Statistique appliquée

Université Pierre et Marie Curie
Maîtrise de Mathématiques

Année 2006/2007

A. Tsybakov

Préambule

Ce polycopié s'adresse aux étudiants ayant suivi un cours d'intégration et un premier cours de probabilités. La Partie 1 contient un bref rappel de quelques notions de base de probabilités, souvent sans démonstration (les manuels de probabilités conseillés sont l'ouvrage de N.Bouleau *Probabilités de l'ingénieur, variables aléatoires et simulation* et le polycopié du cours de J.Lacroix et P.Priouret *Probabilités approfondies*, Chapitres 1 – 3). La Partie 1 présente aussi les résultats probabilistes utilisés dans la Statistique qui généralement ne sont pas exposés dans les cours de probabilités (théorèmes de continuité, régression et corrélation, lois dérivées de la normale multivariée, etc). La Partie 2 introduit les principales notions de la Statistique et décrit quelques méthodes classiques de l'estimation, de tests d'hypothèse et de construction des intervalles de confiance. Enfin, la Partie 3 contient l'application des méthodes statistiques dans les 3 modèles concrets multi-dimensionnels, a savoir, celles de l'analyse en composantes principales, de la régression linéaire multivariée et de l'analyse discriminante (classification).

Les parties marquées par le signe * peuvent être omises en première lecture et ne feront pas l'objet de question aux examens.

Table des matières

Partie 1

Rappels et compléments de probabilités

1

Quelques rappels de probabilités

1.1. Caractéristiques des variables aléatoires

Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité, où $(\Omega,\ \mathcal{A})$ est un espace mesurable et P est une mesure de probabilité sur $\mathcal{A}$. Une *variable aléatoire* (v.a.) X est une fonction mesurable $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ où $\mathcal{B}$ est la tribu borélienne de $\mathbb{R}$. Parfois on écrit $X = X(\omega)$ pour souligner le fait qu'il s'agit d'une fonction de $\omega \in \Omega$.

Définition 1.1. *La fonction de répartition (f.d.r.) d'une variable aléatoire X est la fonction $F : \mathbb{R} \to [0, 1]$ définie par $F(x) = P(X \leq x) = P(\omega : X(\omega) \leq x)$.*

C'est une fonction monotone croissante, continue à droite et telle que $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to\infty} F(x) = 1$. La fonction F sera aussi appelée *la loi* (ou *la distribution*) de X. On va distinguer entre deux principaux types de variables aléatoires : *les variables discrètes* et *les variables continues.*

Variable discrète : X est une variable aléatoire dont les valeurs appartiennent à un ensemble fini ou dénombrable. La variable de Poisson est un exemple de variable discrète dont l'ensemble de valeurs est dénombrable : pour $\theta > 0$ la loi de X est donnée par

$$P(X = k) = \frac{\theta^k}{k!} e^{-\theta}, \quad k = 0, 1, 2, ...$$

On dit alors que X suit la loi de Poisson $\mathcal{P}(\theta)$. La fonction de répartition de X est représentée dans la Figure 1.1. La f.d.r. d'une variable aléatoire discrète est une fonction en escalier.

Variable continue : X est une variable aléatoire dont la loi admet une densité $f \geq 0$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, i.e.

$$F(x) = \int_{-\infty}^{x} f(t) dt,$$

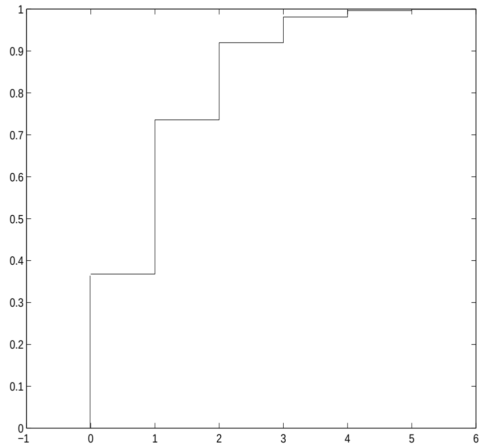

Figure 1.1. La f.d.r. de la loi de Poisson

pour tout $x \in \mathbb{R}$. Dans ce cas la f.d.r. F de X est différentiable presque partout sur $\mathbb{R}$ et la densité de probabilité de X est égale à la dérivée

$$f(x) = F'(x)$$

presque partout. On note que $f(x) \geq 0$ pour tout $x \in \mathbb{R}$ et

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Exemple 1.1. a) *Loi normale (gaussienne)* $\mathcal{N}(\mu, \sigma^2)$ est la loi de densité

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$

où $\mu \in \mathbb{R}$ et $\sigma > 0$. Si $\mu = 0$, $\sigma^2 = 1$, la loi $\mathcal{N}(0,1)$ est dite *loi normale standard*. Dans la suite, l'écriture $X \sim \mathcal{N}(\mu, \sigma^2)$ signifie que la v.a. X suit la loi $\mathcal{N}(\mu, \sigma^2)$.

b) *Loi uniforme sur l'intervalle* $[a, b]$, $-\infty < a < b < \infty$, est la loi notée $U[a, b]$, de densité

$$f(x) = (b-a)^{-1} \mathbb{1}_{[a,b]}(x), \quad x \in \mathbb{R},$$

où $\mathbb{1}_A(\cdot)$ désigne la fonction indicatrice de l'ensemble A :

$$\mathbb{1}_A(x) = I\{x \in A\} = \begin{cases} 1 & \text{si } x \in A, \\ 0 & \text{sinon.} \end{cases}$$

c) *Loi exponentielle* $\mathcal{E}(\theta)$ est la loi de densité

$$f(x) = \theta^{-1} e^{-x/\theta} \mathbb{1}_{[0,+\infty[}(x),$$

où $\theta > 0$. La fonction de répartition de $\mathcal{E}(\theta)$ est

$$F(x) = (1 - e^{-x/\theta}) \mathbb{1}_{[0,+\infty[}(x).$$

Les lois des variables discrètes sont entièrement définies par les probabilités $P(X = \cdot)$, les lois des variables continues – par leur densité $f(\cdot)$. Certaines caractéristiques scalaires de la fonction de répartition (ses fonctionnelles) sont importantes pour la description du comportement des variables aléatoires. Des exemples de telles fonctionnelles sont les moments et les quantiles.

1.1.1. Moments. La *moyenne* (ou l'espérance mathématique) d'une variable aléatoire X est définie par :

$$\mu = E(X) = \int_{-\infty}^{\infty} x dF(x) = \begin{cases} \sum_i iP(X = i) & \text{si } X \text{ est une v.a. discrète,} \\ \int xf(x)dx & \text{si } X \text{ est une v.a. continue.} \end{cases}$$

Le *moment d'ordre* k $(k = 1, 2, ...)$ de X est défini par :

$$\mu_k = E(X^k) = \int_{-\infty}^{\infty} x^k dF(x),$$

ainsi que le *moment centré d'ordre* k :

$$\mu'_k = E((X - \mu)^k) = \int_{-\infty}^{\infty} (x - \mu)^k dF(x).$$

Un cas particulier est la *variance* $\sigma^2 (= \mu'_2 =$ moment centré d'ordre 2) :

$$\sigma^2 = \mathrm{Var}(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2.$$

La racine carrée de la variance s'appelle *écart-type* de X : $\sigma = \sqrt{\mathrm{Var}(X)}$.

Le moment absolu $\bar{\mu}_k$ *d'ordre* k de X est

$$\bar{\mu}_k = E(|X|^k)$$

alors que le *moment absolu centré d'ordre* k est défini par :

$$\bar{\mu}'_k = E(|X - \mu|^k).$$

Bien évidemment, ces définitions supposent l'existence des intégrales respectives : par conséquent, toutes les lois ne possèdent pas nécessairement des moments.

Exemple 1.2. *Non-existence de tous les moments.* Soit X une variable aléatoire de densité de probabilité

$$f(x) = \frac{c}{1 + |x| \log^2 |x|}, \quad x \in \mathbb{R},$$

où la constante $c > 0$ est telle que $\int f = 1$. Alors $E(|X|^a) = \infty$ pour tout $a > 0$.

La proposition suivante s'obtient facilement.

Proposition 1.1. *Soit ξ une variable aléatoire telle que $E(\xi^2) < \infty$. Alors, pour tout c réel,*

$$\begin{aligned} E((\xi - c)^2) &= (E(\xi) - c)^2 + E((\xi - E(\xi))^2) \\ &= (E(\xi) - c)^2 + \mathrm{Var}(\xi). \end{aligned}$$

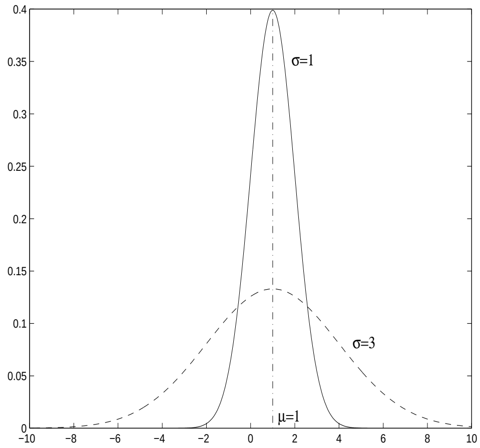

Figure 1.2. La loi normale $\mathcal{N}(\mu, \sigma^2)$
(σ "grand" – beaucoup de dispersion,
σ "petit" – peu de dispersion)

Corollaire 1.1. (Propriété extrémale de la moyenne.) *Soit ξ une variable aléatoire telle que $E(\xi^2) < \infty$. Alors, $\mu = E(\xi)$ si et seulement si*

$$E((\xi - \mu)^2) = \min_{c \in \mathbb{R}} E((\xi - c)^2).$$

La moyenne est utilisée pour caractériser la localisation (position) d'une loi de probabilité. La variance caractérise la dispersion (l'échelle) d'une loi. Une illustration graphique de ces propriétés est donnée dans la Figure 1.2.

Soit F la f.d.r. de la variable aléatoire X dont la moyenne et l'écart-type sont μ et σ. Par transformation affine, on obtient la variable $X_0 = (X - \mu)/\sigma$, telle que $E(X_0) = 0,\ E(X_0^2) = 1$ (la variable *standardisée*). Si F_0 est la f.d.r. de X_0, alors $F(x) = F_0(\frac{x-\mu}{\sigma})$. Si X est une v.a. continue, la densité de X s'écrit

$$f(x) = \frac{1}{\sigma} f_0\Big(\frac{x - \mu}{\sigma}\Big),$$

où f_0 est la densité de X_0. En général, pour définir la loi standardisée F_0 et pour avoir la représentation $F(x) = F_0(\frac{x-\mu}{\sigma})$, il n'est pas nécessaire que la moyenne et la variance existaient. Ceci est fait uniquement pour souligner que F dépend des paramètres *de localisation* (ou de position) μ et *d'échelle* σ. Par exemple, pour la famille des densités de Cauchy dépendant de μ, σ :

$$f(x) = \frac{1}{\pi\sigma(1 + [(x - \mu)/\sigma]^2)},$$

la densité standardisée est $f_0(x) = \frac{1}{\pi(1+x^2)}$. Pourtant, l'espérance et la variance de la loi de Cauchy n'existent pas.

Le problème d'analyse suivant est lié aux moments. Soit F une f.d.r. dont tous les moments sont finis. Étant donnée la suite $\{\mu_k\}$, $k = 1, 2, \ldots$, de tous les moments de F, peut-on reconstituer F ? La réponse est généralement négative. Il existe néanmoins des cas pour lesquels la reconstitution est possible, notamment sous l'hypothèse très forte que

$$\limsup_{k\to\infty} \frac{\bar{\mu}_k^{1/k}}{k} < \infty$$

($\bar{\mu}_k$ étant le k-ème moment absolu). Cette hypothèse est vérifiée, par exemple, si X est une variable aléatoire bornée.

1.1.2. Quantiles. Soit X une variable aléatoire avec la f.d.r. F continue et strictement croissante. *Le quantile d'ordre* p, $0 < p < 1$, de la loi F est alors défini comme solution q_p de l'équation

$$F(q_p) = p. \tag{1.1}$$

On remarque que, pour F strictement croissante et continue, la solution existe et elle est unique, donc dans ce cas le quantile q_p est bien défini par (1.1). Si F n'est pas strictement croissante ou n'est pas continue, on peut modifier la définition (1.1) de la façon suivante.

Définition 1.2. *Soit F une f.d.r. Le quantile q_p d'ordre p de F est la valeur*

$$q_p = \frac{1}{2}\left(\inf\{q : F(q) > p\} + \sup\{q : F(q) < p\}\right).$$

Si p est tel que (1.1) n'a pas de solution (F a un saut), q_p est le point de saut. Si (1.1) admet un intervalle de solutions (p correspond à un "plateau" du graphique de F), alors q_p est le milieu de cet intervalle.

La médiane M de la loi de X est le quantile d'ordre 1/2 :

$$M = q_{1/2}.$$

Notons que $P(X \ge M) \ge 1/2$ et $P(X \le M) \ge 1/2$. Si F est continue, $F(M) = 1/2$.

Les quartiles sont la médiane et les quantiles $q_{1/4}$ et $q_{3/4}$ d'ordre 1/4 et 3/4.

Le pourcentile de $l\,\%$, $0 < l < 100$, de la loi F est le quantile q_p d'ordre $p = l/100$.

La médiane caractérise la position (localisation) d'une loi de probabilités, alors que la différence $\mathcal{I} = q_{3/4} - q_{1/4}$ (dite *intervalle interquartile*) est souvent utilisée comme une caractéristique de l'échelle. Ce sont des analogues à la moyenne μ et à l'écart-type σ respectivement. Mais à la différence de ceux-ci, la médiane et l'intervalle interquartile sont définis *pour toutes les lois F.*

Proposition 1.2. (Propriété extrémale de la médiane.) *Soit ξ une variable aléatoire telle que $E(|\xi|) < \infty$. Alors,*

$$E(|\xi - a|) = \min_{c\in\mathbb{R}} E(|\xi - c|)$$

pour tout $a \in \mathbb{R}$ vérifiant $P(\xi \ge a) \ge 1/2$ et $P(\xi \le a) \ge 1/2$. En particulier,

$$E(|\xi - M|) = \min_{c\in\mathbb{R}} E(|\xi - c|),$$

où M est la médiane de la loi de ξ.

Preuve. Montrons que $E(|\xi - c|) \geq E(|\xi - a|)$ pour tout $c \in \mathbb{R}$. Sans perte de généralité, supposons que $c > a$. On a alors :

$$\begin{array}{ll} |\xi - c| \geq |\xi - a| + (c - a) & \text{si } \xi \leq a, \\ |\xi - c| \geq |\xi - a| & \text{si } a < \xi \leq (a + c)/2, \\ |\xi - c| \geq |\xi - a| - (c - a) & \text{si } \xi > (a + c)/2. \end{array}$$

Par conséquent,

$$E(|\xi - c|) \geq E(|\xi - a|) + (c - a)\Big[P(\xi \leq a) - P(\xi > (a + c)/2)\Big].$$

Il reste à remarquer que $P(\xi \leq a) \geq P(\xi > (a + c)/2)$ pour conclure. En effet, si $P(\xi \leq a) < P(\xi > (a + c)/2)$, en utilisant le fait que $P(\xi \leq a) \geq 1/2$, on obtient $P(\xi \leq a) + P(\xi > (a + c)/2) > 1$, ce qui est impossible. ∎

1.1.3. Mode d'une loi. Si F est une loi discrète, on appelle *mode* de la loi F une valeur k^* telle que

$$P(X = k^*) = \max_k P(X = k).$$

Si F admet une densité f par rapport à la mesure de Lebesgue, le *mode* est défini comme une valeur x^* telle que

$$f(x^*) = \max_x f(x).$$

Evidemment, un mode n'est pas toujours unique. Une densité f est dite *unimodale* si x^* est un unique maximum *local* (et donc global) de f. De façon analogue, on appelle f densité *bimodale* (ou *multimodale*) si elle a deux (respectivement, plusieurs) maxima locaux. Ce lexique n'est pas très précis, car même si le maximum global de la densité f est unique (il y a un seul mode au sens propre), on appelle f multimodale à condition qu'elle possède d'autres maxima locaux. Ainsi que la moyenne et la médiane, le mode renseigne sur la position (la localisation) d'une loi. Le mode peut se révéler intéressant principalement au cas unimodal.

1.1.4. Caractéristiques d'asymétrie et d'aplatissement.

Définition 1.3. *La loi de X (la f.d.r. F) est dite symétrique par rapport à zéro (ou tout simplement symétrique) si $F(x) = 1 - F(-x)$ pour tout $x \in \mathbb{R}$ ($f(x) = f(-x)$ dans le cas continu).*

Définition 1.4. *La loi de X (la f.d.r. F) est dite symétrique par rapport à $\mu \in \mathbb{R}$ si*

$$F(\mu + x) = 1 - F(\mu - x)$$

pour tout $x \in \mathbb{R}$ ($f(\mu + x) = f(\mu - x)$ dans le cas continu). Autrement dit, la f.d.r $F(x + \mu)$ est symétrique par rapport à zéro.

Exercice 1.1. Montrer que si la loi F est symétrique par rapport à μ et $E(|X|) < \infty$, sa médiane et sa moyenne vérifient $M = E(X) = \mu$. Si, en outre, F admet une densité unimodale, alors moyenne = médiane = mode.

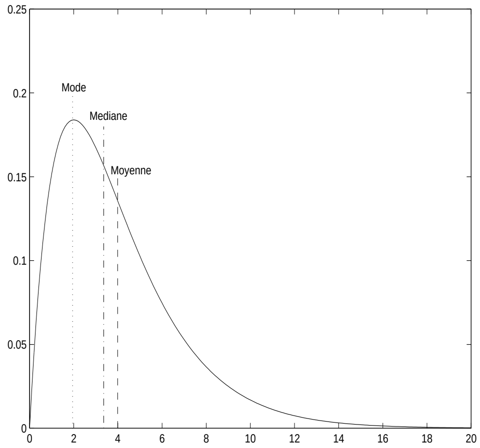

Figure 1.3. Le mode, la médiane et la moyenne d'une loi

Exercice 1.2. Si F est symétrique et tous les moments absolus $\bar{\mu}_k$ existent, alors les moments $\mu_k = 0$ pour tout k impair. Si F est symétrique par rapport à μ et tous les moments absolus $\bar{\mu}_k$ existent, alors $\mu'_k = 0$ pour tout k impair (par exemple, $\mu'_3 = 0$).

On peut qualifier les lois asymétriques comme étant "proches" ou "éloignées" de distributions symétriques. A cette fin, on introduit (pour toute loi de probabilité vérifiant $E(|X|^3) < \infty$) le *coefficient d'asymétrie* (en anglais *"skewness"*)

$$\alpha = \frac{\mu'_3}{\sigma^3}.$$

On remarque que $\alpha = 0$ pour une f.d.r. symétrique avec $E(|X|^3) < \infty$. Notons que le réciproque n'est pas vrai : la condition $\alpha = 0$ n'implique pas la symétrie de la loi.

Exercice 1.3. Donner un exemple de densité non-symétrique avec $\alpha = 0$.

Notons le rôle de σ dans la définition de α : supposons, par exemple, que la densité $f_0(x)$ de X satisfait $\int x f_0(x)dx = 0$ et $\int x^2 f_0(x)dx = 1$ et $\alpha_0 = \mu'_{3,0} = \int x^3 f_0(x)dx$. Pour $\sigma > 0$, $\mu \in \mathbb{R}$, la fonction

$$f(x) = \frac{1}{\sigma} f_0\Big(\frac{x-\mu}{\sigma}\Big),$$

est la densité de la variable $\sigma X + \mu$. Donc $\mathrm{Var}(\sigma X + \mu) = \sigma^2$ et $\mu'_3 = \int (x-\mu)^3 f(x)dx = \sigma^3 \mu'_{3,0}$. En calculant $\alpha = \frac{\mu'_3}{\sigma^3}$ on observe que $\alpha = \alpha_0$. Autrement dit, *le coefficient d'asymétrie α est invariant par rapport aux transformations affines* (d'échelle et de position) de la variable aléatoire X.

Le coefficient α est une mesure controversée : on ne peut pas toujours affirmer que $\alpha > 0$ si la loi est "asymétrique vers la droite" et $\alpha < 0$ si la loi est "asymétrique vers la gauche". Les notions d'asymétrie "vers la droite" ou "vers la gauche" ne sont pas définies rigoureusement.

Coefficient d'aplatissement (en anglais *"kurtosis"*) β est défini de la façon suivante : si le 4ème moment centré μ_4' de la variable aléatoire X existe, alors

$$\beta = \frac{\mu_4'}{\sigma^4} - 3.$$

EXERCICE 1.4. Montrer que, pour la loi normale $\mathcal{N}(\mu, \sigma^2)$, $\mu_4'/\sigma^4 = 3$ et $\beta = 0$.

On note que, comme le coefficient d'asymétrie α , *le kurtosis β est invariant par rapport aux transformations affines.*

Le coefficient β est le plus souvent calculé pour avoir une idée intuitive sur les "queues" de la loi de X. On utilise le vocabulaire suivant : on dit que la loi F a les "queues lourdes" si

$$Q(b) = P(|X| \geq b) \ \ (= \int_{|x|\geq b} f(x)dx \ \text{ dans le cas continu})$$

décroît lentement quand $b \to \infty$, par exemple, de façon polynômiale (comme $1/b^r$ avec $r > 0$). On dit que "les queues sont légères" si $Q(b)$ décroît rapidement (exemple : décroissance exponentielle). Pour la loi normale $\mathcal{N}(0, 1)$, on a : $Q(b) = O(e^{-b^2/2})$, ce qui correspond à $\beta = 0$. Très souvent, si $\beta > 0$, les queues de la loi en question sont plus lourdes que celles de la loi normale et, si $\beta < 0$ (on dit dans ce cas que la loi est *leptokurtique*), elles sont plus légères que celles de la loi normale.

