i) \leqslant n$;

(ii) $P_iP_j = 0$ *si* $i \neq j$, $\forall i, j$;

(iii) $P_i^2 = P_i$, $\forall i$.

Proposition 11.2 — *Soit P un opérateur de projection de $\mathbb{R}^n$ dans $\mathbb{R}^m$, et soit $U_n \rightsquigarrow \mathcal{N}(\zeta,\, I_n)$. Alors $\|PU\|^2$ suit une loi $\chi'^2\big(\mathrm{rg}(P),\, \|P\zeta\|^2\big)$.*

Nota — $\|P\zeta\|^2$ s'appelle le **coefficient de non-centralité**.

Remarque — Réciproquement : si $P \in \mathfrak{L}(\mathbb{R}^n,\mathbb{R}^n)$ (ens. des applications linéaires) et si $X \rightsquigarrow \mathcal{N}(0,\, I_n)$, alors

$$\|PX\|^2 \rightsquigarrow \chi^2(k) \quad \Leftrightarrow \quad P^tP \text{ est un projecteur .}$$

Théorème 11.1 (Cochran 1) — *Soient $P_1, \ldots, P_k$ dans $\mathfrak{L}(\mathbb{R}^n,\, \mathbb{R}^n)$, et $X \rightsquigarrow \mathcal{N}(\zeta,\, I_n)$. On suppose que :*

- *$P_i = P_i^t$;*
- *$I_n = \sum_{i=1}^k P_i$;*

– $\sum_{i=1}^{k} \mathrm{rg}(P_i) \leqslant n$.

Alors $\forall i$, P_i est un projecteur ($P_i^2 = P_i$) et les $(P_i X)_{i=1,\dots,k}$ sont des vecteurs gaussiens de $\mathbb{R}^n$ indépendants et de loi $\mathcal{N}(P_i\zeta, P_i)$.

Théorème 11.2 (Cochran 2) — *Soient $X \rightsquigarrow \mathcal{N}(\zeta, I_n)$ et $Q_1, \dots, Q_k$ des formes quadratiques sur $\mathbb{R}^n$ telles que :*

– $\|X\|^2 = \sum_{i=1}^{k} Q_i(X)\,, \quad \forall x \in \mathbb{R}^n$;

– $\sum_{i=1}^{k} \mathrm{rg}(Q_i) \leqslant n$.

Alors les $(Q_i X)_{i=1,\dots,k}$ sont des $\chi'^2\big(\mathrm{rg}(Q_i), Q_i(\zeta)\big)$ indépendantes.

Remarques — Nous avons :

1) Cochran 1 $\Rightarrow$ Cochran 2 ;
2) $Q_i(X) = \|P_i X\|^2$.

11.4 Propriétés des estimateurs

Le modèle est toujours

$$Y = X\beta + \epsilon\ ,$$

avec Y vecteur $n \times 1$ des observations, X matrice $n \times p$, observée elle aussi, β vecteur paramètre de dimension p (inconnu) et ϵ vecteur aléatoire de dimension n, supposé suivre une $\mathcal{N}(0, \sigma^2 I_n)$.

11.4.1 Estimateur des moindres carrés

$$\mathbb{E}(\hat{\beta}) \ = \ \beta\ ,$$

i.e. $\hat{\beta}$ est sans biais.

$$\mathbb{C}\mathrm{ov}\,(\hat{\beta}) \quad = \quad \sigma^2 (X^t X)^{-1}\ .$$

11.4.2 Résidus

$$\hat{\sigma}^2 \quad = \quad \frac{\mathbb{E}\big(\|Y - X\hat{\beta}\|^2\big)}{n-p}\ .$$

11.4.3 Lois des estimateurs

Proposition 11.3 — *Nous avons :*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\epsilon} \end{pmatrix} \rightsquigarrow \mathcal{N}\left(\begin{pmatrix} \beta \\ 0 \end{pmatrix} \ , \ \sigma^2 \begin{pmatrix} (X^tX)^{-1} & 0 \\ 0 & I_n - X(X^tX)^{-1}X^t \end{pmatrix} \right) .$$

Proposition 11.4 — *Nous avons :*

$$\hat{\sigma}^2 \ \rightsquigarrow \ \frac{\sigma^2}{n-p} \cdot \chi^2(n-p)$$

et cet estimateur est indépendant de $\hat{\beta}$.

D'où la possibilité de construire des intervalles de confiance pour les différents estimateurs.

11.4.4 Test d'une sous-hypothèse linéaire

Soit C une matrice $l \times p$, $0 < l < p$. On fait l'hypothèse que les l lignes de C sont linéairement indépendantes. La question est la suivante : « $C\beta = 0$? »

Soit le sous-espace de dimension $p - 1$

$$\begin{aligned} V_1 &= \left\{ X\beta, \ \beta \in \mathbb{R}^p \ \mid \ C\beta = 0 \right\} \\ &= \left\{ X\beta, \ \beta \in \mathbb{R}^p \ \mid \ \beta_0 = \beta_1 = \ \ldots \ = \beta_{l-1} = 0 \right\} \\ &= \left\{ \tilde{X}\tilde{\beta}, \ \tilde{\beta} = (\beta_l, \cdots, \beta_{p-1}), \ \tilde{X} = p - l \ \text{ dernières colonnes de } \ X \right\} . \end{aligned}$$

La statistique, qui suit une $F(l, n-p)$, est

$$T \ = \ \frac{\| X\hat{\beta} - \tilde{X}\hat{\tilde{\beta}} \|^2 \ / \ l}{\| Y - X\hat{\beta} \|^2 \ / \ (n-p)} .$$

11.5 Théorème de Gauss-Markov et moindres carrés pondérés

On remet en doute l'hypothèse $\mathbb{C}\mathrm{ov}\,(\epsilon) = \sigma^2 I_n$. Maintenant, $\mathbb{C}\mathrm{ov}\,(\epsilon) = \sigma^2 G$, avec G matrice (connue) inversible.

On peut transformer le nouveau modèle pour obtenir un modèle linéaire ordinaire. Soit $G = B^tB$, avec B matrice $n \times n$ inversible. On multiplie le nouveau modèle par B^{-1} :

$$B^{-1}Y \ = \ B^{-1}X\beta + B^{-1}\epsilon \ ,$$

soit

$$Z \quad = \quad X'\beta + \epsilon' \ , \tag{11.1}$$

avec cette fois $\mathbb{C}\mathrm{ov}\,(\epsilon') = \sigma^2 I_n$.

β_z, estimateur des moindres carrés du modèle (2.1), rend minimum la quantité

$$\|Z - X'\beta\|^2 = \|Y - X\beta\|^2_{G^{-1}} \ ,$$

où $\|x\|^2_{G^{-1}} = x^t G^{-1} x$. Ainsi on obtient :

- un premier estimateur $(\hat{\beta})$ qui minimise $\|Y - X\beta\|^2$;
- un second estimateur qui minimise $\|Y - X\beta\|^2_{G^{-1}}$.

On peut définir des estimateurs des moindres carrés associés à une norme arbitraire A, qui rendent minimum $\|Y - X\beta\|^2_A$. Ce sont des **estimateurs des moindres carrés pondérés**.

Si

$$G = \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & v_n \end{pmatrix} \quad i.e. \quad G^{-1} = \begin{pmatrix} \frac{1}{v_1} & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{v_n} \end{pmatrix} \ ,$$

alors le β qui minimise $\|Y - X\beta\|^2_{G^{-1}}$ minimise $\sum_{i=1}^n \frac{1}{v_i}\big(Y_i - (X\beta)_i\big)^2$.

Définition 11.5 — *Si β^* minimise $\|Y-X\beta\|^2_A$, comme $\|Y-X\beta\|^2_A = (Y-X\beta)^t A(Y-X\beta)$, alors $\beta^* = BY$. β^* est dit **estimateur linéaire**.*

Théorème 11.3 (Gauss-Markov) — *Soit le modèle $Y = X\beta + \epsilon$, avec :*

- $\mathrm{rg}(X) = p$;
- *ϵ tel que $\mathbb{E}(\epsilon) = 0$ et $\mathbb{C}ov(\epsilon) = \sigma^2 I_n$.*

Soit $\tilde{\beta}$ un estimateur linéaire et sans biais, i.e. :

- *$\exists S$ tel que $\tilde{\beta} = SY$;*
- *$\forall \beta \in \mathbb{R}^p$, $\mathbb{E}_\beta(\tilde{\beta}) = \beta$.*

Soit $\Sigma_{\tilde{\beta}}$ la matrice de covariance de $\tilde{\beta}$. Alors

$$\Sigma_{\tilde{\beta}} = \Sigma_{\hat{\beta}} + R \ ,$$

où R est une matrice symétrique positive.

Nota — Si l'on veut estimer $a^t\beta$, Gauss-Markov nous dit : « Parmi les estimateurs linéaires et sans biais, dans le modèle standard, l'estimateur des moindres carrés $\hat{\beta}$ donne des estimateurs de $a^t\beta$ de variance minimum, et ce quel que soit a », puisqu'en effet

$$\begin{aligned} \mathbb{V}\mathrm{ar}(a^t\tilde{\beta}) \quad &= \quad a^t \Sigma_{\tilde{\beta}} a \\ &= \quad a^t \Sigma_{\hat{\beta}} a + a^t R a \ , \end{aligned}$$

et $a^t R a \geqslant 0$.

11.6 Coefficient de détermination et coefficients de corrélation

Le carré de le distance entre Y et $\hat{Y}$ s'appelle la **somme des carrés résiduelle** (SCR), car c'est la somme des carrés des résidus :

$$\begin{aligned} \text{SCR} &= \hat{\epsilon}^t\hat{\epsilon} \\ &= \sum_i \hat{\epsilon}_i^2 \ . \end{aligned}$$

On peut utiliser le théorème de Pythagore pour obtenir l'égalité matricielle suivante :

$$Y^tY \ = \ \hat{Y}^t\hat{Y} + (Y-\hat{Y})^t(Y-\hat{Y}) \ ,$$

soit

$$\text{SCT} \ = \ \text{SCM} + \text{SCR} \ ,$$

où SCM est la **somme des carrés due au modèle** (SCR) et SCT la **somme des carrés totale** (SCR). Ce sont ces carrés que l'on retrouve dans les tables d'analyse de la variance (cf. chapitre **??**).

Dans le cas où le modèle comprend un terme constant, on calcule plutôt la somme des carrés totale corrigée SCT_c et la somme des carrés due au modèle sans le terme constant SCM_c :

$$(Y-\bar{Y})^t(Y-\bar{Y}) \ = \ (\hat{Y}-\bar{Y})^t(\hat{Y}-\bar{Y}) + (Y-\hat{Y})^t(Y-\hat{Y}) \ ,$$

soit

$$\text{SCT}_c \ = \ \text{SCM}_c + \text{SCR} \ .$$

Géométriquement, cela revient à prendre comme origine dans $\mathbb{R}^n$, non plus le vecteur de coordonées 0, mais le vecteur dont toutes les coordonnées sont égales à $\bar{Y}$, moyenne de toutes les observations.

On peut donner une mesure de la qualité de l'ajustement du modèle aux observations : il s'agit du **coefficient de détermination** noté R^2

$$R^2 \ = \ \frac{\text{SCM}_c}{\text{SCT}_c} \ .$$

Géométriquement, ce rapport est égal au carré du cosinus de l'angle du vecteur $\overrightarrow{\bar{Y}Y}$ avec le sous-espace V (cf. fig. 11.3).

R^2 s'interprète comme la proportion de variabilité de Y « expliquée » par X. Plus il est proche de 1, meilleure est la qualité de la régression. La quantité $1-R^2$ est la proportion de variabilité de Y qui n'est pas « expliquée » par la régression.

Le test de l'hypothèse « $H_0 \ : \ \beta = 0$ » peut être fait avec la statistique

$$F \ = \ (n-2)\cdot\frac{R^2}{1-R^2}$$

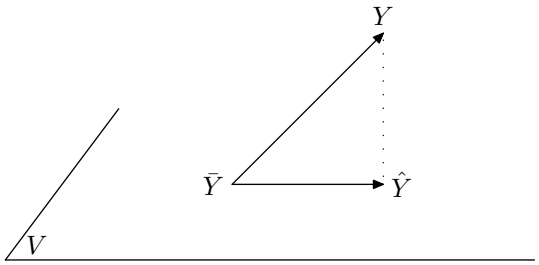


FIGURE 11.3 — Coefficient de détermination.

qui, sous H_0, suit une loi de Fisher $(1, n-2)$.

Remarque — $R = |r|$ où r est le coefficient de corrélation entre X et Y (r n'a pas de sens lorsque X est aléatoire).

On résume souvent l'ensemble des éléments de ce paragraphe dans une table synthétique, appelée **table d'analyse de la variance** de la régression (cf. tab. 11.1).

TABLE 11.1 — Table d'analyse de la variance.

Source de la variation	**Somme des carrés**	**Degrés de liberté**	**Carré moyen**	**Fisher**
Régression	SCM	1	SCM	$(n-2)\frac{\text{SCM}}{\text{SCE}}$
Erreur	SCE	$n-2$	$\frac{\text{SCE}}{(n-2)}$	
Total	SCT	$n-1$	$\frac{\text{SCT}}{(n-1)}$	

11.7 Coefficients multiples, partiels, semi-partiels

Considérons la régression de Y suivant deux variables X et Z :

$$Y \;=\; \beta_0 + \beta_1 X + \beta_2 Z + \epsilon \;.$$

11.7.1 Notations

Les estimateurs de β_0, β_1 et β_2 sont respectivement notés b_0, b_1 et b_2.
Les matrices de variance-covariance et de corrélation de (X,Y,Z) sont notées

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YX} & \sigma_Z^2 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YX} & 1 \end{pmatrix} \;.$$

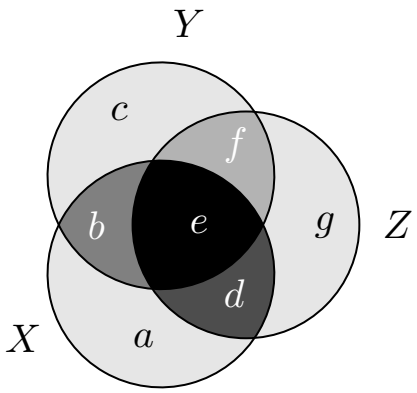


FIGURE 11.4 — Partage de la variance entre Y, X et Z.

Les matrices de variance-covariance et de corrélation de (X,Y,Z) calculées sur l'échantillon de taille n valent

$$\begin{pmatrix} s_X^2 & s_{XY} & s_{XZ} \\ s_{XY} & s_Y^2 & s_{YZ} \\ s_{XZ} & s_{YX} & s_Z^2 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & r_{XY} & r_{XZ} \\ r_{XY} & 1 & r_{YZ} \\ r_{XZ} & r_{YX} & 1 \end{pmatrix} .$$

La figure 13.1 illustre la décomposition de la variance de Y.

11.7.2 Coefficients multiples

Coefficient de détermination multiple R^2 Il mesure la proportion de variance de Y expliquée par X et Z.

$$\begin{aligned} R^2 &= \rho^2_{Y|XZ} \\ &= \rho^2_{XY} + \rho_{YZ|X}(1-\rho^2_{ZX}) \\ &= \rho^2_{XY} + \frac{(\rho_{YZ} - \rho_{YX}\rho_{ZX})^2}{1-\rho^2_{ZX}} \end{aligned} \tag{11.2}$$

Il correspond à la fraction

$$\frac{b+e+f}{b+c+e+f}$$

de la figure 13.1.

Si X et Z indépendante, *i.e.* $\rho_{XZ} = 0$, alors $\rho^2_{Y(X,Z)} = \rho^2_{YX} + \rho^2_{YZ}$.

Coefficient de corrélation multiple R Il s'agit de la racine carrée de R^2.

11.7.3 Coefficients partiels

Coefficient de détermination partielle Le coefficient de détermination partielle entre Y et X mesure la proportion de variance de Y expliquée par X (c.-à-d. que l'effet de Z est maintenu constant). Autrement dit : de la variance de Y qui n'est pas associée à Z, il mesure la proportion expliquée par X.

$$\rho^2_{XY|Z} = \frac{(\rho_{YX} - \rho_{XZ}\rho_{YZ})^2}{(1-\rho^2_{XZ})(1-\rho^2_{YZ})} .$$

Le coefficient de détermination partielle mesure aussi la proportion de variance du résidu de Y par rapport à Z, expliquée par le résidu de X par rapport à Z.

Il correspond à la fraction

$$\frac{b}{b+c}$$

de la figure 13.1.

Coefficient de corrélation partielle Le coefficient de corrélation partielle entre Y et X mesure la liaison entre Y et X lorsque Z est maintenue constante par rapport à X et Y. Il s'agit de la racine carrée du coefficient de détermination partielle.

Facteur d'inflation de la variance Avant introduction d'une seconde covariable dans un modèle de régression linéaire, l'écart-type de b_1 — estimateur de β_1 — vaut

$$\sqrt{\frac{s^2_{YX}}{\sum x^2 - (\sum x)^2/n}} = \sqrt{\frac{s^2_Y(1-r^2_{YX})}{(n-2)s^2_X}} .$$

Après inclusion de Z, la valeur de cet écart-type devient

$$\sqrt{\frac{s^2_Y(1-r^2_{Y|XZ})}{(n-3)s^2_X(1-r^2_{XZ})}} ,$$

où $r_{Y|XZ}$ est l'estimateur de $\rho_{Y|XZ}$. Nous voyons que l'écart-type inclue le terme $1/(1-r^2_{XZ})$, qui est appelé *facteur d'inflation de la variance*, puisqu'il mesure l'impact sur l'écart-type de la corrélation entre X et Z. Si r_{XZ} est proche de 1, l'écart-type de b_1 peut tendre vers l'infini.

De façon similaire, nous avons pour valeur de l'écart-type de b_2

$$\sqrt{\frac{s^2_Y(1-r^2_{Y|XZ})}{(n-3)s^2_Z(1-r^2_{XZ})}} .$$

La proportion sur échantillon de variation de Y expliquée par X et Z est tirée de l'équation (11.2) et vaut

$$r^2_{Y|XZ} = r^2_{XY} + r^2_{YZ|X}(1-r^2_{XZ}) .$$

Notons que cette proportion de variation peut approcher 1 alors même que les coefficients de régression pris individuellement n'ont pas d'effet significatif.

11.7.4 Coefficients semi-partiels

Le coefficient de détermination semi-partielle entre Y et X mesure la proportion de variance de Y expliquée par X seul (c.-à-d. que l'effet de Z est maintenu constant par rapport à X — mais pas par rapport à Y). Autrement dit : de la variance totale de Y, il mesure la proportion expliquée par la seule variable X.

$$\rho^2_{YX(Z)} = \rho^2_{Y|XZ} - \rho^2_{Y|Z} .$$

Il correspond à la fraction

$$\frac{b}{b+c+e+f}$$

de la figure 13.1.

Coefficient de corrélation semi-partielle Le coefficient de corrélation semi-partielle entre Y et X mesure la liaison entre Y et X lorsque Z est maintenue constante par rapport à X. Il s'agit de la racine carrée du coefficient de détermination semi-partielle.

11.7.5 Relation

Le coefficient de détermination partielle s'exprime sous la forme suivante :

$$\rho^2_{YX|Z} = \frac{\rho^2_{YX(Z)}}{1-\rho^2_{Y|Z}} ,$$

c.-à-d. comme le rapport du coefficient de détermination semi-partielle entre Y et X sur 1 moins le coefficient de détermination de Y suivant Z [1].

Nous retrouvons bien la relation à l'aide de la figure 13.1 :

$$\begin{aligned}\frac{\rho^2_{YX(Z)}}{1-\rho^2_{Y|Z}} &= \frac{\frac{b}{b+c+e+f}}{1-\frac{e+f}{b+c+e+f}} \\ &= \frac{b}{b+c+e+f} \times \frac{b+c+e+f}{b+c} \\ &= \frac{b}{b+c} .\end{aligned}$$

Exemple Soit le tableau donné suivant :

1. Ce résultat se généralise sous la forme suivante : il s'agit du rapport du coefficient de détermination semi-partielle entre Y et X sur 1 moins le coefficient de détermination de Y suivant toutes les variables explicatives sauf X.

y	x	z
4	1	8
2	1	7
3	1	7
4	1	9
5	1	5
5	1	4
7	2	3
5	2	6
4	2	7
9	2	2
7	2	3
6	2	2

```
> summary(lm(y~x+z,exrdeux))$r.squared
[1] 0.7266

> cor(exrdeux[, c(1, 2, 3)])
            y          x          z
y   1.0000000  0.6769405 -0.8240243
x   0.6769405  1.0000000 -0.6122400
z  -0.8240243 -0.6122400  1.0000000

> fit1_lm(y~z,exrdeux)
> fit2_lm(x~z,exrdeux)
> fit3_lm(fit1$resid~fit2$resid)
> summary(fit3)$r.squared

[1] 0.1482

> fit4_lm(y~x,exrdeux)
> fit5_lm(z~x,exrdeux)
> fit6_lm(fit4$resid~fit5$resid)
> summary(fit6)$r.squared

[1] 0.4953
```

Nous lisons donc :

- que le coefficient de détermination multiple R^2 égal à 0,7266 ;
- que la contribution de X à l'explication de la variation de Y vaut 0,187 (il s'agit du coefficient de régression centré réduit 0,276 que multiplie le coefficient de corrélation simple 0,6769) ;
- que la contribution de Z à l'explication de la variation de Y vaut 0,540 (il s'agit du coefficient de régression centré réduit $-0{,}655$ que multiplie le coefficient de corrélation simple -0,8240)) ;
- que le coefficient de détermination partiel $r^2_{YX|Z}$ — obtenu en maintenant constante Z par rapport à Y et X — vaut 0,1482 ;
- que le coefficient de détermination partiel $r^2_{YZ|X}$ — obtenu en maintenant constante X par rapport à Y et Z — vaut 0,4953.

Nous pouvons décomposer la proportion de variation de Y de la façon suivante.

Fraction [a] : proportion de variation de Y expliquée par X lorsque l'effet de Z est maintenu constant par rapport à X seulement (et non par rapport à Y) : c'est le r^2 obtenu en régressant Y sur le résidu d'une régression de X par rapport à Z, ou encore le coefficient de détermination semi-partielle de Y sur $X(Z)$:

```
> fit7 <- lm(x ~ z, exrdeux)
> fit8 <- lm(y ~ fit7$resid, exrdeux)
> summary(fit8)$r.squared

[1] 0.04756446
```

Une autre façon de le calculer est :

```
summary(lm(y~x+z,exrdeux))$r.squared - summary(lm(y~z,exrdeux))$r.squared

[1] 0.04756446
```

Fraction [c] : proportion de variation de Y expliquée par Z lorsque l'effet de X est maintenu constant par rapport à Z seulement (et non par rapport à Y) : c'est le r^2 obtenu en régressant Y sur le résidu d'une régression de Z par rapport à X, ou encore le coefficient de détermination semi-partielle de Y sur $Z(X)$:

```
> fit9 <- lm(z ~ x, exrdeux)
> fit10 <- lm(y ~ fit9$resid, exrdeux)
> summary(fit10)$r.squared

[1] 0.2683321
```

Fraction [b] : R^2 de la régression multiple de Y sur X et Z auquel on soustrait [a] et [c] :

$$0{,}7266 - 0{,}0475 - 0{,}2683 \;=\; 0{,}4107\ .$$

Le partitionnement donne donc :

$$\begin{aligned} \text{[a]+[b]+[c]+[d]} &= 0{,}0475 + 0{,}4107 + 0{,}2683 + (1 - 0{,}7266) \\ &= 1{,}0000\ . \end{aligned}$$

Remarque — On peut aussi calculer le r^2 partiel de Y sur X en maintenant constant l'effet de Z comme suit :

$$\begin{aligned} r^2_{XY|Z} &= \frac{\text{[a]}}{\text{[a]+[d]}} \\ &= \frac{0{,}0475}{0{,}0475 + 0{,}2683} \\ &= 0{,}1504 \\ &\approx 0{,}1482 \quad \text{(aux arrondis près).} \end{aligned}$$

Remarque — Nous obtenons également :

- la fraction [a] + [b] en calculant `summary(lm(y~x,exrdeux))$r.squared` ;
- la fraction [c] + [b] en calculant `summary(lm(y~z,exrdeux))$r.squared` ;
- la fraction [d] + [b] en calculant `summary(lm(x~z,exrdeux))$r.squared` ;
- R^2 = [a] + [b] + [c] en calculant `summary(lm(y~x+z,exrdeux))$r.squared`.