Notons aussi que, pour toute loi de probabilité telle que β est bien défini (i.e., $E(|X|^4) < \infty$), on a : $\beta \geq -2$ (voir le paragraphe suivant).

EXEMPLE 1.3. a) Le kurtosis β de la loi uniforme $U[0, 1]$ est égal à $-1, 2$ (queues très légères). C'est une loi leptokurtique.

b) Si la densité de la loi $f(x) \sim |x|^{-5}$ quand $|x|$ tend vers ∞, σ^2 est fini mais $\mu_4' = +\infty$, ce qui implique $\beta = +\infty$ (queues très lourdes). Pour la loi de Cauchy, $\sigma^2 = +\infty$ et $\mu_4' = +\infty$, donc le kurtosis β n'est pas défini.

1.2. Rappel de quelques inégalités

Proposition 1.3. (Inégalité de Markov.) *Soit $h(\cdot)$ une fonction positive croissante et soit X une v.a. telle que $E(h(X)) < \infty$. Alors pour tout $a \in \mathbb{R}$ tel que $h(a) > 0$,*

$$P(X \geq a) \leq \frac{E(h(X))}{h(a)}. \tag{1.2}$$

Preuve. Comme $h(\cdot)$ est une fonction croissante,

$$\begin{aligned} P(X \geq a) \leq P\big(h(X) \geq h(a)\big) &= \int \mathbb{1}_{\{h(x)\geq h(a)\}}\, dF(x) \\ &= E(\mathbb{1}_{\{h(X)\geq h(a)\}}) \leq E\left(\frac{h(X)}{h(a)}\, \mathbb{1}_{\{h(X)\geq h(a)\}}\right) \leq \frac{E(h(X))}{h(a)}. \end{aligned}$$

∎

Corollaire 1.2. (Inégalité de Tchebychev.) *Soit X une v. a. telle que $E(X^2) < \infty$. Alors, pour tout $a > 0$,*

$$P(|X| \geq a) \leq \frac{E(X^2)}{a^2}, \quad P(|X - E(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Preuve. Il suffit de poser $h(t) = t^2$ et d'appliquer (1.2) aux variables aléatoires $|X|$ et $|X - E(X)|$ respectivement. ∎

Proposition 1.4. (Inégalité de Hölder.) *Soit $1 < r < \infty$, $1/r + 1/s = 1$. Soient ξ et η deux variables aléatoires telles que $E(|\xi|^r) < \infty$ et $E(|\eta|^s) < \infty$. Alors $E(|\xi\eta|) < \infty$ et*

$$E(|\xi\eta|) \leq [E(|\xi|^r)]^{1/r}[E(|\eta|^s)]^{1/s}.$$

Preuve. On note d'abord que pour tout $a > 0, b > 0$, par concavité de la fonction $\log t$,

$$(1/r)\log a + (1/s)\log b \leq \log(a/r + b/s),$$

ce qui est équivalent à :

$$a^{1/r}b^{1/s} \leq a/r + b/s.$$

Posons ici $a = |\xi|^r/E(|\xi|^r)$, $b = |\eta|^s/E(|\eta|^s)$ (on suppose pour l'instant que $E(|\xi|^r) \neq 0$, $E(|\eta|^s) \neq 0$), ce qui donne

$$|\xi\eta| \leq [E(|\xi|^r)]^{1/r}[E(|\eta|^s)]^{1/s}\left(|\xi|^r/rE(|\xi|^r) + |\eta|^s/sE(|\eta|^s)\right).$$

On conclut en prenant l'espérance et en utilisant le fait que $1/r + 1/s = 1$. Si $E(|\xi|^r) = 0$ ou $E(|\eta|^s) = 0$, alors $\xi = 0$ (p.s) ou $\eta = 0$ (p.s.), et l'inégalité est triviale. ∎

Corollaire 1.3. (Inégalité de Lyapounov.) *Soit $0 < v < t$ et soit X une variable aléatoire telle que $E(|X|^t) < \infty$. Alors $E(|X|^v) < \infty$ et*

$$[E(|X|^v)]^{1/v} \leq [E(|X|^t)]^{1/t}. \tag{1.3}$$

Preuve. On applique l'inégalité de Hölder avec $\xi = X^v$, $\eta = 1$, $r = t/v$.

En utilisant l'inégalité (1.3) avec $v = 2$, $t = 4$ et $|X - E(X)|$ au lieu de $|X|$ on obtient $\mu_4'/\sigma^4 \geq 1$. Le coefficient d'aplatissement β vérifie donc l'inégalité $\beta \geq -2$.

L'inégalité de Lyapounov implique la chaîne des inégalités entre les moments absolus :

$$E(|X|) \leq [E(|X|^2)]^{1/2} \leq \ldots \leq [E(|X|^k)]^{1/k}.$$

Proposition 1.5. (Inégalité de Jensen.) *Soit $g(\cdot)$ une fonction convexe et soit X une variable aléatoire telle que $E(|X|) < \infty$. Alors*

$$g(E(X)) \leq E(g(X)).$$

Preuve. Par convexité de g, il existe une fonction $g^1(\cdot)$ telle que

$$g(x) \geq g(x_0) + (x - x_0)g^1(x_0)$$

pour tout x, $x_0 \in \mathbb{R}$. On pose $x_0 = E(X)$. Alors

$$g(X) \geq g(E(X)) + (X - E(X))g^1(E(X)).$$

En prenant les espérances on obtient $E(g(X)) \geq g(E(X))$. ∎

Voici un exemple d'application de l'inégalité de Jensen :

$$|E(X)| \leq E(|X|). \tag{1.4}$$

Proposition 1.6. (Inégalité de Cauchy-Schwarz.) *Soient ξ et η deux variables aléatoires telles que $E(\xi^2) < \infty$ et $E(\eta^2) < \infty$. Alors $E|\xi\eta| < \infty$,*

$$(E(\xi\eta))^2 \leq (E|\xi\eta|)^2 \leq E(\xi^2)E(\eta^2) \tag{1.5}$$

et les égalités dans (1.5) sont atteintes si et seulement si il existe $a_1, a_2 \in \mathbb{R}$ tels que $a_1 \neq 0$ ou $a_2 \neq 0$ et, presque sûrement,

$$a_1\xi + a_2\eta = 0. \tag{1.6}$$

Preuve. La deuxième inégalité dans (1.5) est le cas particulier de l'inégalité de Hölder pour $r = s = 2$. La première inégalité dans (1.5) est une conséquence de (1.4). Si (1.6) est vrai, il est évident que

$$(E(\xi\eta))^2 - E(\xi^2)E(\eta^2) = 0. \tag{1.7}$$

Réciproquement, si l'on a (1.7) et $E(\eta^2) \neq 0$, alors $E((\xi - a\eta)^2) = 0$ avec $a = E(\xi\eta)/E(\eta^2)$, ce qui implique $\xi = a\eta$ presque sûrement. Le cas où $E(\eta^2) = 0$ est trivial. ∎

1.3. Suites de variables aléatoires

Soient $\xi_1, \xi_2...$ et ξ des variables aléatoires sur $(\Omega, \mathcal{A}, P)$.

Définition 1.5. *On dit que la suite $(\xi_n)_{n\geq 1}$ converge* **en probabilité** *vers ξ quand $n \to \infty$ (et on écrit $\xi_n \xrightarrow{P} \xi$) si*

$$\lim_{n\to\infty} P(|\xi_n - \xi| \geq \epsilon) = 0$$

pour tout $\epsilon > 0$.

Définition 1.6. *On dit que la suite $(\xi_n)_{n\geq 1}$ converge* **en moyenne quadratique** *vers ξ quand $n \to \infty$ si $E(\xi^2) < \infty$ et*

$$\lim_{n\to\infty} E(|\xi_n - \xi|^2) = 0.$$

Définition 1.7. *On dit que la suite* $(\xi_n)_{n\geq 1}$ *converge* **presque sûrement** *(en abrégé p.s.) vers* ξ *quand* $n \to \infty$ *(et on écrit* $\xi_n \to \xi$ *(p.s.)), si*

$$P(\omega : \ \xi_n(\omega) \not\to \xi(\omega)) = 0.$$

Remarque. La Définition 1.7 est équivalente à la suivante : pour tout $\epsilon > 0$,

$$\lim_{n\to\infty} P(\sup_{k\geq n} |\xi_k - \xi| \geq \epsilon) = 0$$

(voir J.Lacroix, P.Priouret *Probabilités approfondies*, Polycopié du cours, Université Paris 6).

Définition 1.8. *On dit que la suite* $(\xi_n)_{n\geq 1}$ *converge* **en loi** *(ou* **en distribution***) vers* ξ *quand* $n \to \infty$ *(et on écrit* $\xi_n \stackrel{D}{\to} \xi$ *) si*

$$P(\xi_n \leq t) \to P(\xi \leq t) \ \ \textit{quand} \ \ n \to \infty,$$

pour chaque point t *de continuité de la f.d.r.* $F(t) = P(\xi \leq t)$.

Remarque. La convergence en loi est équivalente à la convergence étroite : pour toute fonction f continue et bornée

$$E(f(\xi_n)) \to E(f(\xi)) \ \ \text{quand} \ \ n \to \infty$$

(voir Bouleau N., *Probabilités de l'ingénieur, variables aléatoires et simulation*, Hermann, 1986, Corollaire 3.2.1 et Proposition 3.1.3, p. 178).

Liens entre les différents modes de convergence :

$$\begin{array}{lcccc} \boxed{\text{convergence en moyenne quadratique}} & \Longrightarrow & & & \\ & & \boxed{\text{convergence en probabilité}} & \Longrightarrow & \boxed{\text{convergence en loi}} \\ \boxed{\text{convergence p.s.}} & \Longrightarrow & & & \end{array}$$

Exercice 1.5. Soient $(\xi_n)_{n\geq 1}$ et $(\eta_n)_{n\geq 1}$ deux suites de variables aléatoires. Démontrer les résultats suivants :

1^o. Si $\xi_n \stackrel{P}{\to} a$ et $\eta_n \stackrel{D}{\to} \eta$, où $a \in \mathbb{R}$ est une constante et η est une variable aléatoire, alors

$$\xi_n \eta_n \stackrel{D}{\to} a\eta.$$

Ce résultat reste-t-il vrai si l'on suppose que a est une variable aléatoire ?

2^o. Si $a \in \mathbb{R}$ est une constante, alors

$$\xi_n \stackrel{D}{\to} a \quad \Longleftrightarrow \quad \xi_n \stackrel{P}{\to} a.$$

3^o. (**Théorème de Slutsky**.) Si $\xi_n \stackrel{D}{\to} a$ et $\eta_n \stackrel{D}{\to} \eta$ et $a \in \mathbb{R}$ est une constante, alors

$$\xi_n + \eta_n \stackrel{D}{\to} a + \eta,$$
$$\xi_n \eta_n \stackrel{D}{\to} a\eta.$$

Montrer que si a est une v.a., ces deux relations ne sont pas toujours vérifiées (donner des contre-exemples).

1.4. Indépendance et théorèmes limites

Définition 1.9. *Soient X et Y deux variables aléatoires sur $(\Omega, \mathcal{A}, P)$. On dit que la variable X est indépendante de Y (et on écrit $X \perp\!\!\!\perp Y$) si*

$$P(X \in A,\ Y \in B) = P(X \in A)P(Y \in B)$$

pour tous $A \in \mathcal{B}$ et $B \in \mathcal{B}$.

Si $E(|X|) < \infty,\ E(|Y|) < \infty$, l'indépendance implique

$$E(XY) = E(X)E(Y). \tag{1.8}$$

Important : le réciproque n'est pas vrai ; (1.8) n'est pas équivalent à l'indépendance de X et Y.

Définition 1.10. *Soient $X_1, \ldots, X_n$ des variables aléatoires sur $(\Omega, \mathcal{A}, P)$. On dit que les v.a. $X_1, \ldots, X_n$ sont (mutuellement) indépendantes si, pour tout $A_1, \ldots, A_n \in \mathcal{B}$,*

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n). \tag{1.9}$$

On dit que $(X_n)_{n\geq 1}$ est une suite infinie de variables aléatoires indépendantes si (1.9) est vérifié pour tout $n \geq 1$ entier.

Remarques. 1. Le fait que les X_i soient indépendantes deux à deux (c'est-à-dire $X_i \perp\!\!\!\perp X_j$ pour $i \neq j$) n'implique pas que $X_1, ..., X_n$ soient mutuellement indépendantes. Par contre, l'indépendance mutuelle implique l'indépendance deux à deux. En particulier, si $X_1, ..., X_n$ sont mutuellement indépendantes et $E(|X_i|) < \infty$ pour $i = 1, ..., n$, alors

$$E(X_i X_j) = E(X_i)E(X_j), \quad i \neq j.$$

2. Les transformations mesurables préservent l'indépendance : si $X \perp\!\!\!\perp Y$, alors $f(X) \perp\!\!\!\perp g(Y)$, quelles que soient les fonctions boréliennes $f(\cdot)$ et $g(\cdot)$.

1.4.1. Sommes de variables indépendantes. Considérons la somme $\sum_{i=1}^n X_i$, où les variables aléatoires $X_1, ..., X_n$ sont indépendantes. Si $E(X_i^2) < \infty$ pour $i = 1, ..., n$ (vu l'inégalité de Lyapounov, cela implique que $E(|X_i|) < \infty$), alors

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) \quad \text{(vrai sans hypothèse d'indépendance)} \tag{1.10}$$

et

$$\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \operatorname{Var}(X_i). \tag{1.11}$$

Définition 1.11. *On dit que les variables aléatoires $X_1, ..., X_n$ sont* **i.i.d.** *(indépendantes et identiquement distribuées) si elles sont mutuellement indépendantes et X_i est de même loi que X_j pour tout $1 \leq i, j \leq n$. De façon similaire, $X_1, X_2, ...$ sont appelés i.i.d. si $(X_n)_{n\geq 1}$ est une suite infinie de variables aléatoires indépendantes et de même loi.*

Proposition 1.7. *Soient* $X_1, ..., X_n$ *des v.a. i.i.d. telles que* $E(X_1) = \mu$ *et* $\mathrm{Var}(X_1) = \sigma^2 < \infty$*. Alors la moyenne arithmétique*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

vérifie

$$E(\bar{X}) = \mu \quad et \quad \mathrm{Var}(\bar{X}) = \frac{1}{n}\mathrm{Var}(X_1) = \frac{\sigma^2}{n}.$$

Preuve. On utilise (1.10) et (1.11).

Proposition 1.8. (Loi forte des grands nombres de Kolmogorov.) *Soient* $X_1, X_2, \ldots,$ *des v.a. i.i.d. telles que* $E(|X_1|) < \infty$ *et* $\mu = E(X_1)$*. Alors,*

$$\bar{X} \to \mu \ \ (p.s.) \quad quand \ \ n \to \infty.$$

Preuve. Voir Bouleau N., *Probabilités de l'ingénieur, variables aléatoires et simulation*, Hermann, 1986, Théorème 2.3, p. 170.

EXEMPLE 1.4. Soient X_i des variables i.i.d de loi de Cauchy. La densité de X_1 est

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}.$$

Alors $E(|X_1|) = \infty$, l'espérance $E(X_1)$ n'est pas définie et la moyenne arithmétique $\bar{X}$ n'est pas convergente.

Proposition 1.9. (Théorème central limite.) *Soient* $X_1, X_2, \ldots,$ *des v.a. i.i.d. telles que* $E(X_1^2) < \infty$ *et* $\sigma^2 = \mathrm{Var}(X_1) > 0$*. Alors,*

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \xrightarrow{D} \eta \quad quand \ \ n \to \infty,$$

où $\mu = E(X_1)$ *et* $\eta \sim \mathcal{N}(0,1)$*.*

Preuve. Voir Bouleau N., *Probabilités de l'ingénieur, variables aléatoires et simulation*, Hermann, 1986, Théorème 4.1, p. 181.

1.4.2. Approximation de la loi de $\bar{X}$ par la loi limite normale. Le Théorème central limite (la Proposition 1.9) s'écrit sous la forme équivalente :

$$P\left(\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \le t\right) \to P(\eta \le t) \quad \text{quand} \ \ n \to \infty,$$

pour tout $t \in \mathbb{R}$, où $\eta \sim \mathcal{N}(0,1)$. Notons

$$\Phi(t) = P(\eta \le t)$$

la f.d.r. normale standard. Alors

$$P(\bar{X} \le x) = P\left(\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \le \sqrt{n}\left(\frac{x-\mu}{\sigma}\right)\right) \approx \Phi\left(\sqrt{n}\left(\frac{x-\mu}{\sigma}\right)\right)$$

quand $n \to \infty$. Autrement dit, $P(\bar{X} \leq x)$, la f.d.r. de $\bar{X}$, peut être approximée par la loi normale :

$$P(\bar{X} \leq x) \approx \Phi\left(\sqrt{n}\left(\frac{x-\mu}{\sigma}\right)\right)$$

pour n assez grand.

1.5. Théorèmes de continuité

Proposition 1.10. (Premier théorème de continuité.) *Soit $g(\cdot)$ une fonction continue et soient $\xi_1, \xi_2, \dots$ et ξ des variables aléatoires sur $(\Omega, \mathcal{A}, P)$. Alors,*

$$\begin{array}{llll} (i) & \xi_n \to \xi \ \ (p.s.) & \Rightarrow & g(\xi_n) \to g(\xi) \ \ (p.s.), \\ (ii) & \xi_n \xrightarrow{P} \xi & \Rightarrow & g(\xi_n) \xrightarrow{P} g(\xi), \\ (iii) & \xi_n \xrightarrow{D} \xi & \Rightarrow & g(\xi_n) \xrightarrow{D} g(\xi) \end{array}$$

quand $n \to \infty$.

Preuve. La partie (i) est évidente. Montrons (ii) sous l'hypothèse supplémentaire que $\xi = a$, où a est une constante déterministe. En fait, c'est le seul cas qui présentera un intérêt dans le cadre de ce cours. La continuité de g implique que pour tout $\epsilon > 0$ il existe $\delta > 0$ tel que

$$|\xi_n - a| < \delta \quad \Rightarrow \quad |g(\xi_n) - g(a)| < \epsilon.$$

En particulier, $P(|\xi_n - a| < \delta) \leq P(|g(\xi_n) - g(a)| < \epsilon)$. Comme $\xi_n \xrightarrow{P} a$, on a

$$\lim_{n\to\infty} P(|\xi_n - a| < \delta) = 1 \ \text{ pour tout } \ \delta > 0,$$

ce qui implique

$$\lim_{n\to\infty} P(|g(\xi_n) - g(a)| < \epsilon) = 1 \ \text{ pour tout } \ \epsilon > 0.$$

(iii) Il suffit de démontrer (voir la remarque après la Définition 1.8) que, pour toute fonction continue et bornée $h(x)$, $E(h(g(\xi_n))) \to E(h(g(\xi)))$ quand $n \to \infty$. Comme g est continue, $f = h \circ g$ est aussi continue et bornée. Ceci démontre (iii), car $\xi_n \xrightarrow{D} \xi$ signifie que

$$E(f(\xi_n)) \to E(f(\xi)) \ \text{ quand } \ n \to \infty,$$

pour toute fonction f continue et bornée. ∎

Proposition 1.11. (Deuxième théorème de continuité.) *Soit $g(\cdot)$ une fonction continue et continûment différentiable et soient $X_1, X_2, \dots$ des variables aléatoires i.i.d. telles que $E(X_1^2) < \infty$ avec la variance $\sigma^2 = \mathrm{Var}(X_1) > 0$. Alors*

$$\sqrt{n}\left(\frac{g(\bar{X}) - g(\mu)}{\sigma}\right) \xrightarrow{D} \eta g'(\mu) \ \ \textit{quand} \ \ n \to \infty,$$

où $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\mu = E(X_1)$ et $\eta \sim \mathcal{N}(0,1)$.

Preuve. Sous les hypothèses de la proposition, la fonction

$$h(x) = \begin{cases} \frac{g(x)-g(\mu)}{x-\mu}, & \text{si } x \neq \mu, \\ g'(\mu), & \text{si } x = \mu, \end{cases}$$

est continue. Comme $\bar{X} \xrightarrow{P} \mu$ (vu la Proposition 1.8) et h est continue, on obtient, d'après le Premier théorème de continuité, que

$$h(\bar{X}) \xrightarrow{P} h(\mu) = g'(\mu) \text{ quand } n \to \infty. \tag{1.12}$$

Or,

$$\sqrt{n}\,\frac{g(\bar{X}) - g(\mu)}{\sigma} = \frac{\sqrt{n}}{\sigma}\,h(\bar{X})(\bar{X} - \mu) = h(\bar{X})\eta_n,$$

où $\eta_n = \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)$. La Proposition 1.9 implique que $\eta_n \xrightarrow{D} \eta \sim \mathcal{N}(0,1)$ quand $n \to \infty$. On conclut en utilisant ce fait, ainsi que (1.12) et le résultat 1^o de l'Exercice 1.5. ∎

1.6. Exercices

Exercice 1.6. Soient $\xi_1, ..., \xi_n$ des variables aléatoires indépendantes. Posons

$$\xi_{\min} = \min(\xi_1, ..., \xi_n), \quad \xi_{\max} = \max(\xi_1, ..., \xi_n).$$

1) Montrer que

$$P(\xi_{\min} \geq x) = \prod_{i=1}^{n} P(\xi_i \geq x), \quad P(\xi_{\max} < x) = \prod_{i=1}^{n} P(\xi_i < x).$$

2) Supposons, de plus, que $\xi_1, ..., \xi_n$ sont identiquement distribuées avec la loi uniforme $U[0, \theta]$. Calculer $E(\xi_{\min})$, $E(\xi_{\max})$, $\mathrm{Var}(\xi_{\min})$ et $\mathrm{Var}(\xi_{\max})$.

Exercice 1.7. Soit ξ une variable aléatoire positive avec la f.d.r. F et d'espérance finie. Démontrer que

$$E(\xi) = \int_0^\infty (1 - F(x))dx = \int_0^\infty P(\xi > x)dx.$$

Exercice 1.8. Soient X_1 et X_2 deux v.a. indépendantes de loi exponentielle $\mathcal{E}(\lambda)$. Montrer que $\min(X_1, X_2)$ et $|X_1 - X_2|$ sont des variables aléatoires de lois respectivement $\mathcal{E}(2\lambda)$ et $\mathcal{E}(\lambda)$.

Exercice 1.9. Soit X le nombre d'apparitions de "6" dans 12000 tirages d'un dé. En utilisant le Théorème central limite estimer la probabilité que $1800 < X \leq 2100$. *Indication :* $\Phi(\sqrt{6}) \approx 0.9928$, $\Phi(2\sqrt{6}) \approx 0.999999518$. Utiliser l'inégalité de Tchebychev pour obtenir une autre évaluation de cette probabilité et comparer les résultats.

2

Régression et corrélation

2.1. Couples des variables aléatoires. Lois jointes et marginales

Soit (X, Y) un couple des variables aléatoires. La f.d.r. *jointe* du couple (X, Y) est définie par

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y), \quad x, y \in \mathbb{R}.$$

Les f.d.r. *marginales* sont données par

$$\begin{aligned} F_X(x) &= \lim_{y \to +\infty} F_{X,Y}(x, y) = P(X \leq x), \\ F_Y(y) &= \lim_{x \to +\infty} F_{X,Y}(x, y) = P(Y \leq y). \end{aligned}$$

Dans le cas continu on suppose que $F_{X,Y}$ admet une densité $f_{X,Y} \geq 0$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$, autrement dit

$$\frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y} = f_{X,Y}(x, y) \tag{2.1}$$

presque partout. La densité $f_{X,Y}(x, y)$ vérifie $\int_{\mathbb{R}^2} f_{X,Y}(x, y) dx dy = 1$. Les *densités marginales* de X et Y sont définies par

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx.$$

Dans le cas discret X et Y prennent au maximum un nombre dénombrable de valeurs. La *loi jointe* du couple (X, Y) est définie par les probabilités $P(X = \cdot, Y = \cdot)$. Les *lois marginales* de X et Y sont définies par les probabilités

$$P(X = k) = \sum_m P(X = k, Y = m),$$
$$P(Y = m) = \sum_k P(X = k, Y = m).$$

Important : la connaissance des lois marginales de X et de Y n'est pas suffisante pour la détermination de la loi jointe du couple (X, Y). Considérons l'exemple suivant.

EXEMPLE 2.1. Soient deux densités de probabilité sur $\mathbb{R}^2$:

$$f_1(x, y) = \frac{1}{2\pi} \exp\Big(- \frac{x^2 + y^2}{2} \Big),$$

et

$$f_2(x, y) = \frac{1}{2\pi} \exp\Big(- \frac{x^2 + y^2}{2} \Big) [1 + xy \mathbb{1}_{[-1,1]}(x) \mathbb{1}_{[-1,1]}(y)].$$

Alors les densités marginales de f_1 sont les mêmes que celles de f_2 : elles sont normales standard $\mathcal{N}(0, 1)$.

Les v.a. X et Y sont *indépendantes* si et seulement si

$$F_{X,Y}(x, y) = F_X(x) F_Y(y) \quad \text{pour tout} \quad (x, y) \in \mathbb{R}^2.$$

Dans le cas continu, ceci se traduit par la décomposition

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad \text{pour tout} \quad (x, y) \in \mathbb{R}^2,$$

et dans le cas discret par

$$P(X = k, Y = m) = P(X = k) P(Y = m),$$

pour tous k, m.

2.2. Conditionnement (cas discret)

Soient A et B deux événements aléatoires ($A, B \in \mathcal{A}$) tels que $P(B) \neq 0$. La *probabilité conditionnelle* $P(A|B)$ de A sachant B est définie par

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

Soient X et Y deux v.a. discrètes. Selon cette définition

$$P(Y = k | X = m) = \frac{P(Y = k, X = m)}{P(X = m)}.$$

(Dorénavant on ne considère que les valeurs m telles que $P(X = m) > 0$.) On a alors

$$\sum_k P(Y = k | X = m) = \frac{\sum_k P(Y = k, X = m)}{P(X = m)} = 1.$$

Par conséquent, les probabilités $\{P(Y = k | X = m)\}_k$ définissent une loi discrète de probabilité (appelée *loi conditionnelle de Y sachant que $X = m$*). Si X et Y sont indépendantes,

$$P(Y = k | X = m) = \frac{P(Y = k) P(X = m)}{P(X = m)} = P(Y = k). \tag{2.2}$$

Réciproquement, si la relation (2.2) est vérifiée pour tous k, m, alors $X \perp\!\!\!\perp Y$. *L'espérance conditionnelle de Y sachant que $X = m$* est la **quantité déterministe**

$$E(Y|X = m) = \sum_k kP(Y = k|X = m).$$

La condition $E(|Y|) < \infty$ est suffisante pour assurer l'existence de l'espérance conditionnelle $E(Y|X = m)$, car $P(Y = k, X = m) \le P(Y = k)$.

La *variance conditionnelle* est définie par

$$\mathrm{Var}(Y|X = m) = E(Y^2|X = m) - [E(Y|X = m)]^2.$$

De façon analogue on définit les moments conditionnels, les quantiles conditionnels et autres caractéristiques d'une loi conditionnelle.

Définition 2.1. *Soient X et Y deux variables aléatoires discrètes, telles que $E(|Y|) < \infty$.* **L'espérance conditionnelle $E(Y|X)$ de Y sachant X** *est la variable aléatoire discrète qui ne dépend que de X et qui prend les valeurs*

$$\{E(Y|X = m)\}_m$$

avec les probabilités $P(X = m)$ respectivement.

Important : ne pas confondre la variable aléatoire $E(Y|X)$ avec la quantité déterministe $E(Y|X = m)$.

2.2.1. Propriétés des espérances conditionnelles (cas discret). On suppose ici que toutes les v.a. en question sont discrètes et toutes les espérances mathématiques qu'on considère sont finies.

1^o. *Linéarité.* Pour tout $a \in \mathbb{R}$, $b \in \mathbb{R}$,

$$E(aY_1 + bY_2|X) = aE(Y_1|X) + bE(Y_2|X).$$

2^o. Si X et Y sont indépendantes, alors $E(Y|X) = E(Y)$ (vu (2.2)).

3^o. $E(h(X)|X) = h(X)$ pour toute fonction borélienne h.

4^o. *Théorème de substitution.*

$$E(h(Y, X)|X = m) = E(h(Y, m)|X = m).$$

Preuve. On pose $Y' = h(Y, X)$, c'est une v.a. discrète qui prend les valeurs $h(k, m)$. Donc, la loi conditionnelle de Y' sachant que $X = m$ est donnée par les probabilités

$$\begin{aligned} P(Y' = a|X = m) &= P(h(Y, X) = a|X = m) = \frac{P(h(Y, X) = a, X = m)}{P(X = m)} \\ &= \frac{P(h(Y, m) = a, X = m)}{P(X = m)} = P(h(Y, m) = a|X = m). \end{aligned}$$

Alors, pour tout m fixé,

$$\begin{aligned} E(Y'|X = m) &= \sum_a aP(Y' = a|X = m) = \sum_a aP(h(Y, m) = a|X = m) \\ &= E(h(Y, m)|X = m). \end{aligned}$$

∎

Par conséquent, si $h(x,y) = h_1(y)h_2(x)$, nous avons

$$E(h_1(Y)h_2(X)|X = m) = h_2(m)E(h_1(Y)|X = m),$$

et

$$E(h_1(Y)h_2(X)|X) = h_2(X)E(h_1(Y)|X).$$

5^o. *Théorème de l'espérance itérée.*

$$E(E(Y|X)) = E(Y).$$

Preuve.

$$\begin{aligned}
E(E(Y|X)) &= \sum_m E(Y|X = m)P(X = m) = \sum_m \sum_k kP(Y = k|X = m)P(X = m) \\
&= \sum_{m,k} kP(Y = k, X = m) = \sum_k k \sum_m P(Y = k, X = m) \\
&= \sum_k kP(Y = k) = E(Y).
\end{aligned}$$

∎

EXEMPLE 2.2. Soient ξ et η deux variables aléatoires indépendantes de même loi de Bernoulli, qui prennent les valeurs 1 et 0 avec les probabilités p et $1 - p$. Calculons les espérances conditionnelles $E(\xi + \eta|\eta)$ et $E(\eta|\xi + \eta)$. En utilisant les propriétés 2^o et 3^o on obtient $E(\xi + \eta|\eta) = E(\xi) + \eta = p + \eta$. Cherchons maintenant $E(\eta|\xi + \eta)$. Pour $k = 0, 1, 2$,

$$E(\eta|\xi + \eta = k) = P(\eta = 1|\xi + \eta = k) = \begin{cases} 0, & k = 0, \\ 1/2, & k = 1, \\ 1, & k = 2. \end{cases}$$

Donc $E(\eta|\xi + \eta) = (\xi + \eta)/2$.

2.3. Conditionnement et projection. Meilleure prévision

Considérons l'ensemble de toutes les variables aléatoires ξ sur $(\Omega, \mathcal{A}, P)$ de carré intégrable, i.e. telles que $E(\xi^2) < \infty$. On dit que $\xi \sim \xi'$ si $\xi = \xi'$ (p.s.) par rapport à la mesure P. Ceci définit l'ensemble des classes d'équivalence sur les variables aléatoires telles que $E(\xi^2) < \infty$. On désignera ξ la variable aléatoire de carré intégrable aussi bien que sa classe d'équivalence. En utilisant cette convention, on note $L_2(P) = L_2(\Omega, \mathcal{A}, P)$ l'espace de toutes les variables aléatoires de carré intégrable sur $(\Omega, \mathcal{A}, P)$. C'est un espace de Hilbert muni du produit scalaire

$$\langle X, Y \rangle = E(XY),$$

et de la norme respective $\|X\| = [E(X^2)]^{1/2}$. En effet, $\langle \cdot, \cdot \rangle$ vérifie les axiomes du produit scalaire : pour tous $X, \xi, \eta \in L_2(P)$ et $a, b \in \mathbb{R}$

$$\langle a\xi + b\eta, X \rangle = E([a\xi + b\eta]X) = aE(\xi X) + bE(\eta X) = a\langle \xi, X \rangle + b\langle \eta, X \rangle,$$

et $\langle X, X \rangle \geq 0$; $\langle X, X \rangle = 0$ implique $X = 0$ (p.s.).

Si les variables X et Y sont indépendantes, la connaissance de la valeur prise par X ne donne aucune information sur Y. Mais si X et Y sont dépendantes et si l'on connaît la réalisation de X, ceci nous renseigne aussi sur Y. On pose le *problème de meilleure prévision* de Y étant donné X de façon suivante.

Problème de meilleure prévision. *Soit* $Y \in L_2(P)$ *et soit* X *une variable aléatoire sur* $(\Omega, \mathcal{A}, P)$. *Trouver une fonction borélienne* $g(\cdot)$ *telle que*

$$\|Y - g(X)\| = \min_{h(\cdot)} \|Y - h(X)\|, \tag{2.3}$$

où le minimum est recherché parmi toutes les fonctions boréliennes $h(\cdot)$ *et* $\|\cdot\|$ *est la norme de* $L_2(P)$. *La variable aléatoire* $\widehat{Y} = g(X)$ *est dite* **meilleure prévision de** Y **étant donné** X.

Dans le contexte du problème de meilleure prévision, X est appelée *variable explicative* ou *prédicteur*, Y est appelée *variable expliquée*.

On peut écrire (2.3) sous la forme équivalente :

$$E((Y - g(X))^2) = \min_{h(\cdot)} E((Y - h(X))^2) = \min_{h(\cdot):E(h^2(X))<\infty} E((Y - h(X))^2). \tag{2.4}$$

Il suffit ici de minimiser par rapport à $h(X) \in L_2(P)$, car une solution $g(\cdot)$ de (2.3) est automatiquement dans $L_2(P)$. Notons que (2.4) n'est que la définition de projection orthogonale de Y sur le sous-espace linéaire $L_2^X(P)$ de $L_2(P)$ défini par

$$L_2^X(P) = \{\xi = h(X) : E(h^2(X)) < \infty\}.$$

C'est le sous-espace linéaire de $L_2(P)$ composé de toutes les v.a. de carré intégrable mesurables par rapport à X. Grâce aux propriétés de projection orthogonale, il existe toujours une solution du problème de meilleure prévision : une v.a. $g(X) \in L_2^X(P)$ vérifie (2.3) et (2.4) si et seulement si

$$\langle Y - g(X), h(X)\rangle = 0 \ \text{ pour tout } h(X) \in L_2^X(P),$$

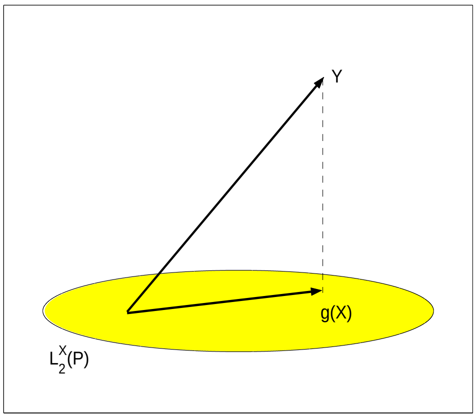

Figure 2.1. La projection orthogonale sur $L_2^X(P)$.

et une telle $g(X)$ est unique à une équivalence près. En passant à la notation avec les espérances, on écrit la formule précédente sous la forme

$$E((Y - g(X))h(X)) = 0 \ \text{ pour tout } h(X) \in L_2^X(P),$$

ou bien,

$$E(Yh(X)) = E(g(X)h(X)) \ \text{ pour tout } h(X) \in L_2^X(P). \tag{2.5}$$

En particulier,

$$E\big(Y\mathbb{1}_A(X)\big) = E\big(g(X)\mathbb{1}_A(X)\big) \text{ pour tout } A \in \mathcal{B} \tag{2.6}$$

où $\mathcal{B}$ est la tribu borélienne sur $\mathbb{R}$.

Remarque. En fait, (2.6) implique (2.5), donc (2.5) et (2.6) sont équivalents. Pour s'en convaincre il suffit d'utiliser le fait que l'espace des fonctions de la forme $\sum_{i=1}^{k} c_i \mathbb{1}_{A_i}(x)$ (fonctions en escaliers) avec $c_i \in \mathbb{R}, A_i \in \mathcal{B}$ est dense dans $L_2(P)$.

On va montrer maintenant que dans le cas discret la seule variable aléatoire $g(X)$ qui vérifie (2.5) (et par conséquent résout le problème de meilleure prévision (2.3)) est unique, à une équivalence près, et égale à l'espérance conditionnelle de Y sachant X.

Proposition 2.1. *Soient X et Y deux v.a.* **discrètes***, telles que $Y \in L_2(P)$. Alors la meilleure prévision $\widehat{Y}$ de Y étant donné X, unique à une équivalence près, est égale à l'espérance conditionnelle*

$$\widehat{Y} = E(Y|X).$$

Preuve. Pour tout $h(X) \in L_2^X(P)$,

$$\begin{aligned} E\left(E(Y|X)h(X)\right) &= \sum_k E(Y|X=k)h(k)P(X=k) \\ &= \sum_k \left[\sum_m mP(Y=m|X=k)\right] h(k)P(X=k) \\ &= \sum_{k,m} m\, h(k)P(Y=m, X=k) = E(Yh(X)). \end{aligned}$$

Donc (2.5) est vérifié avec $g(X) = E(Y|X)$, autrement dit, $E(Y|X)$ est une version de la projection orthogonale de Y sur $L_2^X(P)$. Comme la projection orthogonale dans un espace de Hilbert est unique à une équivalence près, $E(Y|X)$ est une unique solution de (2.5) presque sûrement. ∎

2.4. Probabilité et espérance conditionnelles (cas général)

On peut étendre la définition de l'espérance conditionnelle $E(Y|X)$ au cas de deux variables aléatoires générales X et Y. On utilise la définition suivante.

Définition 2.2. *Soient Y et X deux v. a. telles que $E(|Y|) < \infty$.* **L'espérance conditionnelle** $g(X) = E(Y|X)$ **de** Y **sachant** X *est une variable aléatoire mesurable par rapport à X qui vérifie*

$$E(YI\{X \in A\}) = E(g(X)I\{X \in A\}) \tag{2.7}$$

pour tout ensemble borélien A.

Remarque. On passe ici de l'hypothèse $Y \in L_2(P)$ (i.e. $E(Y^2) < \infty$) à l'hypothèse plus faible $E(|Y|) < \infty$. On peut démontrer (voir J.Lacroix, P.Priouret, *Probabilités approfondies,* Polycopié du cours, Université Paris 6) que la fonction $g(X)$ qui vérifie (2.7) existe et elle est unique (p.s.). C'est une conséquence du Théorème de Radon-Nikodym.

Si $Y \in L_2(P)$, l'existence et l'unicité p.s. de la fonction $g(X)$ vérifiant (2.7) découlent des propriétés de projection orthogonale dans L_2 comme on l'a déjà vu au paragraphe précédent. Comme corollaire, on obtient donc le résultat suivant.

Théorème 2.1. (de meilleure prévision.) *Soient X et Y deux v.a., telles que $Y \in L_2(P)$. Alors la meilleure prévision $\widehat{Y}$ de Y étant donné X, unique à une équivalence près, est égale à l'espérance conditionnelle*

$$\widehat{Y} = E(Y|X).$$

2.4.1. Probabilité et loi conditionnelles. Considérons le cas particulier suivant : on remplace Y par $Y' = I\{Y \in B\}$ où B est un ensemble borélien. Notons que la variable Y' est bornée ($|Y'| \le 1$), donc $E(|Y'|^2) < \infty$. Alors, l'espérance conditionnelle $g(X) = E(Y'|X)$ existe et elle vérifie la relation (cf. (2.7))

$$E\left(I\{Y \in B\}I\{X \in A\}\right) = E(g(X)I\{X \in A\}) \text{ pour tout } A \in \mathcal{B}.$$

Définition 2.3. *Soit $B \in \mathcal{B}$ fixé. La* **probabilité conditionnelle** $P(Y \in B|X)$ *est la variable aléatoire qui vérifie*

$$P(Y \in B, X \in A) = E\left[P(Y \in B|X)I\{X \in A\}\right] \quad \textit{pour tout } A \in \mathcal{B}.$$

Autrement dit, $P(Y \in B|X) = E(I\{Y \in B\}|X)$. La Définition 2.3 implique, en particulier :

$$P(Y \in B|X) = P(Y \in B) \text{ (p.s.) } \forall B \in \mathcal{B} \quad \Longleftrightarrow \quad X \perp\!\!\!\perp Y.$$

Définition 2.4. *Une fonction $(B, x) \mapsto P(Y \in B|X = x)$ de deux variables B et x, où $B \in \mathcal{B}$ et $x \in \mathbb{R}$, est dite* **loi conditionnelle de Y sachant que** $X = x$ *si*

(i) *pour tout B fixé $P(Y \in B|X = x)$ vérifie*

$$P(Y \in B, X \in A) = \int_A P(Y \in B|X = x)dF_X(x) \quad \textit{pour tout } A \in \mathcal{B}, \tag{2.8}$$

(ii) *pour tout x fixé $P(Y \in B|X = x)$ est une mesure de probabilité comme fonction de B.*

Remarque. On sait déjà que, pour tout $B \in \mathcal{B}$, il existe une fonction

$$g_B(x) = P(Y \in B|X = x)$$

telle que (2.8) est vérifié. Est-elle une mesure de probabilité comme fonction de B ? Notons que $g_B(x)$ est définie modulo les valeurs de x dans un ensemble N_B de probabilité nulle. Il est important que, généralement, cet ensemble dépend de B. Il n'est donc pas exclu a priori que l'ensemble $N = \bigcup_{B \in \mathcal{B}} N_B$ soit de probabilité > 0, dans quel cas $P(Y \in B|X = x)$ ne serait plus une mesure de probabilité pour x dans cet ensemble. Par exemple, on ne pourrait pas s'assurer de l'axiome d'additivité de la "probabilité" ainsi définie. Heureusement, dans notre cas où les v.a. en question sont réelles et la tribu est borélienne, on peut choisir une version (dite version régulière, voir M.Loève, *Probability Theory*, 1960, §27.2, Théorème A) de la fonction $g_B(\cdot)$ telle que $P(Y \in B|X = x)$ soit une mesure de probabilité pour tout

$x \in \mathbb{R}$. Dans la suite, on suppose que cette version est choisie dans chaque cas particulier. Si les variables X et Y ont une loi jointe discrète ou une loi jointe continue, il est facile de construire une telle version $P(Y \in B|X = x)$ de façon explicite (voir les Paragraphes 2.2 et 2.5).