En résumé La contribution (au sens de Scherrer) d'une variable explicative n'est égale à la fraction [a] de variation expliquée (au sens du partionnement d ela variation) que dans un seul cas : lorsque toutes les variables explicatives sont orthogonales entre elles (*i.e.* linéairement indépendantes). Dans ce cas, la fraction [b] est nulle. Le coefficient de détermination multiple R^2 se calcule alors comme suit :

- soit en calculant $b_1 r_{YX} + b_2 r_{YZ}$, où b_1 et b_2 sont les coefficients de régression centrés réduits et r_{YX} et r_{YZ} les coefficients de corrélation linéaires simples (Pearson) ;
- soit en additionnant les fractions [a] et [c].

Dans le cas général, c'est-à-dire lorsque les variables explicatives ne sont pas indépendantes, elles expliquent chacune une part de la variation de Y, mais ces fractions se recouvrent plus ou moins. Dans ce cas, la fraction [b] n'est plus nulle ; elle « gruge » une partie des fractions [a] et [c], qui sont donc plus petites que les contributions partielles — ces contributions partielles étant égales à [a] ou [c] plus une partie de [b].

Dans ce cas, le R^2 total de la régression multiple (c.-à-d. la coefficient de détermination multiple) se calcule comme suit :

- soit en calculant $b_1 r_{YX} + b_2 r_{YZ}$, où b_1 et b_2 sont les coefficients de régression centrés réduits et r_{YX} et r_{YZ} les coefficients de corrélation linéaires simples (Pearson) ;
- soit en additionnant les fractions [a], [c] et [b].

Remarque — Il arrive que la fraction [b] soit négative : ceci arrive lorsque deux variables explicatives ont des effets marqués et opposés sur la variable dépendante, tout en étant corrélées entre elles. Dans ce cas, les fractions [a] et [c] sont plus grandes que leurs contributions partielles...

11.8 Sélection de variables

11.8.1 Méthode ascendante (forward)

```
> ozone.lm <- lm(ozone ~ temp, data = x)
> summary(ozone.lm)

Call: lm(formula = ozone ~ temp, data = x)
Residuals:
    Min     1Q  Median    3Q   Max
 -40.92 -17.46 -0.8738 10.44 118.1

Coefficients:
               Value Std. Error   t value  Pr(>|t|)
(Intercept) -147.6461   18.7553    -7.8723    0.0000
       temp    2.4391    0.2393    10.1919    0.0000

Residual standard error: 23.92 on 109 degrees of freedom
Multiple R-Squared: 0.488
F-statistic: 103.9 on 1 and 109 degrees of freedom, the p-value is 0

Correlation of Coefficients:
     (Intercept)
```

```
temp -0.9926

> add1(ozone.lm,  ~ temp + rad + wind)
Single term additions

Model:
ozone ~ temp
       Df Sum of Sq      RSS       Cp
<none>              62367.44 64656.15
   rad  1   2723.08 59644.36 63077.43
  wind  1  11419.45 50947.99 54381.06

> ozone2.lm <- lm(ozone ~ temp + rad, data = x)
> summary(ozone2.lm)

Call: lm(formula = ozone ~ temp + rad, data = x)
Residuals:
    Min     1Q Median    3Q   Max
 -36.61 -15.98 -2.928 12.37 115.6

Coefficients:
               Value Std. Error   t value  Pr(>|t|)
(Intercept) -145.7032   18.4467    -7.8986    0.0000
       temp    2.2785    0.2460     9.2622    0.0000
        rad    0.0571    0.0257     2.2205    0.0285

Residual standard error: 23.5 on 108 degrees of freedom
Multiple R-Squared: 0.5103
F-statistic: 56.28 on 2 and 108 degrees of freedom, the p-value is 0

Correlation of Coefficients:
     (Intercept)    temp
temp -0.9616
 rad  0.0474     -0.2941
```

11.8.2 Méthode descendante (backward

```
> ozone.lm <- lm(ozone ~ temp + rad + wind, data = x)
> drop1(ozone.lm,  ~ temp + rad + wind)
Single term deletions

Model:
ozone ~ temp + rad + wind
       Df Sum of Sq      RSS       Cp
<none>              47964.12 51550.22
  temp  1  19031.74 66995.86 69685.43
   rad  1   2983.87 50947.99 53637.56
  wind  1  11680.24 59644.36 62333.93
```

11.8.3 Méthode pas à pas (stepwise)

```
> ozone.lm <- lm(ozone ~ 1, data = x)
> step(ozone.lm,  ~ temp + rad + wind)
Start:  AIC= 124016.5
 ozone ~ 1

Single term additions

Model:
ozone ~ 1

scale:  1107.29

       Df Sum of Sq      RSS       Cp
<none>              121801.9 124016.5
  temp  1  59434.47  62367.4  66796.6
   rad  1  14779.68 107022.2 111451.4
  wind  1  45762.03  76039.9  80469.0

Step:  AIC= 66796.6
 ozone ~ temp

Single term deletions

Model:
ozone ~ temp

scale:  1107.29

       Df Sum of Sq      RSS       Cp
<none>               62367.4  66796.6
  temp  1  59434.47 121801.9 124016.5
Single term additions

Model:
ozone ~ temp

scale:  1107.29

       Df Sum of Sq      RSS       Cp
<none>              62367.44 66796.60
   rad  1   2723.08 59644.36 66288.10
  wind  1  11419.45 50947.99 57591.73

Step:  AIC= 57591.73
 ozone ~ temp + wind

Single term deletions

Model:
ozone ~ temp + wind
```

```
scale:  1107.29

       Df Sum of Sq      RSS       Cp
<none>              50947.99 57591.73
  temp  1  25091.90 76039.88 80469.04
  wind  1  11419.45 62367.44 66796.60
Single term additions

Model:
ozone ~ temp + wind

scale:  1107.29

       Df Sum of Sq      RSS       Cp
<none>              50947.99 57591.73
   rad  1  2983.867 47964.12 56822.44

Step:  AIC= 56822.44
 ozone ~ temp + wind + rad

Single term deletions

Model:
ozone ~ temp + wind + rad

scale:  1107.29

       Df Sum of Sq      RSS       Cp
<none>              47964.12 56822.44
  temp  1  19031.74 66995.86 73639.60
  wind  1  11680.24 59644.36 66288.10
   rad  1   2983.87 50947.99 57591.73
Call:
lm(formula = ozone ~ temp + wind + rad, data = x)

Coefficients:
 (Intercept)     temp      wind        rad
   -64.23208 1.651208 -3.337598 0.05979717

Degrees of freedom: 111 total; 107 residual
Residual standard error (on weighted scale): 21.17222
```

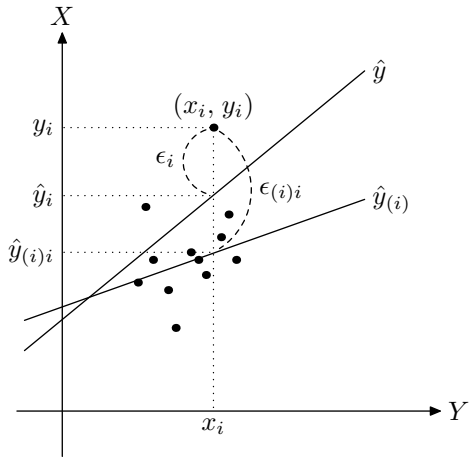
11.9 Adéquation du modèle

Les hypothèses à vérifier sont :
— la normalité des résidus : on trace un « QQ plot » ;
— l'omoscédasticité : ce terme désigne l'indépendance de la variance des résidus vis-à-vis des variables (à expliquer comme explicatives) ;

— l'indépendance entre eux des résidus : test de Durbin-Watson dans le cas de données temporelles.

L'identification des valeurs aberrantes passe par l'étude de la force de levier, et par suite, par le calcul de la distance de Cook.

Il s'avère également intéressant de comparer les résidus standardisés et les résidus studentisés, et d'étudier de près les points pour lesquels ces résidus diffèrent.



11.9.1 Différents types de résidus

Résidus observés Ils sont définis par :

$$\begin{aligned} \epsilon_i &= Y_i - \hat{Y}_i \\ &= Y_i - X_i'\beta \end{aligned}$$

où $\epsilon_i \sim \mathcal{N}\big(0, \sigma^2(1-h_{ii})\big)$ et

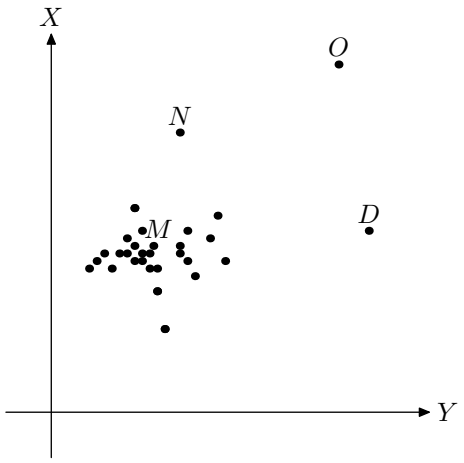
$$\begin{aligned} h_{ii} &= X_i'(X'X)^{-1}X_i' \\ &= \frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} \,. \end{aligned}$$

La matrice $n \times n$ des h_{ii} est appelée **matrice chapeau** (*hat matrix — H matrix*). Nous voyons que plus l'observation x_j est éloignée du centre des données $\bar{x}$, plus la pondération associée à y_j est importante dans le calcul de $\hat{y}_i$, et par conséquent plus cette observation aura un impact fort dans la détermination des valeurs $\hat{y}$.

Il est clair, également, que chaque observation y_j a un impact sur la détermination de $\hat{y}_i$.

Les éléments les plus importants de la matrice H sont ses élements diagonaux, qui mesurent l'impact de y_i sur $\hat{y}_i$. La quantité h_{ii} est appelée **force de levier** (*leverage — h hat value*) : elle mesure l'éloignement de l'observation x_i du centre des données X. Quand cette force h_{ii} est grande, $\hat{y}_i$ est plus sensible aux changements de valeurs de y_i que quand cette force est faible.

La somme des forces de levier $\sum_i h_{ii}$ est égale à $p+1$; aussi, une force de levier supérieure à $2(p+1)/n$ révèle une observation potentiellement aberrante.



Sur la figure — les points O et D sont éloignés du centre du nuage : ils auront donc des des forces de levier importantes. Par contre, le point N aura une force de levier faible. D'un autre côté, si l'on cherche la droite de régression ajustant au mieux les données contenues dans le nuage principal, et que l'on compare cette droite à celle ajustant au mieux le même nuage de points auquel on a adjoint soit N, soit D, soit O, on obtiendra des pentes et un intercept différents. en particulier, on voit que les points N et O auront un impact bien plus grand que D sur les estimations des paramètres de la droite de régression. Ainsi, le force de levier n'est qu'une mesure *partielle* de l'influence de l'observation (x_i, y_i) sur les paramètres de la droite de régression.

ésidus standardisés Pour pouvoir apprécier réellement le rôle des résidus, il est préférable de les standardiser : pour ce faire, on calcule

$$r_i \;=\; \frac{\epsilon_i}{s\sqrt{1-h_{ii}}}$$

qui sont les résidus standardisés : leur moyenne est nulle et leur variance vaut 1.

Ces résidus doivent se trouver dans l'intervalle $[-2,\, 2]$: tout résidu en dehors de cet intervalle indique une valeur aberrante potentielle.

Résidus studentisés Les résidus standardisés sont parfois appelés *résidus intrinsèquement studentisés*, car s^2 n'est pas indépendant de ϵ_i. Une alternative consiste à calculer le *résidu extrinsèquement studentisé* — plus simplement appelé *résidu studentisé* — qui fait intervenir non plus s^2, mais $s^2_{(i)}$, variance estimée sur l'ensemble des points excepté (x_i, y_i). Ainsi l'estimateur $s^2_{(i)}$ est-il indépendant de ϵ_i.

Le résidu studentisé est

$$t_i \;=\; \frac{\epsilon_i}{s_{(i)}\sqrt{1-h_{ii}}}\;,$$

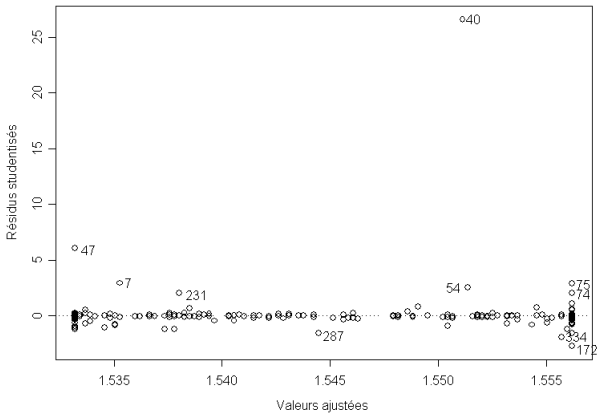
où $s_{(i)}$ est la variance sur l'échantillon auquel on a retiré le i^{e} point.

Ce résidu suit une loi $t(n-p-2)$, où p est le nombre de paramètres (c.-à-d. de variables explicatives). Il peut se réécrire sous la forme

$$t_i \;=\; \frac{\epsilon_i\sqrt{n-p-2}}{\left[(n-p-1)s^2(1-h_{ii}-\epsilon_i^2\right]^{1/2}}\;,$$

qui permet d'**apprécier la mesure de l'influence de l'observation** (x_i, y_i) sur l'ajustement par régression linéaire. En effet, le résidu est important si l'observation a un résidu (ordinaire) important — variabilité dans le sens de l'axe des ordonnées — et/ou s'il a une force de levier importante — variabilité dans le sens de l'axe des abscisses.

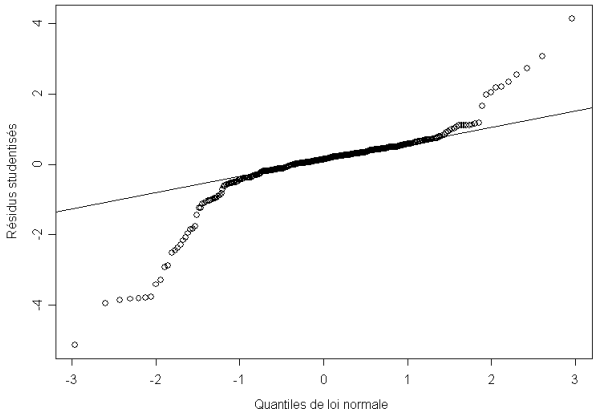
```
> library(MASS)
> fit_lm(log(x[,11])~x[,1])
> plot(fitted(fit),studres(fit),
xlab="Valeurs ajustées",
ylab="Résidus studentisés")
> abline(h=0,lty=2)
> identify(fitted(fit),
studres(fit), row.names(x))
```



11.9.2 Hypothèse de normalité

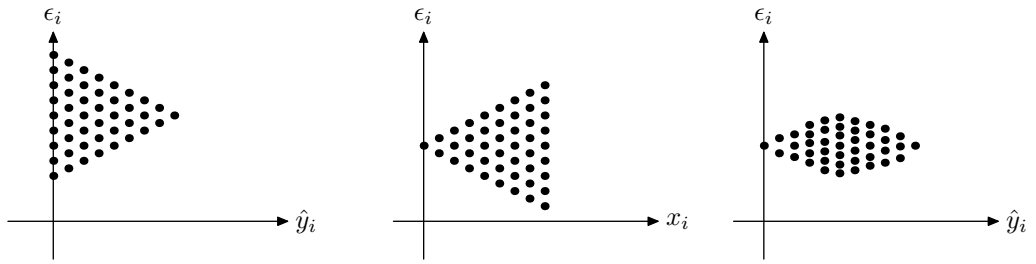
La vérification de cette hypothèse passe habituellement par le tracé d'un **QQ plot** : il s'agit de représenter les couples (ϵ_i, Q_i), où Q_i est la valeur attendue de ϵ_i si la distribution est exactement normale. Si l'hypothèse de normalité est vérifiée, les points du graphes doivent se trouver sur une droite.

```
> qqnorm(studres(fit),
xlab="Quantiles de loi normale",
ylab="Résidus studentisés")
> qqline(studres(fit))
```



11.9.3 Homoscédasticité

La vérification de cette hypothèse passe habituellement par le tracé des résidus en fonction de Y, puis de X, afin de voir si les résidus sont distribués aléatoirement, ou bien s'ils présentent une structure particulière. La figure ci-dessous donne des exemples d'hétéroscédasticité.



11.9.4 Diagnostic d'influence

Une fois la présence de valeurs aberrantes constatée, il convient de les identifier et de mesurer leur influence sur la régression. L'indicateur d'**influence de Cook** prend en compte à la fois la force du levier (variabilité horizontale) et l'importance de l'écart en terme de résidu (variabilité verticale). Il mesure l'influence d'une observation en comparant les estimations obtenues avec et sans cette observation : si les estimations changent peu, alors cette observation est considérée comme peu influente.

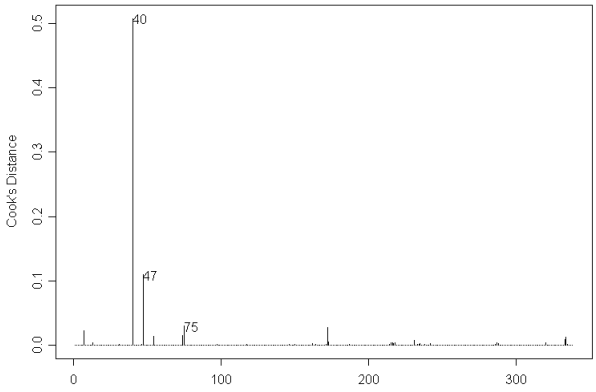
La statistique de Cook consiste en

$$\begin{aligned} D_i &= \sum_{j=1}^{n} \frac{1}{(p+1)s^2}(\hat{y}_{(i)j} - \hat{y}_j)^2 \\ &= \frac{h_{ii}}{(p+1)(1-h_{ii})} r_i^2 \, . \end{aligned}$$

La statistique sera importante si le résidu standardisé (variabilité verticale) est important et/ou si la force de levier (variabilité horizontale) est grande.

Une fois calculée cette statistique, on étudie plus particulièrement les observations pour lesquelles la statistique de Cook excède 1.

```
> plot((fit), which=c(6))
```



11.10 Multicolinéarité

Pour détecter la multicolinéarité, il faut entreprendre deux choses :

— étudier la matrice de corrélation : il y a multicolinéarité si les corrélations entre paires de variables explicatives sont plus importantes que leur corrélations avec la variable dépendante ;
— étudier la tolérance — la tolérante de la variable X_i est $1 - R_i^2$, où R_i est le coefficient de détermination de la variable X_i régressée par toutes les autres variables indépendantes — : si la tolérance est inférieure à 0,10, il y a colinéarité.

Les remèdes sont soit l'obtention de nouvelles données, soit la suppression d'une des variables indépendante corrélée.

Cinquième partie

ACP

12

Introduction

L'**Analyse en Composantes Principales** (**ACP**) fait partie du groupe des méthodes descriptives multidimensionnelles appelées **méthodes factorielles**. De par leur caractère descriptif, ces méthodes ne s'appuient pas sur un modèle probabiliste, mais dépendent d'un modèle géométrique. L'ACP propose, à partir d'un tableau rectangulaire de données comportant les valeurs de p variables quantitatives pour n unités (appelées aussi individus), des représentations géométriques de ces unités et de ces variables. Ces données peuvent être issues d'une procédure d'échantillonnage ou bien de l'observation d'une population tout entière. Les représentations des unités permettent de voir s'il existe une structure, non connue *a priori*, sur cet ensemble d'unités. De façon analogue, les représentations des variables permettent d'étudier les structures de liaisons linéaires sur l'ensemble des variables considérées. Ainsi, on cherchera si l'on peut distinguer des groupes dans l'ensemble des unités en regardant quelles sont les unités qui se ressemblent, celles qui se distinguent des autres, etc. Pour les variables, on cherchera quelles sont celles qui sont très corrélées entre elles, celles qui, au contraire, ne sont pas corrélées aux autres...

Nous verrons après l'exposé de la méthode quelles précautions il faut prendre pour interpréter correctement les représentation obtenues. Dans tous les cas, il ne faut pas oublier d'où sont issues les données utilisées et ce qu'elles représentent et signifient pour le problème que l'on se pose.

Enfin, comme pour toute méthode descriptive, réaliser une ACP n'est pas une fin en soi. L'ACP servira à mieux connaître les données sur lesquelles on travaille, à détecter d'éventuelles données suspectes, et aidera à formuler des hypothèses qu'il faudra étudier à l'aide de modèles et d'études statistiques inférentielles. On pourra aussi, *a posteriori*, se servir des représentations fournies par l'ACP pour illustrer certains résultats dans un but pédagogique.

12.1 Tableau de données

Les données consistent en p mesures, correspondant à des variables quantitatives $\{v_1,v_2,\ldots,v_p\}$, effectuées sur n unités $\{u_1,u_2,\ldots,u_n\}$.

Le tableau de données, noté $\boldsymbol{X}$, est de la forme

$$\boldsymbol{X} \;=\; \begin{array}{c} \\ u_1 \\ u_2 \\ \\ u_i \\ \\ u_n \end{array} \begin{array}{c} \begin{array}{ccccccc} v_1 & v_2 & & v_j & & v_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \end{array}$$

On peut représenter chaque unité par le vecteur de ses mesures sur les p variables :

$$\boldsymbol{U_i}^t \;=\; (x_{i1}, x_{i2}, \ldots, x_{ip}) \;,$$

ce qui donne

$$\boldsymbol{U_i} \;=\; \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{ip} \end{pmatrix} \;.$$

De façon analogue, on peut représenter chaque variable par un vecteur de $\mathbb{R}^n$ dont les composantes sont les valeurs de la variable pour les n unités :

$$\boldsymbol{V_j} \;=\; \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{pmatrix} \;.$$

Pour avoir une image de l'ensemble des unités, on se place dans un espace affine en choisissant comme origine un vecteur particulier de $\mathbb{R}^p$, par exemple le vecteur dont toutes les coordonnées sont nulles. Alors chaque unité sera représentée par un point dans cet espace. L'ensemble des points qui représentent les unités est appelé traditionnellement **nuage des individus**..

En faisant de même dans $\mathbb{R}^n$, chaque variable pourra être représentée par un point de l'espace affine correspondant. L'ensemble de ces points qui représentent les variables est appelé **nuage des variables**.

On constate que ces espaces, qui sont généralement de dimension supérieure ou égale à 2, ne permettent pas de visualiser ces représentations. L'idée générale des méthodes factorielles est de trouver un système d'axes et de plans tels que les projections de ces nuages de points sur ces axes permettent de reconstituer les positions des points les uns par rapport aux autres, c'est-à-dire d'avoir des images le moins déformées possible.

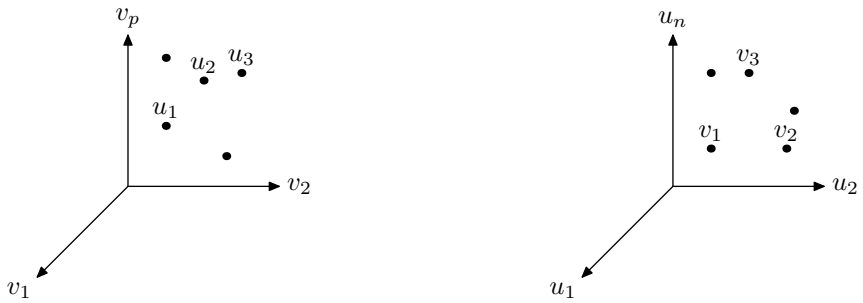


FIGURE 12.1 — Nuage des individus (à gauche) et des variables (à droite).