On peut définir également $F_{Y|X}(\cdot|x)$, la *fonction de répartition conditionnelle de* Y *sachant que* $X = x$: c'est la f.d.r. qui correspond à la mesure de probabilité $P(Y \in \cdot|X = x)$. Pour trouver $F_{Y|X}(\cdot|x)$, il suffit de résoudre l'équation intégrale

$$P(Y \leq y, X \in A) = \int_A F_{Y|X}(y|x)dF_X(x) \quad \text{pour tout } y \in \mathbb{R},\ A \in \mathcal{B}. \tag{2.9}$$

La recherche d'une solution de l'équation intégrale (2.9) est le seul moyen de calculer $F_{Y|X}(\cdot|x)$ dans le cas mixte où Y est discrète et X est continue (ou inversement). Si les variables X et Y ont une loi jointe discrète ou une loi jointe continue, le recours à (2.9) n'est pas nécessaire. Dans ces cas, on caractérise la loi conditionnelle respectivement en termes de probabilités conditionnelles ou de densités : des formules plus simples et directes sont disponibles (voir les Paragraphes 2.2 et 2.5).

L'espérance conditionnelle de Y *sachant que* $X = x$ est la fonction réelle suivante de x :

$$E(Y|X = x) = \int yF_{Y|X}(dy|x).$$

Pour trouver $E(Y|X = x)$, il faut, généralement, résoudre l'équation intégrale :

$$E(YI\{X \in A\}) = \int_A E(Y|X = x)dF_X(x), \quad \text{pour tout } A \in \mathcal{B}.$$

Néanmoins, dans les cas "purement discret" ou "purement continu" le calcul de $E(Y|X = x)$ est beaucoup plus simple (voir les Paragraphes 2.2 et 2.5).

2.4.2. Propriétés de l'espérance conditionnelle. On suppose ici que, pour toutes les variables aléatoires en question, les espérances mathématiques sont finies.

1^o. *Linéarité.* Pour tout $a \in \mathbb{R}$, $b \in \mathbb{R}$,

$$E(aY_1 + bY_2|X) = aE(Y_1|X) + bE(Y_2|X) \ \ \text{(p.s.)}$$

2^o. Si X et Y sont indépendantes, $E(Y|X) = E(Y)$ (p.s.)

Preuve. Vu la définition (2.7) il suffit de montrer que

$$E(YI\{X \in A\}) = E\left(E(Y)I\{X \in A\}\right), \quad \text{pour tout } A \in \mathcal{B}. \tag{2.10}$$

Or,

$$E\left(E(Y)I\{X \in A\}\right) = E(Y)P(X \in A),$$

et on voit que (2.10) est une conséquence de l'indépendance de X et Y. ∎

3^o. $E(h(X)|X) = h(X)$ (p.s.) pour toute fonction borélienne h.

4^o. *Théorème de substitution.*

$$E(h(Y, X)|X = x) = E(h(Y, x)|X = x).$$

Si X et Y sont des v.a. discrètes, ce résultat est prouvé au Paragraphe 2.2. Si X et Y ont la loi jointe continue, la démonstration est aussi facile (Exercice 2.1). La démonstration au cas général n'est pas donnée ici.

5^o. *Théorème de l'espérance itérée.*

$$E(E(Y|X)) = E(Y).$$

Preuve. On pose $A = \mathbb{R}$ dans la définition (2.7), alors $I(X \in A) = 1$, et on obtient le résultat désiré. ∎

2.5. Conditionnement (cas continu)

On suppose maintenant qu'il existe une densité jointe $f_{X,Y}(x,y) \geq 0$ du couple (X,Y) par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$. Définissons

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)}, & \text{si } f_X(x) > 0, \\ f_Y(y), & \text{si } f_X(x) = 0. \end{cases} \tag{2.11}$$

On remarque que $y \mapsto f_{Y|X}(y|x)$ est une densité de probabilité pour tout $x \in \mathbb{R}$, car

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = 1, \qquad f_{Y|X}(y|x) \geq 0. \tag{2.12}$$

Ceci reste vrai si l'on modifie la définition (2.11) en posant $f_{Y|X}(y|x) = \bar{f}(y)$ quand $f_X(x) = 0$, où $\bar{f}(\cdot)$ est une densité de probabilité quelconque.

Notons aussi que, vu (2.11),

$$Y \perp\!\!\!\perp X \qquad \Longleftrightarrow \qquad f_{Y|X}(y|x) = f_Y(y).$$

Proposition 2.2. *Si la densité jointe de (X,Y) existe, alors la loi conditionnelle de Y sachant que $X = x$ est donnée par la formule*

$$P(Y \in B|X = x) = \int_B f_{Y|X}(y|x)dy, \quad B \in \mathcal{B}, \ x \in \mathbb{R}. \tag{2.13}$$

Preuve. Vu (2.12) la partie (ii) de la Définition 2.4 est vérifiée. Il suffit donc de montrer la partie (i) de la Définition 2.4, i.e. que pour tous $A, B \in \mathcal{B}$,

$$P(Y \in B, X \in A) = \int_A \left[\int_B f_{Y|X}(y|x)dy\right] dF_X(x).$$

Comme X possède une densité, $dF_X(x) = f_X(x)dx$. D'après le Théorème de Fubini,

$$\int_A \left[\int_B f_{Y|X}(y|x)dy\right] f_X(x)dx = \int_B \int_A f_{Y|X}(y|x)f_X(x)\,dxdy$$

Mais $f_{Y|X}(y|x)f_X(x) = f_{X,Y}(x,y)$ presque partout par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$ (si $f_X(x) = 0$, alors *a fortiori* $f_{X,Y}(x,y) = 0$). La dernière intégrale est donc égale à

$$\int_B \int_A f_{X,Y}(x,y)dxdy = P(X \in A, Y \in B).$$

∎

De façon similaire on obtient la formule pour l'espérance conditionnelle :

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy.$$

Définissons aussi, comme dans le cas discret, la fonction de *variance conditionnelle* :

$$\begin{aligned}\mathrm{Var}(Y|X=x) &= E(Y^2|X=x) - (E(Y|X=x))^2 \\ &= \int_{-\infty}^{\infty} y^2 f_{Y|X}(y|x)dy - \left[\int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy\right]^2,\end{aligned}$$

ainsi que la variable aléatoire

$$\mathrm{Var}(Y|X) = E(Y^2|X) - (E(Y|X))^2.$$

EXERCICE 2.1. Montrer que le Théorème de substitution est vérifié au cas continu.

REMARQUE. Souvent on définit la densité conditionnelle par

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)}, & \text{si } f_X(x) > 0, \\ 0, & \text{si } f_X(x) = 0. \end{cases} \tag{2.14}$$

Cette définition ne diffère de (2.11) que sur un ensemble de probabilité 0. Notons que la Proposition 2.2 n'est pas vraie si $f_{Y|X}(y|x)$ est donnée par (2.14) : en effet, la partie (ii) de la Définition 2.4 est vérifiée pour presque tout x et non pas pour tout x. Néanmoins, l'espérance conditionnelle est la même dans les deux cas, et la définition (2.14) est souvent tacitement utilisée dans les calculs (cf. les Exemples 2.3 et 2.5 ci-après).

EXEMPLE 2.3. Soient X et Y deux variables aléatoires indépendantes de loi exponentielle de densité $f(u) = \lambda e^{-\lambda u} I\{u > 0\}$ avec $\lambda > 0$. Calculons la densité conditionnelle $f(x|z) = f_{X|X+Y}(x|z)$ et l'espérance conditionnelle $E(X|X+Y)$. Si $z < x$,

$$P(X+Y<z, X<x) = P(X+Y<z, X<z) = \int_0^z \int_0^{z-u} f(u)f(v)dudv,$$

et si $z \geq x$,

$$P(X+Y<z, X<x) = \int_0^x \int_0^{z-u} f(u)f(v)dudv.$$

Par conséquent, pour $z \geq x$ la densité jointe du couple $(X+Y, X)$ est (cf. (2.1))

$$f(z,x) = \frac{\partial^2 P(X+Y<z, X<x)}{\partial x \partial z} = f(z-x)f(x) = \lambda^2 e^{-\lambda z}.$$

Par ailleurs, la densité de $X+Y$ est la convolution de deux densités exponentielles, i.e.

$$f_{X+Y}(z) = \lambda^2 z e^{-\lambda z}.$$

On obtient donc une version de la densité conditionnelle de la forme :

$$f_{X|X+Y}(x|z) = \frac{f(z,x)}{f_{X+Y}(z)} = \frac{1}{z}$$

pour $0 \leq x \leq z$ et $f_{X|X+Y}(x|z) = 0$ pour $x > z$. C'est une densité de la loi uniforme sur $[0,z]$. On obtient donc $E(X|X+Y) = (X+Y)/2$ (p.s.).

Cet exemple est lié au modèle du flux de demandes arrivant vers un système de service. Soit X l'instant où la première demande arrive (l'instant $t = 0$ est marqué par l'arrivée de la demande numéro zéro), Y l'intervalle de temps entre les arrivées de la première et de la deuxième demandes. Alors on cherche la densité de probabilité de l'instant de la première demande sachant que la seconde est arrivée à l'instant z.

2.6. Covariance et corrélation

Soient X et Y deux v.a. de carrés intégrable, i.e. $E(X^2) < \infty$ et $E(Y^2) < \infty$. Par la suite, on notera

$$\sigma_X^2 = \mathrm{Var}(X), \qquad \sigma_Y^2 = \mathrm{Var}(Y).$$

Définition 2.5. *La* **covariance** *entre X et Y est la valeur*

$$\mathrm{Cov}(X,Y) = E\left((X - E(X))(Y - E(Y))\right) = E(XY) - E(X)E(Y).$$

Si $\mathrm{Cov}(X,Y) = 0$ on dit que X et Y sont orthogonales et on écrit $X \perp Y$.

Définition 2.6. *Soit $\sigma_X^2 > 0$ et $\sigma_Y^2 > 0$. La* **corrélation** *(ou le* **coefficient de corrélation***) entre X et Y est la quantité*

$$\mathrm{Corr}(X,Y) = \rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \, \sigma_Y}.$$

2.6.1. Propriétés de la covariance.

1^o. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

2^o. $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X,Y)$, $a, b \in \mathbb{R}$.

3^o. $\mathrm{Cov}(X + a, Y) = \mathrm{Cov}(X,Y)$, $a \in \mathbb{R}$.

4^o.$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

5^o. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(Y,X)$. En effet,

$$\begin{aligned}\mathrm{Var}(X+Y) &= E((X+Y)^2) - (E(X) + E(Y))^2 \\ &= E(X^2) + E(Y^2) + 2E(XY) - E^2(X) - E^2(Y) - 2E(X)E(Y).\end{aligned}$$

6^o. Si X et Y sont indépendantes, $\mathrm{Cov}(X,Y) = 0$.

Important : le réciproque n'est pas vrai. Considérons l'exemple suivant.

Exemple 2.4. Soit $X \sim \mathcal{N}(0,1)$ et $Y = X^2$. Alors,

$$\mathrm{Cov}(X,Y) = E(X^3) - E(X)E(X^2) = E(X^3) = 0.$$

2.6.2. Propriétés de la corrélation.

1^o. $-1 \le \rho_{XY} \le 1$. En effet, d'après l'inégalité de Cauchy-Schwarz,

$$
\begin{aligned}
|\mathrm{Cov}(X,Y)| &= |E\left[(X - E(X))(Y - E(Y))\right]| &(2.15)\\
&\le \sqrt{E[(X - E(X))^2]}\sqrt{E[(Y - E(Y))^2]} = \sigma_X\,\sigma_Y. &(2.16)
\end{aligned}
$$

2^o. Si les v.a. X et Y sont indépendantes, $\rho_{XY} = 0$.

3^o. $|\rho_{XY}| = 1$ si et seulement si il existe un lien linéaire déterministe entre X et Y : il existe $a \ne 0$, $b \in \mathbb{R}$ tels que $Y = aX + b$ (p.s.).
Preuve. On remarque que $|\rho_{XY}| = 1$, si et seulement si l'égalité est atteinte dans l'inégalité de Cauchy-Schwarz (2.16). D'après la Proposition 1.6, ce n'est possible que s'il existe $\alpha,\ \beta \in \mathbb{R}$ tels que $\alpha \ne 0$ ou $\beta \ne 0$ et

$$\alpha(X - E(X)) + \beta(Y - E(Y)) = 0 \ \ \text{(p.s.)}.$$

Ceci est équivalent à l'existence de α, β et $\gamma \in \mathbb{R}$ tels que

$$\alpha X + \beta Y + \gamma = 0 \ \ \text{(p.s.)},$$

avec $\alpha \ne 0$ ou $\beta \ne 0$. Si $\alpha \ne 0$ et $\beta \ne 0$ on a

$$Y = -\frac{\alpha}{\beta}X - \frac{\gamma}{\beta} = aX + b$$

où $a = -\alpha/\beta \ne 0$, $\ \ b = -\gamma/\beta$. La situation quand $\alpha = 0$ ou $\beta = 0$ est impossible : en effet, dans ce cas une des variables Y ou X est constante (p.s.), alors que nous avons supposé que $\sigma_X > 0$ et $\sigma_Y > 0$. ∎

4^o. La corrélation est invariante par rapport aux transformations affines : pour tout $a \ne 0$, $b, d \in \mathbb{R}$,

$$\rho_{aX+b,aY+d} = \rho_{XY}.$$

Si, de plus, , $c \ne 0$,

$$|\rho_{aX+b,cY+d}| = |\rho_{XY}|$$

(vérifiez ceci à titre d'exercice).

On remarque que si $Y = aX + b$, $a, b \in \mathbb{R}$, $a \ne 0$, les variances vérifient

$$\sigma_Y^2 = E((Y - E(Y))^2) = a^2 E((X - E(X))^2) = a^2\sigma_X^2,$$

alors que la covariance vaut

$$\mathrm{Cov}(X,Y) = E\left((X - E(X))a(X - E(X))\right) = a\sigma_X^2,$$

d'où $\rho_{XY} = a/|a|$. On dit que la corrélation entre X et Y est positive si $\rho_{XY} > 0$ et qu'elle est négative si $\rho_{XY} < 0$. La corrélation ci-dessus est donc positive ($= 1$) si $a > 0$ et négative ($= -1$) si $a < 0$.

2.6.3. Interprétation géométrique de la corrélation. Soit $\langle \cdot, \cdot \rangle$ le produit scalaire et $\| \cdot \|$ la norme de $L_2(P)$. Alors,

$$\text{Cov}(X, Y) = \langle X - E(X), Y - E(Y) \rangle$$

et

$$\rho_{XY} = \frac{\langle X - E(X), Y - E(Y) \rangle}{\|X - E(X)\| \, \|Y - E(Y)\|}.$$

Autrement dit, ρ_{XY} est le "cosinus de l'angle" entre $X - E(X)$ et $Y - E(Y)$. Donc, $\rho_{XY} = \pm 1$ signifie que $X - E(X)$ et $Y - E(Y)$ sont colinéaires : $Y - E(Y) = a(X - E(X))$ pour $a \neq 0$.

2.7. Régression

Définition 2.7. *Soient X et Y deux variables aléatoires telles que $E(|Y|) < \infty$. La fonction $g : \mathbb{R} \to \mathbb{R}$ définie par*

$$g(x) = E(Y|X = x)$$

est dite **fonction de régression de Y sur X**.

Il s'agit ici de la régression *simple* (le mot simple signifie que X et Y sont des v.a. réelles). La notion de régression s'étend aux v.a. X et Y multidimensionnelles, il s'agit dans ce cas de la régression *multiple* ou *multivariée* (voir la définition au Chapitre 3).

Exemple 2.5. Soit la densité jointe de X et Y,

$$f_{X,Y}(x, y) = (x + y) I\{0 < x < 1, 0 < y < 1\}.$$

Explicitons la fonction de régression $g(x) = E(Y|X = x)$. La densité marginale de X est

$$f_X(x) = \int_0^1 f_{X,Y}(x, y) dy = (x + 1/2) I\{0 < x < 1\}.$$

Alors, une version de la densité conditionnelle est donnée par

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{x + y}{x + 1/2} I\{0 < x < 1, 0 < y < 1\}$$

et

$$g(x) = E(Y|X = x) = \int_0^1 y f_{Y|X}(y|x) dy = \int_0^1 \frac{y(x + y)}{x + \frac{1}{2}} dy = \frac{\frac{1}{2}x + \frac{1}{3}}{x + \frac{1}{2}}$$

pour $0 < x < 1$. Soulignons que, dans cet exemple, $g(x)$ est une fonction *non-linéaire* de x.

2.8. Variance résiduelle et rapport de corrélation

Dans ce paragraphe, nous supposons que $Y \in L_2(P)$. La variable aléatoire $\xi = Y - g(X)$ représente l'erreur stochastique de l'approximation de Y par sa meilleure prévision $\widehat{Y} = g(X) = E(Y|X)$. On appelle ξ *résidu* de régression. Evidemment,

$$Y = g(X) + \xi. \tag{2.17}$$

Par définition de l'espérance conditionnelle $E(\xi|X) = 0$ (p.s.), donc $E(\xi) = 0$.

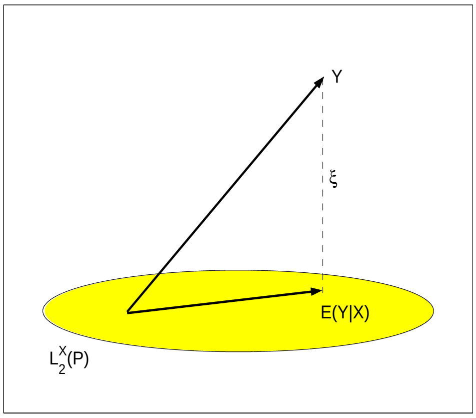

Figure 2.2. Le résidu de régression.

L'erreur quadratique de l'approximation de Y par $g(X)$ est la valeur suivante :

$$\Delta = E((Y - \widehat{Y})^2) = E((Y - g(X))^2) = E\left((Y - E(Y|X))^2\right) = E(\xi^2) = \mathrm{Var}(\xi).$$

On appelle Δ *variance résiduelle*. Elle est plus petite que la variance de Y. En effet, considérons $h(X) \equiv E(Y) = const$. D'après le Théorème de meilleure prévision (Théorème 2.1),

$$\Delta = E\left((Y - g(X))^2\right) \le E\left((Y - h(X))^2\right) = E((Y - E(Y))^2) = \mathrm{Var}(Y).$$

Comme $E(Y)$ est un élément de $L_2^X(P)$, géométriquement cela signifie que la longueur d'une cathète est plus petite que celle de l'hypothénuse (voir la figure suivante).

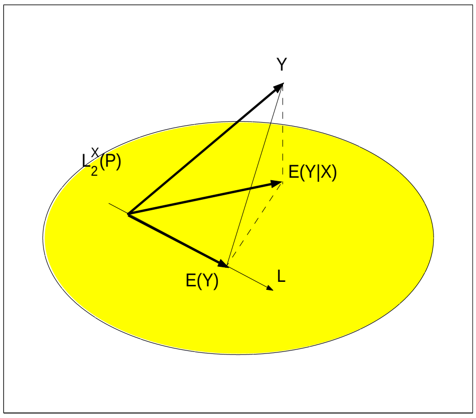

Figure 2.3. Interprétation géométrique de la relation (2.18)

On remarque que l'espace de toutes les v.a. constantes noté L est aussi un sous-espace linéaire de $L_2(P)$. De plus, L est l'intersection de tous les sous-espaces $L_2^X(P)$ pour tout X. Notons que $E(Y)$ est la projection de Y sur L : en effet, d'après le Corollaire 1.1, pour toute constante c,

$$E((Y - c)^2) \ge E((Y - E(Y))^2).$$

Le Théorème de Pythagore (cf. Fig. 2.3) implique

$$\|Y - E(Y)\|^2 = \|E(Y|X) - E(Y)\|^2 + \|Y - E(Y|X)\|^2, \tag{2.18}$$

ce qu'on peut écrire aussi de plusieurs façons équivalentes :

$$\begin{aligned} \mathrm{Var}(Y) = E((Y - E(Y))^2) &= E\left((E(Y|X) - E(Y))^2\right) + E\left((Y - E(Y|X))^2\right) \\ &= \mathrm{Var}\left(E(Y|X)\right) + E\left(\mathrm{Var}(Y|X)\right) \\ &= \mathrm{Var}(g(X)) + \mathrm{Var}(\xi) \qquad\qquad (2.19) \\ &= \mathrm{Var}(g(X)) + \Delta \\ &= \text{variance expliquée par } X \ + \text{ variance résiduelle}, \end{aligned}$$

où la variable aléatoire $\mathrm{Var}(Y|X)$ est définie au § 2.5. On a donc, pour toute variable aléatoire X,

$$\boxed{\mathrm{Var}(Y) = \mathrm{Var}\left(E(Y|X)\right) + E\left(\mathrm{Var}(Y|X)\right).}$$

Exercice 2.2. Montrer (2.18) en utilisant les propriétés de l'espérance conditionnelle données au § 2.4.2.