12.2 Choix d'une distance

Pour faire une représentation géométrique, il faut choisir une distance entre deux points de l'espace. La distance utilisée par l'ACP dans l'espace où sont représentées les unités est la **distance euclidienne** classique. La distance entre deux unités u_i et $u_{i'}$ est égale à

$$d^2(u_i, u_{i'}) \;=\; \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \;.$$

Avec cette distance, toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale. À cette distance on associe un produit scalaire entre deux vecteurs :

$$\begin{aligned} \langle \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} \rangle &= \sum_{j=1}^{p} x_{ij}\, x_{i'j} \\ &= \boldsymbol{U_i}^t\, \boldsymbol{U_{i'}} \;, \end{aligned}$$

ainsi que la norme d'un vecteur

$$\begin{aligned} \|\overrightarrow{ou_i}\|^2 &= \sum_{j=1}^{p} x_{ij}^2 \\ &= \boldsymbol{U_i}^t\, \boldsymbol{U_i} \;. \end{aligned}$$

On peut alors définir l'angle α entre deux vecteurs par son cosinus :

$$\begin{aligned} \cos(\alpha) &= \frac{\langle \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} \rangle}{\|\overrightarrow{ou_i}\| \cdot \|\overrightarrow{ou_{i'}}\|} \\ &= \frac{\sum_{j=1}^{p} x_{ij}\, x_{i'j}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2 \;\times\; \sum_{j=1}^{p} x_{i'j}^2}} \\ &= \frac{\boldsymbol{U_i}^t\, \boldsymbol{U_{i'}}}{\sqrt{(\boldsymbol{U_i}^t\, \boldsymbol{U_i})(\boldsymbol{U_{i'}}^t\, \boldsymbol{U_{i'}})}} \;. \end{aligned}$$

12.3 Choix de l'origine

Le point o correspondant au vecteur de coordonnées toutes nulles n'est pas forcément une origine satisfaisante, car si les coordonnées des points du nuage des individus sont grandes, le nuage est éloigné de cette origine. Il apparaît plus judicieux de choisir une origine liée au nuage lui-même : le centre de gravité du nuage. Pour définir ce centre de gravité, il faut choisir un système de pondération des unités : soit, pour tout $i = 1, \ldots, n$, p_i le poids de l'unité u_i tel que

$$\sum_{i=1}^{n} p_i \;=\; 1 \; .$$

Définition 12.1 — *Le centre de gravité est défini comme étant le point G tel que*

$$\sum_{i=1}^{n} p_i \, \overrightarrow{Gu_i} \;=\; \overrightarrow{0} \; .$$

Pour l'ACP, on choisit de donner le même poids $1/n$ à tous les individus. Le centre de gravité du nuage des individus est alors le point dont les coordonnées sont les valeurs moyennes des variables :

$$\begin{aligned} \boldsymbol{G} \;&=\; \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix} \\ &=\; \begin{pmatrix} x_{\bullet 1} \\ \vdots \\ x_{\bullet j} \\ \vdots \\ x_{\bullet p} \end{pmatrix} \; . \end{aligned}$$

Prendre G comme origine revient à travailler sur le tableau des données centrées :

$$\boldsymbol{X_c} \;=\; \begin{pmatrix} x_{11} - x_{\bullet 1} & \cdots & x_{1j} - x_{\bullet j} & \cdots & x_{1p} - x_{\bullet p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} x_{\bullet 1} & \cdots & x_{ij} - x_{\bullet j} & \cdots & x_{ip} - x_{\bullet p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} x_{\bullet 1} & \cdots & x_{nj} - x_{\bullet j} & \cdots & x_{np} - x_{\bullet p} \end{pmatrix} \; ,$$

et le vecteur des coordonnées centrées de l'individu u_i est

$$\boldsymbol{U_{ci}} \;=\; \begin{pmatrix} x_{i1} - x_{\bullet 1} \\ x_{i2} - x_{\bullet 2} \\ \vdots \\ x_{ij} - x_{\bullet j} \\ \vdots \\ x_{ip} - x_{\bullet p} \end{pmatrix} \; ,$$

celui des coordonnées centrées de la variable v_j étant

$$\boldsymbol{V_{cj}} \;=\; \begin{pmatrix} x_{1j} - x_{\bullet j} \\ x_{2j} - x_{\bullet j} \\ \vdots \\ x_{ij} - x_{\bullet j} \\ \vdots \\ x_{nj} - x_{\bullet j} \end{pmatrix} .$$

12.4 Moments d'inertie

12.4.1 Inertie totale du nuage des individus

Définition 12.2 — *On note I_G le **moment d'inertie du nuage des individus** par rapport au centre de gravité G, et on le définit ainsi :*

$$\begin{aligned} I_G &= \frac{1}{n}\sum_{i=1}^{n} d^2(G, u_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - x_{\bullet j})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U_{ci}}^t\, \boldsymbol{U_{ci}} \ . \end{aligned}$$

Ce moment d'inertie totale est intéressant car c'est une mesure de la dispersion du nuage des indivus autour de son centre de gravité. Si ce moment d'inertie est grand, cela signifie que le nuage est très dispersé, tandis que s'il est petit, alors le nuage est très concentré sur son centre de gravité.

Remarque — On peut voir, en inversant l'ordre des signes sommes, que I_G peut aussi s'écrire sous la forme

$$\begin{aligned} I_G &= \sum_{j=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - x_{\bullet j})^2\right] \\ &= \sum_{j=1}^{p}\mathbb{V}(v_j) \end{aligned}$$

où $\mathbb{V}(v_j)$ est la variance empirique de la variable v_j. Sous cette forme, on constate que l'inertie totale est égale à la trace de la matrice de variance-covariance $\boldsymbol{\Sigma}$ des p variables v_j :

$$I_G \;=\; \mathrm{tr}(\boldsymbol{\Sigma}) \ .$$

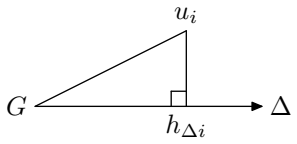


FIGURE 12.2 — Inertie du nuage des individus.

12.4.2 Inertie du nuage des individus par rapport à un axe passant par le barycentre

Définition 12.3 — *L'inertie du nuage des individus par rapport à un axe Δ passant par G est égale, par définition, à*

$$I_\Delta \;=\; \frac{1}{n}\sum_{i=1}^{n} d^2(h_{\Delta i}, u_i)\ ,$$

où $h_{\Delta i}$ est la projection orthogonale de u_i sur l'axe Δ (cf. fig. 12.2). Cette inertie mesure la proximité à l'axe Δ du nuage des individus.

12.4.3 Inertie du nuage des individus par rapport à un sous-espace vectoriel passant par le barycentre

Définition 12.4 — *L'inertie du nuage des individus par raport à un sous-espace vectoriel V passant par G est égale, par définition, à*

$$I_V \;=\; \frac{1}{n}\sum_{i=1}^{n} d^2(h_{Vi}, u_i)\ ,$$

où h_{Vi} est la projection orthogonale de u_i sur le sous-espace V.

12.4.4 Décomposition de l'inertie totale

Si on note $V^\star$ le complémentaire orthogonal de V dans $\mathbb{R}^p$ et $h_{V^\star i}$ la projection orthogonale de u_i sur $V^\star$, en appliquant le théorème de Pythagore, on peut écrire

$$\begin{aligned} d^2(h_{Vi}, u_i) + d^2(h_{V^\star i}, u_i) &= d^2(G, u_i) \\ &= d^2(h_{Vi}, G) + d^2(h_{V^\star i}, G)\ . \end{aligned}$$

On en déduit le résultat suivant.

Théorème 12.1 (Huygens) — *Nous avons*

$$I_V + I_{V^\star} \;=\; I_G\ .$$

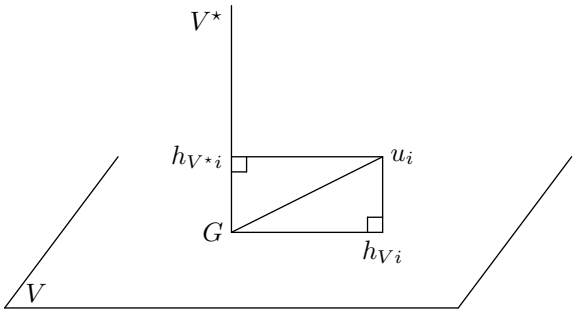


FIGURE 12.3 — Projection du nuage des individus sur un sous-espace.

Dans le cas particulier où le sous-espace est de dimension 1, c.-à-d. est un axe, $I_{V^\star}$ est une mesure de l'allongement du nuage selon cet axe. On emploie pour $I_{V^\star}$ les expressions d'**inertie portée par l'axe** ou bien d'**inertie expliquée par l'axe**..

En projetant le nuage des individus sur un sous-espace V, on perd l'inertie mesurée par I_V, et l'on ne conserve que celle mesurée par $I_{V^\star}$ (fig. 12.3).

De plus, si on décompose l'espace $\mathbb{R}^p$ comme la somme de sous-espaces de dimension 1 et orthogonaux entre eux :

$$\Delta_1 \oplus \Delta_2 \oplus \cdots \oplus \Delta_p \ ,$$

on peut écrire

$$I_G \ = \ I_{\Delta_1^\star} + I_{\Delta_2^\star} + \cdots + I_{\Delta_p^\star} \ .$$

13

Réalisation

13.1 Recherche de l'axe passant par le barycentre et d'inertie minimum

On cherche un axe Δ_1 passant par G d'inertie I_{Δ_1} minimum car c'est l'axe le plus proche de l'ensemble des points du nuage des individus, et donc, si l'on doit projeter ce nuage sur l'axe, c'est lui qui donnera l'image la moins déformée du nuage. Si on utilise la relation entre les inerties donnée au paragraphe précédent, rechercher Δ_1 tel que I_{Δ_1} soit minimum, est équivalent à rechercher Δ_1 tel que $I_{\Delta_1^\star}$ soit maximum.

On définit l'axe Δ_1 par son vecteur directeur unitaire $\overrightarrow{Ga_1}$. Il faut donc trouver $\overrightarrow{Ga_1}$ tel que $I_{\Delta_1^\star}$ soit maximum sous la contrainte que $\|\overrightarrow{Ga_1}\|^2 = 1$.

Nous avons :

$$\begin{aligned} d^2(G,h_{V^\star i}) &= \langle \overrightarrow{Gu_i}, \overrightarrow{Ga_1} \rangle^2 \\ &= a_1^t \boldsymbol{U_{ci}}\, \boldsymbol{U_{ci}}^t a_1 \end{aligned}$$

en utilisant la symétrie du produit scalaire.

Nous en déduisons

$$\begin{aligned} I_{\Delta_1^\star} &= \frac{1}{n} \sum_{i=1}^{n} a_1^t\, \boldsymbol{U_{ci}}\, \boldsymbol{U_{ci}}^t\, a_1 \\ &= a_1^t \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{U_{ci}}\, \boldsymbol{U_{ci}}^t \right] a_1 \ . \end{aligned}$$

Entre crochets, nous reconnaissons la matrice de variance-covariance empirique $\boldsymbol{\Sigma}$ des p variables.

$$I_{\Delta_1^\star} = a_1^t\, \boldsymbol{\Sigma}\, a_1$$

et

$$\|\overrightarrow{Ga_1}\|^2 = a_1^t\, a_1 \ .$$

Le problème à résoudre est donc le suivant : trouver a_1 tel que $a_1^t \mathbf{\Sigma} a_1$ soit maximum avec la contrainte $a_1^t\ a_1 = 1$. C'est le problème de la recherche d'un optimum d'une fonction de plusieurs variables liées par une contrainte — les inconnues étant les composantes de a_1. La méthode des multiplicateurs de Lagrange peut alors être utilisée. Il faut calculer les dérivées partielles de

$$\begin{aligned} g(a_1) &= g(a_{11}, a_{12}, \ldots, a_{1p}) \\ &= a_1^t\ \mathbf{\Sigma}\ a_1 - \lambda_1(a_1^t\ a_1 - 1)\ . \end{aligned}$$

En utilisant la dérivée matricielle, on obtient

$$\begin{aligned} \frac{\partial\ g(a_1)}{\partial a_1} &= 2\mathbf{\Sigma}a_1 - 2\lambda_1 a_1 \\ &= 0\ . \end{aligned}$$

Le système à résoudre est

$$\left\{ \begin{array}{rcll} \mathbf{\Sigma}a_1 - \lambda_1 a_1 &=& 0 & (1) \\ a_1^t a_1 - 1 &=& 0 & (2) \end{array} \right.$$

De l'équation matricielle (1) de ce système, on déduit que a_1 est vecteur propre de la matrice $\mathbf{\Sigma}$ associé à la valeur propre λ_1. En multipliant à gauche par a_1^t les deux membres de l'équation (1), on obtient

$$a_1^t \mathbf{\Sigma} a_1 - \lambda_1 a_1^t a_1\ =\ 0$$

et en utilisant (2) on trouve finalement que

$$a_1^t \mathbf{\Sigma} a_1\ =\ \lambda_1\ .$$

On reconnaît que le premier membre de cette dernière équation est égal à l'inertie $I_{\Delta_1^\star}$, qui doit être maximum. Cela signifie que la valeur propre λ_1 est la plus grande valeur propre de la matrice de variance-covariance $\mathbf{\Sigma}$ et que cette valeur propre est égale à l'inertie portée par l'axe Δ_1.

L'axe Δ_1 pour lequel le nuage des individus a l'inertie minimum a comme vecteur directeur unitaire le premier vecteur propre associé à la plus grande valeur propre de la matrice de variance-covariance $\mathbf{\Sigma}$.

13.2 Recherche des axes suivant

On recherche ensuite un deuxième axe Δ_2 orthogonal au premier et d'inertie minimum. On peut, comme dans le paragraphe précédent, définir l'axe Δ_2 passant par G par son vecteur directeur unitaire a_2. L'inertie du nuage des individus par rapport à son complémentaire orthogonal est égale à

$$I_{\Delta_2^\star}\ =\ a_2^t\ \mathbf{\Sigma}\ a_2$$

et elle doit être maximum avec les deux contraintes suivantes :

$$\left\{ \begin{array}{rcl} a_2^t\ a_2 &=& 1 \\ a_2^t\ a_1 &=& 0\ . \end{array} \right.$$

La deuxième contrainte exprime le fait que le deuxième axe doit être orthogonal au premier, et donc que le produit scalaire des deux vecteurs directeurs est nul. En appliquant

la méthode des multiplicateurs de Lagrange, cette fois avec deux contraintes, on trouve que a_2 est le vecteur propre de $\mathbf{\Sigma}$ correspondant à la deuxième plus grande valeur propre. On peut montrer que le plan défini par les axes Δ_1 et Δ_2 est le sous-espace de dimension 2 qui porte l'inertie maximum.

On peut rechercher de nouveaux axes en suivant la même procédure. Les nouveaux axes sont tous vecteurs propres de $\mathbf{\Sigma}$, et ils correspondent aux valeurs propres ordonnées. La matrice de variance-covariance $\mathbf{\Sigma}$ étant une matrice symétrique réelle, elle possède p vecteurs propres réels, formant une base orthogonale de $\mathbb{R}^p$:

$$\left\{ \begin{array}{l} \Delta_1 \perp \Delta_2 \ldots \perp \Delta_p \ , \\ a_1 \perp a_2 \perp \ldots \perp a_p \ , \\ \lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \ , \\ I_{\Delta_1^\star} \geqslant I_{\Delta_2^\star} \geqslant \ldots \geqslant I_{\Delta_p^\star} \ . \end{array} \right.$$

On passera de la base orthogonale initiale des variables centrées à la nouvelle base orthogonale des vecteurs propres de $\mathbf{\Sigma}$.

Définition 13.1 — *On appelle les nouveaux axes* ***axes principaux****.*

13.3 Contributions des axes à l'inertie totale

En utilisant le théorème de Huygens, on peut décomposer l'inertie totale du nuage des individus :

$$\begin{array}{rcl} I_G & = & I_{\Delta_1^\star} + I_{\Delta_2^\star} + \cdots + I_{\Delta_p^\star} \\ & = & \lambda_1 + \lambda_2 + \cdots + \lambda_p \ . \end{array}$$

Définition 13.2 — *La* ***contribution absolue*** *de l'axe Δ_k à l'inertie totale du nuage des individus est égale à*

$$\mathrm{ca}(\Delta_k \,/\, I_G) \ = \ \lambda_k \ ,$$

qui est la valeur propre qui lui est associée.

Définition 13.3 — *Sa* ***contribution relative*** *est égale à*

$$\mathrm{cr}(\Delta_k \,/\, I_G) \ = \ \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \ .$$

Définition 13.4 — *On emploie souvent l'expression «* ***pourcentage d'inertie expliquée*** *par Δ_k ».*

On peut étendre ces définitions à tous les sous-espaces engendrés par les nouveaux axes. Ainsi, le pourcentage d'inertie expliqué par le plan engendré par les deux premiers axes Δ_1 et Δ_2 est égal à

$$\mathrm{cr}(\Delta_1 \oplus \Delta_2 \ / \ I_G) \ = \ \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \ .$$

Ces pourcentages d'inertie sont des indicateurs qui rendent compte de la variabilité du nuage des individus expliquée par ces sous-espaces. Si les dernières valeurs propres ont des valeurs faibles, on pourra négliger la variabilité qu'expliquent les axes correspondants.

On se contente souvent de faire des représentations du nuage des individus dans un sous-espace engendré par les d premiers axes si ce sous-espace explique un pourcentage d'inertie proche de 1. On peut ainsi réduire l'analyse à un sous-espace de dimension $d < p$.

13.4 Représentation des individus dans les nouveaux axes

Pour faire les représentations des individus dans les plans définis par les nouveaux axes, il suffit de calculer les coordonnées des individus dans les nouveaux axes. Pour obtenir y_{ik}, coordonnée de l'unité i sur l'axe Δ_k, on projette orhtogonalement le vecteur $\overrightarrow{Gu_i}$ sur cet axe et on obtient

$$\begin{aligned} y &= \langle \overrightarrow{Gu_i}, \overrightarrow{a_k} \rangle \\ &= a_k^t \, \boldsymbol{U_{ci}} \end{aligned}$$

et

$$\boldsymbol{Y_i} = \boldsymbol{A}^t \, \boldsymbol{U_{ci}} \ ,$$

où $\boldsymbol{Y_i}$ est le vecteur des coordonnées de l'unité u_i et $\boldsymbol{A}$ est la matrice du changement de base — $\boldsymbol{A}$, matrice des vecteurs propres de norme 1, est orthogonale [1].

Remarque — L'orientation des axes est complètement arbitraire et peut différer d'un logiciel à l'autre. Le signe des coordonnées des individus sur un axe n'a donc pas de signification. En revanche, la comparaison des signes peut s'interpréter. Si deux individus u_i et $u_{i'}$ ont, sur un axe Δ, le premier une coordonnées positive et le second une coordonnée négative, cela signifie qu'ils s'opposent sur cet axe.

13.4.1 Qualité de la représentation des individus

Lorsque des points projections des individus sont éloignés sur un axe (ou sur un plan), on peut assurer que les points représentant ces individus sont éloignés dans l'espace. En revanche, deux individus dont les projections sont proches peuvent ne pas être proches dans l'espace.

Pour interpréter correctement la proximité des projections de deux indiviidus sur un plan, il faut donc s'assurer que ces individus sont bien représentés dans le plan. Pour que l'individu u_i soit bien représenté sur un axe (ou un plan, ou un sous-espace), il faut que l'angle entre le vecteur $\overrightarrow{Du_i}$ et l'axe (resp. le plan, le sous-espace) soit petit. On calcule donc le cosinus de cet angle, ou plutôt le carré de ce cosinus. En effet, en utilisant le théorème de Pythagore, on peut montrer que le carré du cosinus de l'angle d'un vecteur avec un plan engendré par deux vecteurs orthogonaux, est égal à la somme des carrés des cosinus des

1. Son inverse est égale à sa transposée.

angles du vecteur avec chacun des deux vecteurs qui engendrent le plan. Cette propriété se généralise à l'angle d'un vecteur avec un sous-espace de dimension k quelconque.

Si le carré du cosinus de l'angle entre $\overrightarrow{Gu_i}$ et l'axe (resp. le plan, le sous-espace) est proche de 1, alors on pourra dire que l'individu u_i est bien représenté par sa projection sur l'axe (resp. le plan, le sous-espace). Et si deux individus sont bien représentés en projection sur un axe (resp. le plan, le sous-espace) et s'ils ont des projections proches, alors on pourra dire que ces deux individus sont proches dans l'espace. Le carré du cosinus de l'angle α_{ik} entre $\overrightarrow{Gu_i}$ et un axe Δ_k de vecteur directeur unitaire a_k est égal à

$$\begin{aligned}\cos^2(\alpha_{ik}) &= \frac{\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k} \rangle^2}{\|\overrightarrow{Gu_i}\|^2} \\ &= \frac{a_k^t \, \boldsymbol{U_{ci}} \, \boldsymbol{U_{ci}}^t \, a_k}{\boldsymbol{U_{ci}}^t \, \boldsymbol{U_{ci}}} \\ &= \frac{\left[\sum_{j=1}^p (x_{ij} - x_{i\bullet}) \, a_{kj} \right]^2}{\left[\sum_{j=1}^p (x_{ij} - x_{i\bullet}) \right]^2} \; .\end{aligned}$$

En utilisant le théorème de Pythagore, on peut calculer le carré du cosinus de l'angle $\alpha_{ikk'}$ entre $\overrightarrow{Gu_i}$ et le plan engendré par deux axes $\Delta_k \oplus \Delta_{k'}$:

$$\cos^2(\alpha_{ikk'}) \; = \; \cos^2(\alpha_{ik}) + \cos^2(\alpha_{ik'}) \; .$$

Si, après l'étude des pourcentages d'inertie expliquée par les sous-espaces successifs engendrés par les nouveaux axes, on a décidé de ne retenir qu'un sous-espace de dimension $d < p$, on pourra calculer la qualité de la représentation d'un individu u_i en calculant le carré du cosinus de l'angle de $\overrightarrow{Gu_i}$ avec ce sous-espace.

Remarque — Si un individu est très proche du centre de gravité dans l'espace, c.-à-d. si $\|\overrightarrow{Gu_i}\|^2$ est très petit, le point représentant cet individu sur un axe (un plan, un sous-espace) sera bien représenté.

13.4.2 Interprétation des nouveaux axes en fonction des individus

Lorsqu'on calcule l'inertie $I_{\Delta_k^\star}$ portée par l'axe Δ_k, on peut voir quelle est la part de cette inertie due à un individu u_i particulier.

$I_{\Delta_k^\star}$ étant égale à $\dfrac{1}{n} \displaystyle\sum_{i=1}^n d^2(h_{\Delta_{ki}}, G)$, la contribution absolue de u_i à cette inertie est égale à

$$\mathrm{ca}(u_i \,/\, \Delta_k) \; = \; \frac{1}{n} d^2 \left(h_{\Delta_{ki}}, G \right) \; ,$$

puisque tous les individus ont le même poids. Un individu contribuera d'autant plus à la confection d'un axe que sa projection sur cet axe sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribuera faiblement à l'inertie portée par cet axe. On se sert de ces contributions pour interpréter les nouveaux axes de l'ACP en fonction des individus.

On peut aussi, pour un individu particulier u_i, donner sa contribution relative à l'inertie portée par cet axe :

$$\begin{aligned} \mathrm{cr}(\Delta_k \ / \ I_G) &= \frac{\frac{1}{n}\, d^2(h_{\Delta_{ki}}, G)}{I_{\Delta_k^\star}} \\ &= \frac{\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k} \rangle^2}{\lambda_k} \\ &= \frac{a_k^t \, \boldsymbol{U_{ci}} \, \boldsymbol{U_{ci}}^t \, a_k}{\lambda_k} \ . \end{aligned}$$

L'examen de ces contributions permet d'interpréter les axes principaux avec les individus.

13.5 Représentation des variables

On peut envisager le problème de la représentation des variables de façon complètement symétrique de celui des individus. Les raisonnements se font dans $\mathbb{R}^n$ au lieu de $\mathbb{R}^p$. Mais dans l'ACP, au-delà de la symétrie formelle entre les individus et les variables, on peut utiliser la dissymétrie liée à la sémantique : les variables n'ont pas la même signification que les individus. On peut alors faire le raisonnement suivant : on a représenté les individus dans l'espace des anciennes variables, et on a fait un changement de base dans cet espace. Les nouveaux axes sont des combinaisons linéaires des anciens axes et peuvent donc être considérés comme de nouvelles variables combinaisons linéaires des anciennes. On appelle communément ces nouvelles variables **composantes principales**.

On note Z_1,Z_2, ..., Z_p les composantes principales, Z_k étant la nouvelle variable correspondant à l'axe Δ_k :

$$\begin{aligned} Z_k &= \sum_{j=1}^{p} a_{kj} V_{cj} \\ &= \boldsymbol{X_c} \, a_k \end{aligned}$$

et de façon générale

$$\begin{aligned} Z &= (Z_1 \ Z_2 \ \ldots \ Z_k \ \ldots \ Z_p) \\ &= \boldsymbol{X_c} \, \boldsymbol{A} \\ &= \boldsymbol{X_c} \, a_k \ . \end{aligned}$$

Il est alors intéressant de voir comment les anciennes variables sont liées aux nouvelles : pour ce faire, on calcule les corrélations des anciennes variables avec les nouvelles. La représentation des anciennes variables se fera en prenant comme coordonnées des anciennes variables leurs coefficients de corrélation avec les nouvelles variables. On obtient alors ce que l'on appelle communément le **cercle des corrélations** (fig. 13.1), dénomination qui vient du fait qu'un coefficient de corrélation variant entre -1 et 1, les représentations des variables de départ sont des points qui se trouvent à l'intérieur d'un cercle de rayon 1 si on fait la représentation sur un plan.

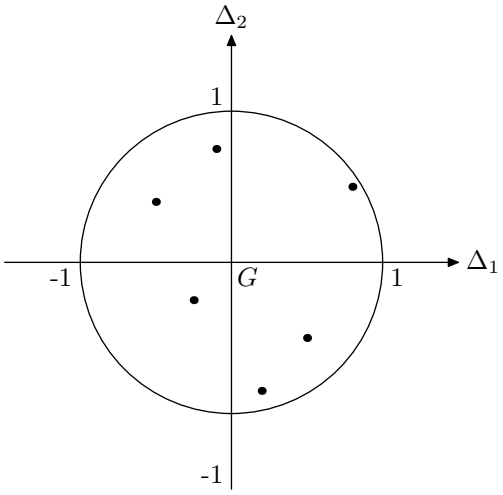


FIGURE 13.1 — Cercle des corrélations.

On peut montrer que

$$\begin{aligned} \mathbb{V}(Z_k) &= \frac{1}{n} a_k^t \, \boldsymbol{X_c^t} \, \boldsymbol{X_c} \, a_k \\ &= a_k^t \, \boldsymbol{\Sigma} \, a_k \\ &= \lambda_k \ , \end{aligned}$$

$$
\begin{aligned}
\mathbb{C}\mathrm{ov}\,(Z_k,\, V_{cj}) &= \frac{1}{n} a_k^t\, \boldsymbol{X_c^t}\, V_{cj} \\
&= a_k^t \boldsymbol{X_c^t}\, \boldsymbol{X_c} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&= \frac{1}{n} a_k^t\, \boldsymbol{\Sigma} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&= \lambda_k\, a_k^t \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&= \lambda_k\, a_{kj}\ .
\end{aligned}
$$

Enfin,

$$
\mathbb{C}\mathrm{orr}(Z_k,\, V_{cj}) \;=\; \sqrt{\lambda_k}\, \frac{a_{kj}}{\sqrt{\mathbb{V}\,(V_j)}}\ ,
$$

où a_{kj} est la j^{e} coordonnée du vecteur directeur unitaire a_k de Δ_k.

De façon générale, la matrice de variance-covariance des composantes principales est égale à $\boldsymbol{\Sigma_Z}$:

$$
\begin{aligned}
\boldsymbol{\Sigma_Z} &= \frac{1}{n} \boldsymbol{A^t}\, \boldsymbol{X_c^t}\, \boldsymbol{X_c}\, \boldsymbol{A} \\
&= \frac{1}{n} \boldsymbol{A^t}\, \boldsymbol{\Sigma}\, \boldsymbol{A} \\
&= \boldsymbol{\Lambda}\ ,
\end{aligned}
$$

où $\boldsymbol{\Lambda}$ est la matrice diagonale des valeurs propres de $\boldsymbol{\Sigma}$:

$$
\boldsymbol{\Lambda} \;=\; \begin{pmatrix} \lambda_1 & & (\mathbf{0}) \\ & \ddots & \\ (\mathbf{0}) & & \lambda_p \end{pmatrix}
$$

et la matrice des covariances entre les composantes principales et les anciennes variables vaut

$$\begin{aligned}\mathbb{C}\mathrm{ov}\,(\boldsymbol{Z}, \boldsymbol{V}) &= \frac{1}{n}\boldsymbol{X_c^t}\,\boldsymbol{X_c}\,\boldsymbol{A} \\ &= \boldsymbol{\Sigma}\,\boldsymbol{A} \\ &= \boldsymbol{A}\,\boldsymbol{\Lambda}\ .\end{aligned}$$

Si l'on remarque que la variance empirique d'une variable est égale au carré de la norme du vecteur qui la représente dans la géométrie euclidienne choisie et que le coefficient de corrélation empirique de deux variables est égal au produit scalaire des deux vecteurs qui les représentent, on pourra interpréter les angles des vecteurs comme des corrélations.

13.5.1 Interprétation des axes en fonction des anciennes variables

On peut interpréter les axes principaux en fonction des anciennes variables. Une ancienne variable V_j expliquera d'autant mieux un axe principal qu'elle sera fortement corrélée avec la composante principale correspondant à cet axe.

13.5.2 Qualité de la représentation des variables

Pour les mêmes raisons qui ont poussé à se préoccuper de la qualié de la représentation des individus, il faut se préoccuper de la qualité de la représentation des variables sur un axe, un plan ou un sous-espace. Une variable sera d'autant mieux représentée sur un axe que sa corrélation avec la composante correspondante sera, en valeur absolue, proche de 1. En effet, le coefficient de corrélation empirique entre une ancienne variable V_{cj} et une nouvelle variable Z_k n'est autre que le cosinus de l'angle du vecteur joignant l'origine du point v_j représentant la variable sur l'axe, avec cet axe.

Une variable sera bien représentée sur un plan si elle est proche du bord du cercle des corrélations, car cela signifie que le cosinus de l'angle du vecteur joignant l'origine au point représentant la variable avec le plan est, en valeur absolue, proche de 1.

Le même raisonnement demeure pour la représentation d'une variable sur un sous-espace.

13.5.3 Étude des liaisons entre variables

Sur le graphique du cercle des corrélations, on peut aussi interpréter les positions des anciennes variables les unes par rapport aux autres en terme de corrélations. Ainsi, soient deux points très proches l'un de l'autre et très proches, également, du cercle des corrélations : les variables correspondant à ces points sont bien représentées dans le plan et très corrélées positivement. Si deux variables sont proches du cercle, mais dans des positions symétriques par rapport à l'origine, elles seront très corrélées négativement.

Deux variables proches du cercle des corrélations et dont les vecteurs qui les joignent à l'origine forment un angle droit, ne seront pas corrélées entre elles.

Il faut, pour interpréter correctement ces graphiques des cercles des corrélations, se souvenir qu'un coefficient de corrélation est une mesure de liaison linéaire entre deux variables, et qu'il peut arriver que deux variables très fortement liées aient un coefficient de corrélation nul ou très faible, si leur liaison n'est pas linéaire.

13.6 Analyse en composantes principales normée

Dans les paragraphes précédents, nous avons étudié l'ACP simple, pour laquelle :
- tous les individus ont le même poids dans l'analyse ;
- toutes les variables sont traitées de façon symétrique (on leur fait jouer le même rôle) ;
- les nouveaux axes sont issus de la matrice de variance-covariance empirique des variables.

Cela pose parfois des problèmes. Le premier reproche fait par des praticiens est que, si les anciennes variables sont hétérogènes, comme par exemple des poids, des tailles et des âges, quel sens peut-on donner aux composantes principales qui sont alors des combinaisons linéaires de variables hétérogènes ? Le deuxième reproche est que, si on change d'unité sur ces variables, on peut changer complètement les résultats de l'ACP. Le dernier reproche vient du fait qu'une variable contribuera d'autant plus à la confection des premiers axes que sa variance sera forte.

Pour échapper à tous ces problèmes, on cherchera à normaliser les variables et à travailler sur des variables sans dimension. Il y a plusieurs façons de normaliser les variables, mais la plus couramment utilisée est celle qui consiste à diviser les valeurs des variables par leur écart-type, c.-à-d. que l'on travaille sur des variables centrées et réduites.

Cela revient à faire la même analyse que pour l'ACP simple, mais à choisir une autre distance euclidienne entre les individus que la distance euclidienne classique. La distance choisie est alors

$$d^2(u_i, u_{i'}) \;=\; \sum_{j=1}^{p} \frac{1}{\sigma_j^2}(x_{ij} - x_{i'j})^2$$

Cette nouvelle distance ne traite plus les variables de façon symétrique, mais elle permet de faire jouer un rôle plus équitable à chacune d'entre elles.

Si on reprend tous les calculs de l'ACP simple, mais en remplaçant les variables de départ par les variables centrées réduites, on voit que ce n'est plus la matrice de variance-covariance, mais la matrice de corrélation $\boldsymbol{R}$ qui intervient dans la recherche de nouveaux axes. Les particularités de l'ACP normée par rapport à l'ACP simple proviennent du fait que la matrice de corrélation $\boldsymbol{R}$ n'a que des 1 sur sa diagonale principale. Cela entraîne que sa trace est toujours égale à p. Or on a vu que la trace de la matrice est égale à l'inertie totale du nuage calculée avec la distance euclidienne que l'on a choisie. L'inertie totale du nuage des individus dans $\mathbb{R}^p$ est donc toujours égale à p dans toute ACP normée.

Cette particularité donne une règle supplémentaire pour choisir le nombre d'axes que l'on va garder pour les interprétations, fondée sur le raisonnement suivant :
- on a p valeurs propres dont la somme vaut p ;
- on peut ne considérer comme significatives que les valeurs propres dont la valeur est supérieure à 1, puisque la valeur moyenne moyenne des valeurs propres vaut 1 et leur somme p.

C'est bien sûr une règle empirique, mais elle peut servir de guide pour le choix de la dimension du sous-espace que l'on veut garder.

Une autre particularité de l'ACP normée est que la représentation des variables avec les cercles de corrélation correspond exactement à la représentation des variables dans $\mathbb{R}^n$ que l'on aurait construite si l'on avait adopté la même démarche que celle qui a servi pour la représentation des individus dans $\mathbb{R}^p$.

13.7 Individus et variables supplémentaires

Il arrive que l'on veuille faire apparaître dans les représentations graphiques certains individus sans qu'ils interviennent dans la détermination des axes. Cela peut être le cas de nouveaux individus que l'on veut simplement positionner par rapport aux autres sans que les positions de ceux-ci soient influencées par les nouveaux. On dit d'eux qu'ils sont des **individus supplémentaires**.

Il en est de même pour les variables. On peut, par exemple, vouloir représenter une variable qui dépend de façon synthétique des p variables choisies pour faire l'ACP, afin de mieux comprendre comment cette variable est liée aux anciennes, mais on ne souhaite pas qu'elle intervienne dans la confection des axes car ses liaisons avec les p variables de départ fausseraient la représentation si elle faisait partie intégrante de l'ACP. Elles sont appelées **variables supplémantaires**.

Pour représenter un individu suplémentaire, il suffit d'exprimer les coordonnées de cet individu dans la nouvelle base des axes principaux. Pour une variable supplémantaire, il suffit de calculer ses coefficients de corrélation empiriques avec les composantes principales. La plupart des logiciels proposent des options permettant de le faire.

Sixième partie

ANOVA

14

Introduction

Dans le cas des expériences en champ, on entend classiquement par **bloc** un ensemble de parcelles voisines et très semblables les unes aux autres, quant aux conditions de croissance et de développement de la végétation.

Ces blocs sont dits **complets** quand tous les objets mis en expérience sont présents dans chacun d'eux, le nombre de parcelles étant alors au moins égal au nombre d'objets.

Ces blocs sont dits **équilibrés** lorsque, $\forall i,j$ — indices de ligne et de colonne —,

$$n_{ij} \;=\; \frac{n_{i.} \times n_{.j}}{n} \;.$$

Parmi les plans d'expérience équilibrés, on trouve les cas de figure où :

- chaque n_{ij} vaut 1 — une seule mesure est faite pour chaque couple de niveaux : il n'y a pas de répétition ;
- tous les couples sont répétés un même (et unique) nombre de fois : $\forall i, j,\ n_{ij} = 4$ par exemple ;
- les blocs sont du type :

		Facteur A			
		1	2	3	4
	1	1	1	1	2
Facteur B	2	1	1	1	2
	3	3	3	3	6

  c.-à-d. que l'on a en quelque sorte répété deux fois la 4$^{\text{e}}$ colonne et trois fois la 3$^{\text{e}}$ ligne.

La répartition des objets au sein des différents blocs se fait normalement de façon complètement aléatoire et indépendamment d'un bloc à l'autre, d'où la notion de **blocs aléatoires complets**, aussi appelés **blocs randomisés**.

Enfin, en présence de plusieurs facteurs, nous parlons d'**expériences factorielles (complètes)** lorsque chacune des modalités d'un facteur est associée à chacune des modalités de l'autre ou des autres facteurs.

Concernant les facteurs étudiés, s'ils sont tous fixes, nous parlons de **modèle d'analyse de la variance (ANOVA)** ; si l'un de ces facteurs est aléatoire, alors nous parlons de **modèle de composantes de la variance**. Un modèle comportant des effets fixe(s) et aléatoire(s) est appelé **modèle mixte**.

15

Sans effet aléatoire

15.1 Un critère de classification (*One-way*)

Soient n observations réparties en g groupes, et y_{ij} l'observation concernant la i^{e} observation du j^{e} groupe. Chaque groupe contient n_j observations. Le modèle s'écrit

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij} \ .$$

où $j = 1, \ldots, g$ et $i = 1, \ldots, n_j$. Dans ce modèle, α_j représente l'effet du groupe j. Par ailleurs, nous supposons que les ϵ_{ij} sont supposés i.i.d. de loi $\mathcal{N}(0, \sigma^2)$, et que $\sum_{j=1}^{g} \alpha_j = 0$, ceci afin d'éviter la surparamétrisation du modèle.

Facteur A

1	2	$\cdots$	g
(1) $\mu + \alpha_1$	(1) $\mu + \alpha_2$	$\cdots$	(1) $\mu + \alpha_g$
$\vdots$	$\vdots$	$\ldots$	$\vdots$
(n_1) $\mu + \alpha_1$	(n_2) $\mu + \alpha_2$	$\cdots$	(n_g) $\mu + \alpha_g$

FIGURE 15.1 — Plan d'expérience à un critère de classification.

On note respectivement

$$\overline{y}_{.j} \ = \ \frac{1}{n_j} \sum_{i=1}^{c} y_{ij} \qquad \text{et} \qquad \overline{y}_{..} \ = \ \frac{1}{n} \sum_{j=1}^{g} \sum_{i=1}^{n_j} y_{ij}$$

la moyenne du groupe j et la moyenne sur l'ensemble de l'échantillon.

On note également

$$\text{SCA} \ = \ \sum_{j=1}^{g} \sum_{i=1}^{n_j} (\overline{y}_{.j} - \overline{y}_{..})^2 \ , \qquad \text{SCE} \ = \ \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_{.j})^2 \ , \qquad \text{SCT} \ = \ \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_{..})^2 \ .$$

L'ANOVA peut alors se résumer au tableau 15.1.

TABLE 15.1 — Table d'ANOVA à un critère de classification.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCE
Erreur	$n-g$	SCE	MCE = SCE / $(n-g)$	
Total	$n-1$	SCT		

La statistique employée pour tester l'effet du facteur A — c.-à-d. pour tester l'hypothèse nulle d'égalité des moyennes [1] — est

$$F_{\mathrm{A}} \;=\; \frac{\mathrm{MCA}}{\mathrm{MCE}} \;\rightsquigarrow\; F(g-1\,,\,n-g)\;.$$

Pour un seuil α fixé, la table donne u_α tel que

$$\mathbb{P}\big[F_{\mathrm{A}}\big(g-1\,,\,(n-g)\big) > u_\alpha\big] \;=\; \alpha\;.$$

Le test au seuil α s'écrit :

$$\text{Rejet de } H_0 \;\Leftrightarrow\; F_{\mathrm{A}} > u_\alpha\;.$$

Estimation de la variance Sous l'hypothèse d'homogénéité des variances, le meilleur estimateur non biaisé de σ^2 est

$$\begin{aligned} s^2 &= \frac{\mathrm{SCE}}{n-g} \\ &= \frac{1}{n-g}\Big[(n_1-1)s_1^2+(n_2-1)s_2^2+\cdots+(n_g-1)s_g^2\Big] \end{aligned}$$

Exemple — Voici la consommation de ménages, enregistrée dans différentes régions, suite à la diffusion de quatre spots de publicité. Le but de l'étude est de comparer l'impact des quatre spots. Le tableau originel est donné en annexe (cf. p. 179).

Nous obtenons :

```
> aov(Conso~factor(Pub), consomenage)

> Terms:
                 factor(Pub) Residuals
 Sum of Squares     4585.68  56187.44
Deg. of Freedom           3       116

Residual standard error: 22.00851
Estimated effects are balanced
```

1. On la note souvent H_0 : $\alpha_1 = \cdots = \alpha_g = 0$.

```
> summary(aov(Conso~factor(Pub), consomenage))

>            Df Sum of Sq  Mean Sq F Value      Pr(F)
factor(Pub)   3   4585.68 1528.560 3.15574 0.02749209
  Residuals 116  56187.44  484.375
```

Au seuil de 5 %, nous en déduisons que les spots publicitaires ont un effet significatif sur la consommation (en effet, 0,027 = 2,7 % < 5 %). Nous obtenons par ailleurs une estimation de la variance égale à 484,37.

15.2 Comparaisons multiples

Lorsque l'hypothèse nulle est rejetée, une question s'ensuit : « quels sont les groupes dont les moyennes diffèrent ? ». Pour répondre à cette question, il faut utiliser l'une des procédures de comparaison multiple établies ; parmi celles-ci, citons :

- la procédure LSD (*Least significant difference*) ;
- la procédure LSD de Fisher ;
- la procédure de Bonferroni ;
- la procédure de Sidak ;
- la procédure de Tukey ;
- la procédure dite GT2 ;
- la procédure de Gabriel ;
- la procédure Dunkan (S-N-K et REGW) ;
- la procédure par contrastes.

Exemple — La comparaison multiple des spots publicitaires, sous S-Plus, avec la procédure de Sidak, est réalisée ainsi :

```
> temp_aov(Conso~factor(Pub), consomenage2)
> multicomp(temp, focus="factor(Pub)", bounds="lower", control=1, plot=T,
> method="sidak")

> 95 % simultaneous confidence bounds for specified
linear combinations, by the Sidak method

critical point: 2.4195000000000002
response variable: Conso

bounds excluding 0 are flagged by '****'

    Estimate Std.Error Lower Bound
1-2    -2.41      5.68      -16.20
1-3     2.29      5.68      -11.50
1-4   -13.80      5.68      -27.50
2-3     4.70      5.68       -9.05
2-4   -11.40      5.68      -25.10
3-4   -16.10      5.68      -29.80
```

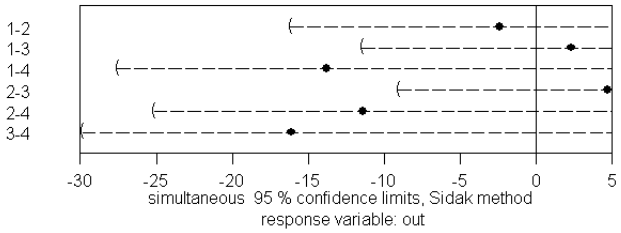


FIGURE 15.2 — Comparaison multiple suivant la procédure de Sidak.

et le graphique obtenu est celui de la figure 15.2.

15.3 Respect de l'hypothèse d'homogénéité des variances

Parmi les tests d'homogénéité des variances, citons :

— le test de Bartlett ;

— le test de Hartley ;

— le test de Cochran ;

Ces tests sont très sensibles à l'hypothèse de normalité. Aussi, d'autres procédures — non paramétriques celles-ci — existent, parmi lesquelles :

— la méthode de Levene ;

— la méthode par *jackknife* ;

— la méthode basée sur les rangs.

15.4 Deux critères de classification (*Two-way*)

Une seconde variable qualitative est introduite dans le plan d'expérience. Distinguons alors :

— le cas où cette variable représente une caractéristique de chaque cellule expérimentale, censée tenir compte d'un manque d'homogénéité entre les diférentes cellules expérimentales : nous parlons alors de **blocs aléatoires** (*randomized block design*) ;

— le cas où cette variable supplémentaire représente un second type de traitement ou facteur : nous parlons alors de **modèle à deux facteurs** (*two factors model*).

15.4.1 Blocs aléatoires (*Randomized block design*)

Les unités expérimentales sont divisées en b blocs de telle sorte que les unités d'un même bloc soient relativement homogènes. Chaque bloc contient g unités expérimentales. Chaque unité expérimentale d'un bloc est affectée aléatoirement à l'un des g groupes. Le but de ce type d'expérimentation est de retirer de la variance intra-groupe la variabilité attribuable aux b blocs.

Le modèle s'écrit

$$y_{ij} = \mu + \alpha_j + \beta_i + \epsilon_{ij}$$

pour $i = 1, \dots, b$ et $j = 1, \dots, g$. On suppose que :

$$\sum_{j=1}^{g} \alpha_j \ = \ 0 \ , \qquad \sum_{i=1}^{b} \beta_i \ = \ 0 \ , \qquad \epsilon_{ij} \ \rightsquigarrow \ \mathcal{N}(0, \sigma^2) \ .$$

		Facteur A			
		1	2	$\dots$	g
Blocs	1	$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_2 + \beta_1$	$\dots$	$\mu + \alpha_g + \beta_1$
	2	$\mu + \alpha_1 + \beta_2$	$\mu + \alpha_2 + \beta_2$	$\dots$	$\mu + \alpha_g + \beta_2$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	b	$\mu + \alpha_1 + \beta_b$	$\mu + \alpha_2 + \beta_b$	$\dots$	$\mu + \alpha_g + \beta_b$

FIGURE 15.3 — Plan d'expérience à deux critères de classification.

On note

$$\overline{y}_{\cdot j} \ = \ \frac{1}{b} \sum_{i=1}^{b} y_{ij} \ , \qquad \overline{y}_{i \cdot} \ = \ \frac{1}{g} \sum_{j=1}^{g} y_{ij} \ , \qquad \overline{y}_{\cdot \cdot} \ = \ \frac{1}{bg} \sum_{i=1}^{b} \sum_{j=1}^{g} y_{ij} \ ,$$

$$\text{SCA} \ = \ b \sum_{j=1}^{g} (\overline{y}_{\cdot j \cdot} - \overline{y}_{\cdot \cdot})^2 \ , \qquad \text{SCB} \ = \ g \sum_{i=1}^{b} (\overline{y}_{i \cdot} - \overline{y}_{\cdot \cdot})^2 \ ,$$

$$\text{SCE} \ = \ \sum_{i=1}^{b} \sum_{j=1}^{g} (y_{ij} - \overline{y}_{\cdot j} - \overline{y}_{i \cdot} + \overline{y}_{\cdot \cdot})^2 \ , \qquad \text{SCT} \ = \ \sum_{i=1}^{b} \sum_{j=1}^{g} (y_{ij} - \overline{y}_{\cdot \cdot})^2 \ .$$

La table de l'ANOVA est donnée par 15.2.

TABLE 15.2 — Table de l'ANOVA à deux critères de classification.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g - 1$	SCA	MCA = SCA / $(g-1)$	MCA / MCE
Blocs	$b - 1$	SCB	MCB = SCB / $(b-1)$	MCB / MCE
Erreur	$(g-1)(b-1)$	SCE	MCE = SCE / $(g-1)(b-1)$	
Total	$gb - 1$	SCT		

La statistique employée pour tester l'effet du facteur A (effet groupe) est

$$F_{\mathrm{A}} \;=\; \frac{\mathrm{MCA}}{\mathrm{MCE}} \;\rightsquigarrow\; F\big(g-1\,,\,(g-1)(b-1)\big).$$

La statistique employée pour tester l'effet « Blocs » est

$$F_{\mathrm{B}} \;=\; \frac{\mathrm{MCB}}{\mathrm{MCE}} \;\rightsquigarrow\; F\big(b-1\,,\,(g-1)(b-1)\big).$$

Exemple — Une partie des données permet d'illustrer l'ANOVA en blocs aléatoires : nous nous intéressons uniquement aux ménages ne contenant qu'une seule personne. L'ANOVA porte donc sur un tableau 4×5, puisqu'il y a un seul ménage (d'une unique personne) par région (et 5 régions), ayant vu les 4 spots publicitaires.

Le résultat est le suivant :

```
> consomenage2_consomenage[consomenage$taille==1,]
> summary(aov(Conso~factor(Pub)+factor(Region), consomenage2))

>              Df Sum of Sq  Mean Sq  F Value      Pr(F)
   factor(Pub)  3  332.8289 110.9430 2.611035 0.09954666
factor(Region)  4  663.6414 165.9104 3.904688 0.02955505
     Residuals 12  509.8804  42.4900
```

L'hypothèse nulle d'égalité des 4 moyennes (correspondant aux spots) peut être rejetée au seuil de 10 % (mais pas au seuil de 5 %), et celle d'égalité des 5 moyennes (correspondant aux régions) au seuil de 3 %.