Définition 2.8. *Soit* $\mathrm{Var}(Y) > 0$. *On appelle* **rapport de corrélation de** Y **sur** X *la quantité positive* $\eta^2_{Y|X}$ *donnée par*

$$\eta^2_{Y|X} = \frac{\mathrm{Var}(g(X))}{\mathrm{Var}(Y)} = \frac{E\left((E(Y) - E(Y|X))^2\right)}{\mathrm{Var}(Y)}.$$

Notons que, vu (2.18),

$$\eta^2_{Y|X} = 1 - \frac{E\left((Y - g(X))^2\right)}{\mathrm{Var}(Y)}.$$

Interprétation géométrique : le rapport de corrélation $\eta^2_{Y|X}$ est le “cosinus carré de l'angle” entre $Y - E(Y)$ et $E(Y|X) - E(Y)$, donc $0 \le \eta^2_{Y|X} \le 1$.

Remarques.

(1) De façon générale, $\eta^2_{X|Y} \neq \eta^2_{Y|X}$ (manque de symétrie).

(2) Les cas extrêmes $\eta^2_{Y|X} = 0$ et $\eta^2_{Y|X} = 1$ correspondent à des valeurs remarquables : $\eta^2_{Y|X} = 1$ implique que $E((Y - E(Y|X))^2) = 0$, donc $Y = g(X)$ (p.s.), autrement dit, Y est lié fonctionnellement à X.

D'autre part, $\eta^2_{Y|X} = 0$ signifie que $E((E(Y) - E(Y|X))^2) = 0$ et $E(Y|X) = E(Y)$ (p.s.), donc la régression est constante. Il est utile de noter que ceci implique l'orthogonalité de X et Y ($\mathrm{Cov}(X, Y) = 0$).

(3) La variance résiduelle peut être exprimée à partir du rapport de corrélation :

$$\Delta = (1 - \eta^2_{Y|X})\mathrm{Var}(Y). \tag{2.20}$$

Proposition 2.3. *Soit* $E(X^2) < \infty$, $E(Y^2) < \infty$ *et* $\mathrm{Var}(X) = \sigma^2_X > 0$, $\mathrm{Var}(Y) = \sigma^2_Y > 0$. *Alors,*

$$\eta^2_{Y|X} \ge \rho^2_{XY}.$$

Preuve. Vu la définition de $\eta^2_{Y|X}$, il suffit de montrer que

$$E\left((E(Y) - E(Y|X))^2\right) \mathrm{Var}(X) \geq [E((X - E(X))(Y - E(Y)))]^2 .$$

D'après le Théorème de l'espérance itérée,

$$\begin{aligned} E((X - E(X))(Y - E(Y))) &= E\left((X - E(X))E([Y - E(Y)]|X)\right) \\ &= E\left((X - E(X))(E(Y|X) - E(Y))\right). \end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwarz, on obtient

$$\begin{aligned} [E((X - E(X))(Y - E(Y)))]^2 &\leq E((X - E(X))^2)E\left((E(Y|X) - E(Y))^2\right) \\ &= \mathrm{Var}(X)E\left((E(Y|X) - E(Y))^2\right). \end{aligned} \tag{2.21}$$

■

Remarque. La condition $\eta^2_{Y|X} = 0$ implique que $\rho_{XY} = 0$, mais généralement le réciproque n'est pas vrai.

2.9. Régression linéaire

Si $E(Y|X = x) = a + bx$ avec $a, b \in \mathbb{R}$, on dit que *la régression de Y sur X est linéaire.* C'est un **cas très particulier**, mais important, de la régression. En utilisant (2.17), on écrit

$$Y = a + bX + \xi$$

où ξ est le résidu, $E(\xi|X) = 0$ (p.s.) ($\Rightarrow E(\xi) = 0$).

Soient $\rho = \rho_{XY}$ et $\sigma_X > 0$, $\sigma_Y > 0$ le coefficient de corrélation entre X et Y et les écart-types de X et Y. Les coefficients de la régression linéaire a et b s'expriment alors à partir de $E(X)$, $E(Y)$, ρ, σ_X et σ_Y. En effet,

$$Y - E(Y) = b(X - E(X)) + \xi.$$

En multipliant cette équation par $X - E(X)$ et en prenant l'espérance, on obtient

$$\mathrm{Cov}(X, Y) = b\mathrm{Var}(X) = b\sigma_X^2,$$

ce qui implique

$$b = \frac{\mathrm{Cov}(X, Y)}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}.$$

Alors,

$$Y = a + \rho\frac{\sigma_Y}{\sigma_X}X + \xi.$$

Or,

$$E(Y) = a + \rho\frac{\sigma_Y}{\sigma_X}E(X),$$

et

$$a = E(Y) - \rho\frac{\sigma_Y}{\sigma_X}E(X).$$

Finalement,

$$Y = E(Y) + \rho\frac{\sigma_Y}{\sigma_X}(X - E(X)) + \xi. \tag{2.22}$$

Proposition 2.4. *Soit* $E(X^2) < \infty$, $E(Y^2) < \infty$ *et* $\mathrm{Var}(X) = \sigma_X^2 > 0$, $\mathrm{Var}(Y) = \sigma_Y^2 > 0$. *Si la fonction de régression* $g(x) = E(Y|X = x)$ *est linéaire, elle s'écrit nécessairement sous la forme*

$$\boxed{E(Y|X = x) = E(Y) + \rho \frac{\sigma_Y}{\sigma_X}\ (x - E(X))} \tag{2.23}$$

et la variance résiduelle vérifie

$$\Delta = (1 - \rho^2)\sigma_Y^2, \tag{2.24}$$

où ρ *est le coefficient de corrélation entre* X *et* Y.

Remarque. On peut également écrire (2.23) sous la forme

$$E(Y|X = x) = E(Y) + \frac{\mathrm{Cov}(X, Y)}{\sigma_X^2}\ (x - E(X)). \tag{2.25}$$

Preuve. L'égalité (2.23) est une conséquence immédiate de (2.22) et du fait que $E(\xi|X = x) = 0$. Montrons (2.24). La variance de

$$g(X) = E(Y) + \rho \frac{\sigma_Y}{\sigma_X}\ (X - E(X))$$

vaut

$$\mathrm{Var}(g(X)) = \mathrm{Var}(\rho \frac{\sigma_Y}{\sigma_X}\ X) = \rho^2 \sigma_Y^2$$

et, d'après (2.19),

$$\Delta = E(\xi^2) = \mathrm{Var}(Y) - \mathrm{Var}(g(X)) = \sigma_Y^2 - \mathrm{Var}(g(X)).$$

∎

Corollaire 2.1. *Soit* $E(X^2) < \infty$, $E(Y^2) < \infty$ *et* $\mathrm{Var}(X) = \sigma_X^2 > 0$, $\mathrm{Var}(Y) = \sigma_Y^2 > 0$. *Si la régression de* Y *sur* X *est linéaire, alors*

$$\eta_{Y|X}^2 = \rho_{XY}^2. \tag{2.26}$$

Le réciproque est aussi vrai : si $\rho_{XY}^2 = \eta_{Y|X}^2$, *alors la régression de* Y *sur* X *est linéaire.*

Preuve. Pour obtenir (2.26), il suffit de comparer (2.20) et (2.24). Pour démontrer la réciproque, on note que si l'égalité est atteinte dans l'inégalité de Cauchy-Schwarz (2.21), alors il existe $\alpha, \beta \in \mathbb{R}$ tels que $\alpha \neq 0$ ou $\beta \neq 0$ et

$$\alpha(X - E(X)) + \beta(E(Y|X) - E(Y)) = 0 \text{ (p.s.)}.$$

Or, $\beta = 0$ est impossible vu la condition $\sigma_X^2 > 0$. On a donc

$$E(Y|X) = E(Y) + a(X - E(X)), \text{ (p.s.) avec } a = -\alpha/\beta.$$

∎

Remarque. Les notions de “lien linéaire *déterministe* entre X et Y” et de “lien linéaire *stochastique* entre X et Y” sont à ne pas confondre. Un lien linéaire déterministe signifie que $Y = aX + b$ (p.s.) avec $a \neq 0$, ,$b \in \mathbb{R}$, tandis qu'un lien linéaire stochastique signifie que la régression de Y sur X est linéaire, i.e. $Y = aX + b + \xi$ (p.s.), où $E(\xi|X) = 0$ (p.s.) et ξ est de variance strictement positive. S'il existe un lien linéaire déterministe, alors $\rho_{XY}^2 = \eta_{Y|X}^2 = \eta_{X|Y}^2 = 1$. S'il existe un lien linéaire stochastique (i.e. seule la régression de Y sur X est linéaire), alors $\rho_{XY}^2 = \eta_{Y|X}^2 \leq 1$ et généralement $\eta_{Y|X}^2 \neq \eta_{X|Y}^2$, car la linéarité de la régression de Y sur X n'implique pas la linéarité de la régression de X sur Y.

Conclusions.

(1) Le coefficient de corrélation ρ_{XY} est particulièrement adapté pour caractériser un lien *linéaire* entre X et Y (la régression linéaire), si un tel lien existe.

(2) Le rapport de corrélation $\eta_{Y|X}^2$ est une mesure de lien entre X et Y plus générale que ρ_{XY}. Elle est utile au cas où la régression de Y sur X est *non-linéaire*.

(3) Si la régression de Y sur X est linéaire, les deux mesures sont identiques : $\eta_{Y|X}^2 = \rho_{XY}^2$.

2.10. Meilleure prévision linéaire

Au lieu de chercher la meilleure prévision de Y parmi *toutes les fonctions boréliennes* $g(X)$, on peut poser un problème moins général : approximer Y par les fonctions de type $a + bX$, où a et b sont des coefficients déterministes. Ce problème (dite de *meilleure prévision linéaire*) est formulé comme suit.

Soit $Y \in L_2(P)$ et soit X une v.a. sur $(\Omega, \mathcal{A}, P)$. Trouver les valeurs déterministes $\widehat{a}$ et $\widehat{b}$ telles que

$$\|Y - \widehat{a} - \widehat{b}X\| = \min_{a,b \in \mathbb{R}} \|Y - a - bX\|, \tag{2.27}$$

où $\|\cdot\|$ est la norme de $L_2(P)$. La variable aléatoire $\widehat{Y}^L = \widehat{a} + \widehat{b}X$ est appelée **meilleure prévision linéaire** *de Y étant donné X.*

Proposition 2.5. *Soit $E(X^2) < \infty$, $E(Y^2) < \infty$ et $\mathrm{Var}(X) = \sigma_X^2 > 0$, $\mathrm{Var}(Y) = \sigma_Y^2 > 0$. Alors*

$$\widehat{a} = E(Y) - \rho\frac{\sigma_Y}{\sigma_X}E(X),$$

$$\widehat{b} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}$$

où $\rho = \mathrm{Corr(X, Y)}$, et la meilleure prévision linéaire de Y étant donné X est

$$\widehat{Y}^L = E(Y) + \rho\frac{\sigma_Y}{\sigma_X}(X - E(X)).$$

Preuve. Notons que (2.27) est équivalent au problème de minimisation

$$E((Y - \widehat{a} - \widehat{b}X)^2) = \min_{a,b \in \mathbb{R}} E((Y - a - bX)^2),$$

ce qui est équivalent, à son tour, au problème

$$E([(Y - E(Y)) - \widehat{a}' - \widehat{b}(X - E(X))]^2) = \min_{a',b\in\mathbb{R}} E([(Y - E(Y)) - a' - b(X - E(X))]^2)$$

(on a fait le changement de variable $a' = a - E(Y) - bE(X)$). Or,

$$E([(Y - E(Y)) - a' - b(X - E(X))]^2) = \sigma_Y^2 + (a')^2 + b^2\sigma_X^2 - 2b\text{Cov}(X, Y),$$

d'où $\widehat{a}' = 0$, $\widehat{b} = \text{Cov}(X, Y)/\sigma_X^2$. ∎

On peut noter la similarité des expressions présentes dans les Propositions 2.4 et 2.5. Néanmoins, les deux résultats sont bien différents : dans la Proposition 2.4 il s'agit d'une fonction de régression exacte (au cas où elle est linéaire), tandis que dans la Proposition 2.5 la variable $\widehat{Y}^L$ n'est qu'une approximation linéaire da la fonction de régression (qui peut être non-linéaire). La différence devient évidente si l'on compare les erreurs quadratiques $\Delta = E((Y - \widehat{Y})^2) = E(\xi^2)$ et $\Delta^L = E((Y - \widehat{Y}^L)^2)$. En effet, d'après le Théorème de Pythagore,

$$\begin{aligned}\Delta^L &= E((Y - g(X))^2) + E((g(X) - \widehat{Y}^L)^2) = E(\xi^2) + E((g(X) - \widehat{Y}^L)^2) \\ &= \Delta + E((g(X) - \widehat{Y}^L)^2),\end{aligned}$$

ce qui implique $\Delta^L \geq \Delta$ avec $\Delta^L = \Delta$ si et seulement si la régression $g(\cdot)$ est linéaire.

2.11. Exercices

EXERCICE 2.3. Soient deux densités de probabilité sur $\mathbb{R}^2$:

$$f_1(t_1, t_2) = I\{0 < t_1, t_2 < 1\}$$

et

$$f_2(t_1, t_2) = [1 + (2t_1 - 1)(2t_2 - 1)]I\{0 < t_1, t_2 < 1\}.$$

Vérifier que f_2 est une densité de probabilité. Montrer que f_1 et f_2 ont les mêmes densités marginales.

EXERCICE 2.4. Soient X et Y deux variables aléatoires indépendantes et de même loi. Utiliser la définition pour démontrer que $E(X|X+Y) = E(Y|X+Y)$ (p.s.). En déduire que $E(X|X+Y) = E(Y|X+Y) = \frac{X+Y}{2}$ (p.s.). Comparer ceci avec les Exemples 2.2 et 2.3.

EXERCICE 2.5. Soient X, Y_1 et Y_2 des variables aléatoires indépendantes, telles que Y_1 et Y_2 sont de loi normale $\mathcal{N}(0, 1)$. On définit la v.a.

$$Z = \frac{Y_1 + XY_2}{\sqrt{1 + X^2}}.$$

Utiliser la loi conditionnelle $P(Z \leq u|X = x)$ pour montrer que $Z \sim \mathcal{N}(0, 1)$.

EXERCICE 2.6. Soient ξ_1 et ξ_2 deux variables aléatoires indépendantes de même loi telle que $0 < \text{Var}(\xi_1) < \infty$. Montrer que les v.a. $\eta_1 = \xi_1 - \xi_2$ et $\eta_2 = \xi_1 + \xi_2$ sont non-corrélées.

EXERCICE 2.7. Soient X, Y, Z des variables aléatoires telles que $E(|Z|) < \infty$. Montrer que $E(E(Z|Y, X)|Y) = E(Z|Y)$.

Exercice 2.8. Soient X et N deux variables aléatoires telles que N prend ses valeurs dans $\{1, 2, \ldots\}$ et $E(|X|) < \infty$, $E(N) < \infty$. On considère la suite $X_1, X_2, \ldots$ des variables indépendantes de même loi que X. Utilisant le conditionnement montrer *l'identité de Wald* : si N est indépendante des X_i, alors

$$E(\sum_{i=1}^{N} X_i) = E(N)E(X).$$

Exercice 2.9. On suppose que $Y = X^3 + \sigma\varepsilon$, où X et ε sont deux variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$ et $\sigma > 0$. Comparer le rapport de corrélation $\eta^2_{Y|X}$ et le carré du coefficient de corrélation ρ^2_{XY} pour ce modèle.

Exercice 2.10. Le salaire désiré d'un individu s'écrit $Y^* = Xb + \sigma\varepsilon$, où $\sigma > 0$, $b > 0$, X est une variable aléatoire telle que $E(X^2) < \infty$ mesurant la capacité de l'individu, ε est indépendante de X et de loi $\mathcal{N}(0,1)$. Si Y^* est plus grand que le SMIC S, alors le salaire reçu Y est Y^*, et S sinon. Calculer $E(Y|X)$. Cette espérance est-elle fonction linéaire de X ?

Exercice 2.11. Soient ξ et η deux variables aléatoires avec $E(\xi) = E(\eta) = 0$, $\mathrm{Var}(\xi) = \mathrm{Var}(\eta) = 1$ et soit le coefficient de corrélation $\mathrm{Corr}(\xi, \eta) = \rho$.
1^o. Montrer que

$$E(\max(\xi^2, \eta^2)) \leq 1 + \sqrt{1-\rho^2}.$$

Indication : on remarque que

$$\max(\xi^2, \eta^2) = \frac{|\xi^2 + \eta^2| + |\xi^2 - \eta^2|}{2}.$$

2^o. Démontrer l'inégalité suivante :

$$P\left(|\xi - E(\xi)| \geq \epsilon\sqrt{\mathrm{Var}(\xi)} \text{ ou } |\eta - E(\eta)| \geq \epsilon\sqrt{\mathrm{Var}(\eta)}\right) \leq \frac{1 + \sqrt{1-\rho^2}}{\epsilon^2}.$$

Exercice 2.12. On considère une suite de variables aléatoires $X_0, \ldots, X_n$ issues du modèle suivant (*modèle d'autorégression*) :

$$X_i = aX_{i-1} + \varepsilon_i, \ i = 1, \ldots, n,$$
$$X_0 = 0,$$

où les ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et $a \in \mathbb{R}$.
1^o. Ecrire X_i en fonction de la série des v.a. $(\varepsilon_1, \ldots, \varepsilon_n)$. En déduire, selon les valeurs du paramètre a, la loi, l'espérance μ et la variance σ_i^2 de X_i.
2^o. Calculer le coefficient de corrélation entre X_i et X_{i+1}.

Exercice 2.13. Soient X et ξ deux variables aléatoires indépendantes et de même loi $U[-1, 1]$ (loi uniforme sur $[-1, 1]$). On pose $Y = I\{X + \xi \geq 0\}$.
1^o. Chercher la fonction de régression de Y sur X.
2^o. Calculer le coefficient de corrélation ρ_{XY}.
3^o. Chercher la loi conditionnelle de Y sachant X et celle de X sachant Y.

Exercice 2.14. Soient X, Y et Z des variables aléatoires telles que X et Y sont de carré intégrable. La *covariance conditionnelle* entre X et Y sachant Z est définie par

$$\mathrm{Cov}(X, Y|Z) = E(XY|Z) - E(X|Z)E(Y|Z).$$

Montrer que

$$\mathrm{Cov}(X, Y) = E(\mathrm{Cov}(X, Y|Z)) + \mathrm{Cov}(E(X|Z), E(Y|Z)).$$

Exercice 2.15. Soit $X \sim \mathcal{N}(0, 1)$ et $Y = X^2$. Quelle est la meilleure prévision de X étant donné Y ? Calculer $\eta^2_{Y|X}$, $\eta^2_{X|Y}$, ρ_{XY}. *Indication* : montrer que les v.a. $|X|$ et $\mathrm{sign}(X)$ sont indépendantes.

Exercice 2.16. Soient X, Y, Z des variables aléatoires telles que $E(|Z|) < \infty$. On considère les espérances conditionnelles $\zeta_1 = E(Z|Y, X)$ et $\zeta_2 = E(E(Z|Y)|X)$.
1^o. On suppose d'abord que $Z = X$ et que la v.a. X est indépendante de Y. Calculer ζ_1 et ζ_2 et remarquer que $\zeta_1 \neq \zeta_2$.
2^o. On suppose ensuite que la loi jointe de $(X,\ Y,\ Z)$ admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^3$. Exprimer ζ_1 et ζ_2. Obtient–on $\zeta_1 = \zeta_2$?
3^o. Soit $E_X(\cdot)$ l'espérance par rapport à la loi marginale de X. Peut–on affirmer que

$$E_X(\zeta_1) = E(Z|Y)?$$

$$E_X(\zeta_1) = E_X(\zeta_2)?$$

Que se passe–t–il si la v.a. X est indépendante de Y ?

3

Vecteurs aléatoires. Loi normale multivariée

3.1. Vecteurs aléatoires

Nous commençons par le rappel sur quelques propriétés de vecteurs aléatoires. Un vecteur aléatoire dans $\mathbb{R}^p$ est un vecteur $\mathbf{x} = (\xi_1, ..., \xi_p)^T$ dont toutes les composantes $\xi_1, ..., \xi_p$ sont des variables aléatoires réelles[1]. De la même façon on définit des matrices aléatoires :

$$\Xi = \begin{pmatrix} \xi_{11} & ... & \xi_{1q} \\ & ... & \\ \xi_{p1} & ... & \xi_{pq} \end{pmatrix},$$

où les ξ_{ij} sont des variables aléatoires réelles. La **fonction de répartition** du vecteur aléatoire $\mathbf{x}$ est définie par

$$F_{\mathbf{x}}(t) = P(\xi_1 \le t_1, ..., \xi_p \le t_p), \quad t = (t_1, ..., t_p)^T \in \mathbb{R}^p.$$

La **fonction caractéristique** du vecteur aléatoire $\mathbf{x}$ est une fonction $\phi_{\mathbf{x}}(\cdot)$ sur $\mathbb{R}^p$ à valeurs complexes définie par

$$\phi_{\mathbf{x}}(t) = E\left(\exp(it^T\mathbf{x})\right), \quad t \in \mathbb{R}^p.$$

Deux vecteurs aléatoires $\mathbf{x} \in \mathbb{R}^p$ et $\mathbf{y} \in \mathbb{R}^q$ sont appelés **indépendants** si, pour tous $A \in \mathcal{B}(\mathbb{R}^p)$ et $B \in \mathcal{B}(\mathbb{R}^q)$, on a : $P(\mathbf{x} \in A, \mathbf{y} \in B) = P(\mathbf{x} \in A)P(\mathbf{y} \in B)$ où $\mathcal{B}(\mathbb{R}^p)$ est la tribu borélienne de $\mathbb{R}^p$. Dans ce cas on écrit $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$. Le résultat suivant donne une caractérisation de l'indépendance : $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$ si et seulement si la fonction caractéristique $\phi_{\mathbf{z}}(u)$ du vecteur $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ se présente, pour tout $u = \begin{pmatrix} a \\ b \end{pmatrix}$ avec $a \in \mathbb{R}^p$ et $b \in \mathbb{R}^q$, sous la forme de produit

$$\phi_{\mathbf{z}}(u) = \phi_{\mathbf{x}}(a)\phi_{\mathbf{y}}(b) \tag{3.1}$$

[1] Par la suite, tout vecteur $\mathbf{x} \in \mathbb{R}^p$ est un vecteur colonne et $\mathbf{x}^T$ désigne le transposé de $\mathbf{x}$.