Une comparaison multiple des 4 moyennes correspondant aux publicités, par la méthode LSD, conclut à une moyenne plus importante pour la publicité 4 au seuil de 5 %. Par contre, la procédure de Tukey conclue à l'absence de différence significative entre ces 4 moyennes.

Une comparaison multiple peut également être conduite pour les 5 moyennes par région.

15.4.2 Deux facteurs (*Two-way factorial design*)

Nous nous intéressons à la relation entre une variable dépendante Y et deux variables qualitatives. Nous supposons que le premier facteur présente g niveaux et le second b. Nous supposons également que c observations sont aléatoirement tirées dans chaque cellule, ce qui donne au total $g \times b \times c$ observations.

Le modèle s'écrit :

$$y_{ij} = \mu + \alpha_j + \beta_i + (\alpha\beta)_{ij} + \epsilon_{ijk}\ ,$$

pour $i = 1, \ldots, b$, $j = 1, \ldots, g$, $k = 1, \ldots, c$. On suppose que

$$\sum_{j=1}^{g} \alpha_j \;=\; \sum_{g=1}^{b} \beta_i \;=\; \sum_{j=1}^{g} (\alpha\beta)_{ij} \;=\; \sum_{i=1}^{b} (\alpha\beta)_{ij} \;=\; 0\ , \qquad \epsilon_{ijk} \;\rightsquigarrow\; \mathcal{N}(0\,,\,\sigma^2)\ .$$

		Facteur A			
		1	2	…	g
Facteur B	1	(1) $\mu+\alpha_1+\beta_1$ ⋮ (c) $\mu+\alpha_1+\beta_1$	(1) $\mu+\alpha_2+\beta_1$ ⋮ (c) $\mu+\alpha_2+\beta_1$	… … …	(1) $\mu+\alpha_g+\beta_1$ ⋮ (c) $\mu+\alpha_g+\beta_1$
	2	(1) $\mu+\alpha_1+\beta_2$ ⋮ (c) $\mu+\alpha_1+\beta_2$	(1) $\mu+\alpha_2+\beta_2$ ⋮ (c) $\mu+\alpha_2+\beta_2$	… … …	(1) $\mu+\alpha_g+\beta_2$ ⋮ (c) $\mu+\alpha_g+\beta_2$
	⋮	⋮	⋮	⋮	⋮
	b	(1) $\mu+\alpha_1+\beta_b$ ⋮ (c) $\mu+\alpha_1+\beta_b$	(1) $\mu+\alpha_2+\beta_b$ ⋮ (c) $\mu+\alpha_2+\beta_b$	… … …	(1) $\mu+\alpha_g+\beta_b$ ⋮ (c) $\mu+\alpha_g+\beta_b$

FIGURE 15.4 — Plan d'expérience à deux facteurs.

On note

$$\text{SCA} = bc\sum_{j=1}^{g}(\overline{y}_{.j.}-\overline{y}_{...})^2\,, \qquad \text{SCB} = gc\sum_{i=1}^{b}(\overline{y}_{i..}-\overline{y}_{...})^2\,, \qquad \text{SCAB} = c\sum_{i=1}^{b}\sum_{j=1}^{g}b(\overline{y}_{ij.}-\overline{y}_{.j.}-\overline{y}_{i..}+\overline{y}_{...})^2\,,$$

$$\text{SCE} = \sum_{i=1}^{b}\sum_{j=1}^{g}\sum_{k=1}^{c}(y_{ijk}-\overline{y}_{ij.})^2 \qquad \text{SCT} = \sum_{i=1}^{b}\sum_{j=1}^{g}\sum_{k=1}^{c}(y_{ijk}-\overline{y}_{...})^2\,.$$

La table de l'ANOVA est écrite ci-dessous (cf. tab. 15.3).

TABLE 15.3 — Table de l'ANOVA à deux facteurs.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCE
Facteur B	$b-1$	SCB	MCB = SCB / $(b-1)$	MCB / MCE
Inter. AB	$(g-1)(b-1)$	SCAB	MCAB = SCAB / $(g-1)(b-1)$	MCAB / MCE
Erreur	$bg(c-1)$	SCE	MCE = SCE / $bg(c-1)$	
Total	$bgc-1$	SCT		

La statistique employée pour tester l'effet du facteur A (effet groupe) est

$$F_{\text{A}} = \frac{\text{MCA}}{\text{MCE}} \rightsquigarrow F\big(g-1\,,\,gb(c-1)\big)\,.$$

La statistique employée pour tester l'effet du facteur B (effet traitement) est

$$F_{\text{B}} = \frac{\text{MCB}}{\text{MCE}} \rightsquigarrow F\big(b-1\,,\,gb(c-1)\big)\,.$$

La statistique employée pour tester l'interaction AB est

$$F_{\text{AB}} \;=\; \frac{\text{MCAB}}{\text{SCE}} \;\rightsquigarrow\; F\big((g-1)(b-1)\,,\,gb(c-1)\big)\;.$$

Exemple — Voici l'ANOVA réalisée sous S-Plus :

```
> summary(aov(out~factor(Pub)+factor(Region)+factor(Pub):factor(Region), consomenage))

>                           Df Sum of Sq  Mean Sq  F Value      Pr(F)
               factor(Pub)   3   4585.68 1528.560 3.606625 0.01600132
            factor(Region)   4   4867.51 1216.878 2.871213 0.02680049
factor(Pub):factor(Region)  12   8937.92  744.826 1.757412 0.06584027
                 Residuals 100  42382.02  423.820
```

Nous en déduisons que les deux effets principaux sont significatifs ; par contre, leur interaction ne l'est pas.

15.4.3 Emboîtement à un niveau (*Two-way nested design*)

Contrairement au paragraphe précédent, il n'est plus ici possible d'assigner aléatoirement les g niveaux du facteur A à chaque bloc. Il s'avère obligatoire de restreindre certains niveaux du facteur A à certains blocs en particulier. Nous parlons ici de **plan d'expérience avec emboîtement** ou encore de **plan d'expérience hiérarchique**.

		Facteur A				
		1	2	3	…	g
Blocs	1	(1) $\mu+\alpha_1+\beta_1$ ⋮ (c) $\mu+\alpha_1+\beta_1$				(1) $\mu+\alpha_g+\beta_1$ ⋮ (c) $\mu+\alpha_g+\beta_1$
	2		(1) $\mu+\alpha_2+\beta_2$ ⋮ (c) $\mu+\alpha_2+\beta_2$			
	⋮	⋮	⋮	⋮	⋮	⋮
	b			(1) $\mu+\alpha_3+\beta_b$ … (c) $\mu+\alpha_3+\beta_b$		

FIGURE 15.5 — Plan d'expérience avec emboîtement.

Le modèle s'écrit :

$$y_{ijk} = \mu + \beta_i + \alpha_{ij} + \epsilon_{ijk}\;,$$

pour $i = 1,\dots,b$, $j = 1,\dots,g$, $k = 1,\dots,c$. On suppose que

$$\sum_{i=1}^{b}\beta_i \;=\; \sum_{i=1}^{b}\sum_{j=1}^{g}\alpha_{ij} \;=\; 0\;, \qquad \epsilon_{ijk} \;\rightsquigarrow\; \mathcal{N}(0,\sigma^2)\;.$$

On note :

$$\text{SCA} \;=\; bc\sum_{j=1}^{g}(\overline{y}_{.j.}-\overline{y}_{...})^2\;, \qquad \text{SCB(A)} \;=\; c\sum_{i=1}^{b}\sum_{j=1}^{g}(\overline{y}_{ij.}-\overline{y}_{.j.})^2\;,$$

$$\text{SCE} \;=\; \sum_{i=1}^{b}\sum_{j=1}^{g}\sum_{k=1}^{c}(\overline{y}_{ijk}-\overline{y}_{ij.})^2 \qquad \text{SCT} \;=\; \sum_{i=1}^{b}\sum_{j=1}^{g}\sum_{k=1}^{c}(\overline{y}_{ijk}-\overline{y}_{...})^2\;.$$

La table de l'ANOVA est donnée ci-dessous.

TABLE 15.4 — Table de l'ANOVA avec emboîtement.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCB(A)
Facteur B emboîté dans A	$g(b-1)$	SCB(A)	MCB(A) = SCB(A) / $g(b-1)$	MCB(A) / MCE
Erreur	$gb(c-1)$	SCE	MCE = SCE / $gb(c-1)$	
Total	$gbc-1$	SCT		

La statistique de Fisher testant l'effet du traitement A vaut

$$F_{\text{A}} = \frac{\text{MCA}}{\text{MCB(A)}} \;\rightsquigarrow\; F\big(g-1\,,\,g(b-1)\big)\;.$$

La statistique de Fisher testant l'effet du traitement B vaut

$$F_{\text{B}} = \frac{\text{MCB(A)}}{\text{MCE}} \;\rightsquigarrow\; F\big(g(b-1)\,,\,gb(c-1)\big)\;.$$

Remarque — L'effet du facteur A est testé en comparant la variance inter-groupe à celle due aux sous-groupes représentés par le facteur B. En effet, on cherche à tester isolément l'effet du facteur A proprement dit : dans ce cas, l'« erreur » due au sous-facteur (B) doit être incorporée au terme d'erreur du rapport de Fisher. Si tel n'était pas le cas, on testerait l'effet du facteur A par rapport aux résidus après avoir ôté l'effet de ce facteur *et* celui du sous-facteur.

Exemple — Considérons des moustiques placés dans des cages, sur chacun desquels sont réalisées deux mesures indépendantes. L'emboîtement est « moustique dans cage ». Les données sont :

Cage 1				Cage 2				Cage 1			
1	2	3	4	1	2	3	4	1	2	3	4
58,5	77,8	84,0	70,1	69,8	56,0	50,7	63,8	56,6	77,8	69,9	62,1
59,5	80,9	83,6	68,3	69,8	54,5	49,3	65,8	57,5	79,2	69,2	64,5

Deux facteurs, l'un emboîté dans l'autre L'analyse par S-Plus donne :

```
> summary(aov(Valeur ~ factor(Cage) + Error(factor(Cage)/factor(Moust)), moustiques))
Error: factor(Cage)
             Df Sum of Sq  Mean Sq
factor(Cage)  2  665.6758 332.8379

Error: factor(Moust) %in% factor(Cage)
          Df Sum of Sq  Mean Sq F Value Pr(F)
Residuals  9  1720.677 191.1864

Error: Within
          Df Sum of Sq  Mean Sq F Value Pr(F)
Residuals 12     15.62 1.301667
```

Pour tester l'« effet cage » sur les mesures, nous écrivons :

```
> 332.8379/191.1864
[1] 1.740908

> 1-pf(1.740908,2,9)
[1] 0.229531
```

qui nous indique que cet effet n'est pas significatif.

Pour tester l'« effet moustique » sur les mesures, nous écrivons :

```
> 191.19/1.30
[1] 147.0692

> 1-pf(147.0692,9,12)
[1] 6.927803e-011
```

qui nous indique que cet effet est hautement significatif.

Concernant les composantes de la variance :

- la variance s^2 due à l'erreur (c.-à-d. entre les deux mesures réalisées sur un même moustique) vaut 1,30 ;
- la variance $s^2_{B(A)}$ entre les sous-groupes (moustiques) emboîtés dans les groupes (cages) vaut $(191{,}19 - 1{,}30)/2 = 94{,}94$;
- la variance s^2_A entre les groupes (cages) vaut $(332{,}84 - 191{,}19)/8 = 17{,}71$;
- la variance totale vaut $1{,}30 + 94{,}94 + 17{,}71 = 113{,}95$.

Nous pouvons donc affirmer que :

- s^2 représente $1{,}30/113{,}95 = 1{,}1$ % de la variance totale ;
- $s^2_{B(A)}$ représente $94{,}94/113{,}95 = 83{,}3$ % de la variance totale ;
- s^2_A représente $17{,}71/113{,}95 = 15{,}6$ % de la variance totale.

Remarquons que :

$$\begin{aligned} \mathbb{V}(Y) &= \sigma^2 + \sigma^2_{B(A)} + \sigma^2_A \ , \\ \mathbb{V}(Y \mid A) &= \sigma^2 + \sigma^2_{B(A)} \ , \\ \mathbb{V}(Y \mid A, B) &= \sigma^2 \ . \end{aligned}$$

Ignorance du facteur « Cages » Les données sont alors de la forme suivante :

1-1	1-2	1-3	1-4	2-1	2-2	2-3	2-4	3-1	3-2	3-3	3-4
58,5	77,8	84,0	70,1	69,8	56,0	50,7	63,8	56,6	77,8	69,9	62,1
59,5	80,9	83,6	68,3	69,8	54,5	49,3	65,8	57,5	79,2	69,2	64,5

En réécrivant le fichier de données comme indiqué en annexe, l'analyse donne :

```
> summary(aov(val ~ factor(Mesure), moustiques))
               Df Sum of Sq  Mean Sq  F Value         Pr(F)
factor(Mesure) 11  2386.353 216.9412 166.6642 2.328582e-011
     Residuals 12    15.620   1.3017
```

Ignorance du facteur « Moustiques » L'analyse donne :

```
> summary(aov(val ~ factor(cage), moustiques))
             Df Sum of Sq  Mean Sq  F Value      Pr(F)
factor(Cage)  2   665.676 332.8379 4.025575 0.03311979
   Residuals 21  1736.297  82.6808
```

15.4.4 Analyse de la covariance (*ANCOVA*)

Lorsqu'une variable intervenant en tant que facteur est quantitative, et non plus qualitative, nous parlons d'analyse de la covariance (ANCOVA). Chaque observation consiste en une paire $(y_{ij},\, z_{ij})$ où y_{ij}, $i = 1, \ldots, n_j$, $j = 1, \ldots, g$ est la i^{e} observation du groupe j portant sur la variable d'intérêt, et z_{ij} est la i^{e} observation du groupe j portant sur la covariable (ou variable explicative, ou variable indépendante). Le modèle s'écrit

$$y_{ij} \;=\; \gamma_{0j} + \gamma_1 z_{ij} + \epsilon_{ij}\ ,$$

où nous supposons que le paramètre de pente γ_1 est le même pour les g groupes. Nous obtenons

$$\hat{y}_{ij} \;=\; \hat{\gamma}_{0j} + \hat{\gamma}_1 z_{ij} + \epsilon_{ij}\ ,$$

et nous construisons la table d'ANOVA comme dans le cas le plus simple (cf. § 15.1) :

Table 15.5 — Table d'ANCOVA.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCE
Covariable	1	SCC	MCC = SCC	MCC / MCE
Erreur	$n-g-1$	SCE	MCE = SCE / $(n-g-1)$	
Total	$n-1$	SCT		

Exemple — Dans l'exemple des spots publicitaires, nous obtenons — en considérant maintenant la variable « Taille » comme une variable quantitative :

```
> summary(aov(Conso ~ factor(Pub) + Taille, consomenage2))
             Df Sum of Sq  Mean Sq  F Value         Pr(F)
factor(Pub)   3   4585.68  1528.56  11.5182 1.169389e-006
     Taille   1  40926.02 40926.02 308.3913 0.000000e+000
  Residuals 115  15261.43   132.71
```

15.5 Trois critères de classification (*Three-way*)

15.5.1 Trois facteurs (*Three-way factorial design*)

Nous sommes en présence de trois facteurs ; chaque cellule (i, j, k) est supposée contenir c observations. Le modèle s'écrit :

$$y_{ijkh} = \mu + \alpha_j + \beta_i + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{ik} + (\alpha\gamma)_{jk} + (\beta\alpha\gamma)_{ijk} + \epsilon_{ijkh} \ ,$$

pour $i = 1, \ldots, b$, $j = 1, \ldots, g$, $k = 1, \ldots, l$ et $h = 1, \ldots, c$.

On suppose que

$$\sum_{j=1}^{g} \alpha_j \ = \ \sum_{g=1}^{b} \beta_i \ = \ \sum_{k=1}^{l} \gamma_k \ = \ 0 \ ,$$

$$\sum_{j=1}^{g} (\alpha\beta)_{ij} \ = \ \sum_{i=1}^{b} (\alpha\beta)_{ij} \ = \ \sum_{j=1}^{g} (\alpha\gamma)_{jk} \ = \ \sum_{k=1}^{l} (\alpha\gamma)_{jk} \ = \ \sum_{i=1}^{b} (\beta\gamma)_{ik} \ = \ \sum_{k=1}^{l} (\beta\gamma)_{ik} \ = \ 0 \ ,$$

$$\sum_{i=1}^{b} (\alpha\beta\gamma)_{ijk} \ = \ \sum_{j=1}^{g} (\alpha\beta\gamma)_{ijk} \ = \ \sum_{k=1}^{l} (\alpha\beta\gamma)_{ijk} \ = \ 0$$

et que

$$\epsilon_{ijkh} \ \rightsquigarrow \ \mathcal{N}(0, \sigma^2) \ .$$

		Facteur A			
		1	2	…	g
	1	(1) $\mu + \alpha_1 + \beta_1$ ⋮ (c) $\mu + \alpha_1 + \beta_1$	(1) $\mu + \alpha_2 + \beta_1$ ⋮ (c) $\mu + \alpha_2 + \beta_1$	… … …	(1) $\mu + \alpha_g + \beta_1$ ⋮ (c) $\mu + \alpha_g + \beta_1$
Facteur B	2	(1) $\mu + \alpha_1 + \beta_2$ ⋮ (c) $\mu + \alpha_1 + \beta_2$	(1) $\mu + \alpha_2 + \beta_2$ ⋮ (c) $\mu + \alpha_2 + \beta_2$	… … …	(1) $\mu + \alpha_g + \beta_2$ ⋮ (c) $\mu + \alpha_g + \beta_2$
	⋮	⋮	⋮	⋮	⋮
	b	(1) $\mu + \alpha_1 + \beta_b$ ⋮ (c) $\mu + \alpha_1 + \beta_b$	(1) $\mu + \alpha_2 + \beta_b$ ⋮ (c) $\mu + \alpha_2 + \beta_b$	… … …	(1) $\mu + \alpha_g + \beta_b$ ⋮ (c) $\mu + \alpha_g + \beta_b$

FIGURE 15.6 — Plan d'expérience à trois critères de classification.

On note :

$$\text{SCA} = bcl\sum_{j=1}^{g}(\overline{y}_{.j..}-\overline{y}_{....})^2\ , \qquad \text{SCB} = gcl\sum_{i=1}^{b}(\overline{y}_{i...}-\overline{y}_{...})^2\ , \qquad \text{SCL} = bgc\sum_{k=1}^{l}(\overline{y}_{..k.}-\overline{y}_{....})\ ,$$

$$\text{SCBA} = cl\sum_{i=1}^{b}\sum_{j=1}^{g}(\overline{y}_{ij..}-\overline{y}_{i...}-\overline{y}_{.j..}+\overline{y}_{....})\ , \qquad \text{SCBL} = cg\sum_{i=1}^{b}\sum_{k=1}^{l}(\overline{y}_{i.k.}-\overline{y}_{i...}-\overline{y}_{..k.}+\overline{y}_{....})\ ,$$

$$\text{SCAL} = cb\sum_{j=1}^{g}\sum_{k=1}^{l}(\overline{y}_{.jk.}-\overline{y}_{.j..}-\overline{y}_{..k.}+\overline{y}_{....})\ ,$$

$$\text{SCABL} = c\sum_{i=1}^{b}\sum_{j=1}^{g}\sum_{k=1}^{l}(\overline{y}_{ijk.}-\overline{y}_{ij..}-\overline{y}_{i.k.}-\overline{y}_{i...}+\overline{y}_{.j..}+\overline{y}_{..k.}-\overline{y}_{....})\ ,$$

$$\text{SCE} = \sum_{i=1}^{b}\sum_{j=1}^{g}\sum_{k=1}^{l}\sum_{h=1}^{c}(y_{ijkl}-\overline{y}_{ijk.})^2\ , \qquad \text{SCT} = \sum_{i=1}^{b}\sum_{j=1}^{g}\sum_{k=1}^{l}\sum_{h=1}^{c}(y_{ijkl}-\overline{y}_{...})^2\ .$$

La table de l'ANOVA est écrite ci-dessous (cf. tab. 15.6).

TABLE 15.6 — Table de l'ANOVA à trois critères de classification.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCE
Facteur B	$b-1$	SCB	MCB = SCB / $(b-1)$	MCB / MCE
Facteur L	$l-1$	SCL	MCB = SCL / $(l-1)$	MCL / MCE
Inter. AB	$(g-1)(b-1)$	SCAB	MCAB = SCAB / $(g-1)(b-1)$	MCAB / MCE
Inter. BL	$(b-1)(l-1)$	SCBL	MCBL = SCBL / $(b-1)(l-1)$	MCBL / MCE
Inter. AL	$(g-1)(l-1)$	SCAL	MCAL = SCAL / $(g-1)(l-1)$	MCAL / MCE
Inter. ABL	$(g-1)(b-1)(l-1)$	SCABL	MCABL = SCABL / $(g-1)(b-1)(l-1)$	MCABL / MCE
Erreur	$bgl(c-1)$	SCE	MCE = SCE / $bgl(c-1)$	
Total	$bglc-1$	SCT		

Les statistiques employées pour tester les différents effets — des trois facteurs comme des interactions — sont semblables à celles vues précédemment.

Exemple — Considérons le jeu de données suivant, qui nous renseigne sur un score calculé chez des personnes, pour lesquelles on distingue le statut par rapport au tabagisme, l'origine ethnique (*race* en anglais) et le sexe :

		Blanc	Non blanc
Homme	Non fumeur	54	52
		54	52
		58	48
	Fumeur	44	18
		40	22
		44	18
Femme	Non fumeuse	44	6
		40	2
		40	2
	Fumeuse	22	24
		18	20
		22	24

Après réécriture du fichier, nous pouvons réaliser l'ANOVA :

```
> summary(aov(conso ~ factor(smoke) * factor(race) * factor(sexe), smoke))
                                        Df Sum of Sq  Mean Sq F Value
                          factor(smoke)  1   770.667  770.667   144.5
                           factor(race)  1  1536.000 1536.000   288.0
                           factor(sexe)  1  2400.000 2400.000   450.0
                 factor(smoke):factor(race)  1   170.667  170.667    32.0
                 factor(smoke):factor(sexe)  1   682.667  682.667   128.0
                  factor(race):factor(sexe)  1    24.000   24.000     4.5
factor(smoke):factor(race):factor(sexe)  1  1290.667 1290.667   242.0
                                  Residuals 16    85.333    5.333

                                             Pr(F)
                          factor(smoke) 0.00000000
                           factor(race) 0.00000000
                           factor(sexe) 0.00000000
                 factor(smoke):factor(race) 0.00003571
                 factor(smoke):factor(sexe) 0.00000000
                  factor(race):factor(sexe) 0.04986461
factor(smoke):factor(race):factor(sexe) 0.00000000
                                  Residuals
```

qui peut se synthétiser sous la forme suivante :

Table 15.7 — Table de l'ANOVA pour l'exemple du score.

Source	**ddl**	**SC**	**MC**	**F**	**P**
Tabagisme	1	770,67	770,67	114,50	0,000
Type	1	1 536,00	1 536,00	288,00	0,000
Sexe	1	2 400,00	2 400,00	450,00	0,000
Tabag.-Type	1	170,67	170,67	32,00	0,000
Tabag.-Sexe	1	682,67	682,67	128,00	0,000
Type-Sexe	1	24,00	24,00	4,50	0,050
Tabag.-Type-Sexe	1	1 290,67	1 290,67	242,00	0,000
Erreur	16	85,33	5,33		
Total	23	6 960,00			

La non-significativité de l'interaction « Type-Sexe » peut être visualisée[1] sur la figure 15.7, dérivée de la commande :

```
> attach(smoke)
> interaction.plot(factor(race), factor(sexe), conso, fun=mean)
```

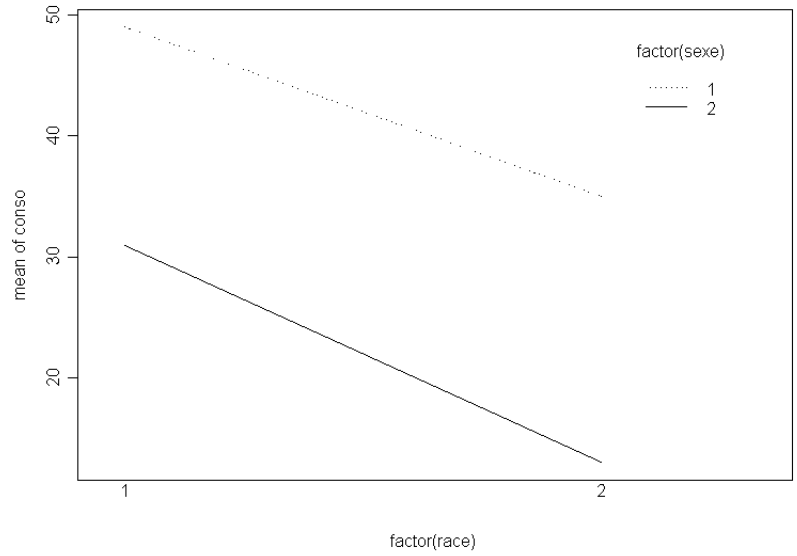


FIGURE 15.7 — Interaction « Type-Sexe ».