(voir Bouleau N., *Probabilités de l'ingénieur, variables aléatoires et simulation*, Hermann, 1986, Proposition 5.12, p. 142). Plus généralement, s'il s'agit de n vecteurs aléatoires, nous avons la définition suivante de l'inépendance.

Définition 3.1. *Soient* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *des vecteurs aléatoires sur* $(\Omega, \mathcal{A}, P)$*, tels que* $\mathbf{x}_i$ *est à valeurs dans* $\mathbb{R}^{p_i}$*. On dit que* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *sont (mutuellement) indépendants si, pour tous* $A_i \in \mathcal{B}(\mathbb{R}^{p_i})$*,* $i = 1, \ldots, n$*,*

$$P(\mathbf{x}_1 \in A_1, \ldots, \mathbf{x}_n \in A_n) = P(\mathbf{x}_1 \in A_1) \cdots P(\mathbf{x}_n \in A_n). \tag{3.2}$$

En utilisant cette définition et (3.1), on obtient facilement que les vecteurs aléatoires $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sont mutuellement indépendants si et seulement si la fonction caractéristique du vecteur composé $\mathbf{x} \stackrel{\text{déf}}{=} (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ est égale au produit des fonctions caractéristiques des vecteurs $\mathbf{x}_i$, $i = 1, \ldots, n$.

3.1.1. Propriétés des vecteurs aléatoires au cas continu. Dans ce chapitre nous considérerons principalement le *cas continu*, c'est-à-dire nous supposerons que la loi de $\mathbf{x}$ admet une densité de probabilité $f_{\mathbf{x}}(\cdot) \geq 0$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$. Cela signifie que

$$F_{\mathbf{x}}(t) = \int_{-\infty}^{t_1} \ldots \int_{-\infty}^{t_p} f_{\mathbf{x}}(u_1, \ldots, u_p) du_1 \ldots du_p$$

pour tout $t = (t_1, \ldots, t_p) \in \mathbb{R}^p$ et

$$f_{\mathbf{x}}(t) = f_{\mathbf{x}}(t_1, \ldots, t_p) = \frac{\partial^p F_{\mathbf{x}}(t)}{\partial t_1 \cdots \partial t_p}.$$

pour presque tout t. Toute densité de probabilité vérifie

$$f_{\mathbf{x}}(t) \geq 0, \quad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{x}}(t_1, \ldots, t_p) dt_1 \ldots dt_p = 1.$$

Soit $\mathbf{x}' = (\xi_1, \ldots, \xi_k)^T$ (où $k < p$) vecteur aléatoire, une partie de $\mathbf{x}$. La **densité marginale** de $\mathbf{x}'$ est

$$f_{\mathbf{x}'}(t_1, \ldots, t_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{x}}(t_1, \ldots, t_p) dt_{k+1} \ldots dt_p.$$

Notons que la connaissance de toutes les densités marginales n'est pas suffisante pour la détermination de la loi du vecteur aléatoire $\mathbf{x}$. Deux vecteurs aléatoires différents peuvent avoir les mêmes lois marginales (voir l'Exemple 2.1 relatif au vecteur de dimension 2).

Soient maintenant $\mathbf{x} = (\xi_1, \ldots, \xi_p)^T$ et $\mathbf{y} = (\eta_1, \ldots, \eta_q)^T$ deux vecteurs aléatoires tels que le couple $(\mathbf{x}, \mathbf{y})$ admet une densité $f_{\mathbf{y},\mathbf{x}}$. La **densité conditionnelle** de $\mathbf{y}$ sachant que $\mathbf{x} = (t_1, \ldots, t_p)^T$, pour un vecteur déterministe $(t_1, \ldots, t_p)$, est définie par

$$f_{\mathbf{y}|\mathbf{x}}(s_1, \ldots, s_q | t_1, \ldots, t_p) = \begin{cases} \frac{f_{\mathbf{y},\mathbf{x}}(s_1, \ldots, s_q, t_1, \ldots, t_p)}{f_{\mathbf{x}}(t_1, \ldots, t_p)}, & \text{si } f_{\mathbf{x}}(t_1, \ldots, t_p) > 0, \\ f_{\mathbf{y}}(s_1, \ldots, s_q), & \text{si } f_{\mathbf{x}}(t_1, \ldots, t_p) = 0. \end{cases} \tag{3.3}$$

Pour $p = q = 1$ on retrouve la définition (2.11) du Paragraphe 2.5 [2]. La **loi conditionnelle de $\mathbf{y}$ sachant que** $\mathbf{x} = a$ (avec $a \in \mathbb{R}^p$ déterministe) est la loi de densité $f_{\mathbf{y}|\mathbf{x}}(\cdot|a)$. Les vecteurs aléatoires $\mathbf{x}$ et $\mathbf{y}$ sont **indépendants** si et seulement si

$$f_{\mathbf{x},\mathbf{y}}(t_1, ..., t_p, s_1, ..., s_q) = f_{\mathbf{x}}(t_1, ..., t_p) f_{\mathbf{y}}(s_1, ..., s_q).$$

L'indépendance signifie que la densité conditionnelle (3.3) n'est fonction que de $(s_1, ..., s_q)$, elle ne dépend pas de la valeur $(t_1, ..., t_p)$ prise par $\mathbf{x}$. Comme dans le cas de deux variables aléatoires réelles, les transformations mesurables des vecteurs aléatoires $\mathbf{x}$ et $\mathbf{y}$ préservent l'indépendance.

3.1.2. Transformations des vecteurs aléatoires. Soit $h = (h_1, ..., h_p)^T$ une transformation, c'est-à-dire une fonction de $\mathbb{R}^p$ dans $\mathbb{R}^p$,

$$h(t_1, ..., t_p) = (h_1(t_1, ..., t_p), ..., h_p(t_1, ..., t_p))^T, \quad t = (t_1, ..., t_p)^T \in \mathbb{R}^p.$$

Le *Jacobien* de la transformation est défini par

$$J_h(t) = \mathrm{Det} \left(\frac{\partial h_i}{\partial t_j}(t) \right)_{i,j},$$

pourvu que les dérivées partielles existent. Rappelons le résultat suivant de l'analyse.

Proposition 3.1. *Supposons que*

(i) les dérivées partielles de $h_i(\cdot)$ sont continues sur $\mathbb{R}^p$ pour $i = 1, ..., p$,

(ii) h est une bijection,

(iii) $J_h(t) \neq 0$ pour tout $t \in \mathbb{R}^p$.

Alors, pour toute fonction $f : \mathbb{R} \to \mathbb{R}^p$ telle que $\int_{\mathbb{R}^p} |f(t)|dt < \infty$ et tout ensemble borélien $K \subseteq \mathbb{R}^p$, on a

$$\int_K f(t)dt = \int_{h^{-1}(K)} f(h(u))|J_h(u)|du.$$

Remarque. D'après le Théorème de fonction inverse, sous les conditions de la Proposition 3.1 la fonction inverse $g(\cdot) = h^{-1}(\cdot)$ existe partout dans $\mathbb{R}^p$ et

$$J_{h^{-1}}(h(u)) = \frac{1}{J_h(u)}, \qquad J_{h^{-1}}(t) = \frac{1}{J_h(h^{-1}(t))}.$$

On voit donc que h vérifie les conditions (i)-(iii) de la Proposition 3.1 si et seulement si $g = h^{-1}$ vérifie ces conditions.

Proposition 3.2. *Soit $\mathbf{y}$ un vecteur aléatoire dans $\mathbb{R}^p$ de densité $f_{\mathbf{y}}$. Soit $g : \mathbb{R}^p \to \mathbb{R}^p$ une transformation vérifiant les hypothèses de la Proposition 3.1. Alors, la densité $f_{\mathbf{x}}$ du vecteur aléatoire $\mathbf{x} = g(\mathbf{y})$ est donnée par :*

$$f_{\mathbf{x}}(u) = f_{\mathbf{y}}(h(u))|J_h(u)|,$$

pour tout $u \in \mathbb{R}^p$, où $h = g^{-1}$.

---

[2] Il est possible d'utiliser aussi la définition un peu différente de (3.3), en modifiant (3.3) sur un ensemble de probabilité 0, par exemple, en posant $f_{\mathbf{y}|\mathbf{x}}(s_1, ..., s_q|t_1, ..., t_p) = 0$ si $f_{\mathbf{x}}(t_1, ..., t_p) = 0$, cf. (2.14).

Preuve. Soit $\mathbf{x} = (\xi_1, ..., \xi_p)^T$, $v = (v_1, ..., v_p)^T$ et $A_v = \{t \in \mathbb{R}^p : \ g_i(t) \leq v_i, \ i = 1, ..., p\}$. D'après la Proposition 3.1 avec $h = g^{-1}$ et $f = f_{\mathbf{y}}$, la f.d.r. de $\mathbf{x}$ s'écrit sous la forme

$$\begin{aligned} F_{\mathbf{x}}(v) &= P(\xi_i \leq v_i, \ i = 1, ..., p) = P(g_i(\mathbf{y}) \leq v_i, \ i = 1, ..., p) \\ &= \int_{A_v} f_{\mathbf{y}}(t)dt = \int_{g(A_v)} f_{\mathbf{y}}(h(u))|J_h(u)|du. \end{aligned}$$

Or,

$$\begin{aligned} g(A_v) &= \{u = g(t) \in \mathbb{R}^p : \ t \in A_v\} = \{u = g(t) \in \mathbb{R}^p : \ g_i(t) \leq v_i, \ i = 1, ..., p\} \\ &= \{u = (u_1, ..., u_p)^T \in \mathbb{R}^p : \ u_i \leq v_i, \ i = 1, ..., p\}, \end{aligned}$$

d'où on obtient

$$F_{\mathbf{x}}(v) = \int_{-\infty}^{v_1} ... \int_{-\infty}^{v_p} f_{\mathbf{y}}(h(u))|J_h(u)|du$$

pour tout $v = (v_1, ..., v_p)^T \in \mathbb{R}^p$. Ceci signifie que la densité de $\mathbf{x}$ est $f_{\mathbf{y}}(h(u))|J_h(u)|$. ∎

Corollaire 3.1. *Si $\mathbf{x} = A\mathbf{y} + b$, où $\mathbf{y}$ est un vecteur aléatoire dans $\mathbb{R}^p$ de densité $f_{\mathbf{y}}$, $b \in \mathbb{R}^p$ est un vecteur déterministe et A est une matrice $p \times p$ telle que $\mathrm{Det}(A) \neq 0$, alors*

$$f_{\mathbf{x}}(u) = f_{\mathbf{y}}(A^{-1}(u - b))|\mathrm{Det}(A^{-1})| = \frac{f_{\mathbf{y}}(A^{-1}(u - b))}{|\mathrm{Det}(A)|}.$$

Pour prouver ce résultat, il suffit d'utiliser la Proposition 3.2 avec $u = g(t) = At + b$ (respectivement, $t = g^{-1}(u) = h(u) = A^{-1}(u - b)$).

3.1.3. Rappel sur quelques propriétés de matrices. Dans la suite, une matrice $p \times q$ est une matrice réelle à p lignes et q colonnes. La notation $\mathrm{Diag}(\lambda_1, \ldots, \lambda_p)$ sera utilisée pour une matrice diagonale dont les éléments diagonaux sont $\lambda_1, \ldots, \lambda_p$. On notera I une matrice unité et $\mathbf{0}$ une matrice nulle (i.e. une matrice dont tous les éléments sont 0), sans indiquer les dimensions lorsqu'il n'y aura pas d'ambiguïté.

Le déterminant et la trace d'une matrice carrée $A = (a_{ij})_{i,j=1,\ldots,p}$ sont définis par

$$\mathrm{Det}(A) = \prod_{i=1}^{p} \lambda_i, \qquad \mathrm{Tr}(A) = \sum_{i=1}^{p} a_{ii} = \sum_{i=1}^{p} \lambda_i,$$

où les λ_i sont les valeurs propres de A. On a

$$\mathrm{Det}(A^T) = \mathrm{Det}(A)$$

où A^T désigne la transposée de A. Si A est inversible,

$$\mathrm{Det}(A^{-1}) = [\mathrm{Det}(A)]^{-1}.$$

Soient deux matrices A, B carrées $p \times p$. Alors

$$\mathrm{Det}(AB) = \mathrm{Det}(A)\,\mathrm{Det}(B).$$

Une matrice carrée $A = (a_{ij})_{i,j=1,\ldots,p}$ est dite **symétrique** si $a_{ij} = a_{ji}$, $i, j = 1, ..., p$ (ou bien $A = A^T$). *Toutes les valeurs propres d'une matrice symétrique sont réelles.*

On dit qu'une matrice symétrique A est **positive** et on écrit $A \geq 0$ si $\mathbf{x}^T A\mathbf{x} \geq 0$ pour tout $\mathbf{x} \in \mathbb{R}^p$. Si, en outre, $\mathbf{x}^T A\mathbf{x} > 0$ pour tout $\mathbf{x} \neq 0$, on appelle A **strictement positive** et on écrit $A > 0$.

Soient deux matrices symétriques A et B. On écrit $A \geq B$ ou $A > B$ si la matrice $A - B$ est positive ou strictement positive respectivement.

Une matrice Γ carrée $p \times p$ est dite **orthogonale** si

$$\Gamma^{-1} = \Gamma^T,$$

ce qui équivaut à

$$\Gamma\Gamma^T = \Gamma^T\Gamma = I,$$

où I est la matrice unité $p \times p$. Les colonnes $\boldsymbol{\gamma}_{(j)}$ de la matrice orthogonale $\Gamma = (\boldsymbol{\gamma}_{(1)}, ..., \boldsymbol{\gamma}_{(p)})$ sont des vecteurs mutuellement orthogonaux de norme 1 :

$$\boldsymbol{\gamma}_{(i)}^T \boldsymbol{\gamma}_{(j)} = \delta_{ij} \text{ pour } i, j = 1, ..., p,$$

de même pour les lignes de Γ. Ici δ_{ij} est le symbole de Kronecker : $\delta_{ij} = 1$ pour $i = j$ et $\delta_{ij} = 0$ pour $i \neq j$. De plus, $|\operatorname{Det}(\Gamma)| = 1$.

Les matrices symétriques sont caractérisées par le **Théorème de décomposition spectrale** :

Soit A une matrice $p \times p$ symétrique. Alors

$$A = \Gamma\Lambda\Gamma^T = \sum_{i=1}^{p} \lambda_i \boldsymbol{\gamma}_{(i)} \boldsymbol{\gamma}_{(i)}^T, \tag{3.4}$$

où

$$\Lambda = \operatorname{Diag}(\lambda_1, \ldots, \lambda_p) = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{pmatrix},$$

les λ_i sont les valeurs propres de A, les $\boldsymbol{\gamma}_{(i)}$ sont les vecteurs propres orthonormés correspondants et $\Gamma = (\boldsymbol{\gamma}_{(1)}, ..., \boldsymbol{\gamma}_{(p)})$ est une matrice $p \times p$ orthogonale.

Un corollaire de ce théorème est :

Une matrice symétrique A est positive (strictement positive) si et seulement si toutes ses valeurs propres sont positives (strictement positives).

Remarques.

(1) Une matrice symétrique peut avoir des valeurs propres multiples, mais tous les vecteurs propres $\boldsymbol{\gamma}_{(i)}$ dans la décomposition spectrale (3.4) sont différents. Les $\boldsymbol{\gamma}_{(i)}$ correspondant aux valeurs propres multiples ne sont pas définis de façon unique.

(2) Sans perte de généralité, on supposera dans la suite que les valeurs propres λ_i d'une matrice symétrique A sont ordonnées :

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p.$$

On appellera $\boldsymbol{\gamma}_{(1)}$ **premier vecteur propre** de A, c'est-à-dire, le vecteur propre correspondant à la valeur propre maximale ; $\boldsymbol{\gamma}_{(2)}$ correspondant à λ_2 sera appelé deuxième vecteur propre, et ainsi de suite.

Fonctions des matrices symétriques. Pour les matrices symétriques, le calcul de fonctions matricielles est simplifié. Par exemple, la puissance A^s, $s > 0$, d'une matrice A symétrique et positive est définie par $A^s = \Gamma\Lambda^s\Gamma^T$. La matrice A^s est aussi symétrique et positive. Le cas particulier est la racine carrée $A^{1/2}$ de la matrice A qui vérifie

$$A = A^{1/2}A^{1/2}, \quad A^{1/2} = (A^{1/2})^T, \quad A^{1/2} \geq 0.$$

Si, de plus, la matrice A est non-dégénérée, la définition de A^s s'étend pour $s < 0$. Notons que $\mathrm{Det}(A^s) = [\mathrm{Det}(A)]^s$, vu la définition de A^s et le fait que $|\mathrm{Det}(\Gamma)| = 1$ pour toute matrice Γ orthogonale.

Projecteurs. Une matrice P telle que

$$P = P^T \quad \text{(matrice symétrique)} \quad \text{et} \quad P^2 = P \quad \text{(matrice idempotente)}$$

est dite **matrice de projection** (ou **projecteur**).

Toutes les valeurs propres d'un projecteur P sont 0 *ou* 1. En effet, soit v un vecteur propre de P, i.e. $Pv = \lambda v$, où λ est la valeur propre de P correspondante. Comme $P^2 = P$,

$$(\lambda^2 - \lambda)v = (\lambda P - P)v = (P^2 - P)v = 0.$$

Ceci implique que $\lambda = 1$ ou $\lambda = 0$. Par conséquent, le rang $\mathrm{Rang}(P)$ d'un projecteur P est le nombre de ses valeurs propres égales à 1 et

$$\mathrm{Rang}(P) = \mathrm{Tr}(P).$$

3.1.4. Matrices de covariance et de corrélation. Un vecteur $\mu = (\mu_1, ..., \mu_p)^T \in \mathbb{R}^p$ est la **moyenne** du vecteur aléatoire $\mathbf{x} = (\xi_1, ..., \xi_p)^T$ si

$$\mu_j = E(\xi_j) = \int ... \int t_j f_{\mathbf{x}}(t_1, ..., t_p) dt_1 ... dt_p. \ \ j = 1, ..., p,$$

On écrit alors $\mu = E(\mathbf{x})$ (ici et par la suite on suppose que toutes les intégrales et toutes les espérances en question sont finies). De la même façon, on définit l'espérance d'une matrice aléatoire. Comme dans le cas de v.a. réelles, l'espérance est une fonctionnelle linéaire : si A est une matrice $p \times q$ et $b \in \mathbb{R}^q$, alors

$$E(A\mathbf{x} + b) = AE(\mathbf{x}) + b = A\mu + b.$$

Cette propriété reste vraie pour les matrices aléatoires : si Ξ est une matrice $p \times m$ aléatoire et A est une matrice $q \times p$ déterministe, alors $E(A\Xi) = AE(\Xi)$, $E(\Xi^T A^T) = E(\Xi^T)A^T$.

La matrice Σ de covariance (ou **la matrice de variance-covariance**) du vecteur aléatoire $\mathbf{x}$ est une matrice $p \times p$ définie par

$$\Sigma \stackrel{\text{déf}}{=} V(\mathbf{x}) = E((\mathbf{x} - \mu)(\mathbf{x} - \mu)^T) = (\sigma_{ij})_{i,j}$$

où

$$\sigma_{ij} = E((\xi_i - \mu_i)(\xi_j - \mu_j)) = \int ... \int (t_i - \mu_i)(t_j - \mu_j) f_{\mathbf{x}}(t_1, ..., t_p) dt_1 ... dt_p.$$

Comme $\sigma_{ij} = \sigma_{ji}$, Σ est une matrice symétrique. On note $\sigma_{ii} \stackrel{\text{déf}}{=} \sigma_i^2$ avec $\sigma_i \geq 0$.

Définissons également la **matrice de covariance** (ou **la matrice des covariances croisées**) des vecteurs aléatoires $\mathbf{x} \in \mathbb{R}^p$ et $\mathbf{y} \in \mathbb{R}^q$:

$$C(\mathbf{x}, \mathbf{y}) = E((\mathbf{x} - E(\mathbf{x}))(\mathbf{y} - E(\mathbf{y}))^T).$$

C'est une matrice $p \times q$. On dit que $\mathbf{x}$ est **orthogonal** à $\mathbf{y}$ (ou bien que $\mathbf{x}$ et $\mathbf{y}$ sont **non-corrélés**) et on écrit $\mathbf{x} \perp \mathbf{y}$ si $C(\mathbf{x},\mathbf{y}) = \mathbf{0}$ (matrice $p \times q$ nulle).

3.1.5. Propriétés des matrices de covariance. Soient $\mathbf{x} \in \mathbb{R}^p$ et $\mathbf{y} \in \mathbb{R}^q$ deux vecteurs aléatoires. Alors :

(C1) $\Sigma = E(\mathbf{x}\mathbf{x}^T) - \mu\mu^T$, où $\mu = E(\mathbf{x})$.

(C2) Pour tout $a \in \mathbb{R}^p$, $\mathrm{Var}(a^T\mathbf{x}) = a^T V(\mathbf{x})a$.

Preuve. Par linéarité de l'espérance,

$$\begin{aligned}\mathrm{Var}(a^T\mathbf{x}) &= E((a^T\mathbf{x} - E(a^T\mathbf{x}))^2) = E\left([a^T(\mathbf{x} - E(\mathbf{x}))]^2\right) = E\left(a^T(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T a\right)\\ &= a^T E\left((\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\right) a = a^T V(\mathbf{x})a.\end{aligned}$$

∎

(C3) $\Sigma = V(\mathbf{x})$ est une matrice symétrique et positive. En effet, $\mathrm{Var}(a^T\mathbf{x}) \geq 0$ (variance d'une v.a. réelle) et, vu (C2), $V(\mathbf{x}) \geq 0$.

(C4) Soit A une matrice $q \times p$ et $b \in \mathbb{R}^q$. Alors $V(A\mathbf{x}+b) = AV(\mathbf{x})A^T$.

Preuve. Soit $\mathbf{y} = A\mathbf{x} + b$, alors par linéarité de l'espérance,

$$E(\mathbf{y}) = E(A\mathbf{x}+b) = A\mu + b \quad \text{et} \quad \mathbf{y} - E(\mathbf{y}) = A(\mathbf{x}-\mu).$$

Par linéarité de l'espérance pour les matrices aléatoires,

$$V(\mathbf{y}) = E(A(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T A^T) = AV(\mathbf{x})A^T.$$

∎

(C5) $C(\mathbf{x},\mathbf{x}) = V(\mathbf{x})$.

(C6) $C(\mathbf{x},\mathbf{y}) = C(\mathbf{y},\mathbf{x})^T$.

(C7) Pour deux vecteurs aléatoires $\mathbf{x}_1 \in \mathbb{R}^p$ et $\mathbf{x}_2 \in \mathbb{R}^p$, on a $C(\mathbf{x}_1+\mathbf{x}_2,\mathbf{y}) = C(\mathbf{x}_1,\mathbf{y}) + C(\mathbf{x}_2,\mathbf{y})$.

(C8) Si A est une matrice $m \times p$ et B est une matrice $k \times q$, alors $C(A\mathbf{x}, B\mathbf{y}) = AC(\mathbf{x},\mathbf{y})B^T$.

(C9) Si $\mathbf{x}$ et $\mathbf{y}$ sont deux vecteurs aléatoires de même dimension p,

$$V(\mathbf{x}+\mathbf{y}) = V(\mathbf{x}) + C(\mathbf{x},\mathbf{y}) + C(\mathbf{y},\mathbf{x}) + V(\mathbf{y}) = V(\mathbf{x}) + C(\mathbf{x},\mathbf{y}) + C(\mathbf{x},\mathbf{y})^T + V(\mathbf{y}).$$

(C10) Si $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$, alors $C(\mathbf{x},\mathbf{y}) = \mathbf{0}$ (matrice $p \times q$ nulle). L'implication inverse n'est pas vraie. Un contre-exemple est déjà donné dans le cas $p = q = 1$ (Exemple 2.4).

La **matrice de corrélation** P du vecteur aléatoire $\mathbf{x}$ est définie par $P = (\rho_{ij})_{1 \leq i,j \leq p}$ avec

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j},$$

pourvu que tous les σ_i soient strictement positifs. On remarque que :

– Les éléments diagonaux $\rho_{ii} = 1$, $i = 1, ..., p$.

– La matrice P est positive. En effet, $P = \Delta^{-1}\Sigma\Delta^{-1}$ avec la matrice diagonale $\Delta = \mathrm{Diag}(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{pp}})$, donc la positivité de Σ implique que $P \geq 0$.

3.2. Loi normale multivariée

3.2.1. Loi normale dans $\mathbb{R}$. On rappelle que la loi normale $\mathcal{N}(\mu, \sigma^2)$ dans $\mathbb{R}$ est la loi de densité

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big),$$

où $\mu \in \mathbb{R}$ est la moyenne et $\sigma^2 > 0$ est la variance. La fonction caractéristique de la loi normale $\mathcal{N}(\mu, \sigma^2)$ vaut

$$\phi(t) = \exp\Big(i\mu t - \frac{\sigma^2 t^2}{2}\Big),$$

en particulier, $\phi(t) = e^{-t^2/2}$ pour la loi $\mathcal{N}(0, 1)$. Par convention, nous allons inclure les lois dégénérées (lois de Dirac) dans la famille des lois normales. Soit $\mu \in \mathbb{R}$. La v.a. X suit la loi de Dirac en μ si $P(X = \mu) = 1$, $\phi(t) = e^{i\mu t}$.

3.2.2. La loi $\mathcal{N}_p(0, I)$. Notons $\mathcal{N}_p(0, I)$, où I désigne la matrice unité $p \times p$, la loi du vecteur aléatoire $\mathbf{x} = (\xi_1, ..., \xi_p)^T$ dont les composantes ξ_i, $i = 1, ..., p$, sont des variables aléatoires i.i.d. de loi $\mathcal{N}(0, 1)$.

Propriétés de la loi $\mathcal{N}_p(0, I)$.

1^o. La moyenne et la matrice de covariance de $\mathbf{x} \sim \mathcal{N}_p(0, I)$ sont : $E(\mathbf{x}) = 0$, $V(\mathbf{x}) = I$.

2^o. La loi $\mathcal{N}_p(0, I)$ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$ de densité

$$f_{\mathbf{x}}(u) = (2\pi)^{-p/2} \exp\Big(- \frac{1}{2} u^T u\Big), \quad u \in \mathbb{R}^p.$$

En effet,

$$f_{\mathbf{x}}(u) = \prod_{i=1}^p \varphi(u_i) = (2\pi)^{-p/2} \prod_{i=1}^p \exp\Big(- \frac{u_i^2}{2}\Big) = (2\pi)^{-p/2} \exp\Big(- \frac{1}{2} u^T u\Big)$$

où $u = (u_1, ..., u_p)^T$ et $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ est la densité de la loi $\mathcal{N}(0, 1)$.

3^o. La fonction caractéristique de $\mathcal{N}_p(0, I)$ vaut

$$\begin{aligned}\phi_{\mathbf{x}}(a) &= E\left(e^{ia^T\mathbf{x}}\right) = E\left(\prod_{j=1}^p e^{ia_j\xi_j}\right) \\ &= \prod_{j=1}^p E\left(e^{ia_j\xi_j}\right) = \prod_{j=1}^p e^{-a_j^2/2} = \exp\Big(- \frac{1}{2} a^T a\Big),\end{aligned}$$

où $a = (a_1, ..., a_p)^T \in \mathbb{R}^p$.

3.2.3. Loi normale dans $\mathbb{R}^p$. On dit que deux vecteurs aléatoires $\mathbf{x} \in \mathbb{R}^p$ et $\mathbf{y} \in \mathbb{R}^p$ sont de même loi et on écrit

$$\mathbf{x} \stackrel{D}{=} \mathbf{y}$$

si et seulement si $P(\mathbf{x} \in B) = P(\mathbf{y} \in B)$ pour tout $B \in \mathcal{B}(\mathbb{R}^p)$, où $\mathcal{B}(\mathbb{R}^p)$ est la tribu borélienne de $\mathbb{R}^p$.

On se rappelle le résultat suivant du cours de probabilités (cf. J.Lacroix, P.Priouret *Probabilités approfondies*, Polycopié du cours, Université Paris 6).

Lemme 3.1. *Soient* $\mathbf{x} \in \mathbb{R}^p$ *et* $\mathbf{y} \in \mathbb{R}^p$ *deux vecteurs aléatoires. Alors* $\mathbf{x} \stackrel{D}{=} \mathbf{y}$ *si et seulement si* $\phi_{\mathbf{x}}(a) = \phi_{\mathbf{y}}(a)$ *pour tout* $a \in \mathbb{R}^p$*.*

Définition 3.2. *(Première définition de la loi normale multivariée.) Un vecteur aléatoire* $\mathbf{x}$ *suit une loi normale dans* $\mathbb{R}^p$ *si et seulement si il existe* A*, une matrice* $p \times p$*, et un vecteur* $\mu \in \mathbb{R}^p$ *tels que*

$$\mathbf{x} \stackrel{D}{=} A\mathbf{y} + \mu \quad \textit{avec} \quad \mathbf{y} \sim \mathcal{N}_p(0, I). \tag{3.5}$$

Remarque. On dit aussi que $\mathbf{x}$ est un *vecteur normal dans* $\mathbb{R}^p$ ou bien un *vecteur gaussien dans* $\mathbb{R}^p$. Une loi normale est parfois appelée *loi de Laplace-Gauss.*

Les propriétés suivantes découlent facilement de la Définition 3.2 :

1^o. $E(\mathbf{x}) = \mu$.

2^o. $V(\mathbf{x}) = AV(\mathbf{y})A^T = AA^T$. On désigne $\Sigma \stackrel{\text{déf}}{=} AA^T$.

3^o. La fonction caractéristique d'un vecteur normal $\mathbf{x}$ vaut

$$\begin{aligned}\phi_{\mathbf{x}}(a) &= E\left(e^{ia^T\mathbf{x}}\right) = E\left(e^{ia^T(A\mathbf{y}+\mu)}\right) = e^{ia^T\mu}E\left(e^{ib^T\mathbf{y}}\right) \quad (\text{avec } b = A^Ta) \\ &= e^{ia^T\mu - \frac{1}{2}b^Tb} = e^{ia^T\mu - \frac{1}{2}a^T\Sigma a}. \end{aligned} \tag{3.6}$$

De plus, seules les lois normales peuvent avoir une fonction caractéristique de cette forme, comme le montre le résultat suivant.

Théorème 3.1. *Soit* $\phi : \mathbb{R}^p \to \mathbb{C}$ *une fonction à valeurs complexes. Alors,* ϕ *est la fonction caractéristique d'une loi normale si et seulement si il existe* $\mu \in \mathbb{R}^p$ *et une matrice* $p \times p$ *symétrique positive* Σ *tels que*

$$\phi(a) = e^{ia^T\mu - \frac{1}{2}a^T\Sigma a}, \quad a \in \mathbb{R}^p. \tag{3.7}$$

Preuve. La nécessité est démontrée dans (3.6). Pour prouver la suffisance, il faut montrer qu'il existe un vecteur normal $\mathbf{x}$ dans $\mathbb{R}^p$ tel que $\phi(\cdot)$ soit sa fonction caractéristique. Considérons le vecteur aléatoire $\mathbf{x} = \Sigma^{1/2}\mathbf{y} + \mu$, où $\mathbf{y} \sim \mathcal{N}_p(0, I)$. Par Définition 3.2, $\mathbf{x}$ est un vecteur normal dans $\mathbb{R}^p$. La moyenne et la matrice de covariance de $\mathbf{x}$ sont $E(\mathbf{x}) = \mu$ et $V(\mathbf{x}) = \Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma$, vu la propriété (C4) des matrices de covariance. D'après (3.6) la fonction caractéristique de $\mathbf{x}$ coïncide avec la fonction ϕ donnée dans (3.7). ∎

Le Théorème 3.1 et le Lemme 3.1 entraînent la conséquence suivante : *toute loi normale dans* $\mathbb{R}^p$ *est entièrement déterminée par la donnée de sa moyenne et de sa matrice de covariance.* Ceci explique que par la suite on utilisera la notation

$$\mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma)$$

pour un vecteur aléatoire normal $\mathbf{x}$ de moyenne μ et de matrice de covariance Σ. Une autre conséquence du Théorème 3.1 et du Lemme 3.1 est que l'on peut aussi définir la loi normale de façon suivante.

Définition 3.3. *(Deuxième définition équivalente de la loi normale multivariée.) Un vecteur aléatoire* $\mathbf{x}$ *suit une loi normale dans* $\mathbb{R}^p$ *si et seulement si la fonction caractéristique de* $\mathbf{x}$ *est de la forme*

$$\phi_{\mathbf{x}}(a) = e^{ia^T\mu - \frac{1}{2}a^T\Sigma a}$$

où $\mu \in \mathbb{R}^p$ *et* Σ *est une matrice* $p \times p$ *symétrique positive. Dans ce cas* μ *est la moyenne et* Σ *est la matrice de covariance de* $\mathbf{x}$*, et on écrit :*

$$\mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma).$$

Proposition 3.3. *Soit* $\mu \in \mathbb{R}^p$ *et soit* Σ *une matrice* $p \times p$ *symétrique positive. Alors*

$$\mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma) \quad \Longleftrightarrow \quad \mathbf{x} \stackrel{D}{=} \Sigma^{1/2}\mathbf{y} + \mu, \ \textit{où } \mathbf{y} \sim \mathcal{N}_p(0, I).$$

Remarque. Pour une matrice symétrique positive Σ il existe, en général, plusieurs matrices carrées A telles que $\Sigma = AA^T$. Alors, la matrice A dans (3.5) n'est pas définie de façon unique : on peut obtenir la même loi normale par plusieurs transformations équivalentes A du vecteur $\mathbf{y} \sim \mathcal{N}_p(0, I)$. On peut prendre, par exemple, la matrice symétrique $A = \Sigma^{1/2}$, mais aussi des matrice non-symétriques A. En effet, d'après le Théorème de décomposition spectrale, on peut écrire

$$\Sigma = \Gamma\Lambda\Gamma^T = \sum_{j=1}^{p} \lambda_j \boldsymbol{\gamma}_{(j)}\boldsymbol{\gamma}_{(j)}^T = \sum_{j=1}^{k} \mathbf{a}_{(j)}\mathbf{a}_{(j)}^T = AA^T$$

où Γ est une matrice $p \times p$ orthogonale, $\Lambda = \mathrm{Diag}(\lambda_i)$ est une matrice $p \times p$ diagonale de rang $\mathrm{Rang}(\Lambda) = k \le p$, les $\boldsymbol{\gamma}_{(j)}$ sont les colonnes de Γ, $\mathbf{a}_{(j)} = \sqrt{\lambda_j}\boldsymbol{\gamma}_{(j)}$ et A est une matrice $p \times p$ définie par $A = (\mathbf{a}_{(1)}, ..., \mathbf{a}_{(k)}, 0, ..., 0)$. Encore un autre exemple de matrice non-symétrique (triangulaire) A vérifiant $\Sigma = AA^T$ est donné dans l'Exercice 3.6.

Nous allons distinguer entre deux types de lois normales dans $\mathbb{R}^p$: lois normales non-dégénérées et lois normales dégénérées.

3.2.4. Loi normale non-dégénérée dans $\mathbb{R}^p$. C'est une loi normale dont la matrice de covariance Σ est strictement positive, $\Sigma > 0$ ($\Leftrightarrow$ $\mathrm{Det}(\Sigma) > 0$). Alors, il existe une matrice symétrique et positive $A_1 = \Sigma^{1/2}$ (racine carrée de Σ, voir la définition au Paragraphe 3.1.3), telle que $\Sigma = A_1^2 = A_1^T A_1 = A_1 A_1^T$. Par ailleurs, $[\mathrm{Det}(A_1)]^2 = \mathrm{Det}(\Sigma) > 0$, donc $\mathrm{Det}(A_1) > 0$ et A_1 est inversible. D'après la Définition 3.3, si le vecteur aléatoire $\mathbf{x}$ suit la loi normale $\mathcal{N}_p(\mu, \Sigma)$, sa fonction caractéristique vaut

$$\phi_{\mathbf{x}}(a) = e^{ia^T\mu - \frac{1}{2}a^T\Sigma a}, \quad a \in \mathbb{R}^p,$$

et d'après (3.6) on a

$$\phi_{\mathbf{x}}(a) = E\left(e^{ia^T(A_1\mathbf{y}+\mu)}\right) = \phi_{A_1\mathbf{y}+\mu}(a),$$

où $\mathbf{y} \sim \mathcal{N}_p(0, I)$. Alors, vu le Lemme 3.1,

$$\mathbf{x} \stackrel{D}{=} A_1\mathbf{y} + \mu.$$

D'après le Corollaire 3.1, la densité de $\mathbf{x}$ est

$$\begin{aligned} f_{\mathbf{x}}(u) &= \mathrm{Det}(A_1^{-1}) f_{\mathbf{y}}(A_1^{-1}(u-\mu)) = \frac{1}{\mathrm{Det}(A_1)} f_{\mathbf{y}}(A_1^{-1}(u-\mu)) \\ &= \frac{1}{(2\pi)^{p/2}\sqrt{\mathrm{Det}(\Sigma)}} \exp\left(-\frac{1}{2}(u-\mu)^T\Sigma^{-1}(u-\mu)\right), \quad u \in \mathbb{R}^p. \end{aligned}$$

Nous avons donc démontré le résultat suivant.

Corollaire 3.2. *La loi normale non-dégénérée $\mathcal{N}_p(\mu, \Sigma)$, où $\mu \in \mathbb{R}^p$ et $\Sigma > 0$, est une loi admettant la densité f par rapport à la mesure de Lebesgue dans $\mathbb{R}^p$ de la forme*

$$\boxed{f(t) = \frac{1}{(2\pi)^{p/2}\sqrt{\mathrm{Det}(\Sigma)}} \exp\left(-\frac{1}{2}(t-\mu)^T\Sigma^{-1}(t-\mu)\right), \quad t \in \mathbb{R}^p.}$$

3.2.5. Loi normale dégénérée dans $\mathbb{R}^p$. C'est une loi normale dont la matrice de covariance Σ est dégénérée : $\mathrm{Det}(\Sigma) = 0$ (autrement dit, $\mathrm{Rang}(\Sigma) = k < p$). Par exemple, on peut considérer $\Sigma = \mathbf{0}$ (matrice nulle), alors la fonction caractéristique de $\mathbf{x} \sim \mathcal{N}_p(\mu, \mathbf{0})$ est $\phi_{\mathbf{x}}(a) = e^{ia^T\mu}$ et $\mathbf{x}$ suit la loi de Dirac en μ.

D'après la Proposition 3.3, si $\mathrm{Rang}(\Sigma) = k \le p$ (et donc $\mathrm{Rang}(\Sigma^{1/2}) = k$), la loi de $\mathbf{x}$ concentre toute sa masse sur un sous-espace affine de $\mathbb{R}^p$ de dimension k.

Proposition 3.4. *Soit $\mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma)$ et $\mathrm{Rang}(\Sigma) = k < p$. Alors, il existe un sous-espace linéaire $H \subset \mathbb{R}^p$ de dimension $p-k$ tel que, pour tout vecteur $a \in H$, la combinaison linéaire $a^T\mathbf{x}$ suit une loi de Dirac.*

Preuve. Soit $H = \mathrm{Ker}(\Sigma)$, alors $\dim(H) = p-k$. Si $a \in H$ (i.e. $\Sigma a = 0$), la fonction caractéristique de la v.a. $a^T\mathbf{x}$ est

$$\phi(u) = E\left(e^{i(a^T\mathbf{x})u}\right) = E\left(e^{i(ua)^T\mathbf{x}}\right) = e^{i(ua)^T\mu - \frac{1}{2}(ua)^T\Sigma(ua)} = e^{iu(a^T\mu)}.$$

La loi de $a^T\mathbf{x}$ est donc la loi de Dirac $\mathcal{N}(a^T\mu, 0)$. ∎

Par conséquent, pour tout $a \in H$, la variable aléatoire $a^T(\mathbf{x} - E(\mathbf{x}))$ suit la loi de Dirac en 0. La Proposition 3.4 montre alors que toute la masse d'une loi normale dégénérée est concentrée sur un sous-espace affine de $\mathbb{R}^p$ de dimension $k < p$. Une loi normale dégénérée n'est pas absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$.

3.2.6. Lois des combinaisons linéaires.

Théorème 3.2. *(Troisième définition équivalente de la loi normale multivariée.) Un vecteur aléatoire $\mathbf{x} \in \mathbb{R}^p$ suit une loi normale si et seulement si, pour tous $a \in \mathbb{R}^p$, les combinaisons linéaires $a^T\mathbf{x}$ sont des variables aléatoires normales réelles.*

Remarque. Rappelons que, par convention, une loi de Dirac est un cas particulier de loi normale correspondant à la variance nulle.