15.5.2 Emboîtement à deux niveaux (*Three-way nested design*)

Considérons un modèle à deux niveaux d'emboîtement : il s'agit d'étudier deux facteurs suivant un plan d'expériences incluant des blocs randomisés, blocs destinés à contrôler un possible facteur de nuisance.

	A_1							⋯⋯	A_a						
	B_1			⋯	B_b			⋯⋯	B_1			⋯	B_b		
1	C_1	⋯	C_c		C_1	⋯	C_c		C_1	⋯	C_c	⋯	C_1	⋯	C_c
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r															

FIGURE 15.8 — Plan d'expérience à deux niveaux d'emboîtement.

1. Une interaction significative se traduirait par l'intersection des deux droites.

Le modèle s'écrit :

$$y_{ijk} = \mu + \beta_i + \alpha_j + \gamma_k + (\alpha\gamma)_{jk} + \epsilon_{ijk} \ ,$$

pour $i = 1, \ldots, b$, $j = 1, \ldots, g$, $k = 1, \ldots, l$.

On suppose que

$$\alpha_j \rightsquigarrow \mathcal{N}(0, \sigma_\alpha^2) \ , \qquad (\alpha\beta)_{ij} \rightsquigarrow \mathcal{N}(0, \sigma_{\alpha\beta}^2) \ , \qquad (\alpha\beta\gamma)_{ijk} \rightsquigarrow \mathcal{N}(0, \sigma_{\alpha\beta\gamma}^2) \ , \qquad \epsilon_{ijk} \rightsquigarrow \mathcal{N}(0, \sigma^2) \ .$$

On note :

$$\text{SCA} = bcr \sum_{j=1}^{g} (\overline{y}_{.j.} - \overline{y}_{...})^2 \ , \qquad \text{SCB(A)} = cr \sum_{i=1}^{a} \sum_{j=1}^{b} (\overline{y}_{ij.} - \overline{y}_{i..})^2 \ ,$$

$$\text{SCC(BA)} = c \sum_{j=1}^{g} \sum_{i=1}^{b} \sum_{k=1}^{r} (\overline{y}_{ijk.} - \overline{y}_{ij.})^2 \ ,$$

$$\text{SCE} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \sum_{l=1}^{r} (\overline{y}_{ijkl} - \overline{y}_{ijk.})^2 \ , \qquad \text{SCT} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \sum_{l=1}^{r} (\overline{y}_{ijkl} - \overline{y}_{....})^2 \ .$$

La table de l'ANOVA est écrite ci-dessous (cf. tab. 15.8).

TABLE 15.8 — Table de l'ANOVA à deux niveaux d'emboîtement.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCB(A)
Facteur B (dans A)	$g(b-1)$	SCB(A)	MCB(A) = SCB(A) / $g(b-1)$	MCB(A) / MCC(AB)
Facteur C (dans B)	$gb(r-1)$	SCC(BA)	MCC(AB) = SCC(AB) / $gb(c-1)$	MCC(AB) / MCE
Erreur	$abr(c-1)$	SCE		
Total	$abcr-1$	SCT		

La statistique de Fisher testant l'effet du traitement A vaut

$$F_{\text{A}} = \frac{\text{MCA}}{\text{MCB(A)}} \rightsquigarrow F\big(g-1 \ , \ g(b-1)\big) \ .$$

La statistique de Fisher testant l'effet du traitement B vaut

$$F_{\text{B}} = \frac{\text{MCB(A)}}{\text{MCC(AB)}} \rightsquigarrow F\big(g(b-1) \ , \ gb(c-1)\big) \ .$$

La statistique de Fisher testant l'effet du traitement C vaut

$$F_{\text{C}} = \frac{\text{MCC(AB)}}{\text{MCE}} \rightsquigarrow F\big(gb(c-1) \ , \ gbc(r-1)\big) \ .$$

15.5.3 Carré latin (*Latin Square design*)

Nous sommes en présence d'un facteur d'intérêt A (traitement) et de deux facteurs de nuisance (blocs). Le dispositif est constitué d'un nombre de parcelles qui est un carré (9, 16, 25, ...), et il comporte autant de lignes de parcelles que de colonnes de parcelles ; au sein de ce dispositif, chaque objet est présent une et une seule fois dans chaque ligne et dans chaque colonne.

Exemple — Nous considérons un essai de chauffage du sol sur une variété de plante, et nous relevons les accroissements moyens en hauteur. Chaque température n'apparaît qu'une seule fois par ligne et par colonne.

20 °C 185	30 °C 242	15 °C 177	25 °C 214
15 °C 117	25 °C 229	20 °C 209	30 °C 238

Serre A

30 °C 200	20 °C 200	25 °C 222	15 °C 154
25 °C 218	15 °C 174	30 °C 247	20 °C 205

Serre B

FIGURE 15.9 — Carré latin.

Le modèle s'écrit :

$$y_{ik(j)} \quad = \quad \mu + \beta_i + \gamma_k + \alpha_j + \epsilon_{ik(j)} \ ,$$

pour $i = 1, \ldots, b$, $j = 1, \ldots, b$, $k = 1, \ldots, b$. On suppose que $\epsilon_{ik(j)} \rightsquigarrow \mathcal{N}(0, \sigma^2)$. On suppose de plus que :

$$\sum_{i=1}^{b} \beta_i \ = \ \sum_{k=1}^{b} \gamma_k \ = \ \sum_{j=1}^{b} \alpha_j \ = \ 0 \ .$$

Les deux effets « bloc » — c.-à-d. dûs aux lignes et aux colonnes — sont représentés par β_i et γ_k, tandis que l'effet du facteur d'intérêt A (traitement) est représenté par α_j.

Par la suite, les notations L et C désigneront les termes *ligne* et *colonne*. On note :

$$\text{SCA} = b\sum_{j=1}^{b}(\overline{y}_{..j} - \overline{y}_{...})^2 , \qquad \text{SCL} = b\sum_{i=1}^{b}(\overline{y}_{i..} - \overline{y}_{...})^2 ,$$

$$\text{SCC} = b\sum_{k=1}^{b}\sum_{k=1}^{b}(\overline{y}_{.k.} - \overline{y}_{...})^2 ,$$

$$\text{SCE} = \sum_{i=1}^{b}\sum_{k=1}^{b}(y_{ik(j)} - \overline{y}_{i..} - \overline{y}_{.k.} + 2\overline{y}_{...})^2 , \qquad \text{SCT} = \sum_{i=1}^{b}\sum_{k=1}^{b}(y_{ik(j)} - \overline{y}_{...})^2 .$$

La table de l'ANOVA est écrite ci-dessous (cf. tab. 15.9).

TABLE 15.9 — Table de l'ANOVA pour le carré latin.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$b-1$	SCA	MCA = SCA / $(b-1)$	MCA / MCE
Facteur *lignes*	$b-1$	SCL	MCL = SCL / $(b-1)$	MCL / MCE
Facteur *colonnes*	$b-1$	SCC	MCC = SCC / $(b-1)$	MCC / MCE
Erreur	$(b-1)(b-2)$	SCE	MCE = SCE / $(b-1)(b-2)$	
Total	b^2-1	SCT		

Exemple — Dans l'essai sur le chauffage du sol, nous obtenons la table 15.10.

TABLE 15.10 — ANOVA sur l'essai de chauffage du sol.

Source	**ddl**	**SC**	**MC**	**F**	**P**
Températures	3	13,61	4,54	43,00	0,000
Lignes	3	661,00	220,00	2,09	0,203
Colonnes	3	2 832,00	944,00	8,95	0,012
Erreur	6	633,00	105,50		
Total	15	17,74			

15.6 Plus de trois critères de classification (2^p *design*)

S'il y a p facteurs présentant $b_1, b_2, \ldots, b_p$ niveaux,

16

Avec effet(s) aléatoire(s)

16.1 Un facteur aléatoire (*One-way random effect model*)

Ce cas se traite exactement comme si le facteur était non pas élatoire, mais fixe. Soient n observations réparties en g groupes, où les g groupes sont supposés être tirés au sort parmi un grand nombre de groupes. On note y_{ij} l'observation concernant la i^{e} observation du groupe j. Chaque groupe j contient n_j observations. Le modèle s'écrit

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij} \ .$$

α_j représente l'effet du groupe j. On suppose que :

$$\epsilon_{ij} \overset{\text{i.i.d.}}{\rightsquigarrow} \mathcal{N}(0, \sigma^2) \ , \qquad \alpha_j \overset{\text{i.i.d.}}{\rightsquigarrow} \mathcal{N}(0, \sigma_\alpha^2) \ .$$

L'hypothèse nulle d'égalité entre tous les groupes se traduit par la nullité de σ_α^2. En pratique, le tableau de l'ANOVA et la statistique de Fisher sont exactement celles de la section 15.1. Dans le cas présent, les observations y_{ij} et y_{kj} sont corrélées : puisque ces deux observations contiennent une variable aléatoire commune α_j, nous avons

$$\mathbb{C}\text{ov}\left(y_{ij}, y_{kj}\right) \ = \ \sigma_\alpha^2$$

et

$$\mathbb{V}\left(y_{ij}\right) \ = \ \sigma_\alpha^2 + \sigma^2 \ .$$

À partir de ces constatations, nous pouvons définir le **coefficient de corrélation intraclasse** — qui mesure la corrélation entre deux observations appartenant à un même groupe — comme étant :

$$\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma^2} \ .$$

Si l'hypothèse nulle est rejetée — c.-à-d. que $\sigma_\alpha^2 \neq 0$ — nous pouvons estimer σ_α par

$$\frac{\text{MCA} - \text{MCE}}{\frac{1}{g-1}\left[n - \frac{1}{n}\sum_{j=1}^{g} n_j^2\right]} \ .$$

Exemple — Concernant la campagne de spots publicitaires, nous pouvons utiliser les données concernant le spot n° 2 : nous considérons que les 5 régions constituent un échantillon aléatoire d'une population incluant toutes les régions existantes. L'analyse avec un effet aléatoire « Région » s'écrit :

```
> consomenage3_consomenage[consomenage$Pub==2,]
> raov(Conso~factor(Region), consomenage3)

Call:
   raov(formula = Conso ~ factor(Region), data = consomenage3)

Terms:
                factor(Region) Residuals
 Sum of Squares        4790.53  11090.12
Deg. of Freedom              4        25

Residual standard error: 21.06193
Estimated effects are balanced
```

La statistique de Fisher et la *p-value* sont calculées ci-dessous :

```
> (4790.53/4)/(11090.12/25)
[1] 2.699774

> 1-pf(2.699774,4,25)
[1] 0.05363328
```

d'où nous concluons que l'effet « Région » n'est pas significatif au seuil de 5 %. Considérons toutefois sa significativité au seuil de 10 % : l'estimation de la variance de cet effet σ_α^2 — variance inter-région — est

$$\frac{(4\,790{,}53)/4 - (11\,090{,}12)/25}{1/4\Big(30 - 180/30\Big)} \ = \ 754{,}03 \ .$$

Une estimation de la variance intra-région est

$$\frac{(11\,090{,}12)/25}{6} \ = \ 73{,}93 \ .$$

Nous constatons que la variance inter-région est environ dix fois supérieure à la variance intra-région : ainsi, la variance de l'effet « Région » ne peut être négligée, ce qu'affirme la significativité de la statistique au seuil de 10 %.

16.2 Deux facteurs aléatoires (*Two-way random effects model*)

En reprenant exactement les mêmes notations que la section 15.4.2, le modèle s'écrit :

$$y_{ij} = \mu + \alpha_j + \beta_i + (\alpha\beta)_{ij} + \epsilon_{ijk} \ ,$$

pour $i = 1, \ldots, b$, $j = 1, \ldots, g$, $k = 1, \ldots, c$, avec les hypothèses suivantes :

$$\alpha_j \rightsquigarrow \mathcal{N}(0, \sigma_\alpha^2) , \qquad \beta_i \rightsquigarrow \mathcal{N}(0, \sigma_\beta^2) , \qquad (\alpha\beta)_{ij} \rightsquigarrow; \mathcal{N}(0, \sigma_{\alpha\beta}^2) , \qquad \epsilon_{ijk} \rightsquigarrow \mathcal{N}(0, \sigma^2) .$$

On appelle **coefficient de corrélation intraclasse** concernant les réponses au traitement A la quantité

$$\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2} .$$

Un coefficient similaire peut être calculé concernant les réponses au traitement B, ainsi que concernant les observations d'une même cellule (même niveau de traitement pour A et même niveau de traitement pour B).

La table de l'ANOVA est écrite ci-dessous (cf. tab. 16.1).

TABLE 16.1 — Table de l'ANOVA à deux facteurs aléatoires.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCAB
Facteur B	$b-1$	SCB	MCB = SCB / $(b-1)$	MCB / MCAB
Inter. AB	$(g-1)(b-1)$	SCAB	MCAB = SCAB / $(g-1)(b-1)$	MCAB / MCE
Erreur	$bg(c-1)$	SCE	MCE = SCE / $bg(c-1)$	
Total	$bgc-1$	SCT		

La statistique employée pour tester l'effet du facteur A (effet groupe) est

$$F_{\text{A}} = \frac{\text{MCA}}{\text{MCAB}} \rightsquigarrow F\big(g-1 \, , \, (g-1)(b-1)\big) .$$

La statistique employée pour tester l'effet du facteur B (effet traitement) est

$$F_{\text{B}} = \frac{\text{MCB}}{\text{MCAB}} \rightsquigarrow F\big(b-1 \, , \, (g-1)(b-1)\big) .$$

La statistique employée pour tester l'interaction AB est

$$F_{\text{AB}} = \frac{\text{MC(AB)}}{\text{MCE}} \rightsquigarrow F\big((g-1)(b-1) \, , \, gb(c-1)\big) .$$

Exemple — Dans l'étude de l'impact des spots publicitaires, nous obtenons le modèle voulu et l'analyse par les commandes suivantes :

```
> raov(formula = Conso ~ factor(Pub) * factor(Region), data = consomenage2)

Terms:
                 factor(Pub) factor(Region) factor(Pub):factor(Region)
 Sum of Squares      4585.68        4867.51                    8937.92
Deg. of Freedom            3              4                         12
```

```
                Residuals
 Sum of Squares  42382.02
Deg. of Freedom       100

Residual standard error: 20.58689
Estimated effects are balanced
```

Les degrés de significativité des tests des effets « Pub », « Régions », « Pub-Régions » sont respectivement :

```
> 1 - pf((4585.68/3)/(8937.92/12), 3, 12)
[1] 0.160266

> 1 - pf((4867.51/4)/(8937.92/12), 4, 12)
[1] 0.2294265

> 1 - pf((8937.92/12)/(42382.02/100), 12, 100)
[1] 0.06584021
```

16.3 Modèle mixte (*Two-way mixed effects model*)

Elle comporte un effet fixe et un effet aléatoire. En reprenant les notations de la section précédente, nous supposons que l'effet de A est fixe, tandis que celui de B est aléatoire — celui de l'interaction étant aléatoire.

La table de l'ANOVA est écrite ci-dessous (cf. tab. 16.2).

TABLE 16.2 — Tableau de l'ANOVA.

Source	**ddl**	**SC**	**MC**	**F**
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCAB
Facteur B	$b-1$	SCB	MCB = SCB / $(b-1)$	MCB / MCE
Inter. AB	$(g-1)(b-1)$	SCAB	MCAB = SCAB / $(g-1)(b-1)$	MCAB / MCE
Erreur	$bg(c-1)$	SCE	MCE = SCE / $bg(c-1)$	
Total	$bgc-1$	SCT		

La statistique employée pour tester l'effet du facteur A (effet groupe) est

$$F_{\text{A}} \quad = \quad \frac{\text{MCA}}{\text{MCAB}} \ \rightsquigarrow \ F\big(g-1\,,\,(g-1)(b-1)\big)\ .$$

La statistique employée pour tester l'effet du facteur B (effet traitement) est

$$F_{\mathrm{B}} \;=\; \frac{\mathrm{MCB}}{\mathrm{MCE}} \;\rightsquigarrow\; F\big(b-1\,,\,gb(c-1)\big)\;.$$

La statistique employée pour tester l'interaction AB est

$$F_{\mathrm{AB}} \;=\; \frac{\mathrm{MCAB}}{\mathrm{MCE}} \;\rightsquigarrow\; F\big((g-1)(b-1)\,,\,gb(c-1)\big)\;.$$

16.4 Blocs aléatoires avec subdivisions (*Split Plot*)

Ce type de plan expérimental provient de la recherche agronomique. On désire tester un traitement A sur plusieurs bandes de terre ; ces bandes peuvent elles-mêmes être divisées en sous-unités (ou *split plots*), afin de tester un second traitement.

Pour une expérience à deux facteurs, de type gb (g niveaux du premier facteur et b niveau du second) et comportant c blocs, la première étape consiste en une répartition classique des g variantes du premier facteur au sein des c blocs, conduisant à la délimitation de gc **parcelles primaires** (*whole plots*). La seconde étape consiste ensuite en une répartition aléatoire et indépendante des b variantes du second facteur à l'intérieur des gc parcelles principales, de manière à constituer gbc **parcelles secondaires** (*subplots*).

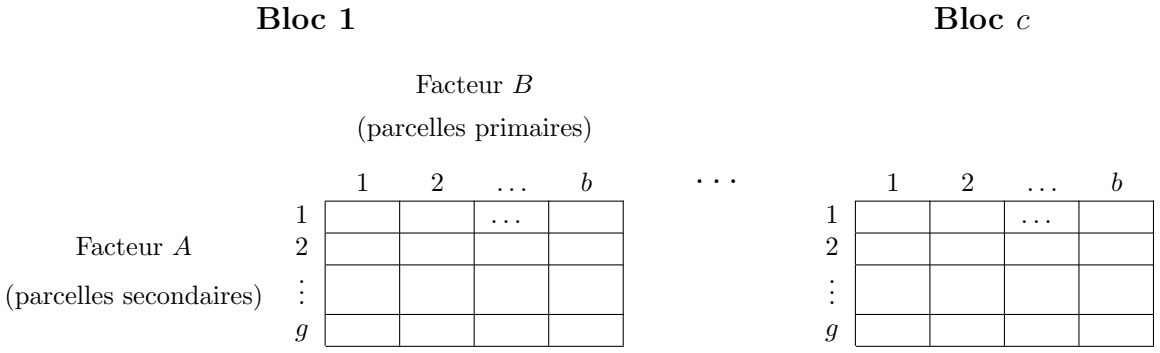


FIGURE 16.1 — Plan d'expérience avec parcelles divisées.

Remarque — La **variable d'intérêt est celle délimitant les parcelles secondaires**.

Le modèle s'écrit :

$$y_{ijk} \;=\; \mu + \alpha_j + \beta_i + (\alpha\beta)_{ij} + \rho_k + (\rho\alpha)_{ik} + \epsilon_{ijk}\;,$$

pour $i = 1, \ldots, g$, $j = 1, \ldots, b$, $k = 1, \ldots, c$, et où μ désigne la moyenne générale (*grand mean*), α_i l'effet du traitement A au niveau i, β_j l'effet du traitement B au niveau j et ρ_k

l'effet de la duplication. On suppose que les ρ_k sont i.i.d. de loi $\mathcal{N}(0, \sigma_\rho^2)$, que les $(\rho\alpha)_{ik}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma_{\rho\alpha}^2)$, et que

$$\sum_{k=1}^{r}(\rho\alpha)_{ik} = 0 \quad \forall i \ .$$

On suppose aussi que les ϵ_{ijk} sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et que les ρ_i, $(\rho\alpha)_{ij}$ et ϵ_{ijk} sont indépendants. Concernant les effets fixes, on suppose que

$$\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = 0 \ , \qquad \sum_{i=1}^{a}(\alpha\beta)_{ij} = 0 \quad \forall j \ , \qquad \sum_{j=1}^{b}(\alpha\beta)_{ij} = 0 \quad \forall i \ .$$

On a

$$\mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \ .$$

Toutes les observations ont la même variance :

$$\mathbb{V}(y_{ijk}) = \sigma_\rho^2 + \sigma_{\rho\alpha}^2 + \sigma^2 \ .$$

Les observations à l'intérieur d'un même terrain ont une corrélation constante de

$$\frac{\sigma^2 + \sigma_{\rho\alpha}^2}{\sigma^2 + \sigma_\rho^2 + \sigma_{\rho\alpha}^2} \ .$$

Dans l'écriture des carrés, les abbréviations sont P pour primaire et S pour secondaire. On note :

$$\text{SCA} = \sum_{j=1}^{g}(\overline{y}_{.j.} - \overline{y}_{...})^2 \ , \qquad \text{SCB} = \sum_{i=1}^{b}(\overline{y}_{i..} - \overline{y}_{...})^2 \ , \qquad \text{SCR} = \sum_{k=1}^{c}(\overline{y}_{..k} - \overline{y}_{...})^2 \ ,$$

$$\text{SCEP} = \sum_{i=1}^{b}\sum_{k=1}^{c}(\overline{y}_{i.k} - \overline{y}_{...})^2 - \text{SCR} - \text{SCB} = \text{erreur liée aux parcelles primaires} \ ,$$

$$\text{SCES} = \text{SCT} - \text{SCA} - \text{SCB} - \text{SCR} - \text{SCAB} - \text{SCEP} = \text{erreur liée aux parcelles secondaires} \ ,$$

$$\text{SCAB} = \sum_{i=1}^{b}\sum_{j=1}^{g}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 \ , \qquad \text{SCT} = \sum_{i=1}^{b}\sum_{j=1}^{g}\sum_{k=1}^{c}(\overline{y}_{ijk} - \overline{y}_{...})^2 \ .$$

La table de l'ANOVA est écrite ci-dessous (cf. tab. 16.3).

TABLE 16.3 — Table de l'ANOVA pour le modèle avec blocs aléatoires contenant des subdivisions.

Source	ddl	SC	MC	F
Entre les parcelles primaires (*whole plots*)				
Blocs	$c-1$	SCR		
Facteur B	$b-1$	SCB	MCB = SCB / $(b-1)$	MCB / MCEP
Erreur I	$(b-1)(c-1)$	SCEP	MCEP = SCEP / $(b-1)(c-1)$	
(inter. bloc - A)				
Entre les parcelles secondaires (*subplots*)				
Facteur A	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCES
Inter. AB	$(g-1)(b-1)$	SCAB	MCAB = SCAB / $(g-1)(b-1)$	MCAB / MCES
Erreur II	$g(b-1)(c-1)$	SCES	MCES = SCES / $g(b-1)(c-1)$	
(inter. bloc - B)	$((g-1)(c-1))$			
(inter. bloc - A - B)	$((g-1)(b-1)(c-1))$			
Total	$gbc-1$	SCT		

La statistique de Fisher testant l'effet du traitement A vaut

$$F_{\mathrm{A}} = \frac{\mathrm{MCA}}{\mathrm{MCES}} \rightsquigarrow F\big(g-1\ ,\ b(g-1)(c-1)\big)\ .$$

La statistique de Fisher testant l'effet du traitement B vaut

$$F_{\mathrm{B}} = \frac{\mathrm{MCB}}{\mathrm{MCEP}} \rightsquigarrow F\big(b-1\ ,\ (b-1)(c-1)\big)\ .$$

La statistique de Fisher testant l'interaction vaut

$$F_{\mathrm{AB}} = \frac{\mathrm{SCAB}}{\mathrm{SCES}} \rightsquigarrow F\big((g-1)(b-1)\ ,\ b(g-1)(c-1)\big)\ .$$

Exemple — Nous supposons que, dans l'exemple du spot publicitaire, chaque région représente un bloc. Dans chaque région, il y a 4 parcelles primaires, correspondant aux 4 spots. Pour chaque spot, nous avons 6 familles différentes, suivant le critère de leur taille — ainsi le critère « Taille » constitue-t-il les parcelles secondaires. Le but est de déterminer s'il y a une différence significative de consommation entre des ménages de tailles différentes.

La procédure S-Plus est la suivante [1]

```
> summary(aov(Conso ~ factor(Taille) * factor(Pub)
                        + Error(factor(Region) * factor(Pub)), consomenage))

Error: factor(Pub)
            Df Sum of Sq Mean Sq
factor(Pub)  3   4585.68 1528.56

Error: factor(Region):factor(Pub)
          Df Sum of Sq  Mean Sq F Value Pr(F)
Residuals 12  8937.917 744.8265

Error: Within
```

1. Elle n'est pas explicitement donnée dans le chapitre « Split Plot » du Guide de l'Utilisateur, Tome 1.

```
                           Df Sum of Sq  Mean Sq  F Value      Pr(F)
             factor(Taille)  5  40967.65 8193.529 654.4706 0.00000000
factor(Taille):factor(Pub) 15    412.82   27.522   2.1983 0.01284298
                 Residuals 80   1001.55   12.519
```

Nous en concluons que l'effet « Taille » est grandement significatif.

Quelques vérifications, concernant les calculs des statistiques :

```
> (4585.68/3)/(8937.92/12)
[1] 2.052236
> (40967.65/5)/(1001.55/80)
[1] 654.468
> (4585.68/3)/(8937.91/12)
[1] 2.052238
```

Pour tester la significativité de l'effet attribué au type de spot publicitaire, nous calculons :

```
> 1-pf((4585.68/3)/(8937.91/12),3,12)
[1] 0.1602657
```

Nous en concluons que cet effet n'est pas significatif ($p > 16$ %).
Nous pouvons récapituler l'ANOVA au travers du tableau 16.4.

TABLE 16.4 — Résumé de l'ANOVA.

Source	**ddl**	**SC**	**F**	***P***
Entre les spots publicitaires (parcelles primaires)				
Région (bloc)	4	4 867,51		
Spot (primaire)	3	4 585,68	2,05	0,16
Erreur I	12	8 937,92		
(inter. région - spot)				
Entre les types de familles (parcelles secondaires)				
Taille	5	40 967,65	654,47	0,00
Inter. Taille - Spot	15	412,82	2,20	0,01
Erreur II	80	1 001,55		
(inter. Région - Taille)	(20)			
(inter. Région - Taille - Spot)	(60)			
Total	119	60 773,13		

16.5 Blocs aléatoires avec subdivisions sur des mesures répétées (*Repeated measures Split Plot design*)

Les sujets sont échantillonnés aléatoirement par groupes de sujets (parcelles primaires). IL y a b groupes contenant chacun c sujets. Chaque groupe de sujets reçoit une certaine

dose d'un traitement (parcelle secondaire), parmi g doses. Ensuite, chaque sujet est ré-échantillonné, de manière à ce qu'il reçoive chaque dose du traitement. La répétition porte donc sur les sujets : il est impératif d'introduire un effet aléatoire « Sujet » dans le modèle pour prnedre en compte la corrélation des données issues d'un même sujet. Le modèle s'écrit :

$$y_{ijk} \;=\; \mu + \alpha_j + \beta_i + (\alpha\beta)_{ij} + \gamma_{ik} + \epsilon_{ijk} \;, \tag{16.1}$$

pour $i = 1, \ldots, b$, $j = 1, \ldots, g$, $k = 1, \ldots, c$, où μ est la moyenne générale (*grand mean*), α_j représente l'effet du traitement à la dose j, β_i représente l'effet des groupes de sujets, et γ_k est l'effet aléatoire « Sujet ». Concernant les effets fixes, on suppose que

$$\sum_{j=1}^{g} \alpha_g \;=\; \sum_{i=1}^{b} \beta_i \;=\; \sum_{j=1}^{g} (\alpha\beta)_{ij} \;=\; \sum_{i=1}^{b} (\alpha\beta)_{ij} \;=\; 0 \;.$$

On suppose par ailleurs que les γ_k sont i.i.d. de loi $\mathcal{N}(0, \sigma_\gamma^2)$, et que les ϵ_{ijk} sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

On note :

$$\text{SCB} \;=\; gc \sum_{i=1}^{b} (\overline{y}_{i..} - \overline{y}_{...})^2 \;, \qquad \text{SCS} \;=\; g \sum_{j=1}^{b} \sum_{k=1}^{c} (\overline{y}_{i.k} - \overline{y}_{i..})^2 \;, \qquad \text{SCA} \;=\; cb \sum_{j=1}^{g} (\overline{y}_{.j.} - \overline{y}_{...})^2 \;,$$

$$\text{SCAB} \;=\; c \sum_{i=1}^{b} \sum_{j=1}^{g} (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 \;,$$

$$\text{SCE} \;=\; \sum_{i=1}^{b} \sum_{j=1}^{g} \sum_{k=1}^{c} (\overline{y}_{ijk} - \overline{y}_{i.k} - \overline{y}_{ij.} + \overline{y}_{i..})^2 \;, \qquad \text{SCT} \;=\; \sum_{i=1}^{b} \sum_{j=1}^{g} \sum_{k=1}^{c} (\overline{y}_{ijk} - \overline{y}_{...})^2 \;.$$

La table de l'ANOVA est donnée ci-dessous.

Table 16.5 — Table de l'ANOVA pour le modèles avec blocs aléatoires contenant des subdividions sur des mesures répétées.

Source	**ddl**	**SC**	**MC**	**F**
Facteur B (groupes) (*parcelles primaires*)	$b-1$	SCB	MCB = SCB / $(b-1)$	MCB / MCS
Facteur S (sujets)	$b(c-1)$	SCS	MCS = SCS / $b(c-1)$	MCS/MCE
Facteur A (traitement) (*parcelles secondaires*)	$g-1$	SCA	MCA = SCA / $(g-1)$	MCA / MCE
Inter. AB	$(g-1)(b-1)$	SCAB	MCAB = SCAB / $(g-1)(b-1)$	MCAB / MSE
Erreur	$b(g-1)(c-1)$	SCE	MCE = SCE / $b(g-1)(c-1)$	
Total	$gbc-1$	SCT		

La statistique de Fisher testant l'effet de la constitution des groupes (facteur B) vaut

$$F_{\text{B}} = \frac{\text{MCB}}{\text{MCS}} \;\rightsquigarrow\; F\big(b-1 \,,\, b(c-1)\big) \;.$$

La statistique de Fisher testant l'effet du traitement (facteur A) vaut

$$F_{\mathrm{A}} = \frac{\mathrm{MCA}}{\mathrm{MCE}} \rightsquigarrow F\big(g-1\ ,\ b(g-1)(c-1)\big)\ .$$

La statistique de Fisher testant l'interaction vaut

$$F_{\mathrm{AB}} = \frac{\mathrm{MCAB}}{\mathrm{MCE}} \rightsquigarrow F\big((g-1)(b-1)\ ,\ b(g-1)(c-1)\big)\ .$$

La statistique de Fisher testant l'effet « Sujet » vaut

$$F_{\mathrm{S}} = \frac{\mathrm{MCS}}{\mathrm{MCE}} \rightsquigarrow F\big(b(c-1)\ ,\ b(g-1)(c-1)\big)\ .$$

Septième partie

RÉÉCHANTILLONNAGE

17

Jackknife

17.1 Définitions

17.1.1 Cas unidimensionnel

L'objectif initial du *jackknife* est de réduire le biais d'un estimateur. Soient n réalisations indépendantes $(X_1, \ldots, X_n)$ d'une variable X de loi $\mathbb{P}_\theta$ dépendant d'un paramètre réel θ ; on possède un estimateur T_n biaisé de θ :

$$\mathbb{E}\,(T_n) = \theta + B(n,\theta)$$

On note $\mathcal{E}_i$ le sous-échantillon $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ obtenu à partir de l'échantillon initial en supprimant la i^{e} observation ; cela revient à dire que l'on fait un sondage dans l'échantillon de base en tirant $n-1$ observations sans remise. T^i_{n-1} désigne la statistique fondée sur $\mathcal{E}_i$ selon la même règle de décision que celle de T_n.

Définition 17.1 — *On appelle* **pseudo-valeur d'ordre i** *de T_n la statistique*

$$J_i(T) = n\,T_n - (n-1)\,T^i_{n-1}\ .$$

Définition 17.2 — *On appelle* **jackknife** *de T_n la statistique $J(T_n)$ moyenne des pseudo-valeurs :*

$$\begin{aligned} J(T_n) &= \frac{1}{n}\sum_{i=1}^{n} J_i(T_n) \\ &= n\,T_n - \frac{n-1}{n}\sum_{i=1}^{n} T^i_{n-1} \\ &= T_n - \frac{n-1}{n}\sum_{i=1}^{n}(T^i_{n-1} - T_n)\ . \end{aligned}$$

$J(T_n)$ est appelé estimateur du jackknife de T_n, ou « jackknifé » de T_n.

Nota — Si T_n est sans biais, alors $J(T_n)$ l'est aussi.

Calculons la variance empirique $S^2_{PJ}(T)$ des n pseudo-valeurs $J_i(T)$:

$$\begin{aligned} S^2_{PJ}(T) &= \frac{1}{n-1}\sum_{i=1}^{n}\left[J_i(T_n) - J(T_n)\right]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\Big(n\,T_n - (n-1)\,T^i_{n-1} - \frac{1}{n}\sum_{i=1}^{n}\left[n\,T_n - (n-1)\,T^i_{n-1}\right]\Big)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\Big[-(n-1)\,T^i_{n-1} + \frac{n-1}{n}\sum_{i=1}^{n}T^i_{n-1}\Big]^2 \ . \end{aligned}$$

D'où

$$S^2_{PJ}(T) = (n-1)\sum_{i=1}^{n}(T^i_{n-1} - \overline{T}^i_{n-1})^2$$

avec

$$\overline{T}^i_{n-1} = \frac{1}{n}\sum_{i=1}^{n}T^i_{n-1} \ .$$

D'autre part,

$$\mathbb{V}\left(J(T_n)\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbb{V}\left(J_i(T_n)\right) + \sum_{i\neq j}\mathbb{C}\mathrm{ov}\,\left(J_i(T_n),\,J_j(T_n)\right)\right] \ .$$

Les $J_i(T_n)$ peuvent être considérés comme i.i.d. ; sous cette conjecture,

$$\mathbb{V}\left(J(T_n)\right) = \frac{1}{n}\,\mathbb{V}\left(J_1(T_n)\right) \ .$$

On peut estimer $\mathbb{V}\left(J(T_n)\right)$ par $\dfrac{1}{n}\,S^2_{PJ}(T)$, c.-à-d.

$$\begin{aligned} \widehat{\mathbb{V}}\left(J(T_n)\right) &= \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\left[J_i(T_n) - J(T_n)\right]^2 \\ &= \frac{n-1}{n}\sum_{i=1}^{n}\left(T^i_{n-1} - \overline{T}^i_{n-1}\right)^2 \ . \end{aligned}$$

Par la suite, on notera

$$JV(T_n) = \frac{1}{n}\,S^2_{PJ}(T) \ .$$

17.1.1.1 Cas multidimensionnel

Soient $\theta = (\theta_j)_{j=1,\ldots,p}$ et $T_n = (T^j_n)_{j=1,\ldots,p}$ une statistique à valeurs dans $\mathbb{R}^p$. On définit le vecteur T^i_{n-1} de coordonnées $(T^{i,j}_{n-1})_{j=1,\ldots,p}$, où $T^{i,j}_{n-1}$ est construit à partir de T^j_n comme T^i_{n-1} à partir de T_n dans le cas unidimensionnel.

La pseudo-valeur d'ordre i est un vecteur de $\mathbb{R}^p$:

$$\begin{aligned} J_i(T) &= n\,T_n - (n-1)\,T^i_{n-1} \\ &= T_n - (n-1)\ (T^i_{n-1} - T_n) \ . \end{aligned}$$

Posons, pour $i = 1, \ldots, n$,

$$e_i = T_{n-1}^i - T_n$$

et

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i \ .$$

Le jackknife de T_n est

$$\begin{aligned} J(T_n) &= \frac{1}{n} \sum_{i=1}^{n} J_i(T_n) \\ &= n\, T_n - \frac{n-1}{n} \sum_{i=1}^{n} T_{n-1}^i \\ &= T_n - \frac{n-1}{n} \sum_{i=1}^{n} e_i \\ &= T_n - (n-1)\bar{e} \ . \end{aligned}$$

L'estimateur de $\mathbb{V}(T_n)$ est, par analogie avec le cas unidimensionnel, la matrice carrée d'ordre p

$$\begin{aligned} JV(T_n) &= \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{n-1}^i - \overline{T}_{n-1}^i\right) \left(T_{n-1}^i - \overline{T}_{n-1}^i\right)^t \\ &= \frac{n-1}{n} \sum_{i=1}^{n} (e_i - \bar{e}) \ (e_i - \bar{e})^t \ . \end{aligned}$$

Remarque — Pour un paramètre multidimensionnel, on préfèrera utiliser — dans la pratique — la méthode du jackknife unidimensionnel coordonnée par coordonnée.

17.1.2 Propriétés

Théorème 17.1 — *Si le biais de T_n est de la forme*

$$B(n, \theta) = \sum_{k=1}^{\infty} \frac{a_k}{n^k} \ ,$$

alors $J(T_n)$ est un estimateur biaisé en θ d'ordre supérieur ou égal à 2 en $\dfrac{1}{n}$.

En effet, si $B(J)$ est le biais de $J(T_n)$, on a

$$\begin{aligned} B(J) &= n\, B(n, \theta) - (n-1)\, B(n-1, \theta) \\ &= n \sum_{k=1}^{\infty} \frac{a_k}{n^k} - (n-1) \sum_{k=1}^{\infty} \frac{a_k}{(n-1)^k} \\ &= -\frac{a_2}{n\,(n-1)} + \sum_{k=3}^{\infty} a_k \left(\frac{1}{n^{k-1}} - \frac{1}{(n-1)^{k-1}} \right) \ . \end{aligned}$$

Donc, si $\mathbb{E}\,(T_n - \theta) = \mathrm{O}(\frac{1}{n})$, alors $\mathbb{E}\,\big(J(T_n)\big) = \mathrm{O}(\frac{1}{n^2})$. On constate que si $a_k = 0$ pour tout $k \geqslant 2$, $J(T_n)$ est un **estimateur sans biais** de θ. La méthode du jackknife est donc un moyen **robuste**[1] pour diminuer le biais d'un estimateur.

Ce théorème peut être étendu : si l'ordre du premier terme en $\dfrac{1}{n}$ de $B(n, \theta)$ est α ($\alpha \geqslant 2$), *i.e.*

$$\mathbb{E}\,(T_n - \theta) = \frac{a_0}{n^\alpha} + \sum_{k=1}^{\infty} \frac{a_k}{n^{\alpha+k}} \ ,$$

le terme en $\frac{1}{n^\alpha}$ sera éliminé en considérant la statistique

$$J(T_n) = \frac{n^\alpha\, T_n - (n-1)^\alpha\, \overline{T}^{\,i}_{n-1}}{n^\alpha - (n-1)^\alpha} \ .$$

Théorème 17.2 — *Soit T_n un estimateur tel que*

$$T_n \ \xrightarrow{\ \mathbb{P}\ } \ \theta \ ,$$

alors

$$J(T_n) \ \xrightarrow{\ \mathbb{P}\ } \ \theta \ .$$

17.1.3 Généralisation du jackknife

Soient deux estimateurs biaisés $\widehat{\theta}_1$ et $\widehat{\theta}_2$ du paramètre θ tels que

$$\begin{aligned} \mathbb{E}\,(\widehat{\theta}_1 - \theta) &= B_1(n, \theta) \ , \\ \mathbb{E}\,(\widehat{\theta}_2 - \theta) &= B_2(n, \theta) \ , \end{aligned}$$

avec

$$B_1(n, \theta) \neq B_2(n, \theta) \ .$$

On pose

$$R = \frac{B_1(n, \theta)}{B_2(n, \theta)} \ .$$

L'estimateur $\widehat{\theta} = G(\widehat{\theta}_1, \widehat{\theta}_2)$ défini par

$$\widehat{\theta} = \frac{\widehat{\theta}_1 - R \cdot \widehat{\theta}_2}{1 - R}$$

est sans biais pour θ.

Remarques — Notons :

1° qu'une situation fréquente est celle où $B_1(n,\theta) = b(\theta) \cdot f_1(n)$ et $B_2(n, \theta) = b(\theta) \cdot f_2(n)$; dans ce cas-là,

$$R = \frac{f_1(n)}{f_2(n)} \ ;$$

1. « Robuste » au sens où la loi de T_n n'intervient pas explicitement.

2° le jackknife $J(T_n)$ d'une statistique T_n en est un cas particulier, avec $\widehat{\theta}_1 = T_n$, $\widehat{\theta}_2 = \overline{T}^i_{n-1}$ et $R = (n-1)/n$.

Dans le cas particulier de la première remarque, l'estimateur $G(\widehat{\theta}_1, \widehat{\theta}_2)$ peut être exprimé de la façon suivante :

$$\widehat{\theta} = \frac{\begin{vmatrix} \widehat{\theta}_1 & \widehat{\theta}_2 \\ f_1(n) & f_2(n) \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ f_1(n) & f_2(n) \end{vmatrix}} .$$

Cette expression permet une généralisation plus large de la procédure du jackknife.

Définition 17.3 — *Soient $k+1$ estimateurs $\widehat{\theta}_1, \ldots, \widehat{\theta}_{k+1}$ du paramètre θ tels que*

$$\mathbb{E}\,(\widehat{\theta}_i - \theta) = \sum_{j=1}^{\infty} b_j(\theta) \cdot f_{ij}(n) \qquad i = 1, \ldots, k+1$$

On appelle **jackknife généralisé d'ordre k** *la statistique $\widehat{\theta}^{(k)}$ définie par*[1]

$$\begin{aligned} \widehat{\theta}^{(k)} &= \frac{\begin{vmatrix} \widehat{\theta}_1 & \cdots & \widehat{\theta}_{k+1} \\ f_{11}(n) & \cdots & f_{k+1,k}(n) \\ \vdots & & \vdots \\ f_{1k}(n) & \cdots & f_{k+1,k}(n) \end{vmatrix}}{\begin{vmatrix} 1 & \cdots & 1 \\ f_{11}(n) & \cdots & f_{k+1,k}(n) \\ \vdots & & \vdots \\ f_{1k}(n) & \cdots & f_{k+1,k}(n) \end{vmatrix}} \\ &= \frac{D(\widehat{\theta}, k+1)}{D(1, k+1)} . \end{aligned}$$

Théorème 17.3 — *Si $\widehat{\theta}^{(k)}$ existe, et si*

$$\mathbb{E}\,(\widehat{\theta}_i - \theta) = \sum_{j=1}^{k} b_j(\theta) \cdot f_{ij}(n)$$

pour $i = 1, \ldots, k+1$, alors

$$\mathbb{E}\,(\widehat{\theta}^{(k)}) = \theta .$$

Corollaire 17.1 — *Dans le cas où*

$$f_{i+1,j}(n) = \frac{1}{(n-1)^j}$$

alors

$$\mathbb{E}\,(\widehat{\theta}^{(k)} - \theta) = O(n^{-(k+1)}) .$$

1. À condition que le dénominateur ne soit pas nul...

18

Bootstrap

La méthode du **bootstrap** est considérée comme la forme la plus évoluée de jackknife. C'est une procédure de rééchantillonnage dont l'objectif est d'étudier les propriétés d'une statistique $T(X_1,..,X_n,\mathbb{P})$ fondée sur un échantillon $(X_1,..,X_n)$ d'une v.a. X de loi $\mathbb{P}$.

18.1 Principe du bootstrap

Soit une v.a.r. X de loi $\mathbb{P}$, de fonction de répartition F dépendant d'un paramètre θ, dont on possède un échantillon indépendant $\mathcal{E} = (X_1,..,X_n)$. L'idée est de rééchantillonner de façon indépendante dans $\mathcal{E}$ et d'étudier le comportement de la statistique $T(X_1,..,X_n,F)$. L'algorithme du bootstrap peut être résumé comme suit :

Phase 1 $\mathcal{E}$ sert de population de base et est munie de la loi de probabilité empirique

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

de fonction de répartition F_n ;

Phase 2 conditionnellement à $\mathbb{P}_n$, on procède dans $\mathcal{E}$ à N tirages équiprobables avec remise ; $\mathcal{E}^\star = (X_1^\star,..,X_N^\star)$ est l'échantilon ainsi obtenu et tel que

$$\forall i \in \{1,\ldots,N\},\ \ \exists j,\, 1 \leqslant j \leqslant n \ \text{ t.q. } \ X_i^\star = X_j \ ;$$

Phase 3 on approche le comportement de $T(\mathcal{E},\, F)$ par celui de $T(\mathcal{E}^\star,\, F_n) = T^\star$; $T^\star$ est la statistique *bootstrappée*.

Cette dernière phase sera souvent itérée pour donner lieu à approximation par la méthode de Monte-Carlo. Dans ce cas, la phase 2 est répétée B fois (B relativement grand), engendrant B échantillons $\mathcal{E}_k^\star \ \ k = 1,\ldots,B$, avec $\mathcal{E}_k^\star = (X_{1k},\ldots,X_{Nk})$. On observe donc B valeurs $T_k^\star = T(\mathcal{E}_k^\star,\, F_n)$ de T.

18.2 Exemples d'application

La partie la plus importante de la méthode du bootstrap concerne la détermination de la loi de $T^\star$, ou tout au moins de son espérance et de sa variance. Trois cas sont à envisager :

1° un calcul direct permet d'établir les éléments de la loi de $T^\star$;
2° on itère un très grand nombre de fois les phases 2 et 3 de l'algorithme bootstrap ;
3° on « linéarise » en recourant à un développement en série de Taylor.

18.2.1 Loi de Bernouilli

Soit X une v.a. de loi de Bernouilli $\mathcal{B}(p)$ avec $p = \mathbb{P}(X = 1)$. Soit f_n la fréquence empirique de 1 dans l'échantillon initial $\mathcal{E} = (X_1, \ldots, X_n)$, c.-à-d. l'estimateur optimal usuel de p. On considère la statistique

$$T(\mathcal{E}, F_n) = f_n - p \ .$$

Un échantillon bootstrap de taille N, $\mathcal{E}^\star = (X_1^\star, \ldots, X_n^\star)$ est une suite de N tirages équiprobables avec remise dans $\mathcal{E}$; la loi « bootstrap » de $X_i^\star$ est donc, conditionnellement à $\mathcal{E}$, une loi de Bernouilli $\mathcal{B}(f_n)$. La statistique bootstrappée sera

$$\begin{aligned} T^\star &= T(\mathcal{E}^\star, F_n) \\ &= \frac{1}{N} \sum_{i=1}^{N} X_i^\star - f_n \ . \end{aligned}$$

En notant $E_\star$ (respectivement $V_\star$) l'espérance (resp. la variance) prise par rapport à la loi bootstrap $\mathcal{B}(f_n)$, on a par un calcul direct élémentaire

$$\begin{aligned} E_\star(T^\star) &= 0 \ , \\ V_\star(T^\star) &= \frac{f_n(1-f_n)}{N} \ . \end{aligned}$$

18.2.2 Loi binomiale

Soit X une v.a. de loi de binomiale $\mathcal{B}(k,\ p)$ avec $p = \mathbb{P}(X = 1)$. On considère un échantillon $\mathcal{E} = (X_1, \ldots, X_n)$. L'estimateur $\hat{p}$ de p fondé sur $\mathcal{E}$ est

$$\begin{aligned} \hat{p} &= \frac{1}{nk} \sum_{i=1}^{k} X_i \\ &= \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i \end{aligned}$$

avec

$$\hat{p}_i = \frac{X_i}{k} \qquad i = 1, \ldots, n \ .$$

La loi « bootstrap » est une loi binomiale $\mathcal{B}(k, \hat{p})$. Si la statistique $T(\mathcal{E}, F_n)$ est $\hat{p} - p$, alors la statistique bootstrappée est

$$T^\star = \frac{1}{nk} \sum_{i=1}^{N} X_i^\star - \hat{p}$$

et

$$\begin{aligned} E_\star(T^\star) &= 0 \ , \\ V_\star(T^\star) &= \frac{\hat{p}\,(1-\hat{p})}{Nk} \ . \end{aligned}$$

18.2.3 Variance

Soit X une var de loi $\mathbb{P}$, de variance finie $V_{\mathbb{P}}(X) = \sigma^2$. Un échantillon $\mathcal{E}$ étant donné, on s'intéresse à la statistique

$$T(\mathcal{E}, F_n) \ = \ S_n^2 - \sigma^2$$

avec

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

La valeur bootstrappée de T sera la différence entre l'estimateur de la variance calculée sur $\mathcal{E}^\star$ et la vraie variance dans $\mathcal{E}$, soit

$$T^\star = \frac{1}{N-1} \sum_{i=1}^{N} (X_i^\star - \overline{X}^\star)^2 - \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

avec

$$\overline{X}^\star = \frac{1}{N} \sum_{i=1}^{N} X_i^\star \ .$$

On a alors

$$T^\star = {S_N^\star}^2 - \frac{n-1}{n} \, S_n^2 \ .$$

18.2.4 Dispersion d'une moyenne empirique

Prenons pour paramètre $\sigma(\overline{X})$, écart-type sous la loi $\mathbb{P}$ de la moyenne empirique de l'échantillon $\mathcal{E}$ extrait de cette même loi. Puisque

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} \ ,$$

σ^2 étant la variance de la v.a. X sous la loi $\mathbb{P}$, la statistique d'intérêt est

$$T(\mathcal{E}, F_n) = \frac{S_n}{\sqrt{n}} \ .$$

$\mathcal{E}^\star$ de taille n étant tiré, le bootstrap de T est

$$T^\star = \frac{S'_n}{\sqrt{n}}$$

où ${S'_n}^2$ est la vraie variance sur $\mathcal{E}$, population finie :

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

Si on prend pour T la variance de $\overline{X}$, on obtient

$$ T^\star = \frac{{S'_n}^2}{n} \ . $$

La variance bootstrappée est alors un estimateur (négativement) biaisé de $V_{\mathbb{P}}(\overline{X})$ sous la loi $\mathbb{P}$:

$$ \begin{aligned} \mathbb{E}\,(T^\star) &= \frac{1}{n} \cdot \frac{n-1}{n}\, \sigma^2 \\ &= \frac{n-1}{n}\, V_{\mathbb{P}}(\overline{X}) \\ &\leqslant V_{\mathbb{P}}(\overline{X}) \ . \end{aligned} $$

Remarque — Dans cet exemple, on peut itérer la phase 2 de l'algorithme du bootstrap. Puisque B échantillons bootstrap ont été engendrés, et puisque l'on veut connaître l'écart-type de $\overline{X}$, on va calculer la suite des $\{X_k^\star\}_{k=1,\ldots,\,B}$, $X_k^\star$ étant la moyenne empirique des observations de l'échantillon bootstrap numéro k, puis utiliser l'écart-type de cette suite pour approcher $\sigma/\sqrt{n}$, soit

$$ \left| \frac{1}{B-1} \sum_{k=1}^{B} (\overline{X}_k^\star - \overline{X}^\star)^2 \right|^{\frac{1}{2}} $$

avec

$$ \overline{X}^\star \ = \ \frac{1}{B} \sum_{k=1}^{B} \overline{X}_k^\star \ . $$

18.2.5 Coefficient de corrélation

On s'intéresse à la corrélation linéaire ρ existant entre deux variables X et Y sur la base d'un échantillon de taille n.

La valeur numérique trouvée sur $\mathcal{E}$ est $\hat{p}_0$. On désire connaître une caractéristique de précision de cet estimateur, par exemple son écart-type $\sigma(\hat{p}_0)$, sans faire référence à une loi quelconque pour le couple (X,Y). On engendre B échantillons de taille n, indépendamment et avec remise, à partir de $\mathcal{E}$, et on calcule $\rho_k^\star$, pour $k = 1, \ldots, B$. L'estimateur bootstrap de $\sigma(\hat{p}_0)$ est

$$ \left| \frac{1}{B-1} \sum_{k=1}^{B} (\hat{\rho}_k^\star - \overline{\hat{\rho}}^\star)^2 \right|^{\frac{1}{2}} \ . $$

On peut en outre tracer l'histogramme de l'échantillon $(\rho_1^\star, \ldots, \rho_B^\star)$, considérer celui-ci comme une approximation de la loi de $\hat{\rho}_0$ et comparer au graphe obtenu sous l'hypothèse de normalité. Il est également possible de déterminer un intervalle de confiance approché pour ρ.

Notons $F^\star$ la fonction de répartition de la loi bootstrap de $\rho^\star$; par exemple, si la loi bootstrap est évaluée par itération d'échantillons, on peut approcher $F^\star(x)$ par

$$ \frac{1}{B} \sum_{k=1}^{B} \mathbb{1}\!\mathrm{l}_{]-\infty,\ \rho_k^\star[}(x) \ . $$

Pour $0 \leqslant \alpha \leqslant 1$, on définit

$$\hat{\rho}_m = (F^\star)^{-1}\Big(\frac{\alpha}{2}\Big)$$

et

$$\hat{\rho}_M = (F^\star)^{-1}\Big(1 - \frac{\alpha}{2}\Big)$$

L'intervalle $[\hat{\rho}_m,\ \hat{\rho}_M]$ est un intervalle de confiance approché de niveau $1-\alpha$. Cette procédure porte le nom de **méthode des fractiles**.

Il existe une procédure dérivée, dite **méthode des fractiles corrigée du biais**. Φ étant la fonction de répartition de la loi normale centrée réduite, soit

$$\rho_0 = \Phi^{-1}\big(F^\star(\hat{\rho}_0)\big)$$

et u le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. On prend alors comme intervalle de confiance approché

$$\Big[\ (F^\star)^{-1}\big[\Phi(2\rho_0 - u)\big]\ ,\ (F^\star)^{-1}\big[\Phi(2\rho_0 + u)\big]\ \Big]\ .$$

18.3 Propriétés asymptotiques du bootstrap

Soit un rééchantillonnage de taille $N = n$. Soit $T^\star = T(\mathcal{E}^\star,\ F_n)$ où T vaut successivement

$$T_1 = \overline{X}_n - \mathbb{E}_{\,\mathbb{P}}(X)$$
$$T_2 = \frac{\overline{X}_n - \mathbb{E}_{\,\mathbb{P}}(X)}{\sigma_{\mathbb{P}}(X)}$$

et

$$T_3 = F_n^{-1}(x) - F^{-1}(x)\ .$$

Théorème 18.1 — *Nous avons :*

(i) Si $\mathbb{E}\,(X^2) < \infty$*, alors*

$$A_n = \sup_{t\in\mathbb{R}} \Big|\mathbb{P}(\sqrt{n}\,T_1 \leqslant t) - \mathbb{P}(\sqrt{n}\,T_1^\star \leqslant t)\Big| \quad \xrightarrow{\ ps\ } \quad 0 \qquad (n \to \infty)$$

avec

$$T_1^\star = \overline{X}_n^\star - \overline{X}_n\ .$$

La vitesse de convergence est donnée par

$$\limsup \frac{\sqrt{n}}{\sqrt{\log\big(\log(n)\big)}} A_n\ ,$$

qui est constante — et que nous notons c_1*.*

(ii) *Si* $\mathbb{E}\left(|X|^3\right) < \infty$, *alors*

$$B_n = \sup_{t\in\mathbb{R}} \left| \mathbb{P}(\sqrt{n}\, T_2 \leqslant t) - \mathbb{P}(\sqrt{n}\, T_2^\star \leqslant t) \right| \quad \xrightarrow{ps} \quad 0 \qquad (n \to \infty) \ ,$$

avec

$$T_2^\star = \frac{\overline{X}_n^\star - \overline{X}_n}{S_n'} \ .$$

La vitesse de convergence de B_n *est supérieure à* $1/\sqrt{n}$, *au sens où*

$$\limsup \sqrt{n}\, B_n \leqslant c_2 \ ,$$

avec c_2 *constante.*

(iii) *Si* F'' *existe au voisinage de* $F^{-1}(t)$ *et si* $F'\big(F^{-1}(t)\big) > 0$, *alors*

$$C_n = \sup_{t\in\mathbb{R}} \left| \mathbb{P}(\sqrt{n}\, T_3 \leqslant t) - \mathbb{P}(\sqrt{n}\, T_3^\star \leqslant t) \right| \quad \xrightarrow{ps} \quad 0 \qquad (n \to \infty) \ ,$$

avec

$$\limsup \frac{\sqrt[4]{n}}{\sqrt{\log\big(\log(n)\big)}}\ C_n \ .$$

Théorème 18.2 — *Soit* $\mathcal{E} = (X_i)_{i=1,\dots,\,n}$ *une suite de var i.i.d. de loi* $\mathbb{P}$, *d'espérance* m *et de variance* σ^2. *Soit* $\mathcal{E}^\star = (X_i^\star)_{i=1,\dots,\,N}$ *un échantillon bootstrap extrait de* $\mathcal{E}$. *Conditionnellement à* $\mathcal{E}$, *pour* $N \to \infty$ *et* $n \to \infty$ *:*

$$\sqrt{N}\,(\overline{X}_n^\star - \overline{X}_n) \ \xrightarrow{\mathcal{L}}\ \mathcal{N}(0,\sigma)$$

et

$$\mathbb{P}\big(|S_N^\star - \sigma| > \epsilon \mid \mathcal{E}\big) \ \xrightarrow{ps}\ 0$$

avec

$$S_N^\star = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i^\star - \overline{X}_N^\star)^2} \ .$$

Théorème 18.3 — *On suppose que la fonction de répartition* F *d'une v.a.* X *possède une unique médiane* Md *et une dérivée* f *positive et continue sur un voisinage de* Md. *Alors*

$$\sqrt{n}\,(Md_n^\star - Md_n) \ \xrightarrow{\mathcal{L}}\ \mathcal{N}\left(0\ ,\ \frac{1}{2\,f(Md)}\right) \ .$$

19

Lien

19.1 Le jackknife infinitésimal

Nous reprenons les notations du chapitre sur le jackknife. Le paramètre θ est estimé par une statistique T_n mise sous la forme

$$T\big(X_1, \ldots, X_n \; ; \; \frac{1}{n}, \ldots, \frac{1}{n}\big) \; ,$$

où l'on fait apparaître les « poids » affectant les observations de l'échantillon $\mathcal{E}(X_1, \ldots, X_n)$.

Plus généralement, on peut définir la valeur de la statistique T pour des poids quelconques

$$\omega \; = \; \big\{\omega_i \; , i = 1, \ldots, n \; ; \; \omega_i \geqslant 0 \; ; \; \sum_{i=1}^{n} \omega_i = 1\big\}$$

et on la notera $T(\mathcal{E},\omega)$.

Hypothèses — Nous supposons que :

1° $T(\mathcal{E}, \omega)$ est au moins deux fois dérivable par rapport aux poids ω_i ; u désignant les poids uniformes, $u = (1/n \; , \ldots, \; 1/n)$, on note

$$G_i = \frac{\partial \, T(\mathcal{E},\omega)}{\partial \omega_i}\bigg|_u$$

et

$$G_{ii} = \frac{\partial^2 \, T(\mathcal{E},\omega)}{\partial \omega_i^2}\bigg|_u \; .$$

2° $T(\mathcal{E}, \omega)$ est homogène, *i.e.* $\forall \lambda > 0$,

$$T(\mathcal{E}, \, \lambda\omega) = T(\mathcal{E}, \, \omega) \; .$$

L'homogénéité de $T(\mathcal{E}, \omega)$ permet de ne plus imposer la contrainte $\sum_{i=1}^{n} \omega_i = 1$, puisque dans ce cas $T(\mathcal{E}, \omega_1, \ldots, \omega_n)$ est identique à

$$T\left(\mathcal{E}, \frac{\omega_1}{\sum \omega_i}, \ldots, \frac{\omega_n}{\sum \omega_i}\right) .$$

Ces deux hypothèses impliquent

$$\sum_{i=1}^{n} G_i = 0 .$$

Soit $u_i(\epsilon)$ le vecteur de poids défini par $\omega_j = 1/n$, $j = 1, \ldots, n$ et $j \neq i$, et $\omega_i = 1/n - \epsilon$ avec $0 \leqslant \epsilon \leqslant 1/n$:

$$T\big(\mathcal{E}, u_i(\epsilon)\big) \;=\; T_i(\epsilon) .$$

Le vecteur normalisé correspondant serait

$$\omega_i \;=\; \begin{cases} \frac{1}{n(1-\epsilon)} & \text{pour} j = 1, \ldots, n,\ j \neq i \\ \\ \frac{1-n\epsilon}{n(1-\epsilon)} & \text{sinon.} \end{cases}$$

Sous cette forme, il est évident que $u_i(0) = u$ et que

$$u_i\left(\frac{1}{n}\right) = \left(\frac{1}{n-1}, \ldots, \frac{1}{n-1}, 0\ , \frac{1}{n-1}, \ldots, \frac{1}{n-1}\right) .$$

Partant,

$$\begin{aligned} T_i(0) &= T(\mathcal{E}, u) \\ &= T_n\ , \\ T_i\left(\frac{1}{n}\right) &= T_{n-1}^{i}\ . \end{aligned}$$

On sait que l'estimateur du jackknife de $T_n = T(\mathcal{E}, u)$ est

$$J(T_n) = n\, T_n - \frac{n-1}{n} \sum_{i=1}^{n} T_{n-1}^{i}$$

ou

$$T_n - J(T_n) = (n-1)\left(\frac{1}{n} \sum_{i=1}^{n} T_{n-1}^{i} - T_n\right) .$$

Par analogie, considérons la quantité $B(\epsilon)$ définie comme le biais précédent $T_n - J(T_n)$ par

$$B(\epsilon) = \frac{1-\epsilon}{n\epsilon^2}\left(\frac{1}{n} \sum_{i=1}^{n} T_i(\epsilon) - T_n\right) .$$

Par un développement de Taylor à l'ordre 2 de $T_i(\epsilon)$ autour de $\epsilon = 0$, on obtient

$$T_i(\epsilon) = T_i(0) + \epsilon\left(\frac{\partial\, T_i(\epsilon)}{\partial \epsilon}\right)_{\epsilon=0} + \frac{\epsilon^2}{2}\left(\frac{\partial^2\, T_i(\epsilon)}{\partial \epsilon^2}\right)_{\epsilon=0} + \mathrm{o}(\epsilon^2) .$$

En remarquant que

$$\begin{aligned}\left(\frac{\partial\, T_i(\epsilon)}{\partial\epsilon}\right)_{\epsilon=0} &= -G_i\ ,\\ \left(\frac{\partial^2\, T_i(\epsilon)}{\partial\epsilon^2}\right)_{\epsilon=0} &= G_{ii}\ .\end{aligned}$$

on a une approximation du biais $B(\epsilon)$:

$$B(\epsilon) = \frac{1-\epsilon}{n\epsilon^2}\Big(-\frac{\epsilon}{n}\sum_i G_i + \frac{\epsilon^2}{2n}\sum_i G_{ii}\Big) + \mathrm{o}(\epsilon^2)\ ,$$

soit, puisque $\sum_i G_i = 0$,

$$B(\epsilon) = \frac{1-\epsilon}{n\epsilon^2}\sum_i G_{ii}\big(1+\mathrm{o}(1)\big)$$

et

$$\begin{aligned}\lim_{\epsilon\to 0} B(\epsilon) &= \frac{1}{2n}\sum_i G_{ii}\\ &= B(0)\ .\end{aligned}$$

Définition 19.1 — *On appelle* **jackknife infinitésimal** *de T_n la statistique $JI(T_n)$ définie par*

$$JI(T_n) = T_n - B(0)\ .$$

De façon analogue, on peut définir une variance infinitésimale. En effet,

$$\frac{S^2(\epsilon)}{n} = \frac{1-\epsilon}{(n\epsilon)^2}\sum_{i=1}^n \big(T_i(\epsilon) - \overline{T}(\epsilon)\big)^2\ ,$$

où

$$\overline{T}(\epsilon) = \frac{1}{n}\sum_{i=1}^n T_i(\epsilon)\ .$$

Cette expression coïncide avec celle donnée pour $JV(T_n)$ lorsque $\epsilon = 1/n$ (jackknife classique). Un calcul simple permet d'établir que

$$\lim_{\epsilon\to 0}\frac{S^2(\epsilon)}{n} = \frac{1}{n^2}\sum_{i=1}^n G_i^2\ .$$

Définition 19.2 — *On appelle* **estimateur du jackknife infinitésimal de la variance** *de T_n la statistique*

$$JIV(T_n) = \frac{1}{n^2}\sum_{i=1}^n G_i^2\ .$$

19.2 Linéarisation

Comme précédemment, soit $\mathcal{E} = (X_1, \ldots, X_n)$ l'échantillon initial et $\mathcal{E}^\star = (X_1^\star, \ldots, X_n^\star)$ l'échantillon bootstrappé. On a vu que la valeur de la statistique $T(\mathcal{E}^\star, F_n)$ peut être écrite sous la forme

$$T^\star = \phi(\mathbb{P}^\star) \ ,$$

où $\mathbb{P}^\star$ est issu d'une observation de la loi multinomiale

$$\mathcal{M}\left(n \ ; \ \frac{1}{n}, \ldots, \frac{1}{n}\right) \ .$$

$\mathbb{P}^\star$ suit symboliquement une loi

$$\frac{1}{n} \ \mathcal{M}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right) \ ,$$

et $\mathbb{P}^\star$ définit un vecteur de poids $(\mathbb{P}_1^\star, \ldots, \mathbb{P}_n^\star)$, où $\mathbb{P}_i$ est la fréquence d'apparition de X_i dans l'échantillon $\mathcal{E}^\star$.

Les propriétés de le loi multinomiale donnent

$$\begin{array}{rcl} \mathbb{E}_\star(\mathbb{P}^\star) & = & \dfrac{1}{n} \, e \\ & = & u \ , \end{array}$$

où $e = (1, \ldots, 1)$ est le vecteur ligne $1 \times n$ unitaire, et

$$\mathbb{V}_\star(\mathbb{P}^\star) = \frac{1}{n^2} \ I_n - \frac{1}{n^3} \ e^t e$$

où I_n est ma matrice identité d'ordre n.

Effectuons un développement de Taylor de $T^\star = \phi(\mathbb{P}^\star)$ au voisinage de u :

$$\phi(\mathbb{P}^\star) = \phi(u) + (\mathbb{P}^\star - u) \ D^t + \frac{1}{2}(\mathbb{P}^\star - u) \ H \ (\mathbb{P}^\star - u)^t \ ,$$

où D est le vecteur-ligne gradient d'élément courant

$$D_i = \left(\frac{\partial \, \phi(\mathbb{P}^\star)}{\partial \mathbb{P}_i^\star} \right)_{\mathbb{P}^\star = u} \ ,$$

et H est la matrice des dérivées secondes d'élément H_{ij} égal à

$$H_{ij} = \left(\frac{\partial^2 \, \phi(\mathbb{P}^\star)}{\partial \mathbb{P}_i^\star \ \partial \mathbb{P}_j^\star} \right)_{\mathbb{P}^\star = u} \ .$$

En imposant l'homogénéité de $\phi(\mathbb{P}^\star)$, on a

$$e \, H \, e^t = 0 \ .$$

On obtient, puisque $\mathbb{E}_\star(\mathbb{P}^\star - u) = 0$,

$$\begin{array}{rcl} \mathbb{E}_\star\big[\phi(\mathbb{P}^\star) - \phi(u)\big] & = & \dfrac{1}{2}\mathbb{E}\left[(\mathbb{P}^\star - u) \ H \ (\mathbb{P}^\star - u)^t\right] \\ & = & \dfrac{1}{2}\mathrm{tr}\big[H \, V_\star(\mathbb{P}^\star)\big] \ . \end{array}$$

Soit

$$\mathbb{E}_\star\big[\phi(\mathbb{P}^\star) - \phi(u)\big] \approx \frac{1}{2n^2} \sum_{i=1}^{n} H_{ii} \ .$$

Remarque — $\phi(\mathbb{P}^\star) - \phi(u)$ est égal à $T^\star - T_n$, différence entre la statistique bootstrappée et la statistique d'origine.

De même, le calcul de $V_\star\big(\phi(\mathbb{P}^\star)\big)$ fournit

$$\begin{aligned} V_\star\big(\phi(\mathbb{P}^\star)\big) &\approx D\, V_\star(\mathbb{P}^\star)\, D^t \\ &= \frac{1}{n^2}\sum_{i=1}^n D_i^2 \ . \end{aligned}$$

Les deux relations précédentes écrites en faisant intervenir $T^\star$ et T_n s'expriment de la façon suivante :

$$\begin{aligned} \mathbb{E}_\star(T^\star - T_n) &\approx \frac{1}{2n^2}\sum_{i=1}^n H_{ii} \ , \\ V_\star(T^\star - T_n) &\approx \frac{1}{n^2}\sum_{i=1}^n D_i^2 \ , \end{aligned}$$

expressions similaires à celles qui ont été établies précédemment pour le jackknife infinitésimal.

En outre, la logique du bootstrap conduit à approximer le biais $\mathbb{E}_{\mathbb{P}}(T_n - \theta)$ par

$$\frac{1}{2n^2}\sum_{i=1}^n H_{ii}$$

et la variance $\mathbb{V}_{\mathbb{P}}(T_n - \theta)$ par

$$\frac{1}{n^2}\sum_{i=1}^n D_i^2 \ .$$

Annexe A

Jeux de données

A.1 Spots publicitaires

Table A.1 — Consommation des ménages soumis à une campagne publicitaire.

Region	Pub1	Pub2	Pub3	Pub4	Taille
1	12.35	21.86	14.43	21.44	1
1	20.52	42.17	22.26	31.21	2
1	30.85	49.61	23.99	40.09	3
1	39.35	63.65	36.98	55.68	4
1	48.87	73.75	42.13	65.81	5
1	58.01	85.95	54.19	76.61	6
2	28.26	13.76	14.44	30.78	1
2	37.67	24.59	29.63	45.75	2
2	44.70	37.30	38.27	56.37	3
2	57.54	49.53	51.59	70.19	4
2	67.57	59.25	59.09	79.81	5
2	77.70	67.68	71.69	94.23	6
3	10.97	0.00	2.90	6.46	1
3	26.70	2.41	17.28	18.61	2
3	36.81	16.10	19.62	30.14	3
3	51.34	22.71	29.53	39.12	4
3	62.69	30.19	38.57	51.15	5
3	72.68	41.64	48.20	59.11	6
4	0.00	11.90	4.48	27.62	1
4	4.52	27.75	18.01	42.63	2
4	13.71	42.22	21.96	59.20	3
4	27.91	56.06	34.42	74.92	4
4	38.57	66.16	40.14	92.37	5
4	42.71	78.71	57.06	98.02	6
5	13.11	8.00	10.90	14.36	1
5	16.89	18.27	28.22	26.37	2
5	27.99	27.72	38.62	34.15	3
5	36.35	42.04	48.31	54.02	4
5	48.85	48.50	60.23	59.90	5
5	61.97	59.92	71.39	74.79	6

Pour le traitement par S-Plus, le tableau doit être réécrit sous la forme A.2, ce qui peut se faire grâce au code suivant :

```
> temp <- consomenage[rep(1:30, rep(4,30)), c(1,6)]
> ymat <- data.matrix(consomenage[, paste("pub",1:4, sep="")])
> consomenage2 <- cbind(temp, Pub = ordered(rep(paste("pub", 1:4, sep = ""), 30)),
                                Conso = as.vector(t(ymat)))
```

TABLE A.2 — Données pour S-Plus.

Obs.	Region	Taille	Pub	Conso
1	1	1	pub1	12.35
2	1	2	pub1	20.52
3	1	3	pub1	30.85
4	1	4	pub1	39.35
5	1	5	pub1	48.87
6	1	6	pub1	58.01
⋮	⋮	⋮	⋮	⋮
115	5	1	pub4	14.36
116	5	2	pub4	26.37
117	5	3	pub4	34.15
118	5	4	pub4	54.02
119	5	5	pub4	59.90
120	5	6	pub4	74.79

A.2 Moustiques

TABLE A.3 — Données concernant des moustiques placés dans des cages.

Obs.	Cage	Moust	Valeur	Mesure
1	1	1	58.5	1
2	1	1	59.5	1
3	1	2	77.8	2
4	1	2	80.9	2
5	1	3	84.0	3
6	1	3	83.6	3
7	1	4	70.1	4
8	1	4	68.3	4
9	2	1	69.8	5
10	2	1	69.8	5
11	2	2	56.0	6
12	2	2	54.5	6
13	2	3	50.7	7
14	2	3	49.3	7
15	2	4	63.8	8
16	2	4	65.8	8
17	3	1	56.6	9
18	3	1	57.5	9
19	3	2	77.8	10
20	3	2	79.2	10
21	3	3	69.9	11
22	3	3	69.2	11
23	3	4	62.1	12
24	3	4	64.5	12