Preuve. Tout d'abord, on observe que, pour tout $a \in \mathbb{R}^p$ et tout $u \in \mathbb{R}$, la fonction caractéristique $\phi_{a^T\mathbf{x}}(u)$ de la variable aléatoire réelle $a^T\mathbf{x}$ est liée avec celle du vecteur $\mathbf{x}$:

$$\phi_{a^T\mathbf{x}}(u) = E\left(e^{ia^T\mathbf{x}u}\right) = \phi_{\mathbf{x}}(ua). \tag{3.8}$$

Nécessité. Soit $\mathbf{x}$ un vecteur normal dans $\mathbb{R}^p$. Montrons que $a^T\mathbf{x}$ est une variable aléatoire normale pour tout $a \in \mathbb{R}^p$. On utilise (3.8) pour obtenir, pour tout $u \in \mathbb{R}$,

$$\phi_{a^T\mathbf{x}}(u) = e^{iua^T\mu - \frac{1}{2}u^2a^T\Sigma a},$$

où μ et Σ sont la moyenne et la matrice de covariance de $\mathbf{x}$. Alors,

$$\phi_{a^T\mathbf{x}}(u) = e^{i\mu_0 u - \frac{1}{2}u^2\sigma_0^2}$$

avec $\mu_0 = a^T\mu$ et $\sigma_0^2 = a^T\Sigma a$. Par conséquent,

$$a^T\mathbf{x} \sim \mathcal{N}(\mu_0, \sigma_0^2) = \mathcal{N}(a^T\mu, a^T\Sigma a).$$

Suffisance. Réciproquement, montrons que si $a^T\mathbf{x}$ est une variable normale pour tout $a \in \mathbb{R}^p$, alors $\mathbf{x}$ est un vecteur normal dans $\mathbb{R}^p$. On remarque d'abord que si $a^T\mathbf{x}$ est une variable normale pour tout $a \in \mathbb{R}^p$, alors $E(\|\mathbf{x}\|^2) < \infty$ où $\|\mathbf{x}\|$ est la norme Euclidienne de $\mathbf{x}$ (pour le voir il suffit de prendre successivement comme a les vecteurs de la base canonique de $\mathbb{R}^p$). Donc, la moyenne $\mu = E(\mathbf{x})$ et la matrice de covariance $\Sigma = V(\mathbf{x})$ sont bien définies.

On fixe maintenant $a \in \mathbb{R}^p$. Par hypothèse, il existe $m \in \mathbb{R}$ et $s^2 \geq 0$ tels que $a^T\mathbf{x} \sim \mathcal{N}(m, s^2)$. Vu la linéarité de l'espérance et la propriété (C2) des matrices de covariance,

$$m = E(a^T\mathbf{x}) = a^T\mu, \quad s^2 = \mathrm{Var}(a^T\mathbf{x}) = a^T\Sigma a.$$

Or, la fonction caractéristique de $a^T\mathbf{x}$ est

$$\phi_{a^T\mathbf{x}}(u) = e^{imu - \frac{1}{2}s^2u^2} = e^{iua^T\mu - \frac{1}{2}u^2a^T\Sigma a}.$$

En utilisant (3.8) on obtient

$$\phi_{\mathbf{x}}(a) = \phi_{a^T\mathbf{x}}(1) = e^{ia^T\mu - \frac{1}{2}a^T\Sigma a}.$$

Puisque $a \in \mathbb{R}^p$ est arbitraire, on en déduit (vu la Définition 3.3) que $\mathbf{x}$ est un vecteur aléatoire normal dans $\mathbb{R}^p$ de moyenne μ et de matrice de covariance Σ. ∎

3.2.7. Propriétés de la loi normale multivariée. Soit $\mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma)$, où $\mu \in \mathbb{R}^p$ et Σ est une matrice $p \times p$ symétrique et positive ($\Sigma \geq 0$). Notons quelques propriétés de $\mathbf{x}$ que l'on va utiliser par la suite et dont certaines ont été démontrées dans les paragraphes précédents :

(N1) Soit $\Sigma > 0$, alors le vecteur aléatoire $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \mu)$ vérifie

$$\mathbf{y} \sim \mathcal{N}_p(0, I).$$

(N2) *Les combinaisons linéaires $a^T\mathbf{x}$, pour tout $a \in \mathbb{R}^p$, sont des variables aléatoires normales réelles :*

$$a^T\mathbf{x} \sim \mathcal{N}(a^T\mu, a^T\Sigma a).$$

En particulier, les densités marginales de la loi $\mathcal{N}_p(\mu, \Sigma)$ sont normales. Le réciproque n'est pas vrai (voir l'Exemple 2.1).

(N3) *Toute transformation affine d'un vecteur normal est un vecteur normal* : si $\mathbf{y} = B\mathbf{x} + c$, où B est une matrice déterministe $q \times p$ et $c \in \mathbb{R}^q$ est un vecteur déterministe, alors

$$\mathbf{y} \sim \mathcal{N}_q(B\mu + c, B\Sigma B^T).$$

Preuve. La loi de $\mathbf{y}$ est normale car toute combinaison linéaire de $a^T\mathbf{y}$ est une v.a. normale réelle. En effet, pour tout $a \in \mathbb{R}^q$,

$$a^T\mathbf{y} = a^T B\mathbf{x} + a^T c = b^T\mathbf{x} + d$$

où $b = B^T a \in \mathbb{R}^p$ et $d = a^T c$. D'après le Théorème 3.2 on obtient que les combinaisons linéaires $b^T\mathbf{x}$ sont des v.a. normales pour tout $b \in \mathbb{R}^p$. Il s'ensuit que les combinaisons linéaires $a^T\mathbf{y}$ sont normales pour tout $a \in \mathbb{R}^q$ et alors, d'après ce même théorème, $\mathbf{y}$ est un vecteur normal dans $\mathbb{R}^q$. Sa moyenne et sa matrice de covariance sont données par

$$E(\mathbf{y}) = B\mu + c, \quad V(\mathbf{y}) = B\Sigma B^T,$$

vu la propriété (C4) des matrices de covariance. ∎

(N4) *La loi $\mathcal{N}_p(0, \sigma^2 I)$ est invariante par rapport aux transformations orthogonales* : si Γ est une matrice orthogonale, $\sigma^2 \geq 0$ et $\mathbf{x} \sim \mathcal{N}_p(0, \sigma^2 I)$, alors $\Gamma\mathbf{x} \sim \mathcal{N}_p(0, \sigma^2 I)$.
Preuve. On utilise (N3) avec $B = \Gamma$, $c = 0$.

(N5) *Tout sous-ensemble de coordonnées d'un vecteur normal est un vecteur normal* : soit $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$, où $\mathbf{x}_1 \in \mathbb{R}^k$ et $\mathbf{x}_2 \in \mathbb{R}^{p-k}$, alors $\mathbf{x}_1$ et $\mathbf{x}_2$ sont des vecteurs normaux.
Preuve. On utilise (N3) avec $c = 0$, en prenant comme B la matrice $k \times p$ de la forme $B = (I_k, \mathbf{0})$, où I_k est la matrice unité $k \times k$. On en déduit que $\mathbf{x}_1$ est normal. Pour $\mathbf{x}_2$ on prend la matrice $(p-k) \times p$ définie par $B = (\mathbf{0}, I_{p-k})$. ∎

(N6) *Deux vecteurs aléatoires $\mathbf{x}$ et $\mathbf{y}$ tels que* **la loi jointe de $(\mathbf{x}, \mathbf{y})$ est normale** *sont indépendants si et seulement si $C(\mathbf{x}, \mathbf{y}) = \mathbf{0}$.*
Preuve. La nécessité de la condition $C(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ découle de la propriété (C10) des matrices de covariance.
Suffisance. Soit $\mathbf{z}$ un vecteur normal dans $\mathbb{R}^{q+p}$ tel que $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$, où $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^q$, et $C(\mathbf{x}, \mathbf{y}) = \mathbf{0}$. Pour prouver que $\mathbf{x}$ et $\mathbf{y}$ sont indépendants, il suffit de montrer (vu (3.1)) que la fonction caractéristique $\phi_{\mathbf{z}}(u)$ de $\mathbf{z}$ peut être décomposée comme

$$\phi_{\mathbf{z}}(u) = \phi_{\mathbf{x}}(a)\phi_{\mathbf{y}}(b) \quad \text{avec } u = \begin{pmatrix} a \\ b \end{pmatrix}, \quad a \in \mathbb{R}^p, \quad b \in \mathbb{R}^q.$$

Notons que

$$E(\mathbf{z}) = \begin{pmatrix} E(\mathbf{x}) \\ E(\mathbf{y}) \end{pmatrix}, \quad V(\mathbf{z}) = \begin{pmatrix} V(\mathbf{x}) & C(\mathbf{x}, \mathbf{y}) \\ C(\mathbf{y}, \mathbf{x}) & V(\mathbf{y}) \end{pmatrix} = \begin{pmatrix} V(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & V(\mathbf{y}) \end{pmatrix}.$$

La fonction caractéristique $\phi_{\mathbf{z}}(u)$ est donc

$$\begin{aligned}\phi_{\mathbf{z}}(u) &= \phi_{\mathbf{z}}(a, b) = \exp\left[i(a^T E(\mathbf{x}) + b^T E(\mathbf{y})) - \frac{1}{2}(a^T, b^T)V(\mathbf{z})\begin{pmatrix} a \\ b \end{pmatrix}\right] \\ &= \exp\left[ia^T E(\mathbf{x}) - \frac{1}{2}a^T V(\mathbf{x})a\right]\exp\left[ib^T E(\mathbf{y}) - \frac{1}{2}b^T V(\mathbf{y})b\right] = \phi_{\mathbf{x}}(a)\phi_{\mathbf{y}}(b).\end{aligned}$$

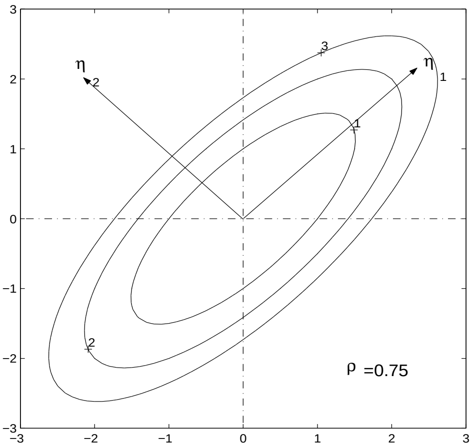

Figure 3.1. Exemple d'ellipses de concentration d'une loi normale.

pour tout $u = \begin{pmatrix} a \\ b \end{pmatrix}$. ∎

On a aussi le résultat plus général suivant dont la preuve est analogue à celle de la propriété (N6).

Proposition 3.5. *Soient* $\mathbf{x}_1, \ldots, \mathbf{x}_J$ *des vecteurs aléatoires tels que :*
(i) la loi **jointe** *de* $(\mathbf{x}_1, \ldots, \mathbf{x}_J)$ *est normale,*
(ii) les matrices de covariance $C(\mathbf{x}_k, \mathbf{x}_j) = \mathbf{0}$*,* $k \neq j$*, pour* $k, j = 1, \ldots, J$*.*
Alors, les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_J$ *sont mutuellement indépendants.*

Géométrie de la loi normale multivariée. Soit $\Sigma > 0$. La densité de $\mathcal{N}_p(\mu, \Sigma)$ est constante sur les surfaces

$$\{x : \ (x-\mu)^T \Sigma^{-1} (x-\mu) = C\},$$

où $C \geq 0$. Généralement, pour une densité de probabilité quelconque f, les ensembles $\{x : f(x) \geq c\}$ avec $c > 0$ sont appelés *ensembles de niveau* de la loi correspondante. Pour une loi normale, les ensembles de niveau sont des ellipsoïdes. On les appelle *ellipsoïdes de concentration.*

3.3. Espérance conditionnelle d'un vecteur aléatoire

Soient $\mathbf{x} = (\xi_1, ..., \xi_p)^T$ et $\mathbf{y} = (\eta_1, ..., \eta_q)^T$ deux vecteurs aléatoires. Dans ce paragraphe, nous supposerons que la densité jointe $f_{\mathbf{x},\mathbf{y}}(t_1, ..., t_p, s_1, ..., s_q)$ de $(\mathbf{x}, \mathbf{y})$ existe. **L'espérance conditionnelle de y sachant x** est définie alors comme un vecteur aléatoire dans $\mathbb{R}^q$ de la forme

$$E(\mathbf{y}|\mathbf{x}) = (E(\eta_1|\mathbf{x}), \ldots, E(\eta_q|\mathbf{x}))^T$$

avec $E(\eta_j|\mathbf{x}) = g_j(\mathbf{x})$ où, pour tout $t = (t_1, ..., t_p) \in \mathbb{R}^p$,

$$g_j(t) = E(\eta_j|\mathbf{x} = t) = \int_{-\infty}^{\infty} s f_{\eta_j|\mathbf{x}}(s|t) ds = \int_{-\infty}^{\infty} s f_{\eta_j|\mathbf{x}}(s|t_1, ..., t_p) ds$$

et (cf. (3.3))

$$f_{\eta_j|\mathbf{x}}(s|t_1,...,t_p) = \begin{cases} \frac{f_{\eta_j,\mathbf{x}}(s,t_1,...,t_p)}{f_{\mathbf{x}}(t_1,...,t_p)}, & \text{si } f_{\mathbf{x}}(t_1,...,t_p) > 0, \\ f_{\eta_j}(s), & \text{si } f_{\mathbf{x}}(t_1,...,t_p) = 0. \end{cases}$$

Il est facile de voir que $E(\eta_j|\mathbf{x})$ est fini si $E(|\eta_j|) < \infty$. Toutes les propriétés de l'espérance conditionnelle établies au Chapitre 2 restent vraies pour des vecteurs aléatoires de dimension quelconque, en particulier, le Théorème de l'espérance itérée :

$$E(E(\mathbf{y}|\mathbf{x})) = E(\mathbf{y})$$

et le Théorème de substitution (sous la forme matricielle) :

$$E(h(\mathbf{x})\mathbf{y}^T|\mathbf{x}) = h(\mathbf{x})E(\mathbf{y}^T|\mathbf{x}) \quad \text{(p.s.)}$$

pour toute fonction borélienne $h : \mathbb{R}^p \to \mathbb{R}^q$. La *matrice de covariance conditionnelle* est définie par :

$$\begin{aligned} V(\mathbf{y}|\mathbf{x}) &= E\Big((\mathbf{y} - E(\mathbf{y}|\mathbf{x}))(\mathbf{y} - E(\mathbf{y}|\mathbf{x}))^T|\mathbf{x}\Big) \\ &= E(\mathbf{y}\mathbf{y}^T|\mathbf{x}) - E(\mathbf{y}|\mathbf{x})E(\mathbf{y}|\mathbf{x})^T. \end{aligned}$$

3.3.1. Théorème de meilleure prévision. Notons $\|a\| = \sqrt{a_1^2 + ... + a_p^2}$ la norme Euclidienne du vecteur $a = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$. Soit $L_2(P, \mathbb{R}^q)$ l'espace de Hilbert de tous les vecteurs aléatoires $\mathbf{x}$ dans $\mathbb{R}^q$ de carré intégrable, i.e. tels que $E(\|\mathbf{x}\|^2) < \infty$ (cf. la définition de l'espace $L_2(P) = L_2(P, \mathbb{R}^1)$ au Chapitre 2).

Définition 3.4. *Soient* $\mathbf{x} \in \mathbb{R}^p$ *et* $\mathbf{y} \in L_2(P, \mathbb{R}^q)$ *deux vecteurs aléatoires et soit* G *une fonction borélienne de* $\mathbb{R}^p$ *dans* $\mathbb{R}^q$*. Un vecteur aléatoire* $G(\mathbf{x})$ *est appelé* **meilleure prévision** *de* $\mathbf{y}$ *étant donné* $\mathbf{x}$ *si*

$$E\left((\mathbf{y} - G(\mathbf{x}))(\mathbf{y} - G(\mathbf{x}))^T\right) \leq E\left((\mathbf{y} - H(\mathbf{x}))(\mathbf{y} - H(\mathbf{x}))^T\right) \tag{3.9}$$

pour toutes les fonctions boréliennes H *de* $\mathbb{R}^p$ *dans* $\mathbb{R}^q$*.*

EXERCICE 3.1. Montrer que (3.9) implique

$$E(\|\mathbf{y} - G(\mathbf{x})\|^2) = \min_{H(\cdot)} E(\|\mathbf{y} - H(\mathbf{x})\|^2),$$

où le minimum est pris sur toutes les fonctions boréliennes H de $\mathbb{R}^p$ dans $\mathbb{R}^q$.

Comme dans le cas $p = q = 1$, on obtient le Théorème de meilleure prévision :

Théorème 3.3. *Soient* $\mathbf{x} \in \mathbb{R}^p$ *et* $\mathbf{y} \in L_2(P, \mathbb{R}^q)$ *deux vecteurs aléatoires tels que la loi jointe de* $(\mathbf{x}, \mathbf{y})$ *admet une densité par rapport à la mesure de Lebesgue sur* $\mathbb{R}^{p+q}$*. Alors, la meilleure prévision de* $\mathbf{y}$ *étant donné* $\mathbf{x}$*, unique à une équivalence près, est égale à l'espérance conditionnelle*

$$G(\mathbf{x}) = E(\mathbf{y}|\mathbf{x}).$$

Preuve. Il suffit de chercher le minimum parmi les fonctions $H(\cdot)$ telles que $E(\|H(\mathbf{x})\|^2) < \infty$. Pour une telle fonction $H(\mathbf{x})$:

$$\begin{aligned}
&E\left((H(\mathbf{x}) - \mathbf{y})(H(\mathbf{x}) - \mathbf{y})^T)\right) \\
&= E\left([(H(\mathbf{x}) - G(\mathbf{x})) + (G(\mathbf{x}) - \mathbf{y})][(H(\mathbf{x}) - G(\mathbf{x})) + (G(\mathbf{x}) - \mathbf{y})]^T\right) \\
&= E\left((H(\mathbf{x}) - G(\mathbf{x}))(H(\mathbf{x}) - G(\mathbf{x}))^T\right) + E\left((H(\mathbf{x}) - G(\mathbf{x}))(G(\mathbf{x}) - \mathbf{y})^T\right) \\
&\quad + E\left((G(\mathbf{x}) - \mathbf{y})(H(\mathbf{x}) - G(\mathbf{x}))^T\right) + E\left((G(\mathbf{x}) - \mathbf{y})(G(\mathbf{x}) - \mathbf{y})^T\right).
\end{aligned}$$

En utilisant le Théorème de l'espérance itérée et le Théorème de substitution, on obtient

$$\begin{aligned}
E\left((H(\mathbf{x}) - G(\mathbf{x}))(G(\mathbf{x}) - \mathbf{y})^T\right) &= E\left[E\left((H(\mathbf{x}) - G(\mathbf{x}))(G(\mathbf{x}) - \mathbf{y})^T|\mathbf{x}\right)\right] \\
&= E\left[(H(\mathbf{x}) - G(\mathbf{x}))E\left((G(\mathbf{x}) - \mathbf{y})^T|\mathbf{x}\right)\right] = \mathbf{0},
\end{aligned}$$

d'où découle le résultat du théorème. ∎

3.4. Théorème de corrélation normale

Les propriétés établies au Paragraphe 3.2.7 nous permettent d'obtenir le résultat suivant qui joue un rôle fondamental.

Théorème 3.4. (Théorème de corrélation normale.) *Soit un vecteur normal* $\mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma)$ *tel que*

$$\mathbf{x} = \begin{pmatrix} \xi \\ \theta \end{pmatrix}, \quad \xi \in \mathbb{R}^k, \quad \theta \in \mathbb{R}^l, \quad p = k + l, \quad \mu = \begin{pmatrix} \mu_\xi \\ \mu_\theta \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\theta} \\ \Sigma_{\theta\xi} & \Sigma_{\theta\theta} \end{pmatrix},$$

où $\Sigma_{\xi\xi}$ *est une matrice* $k \times k$, $\Sigma_{\theta\theta}$ *est une matrice* $l \times l$, *et* $\Sigma_{\theta\xi} = \Sigma_{\xi\theta}^T$ *est une matrice* $l \times k$. *Supposons que* $\Sigma > 0$. *Alors :*

(i) Presque sûrement,

$$\boxed{\begin{aligned} E(\theta|\xi) &= \mu_\theta + \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}(\xi - \mu_\xi), \\ V(\theta|\xi) &= \Sigma_{\theta\theta} - \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}\Sigma_{\xi\theta}. \end{aligned}} \tag{3.10}$$

(ii) La loi conditionnelle de θ *sachant que* $\xi = a$ *(avec* $a \in \mathbb{R}^k$ *déterministe) est normale :*

$$\mathcal{N}_l(\mu_\theta + \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}(a - \mu_\xi), V^*), \quad \text{où} \quad V^* \stackrel{\text{déf}}{=} \Sigma_{\theta\theta} - \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}\Sigma_{\xi\theta}.$$

(iii) Les vecteurs aléatoires ξ *et*

$$\eta = \theta - \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}\xi$$

sont indépendants.

REMARQUES.

(1) La **fonction de régression de θ sur ξ** est définie comme une fonction déterministe $a \mapsto E(\theta|\xi = a)$, $a \in \mathbb{R}^k$. Vu le Théorème de corrélation normale, on peut formuler la conclusion suivante.

Si la **loi jointe** *de* (ξ, θ) *est normale, la régression de* θ *sur* ξ *est linéaire.*

Important : soulignons qu'on a besoin ici de la **normalité de la loi jointe de** (ξ, θ). La normalité de ξ et de θ ne suffit pas : le fait que ξ et θ soient deux vecteurs normaux n'implique pas que la loi jointe de (ξ, θ) est normale (cf. Exemple 2.1).

(2) Si $\Sigma > 0$, la matrice V^* est aussi strictement positive : $V^* > 0$. En effet, comme $\Sigma > 0$, pour tous $a \in \mathbb{R}^k$, $b \in \mathbb{R}^l$, on a l'inégalité

$$(a^T \, b^T)\Sigma \begin{pmatrix} a \\ b \end{pmatrix} = (a^T \, b^T) \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\theta} \\ \Sigma_{\theta\xi} & \Sigma_{\theta\theta} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} > 0,$$

ce qui équivaut à

$$a^T \Sigma_{\xi\xi} a + a^T \Sigma_{\xi\theta} b + b^T \Sigma_{\theta\xi} a + b^T \Sigma_{\theta\theta} b > 0. \tag{3.11}$$

Si l'on choisit

$$a = -\Sigma_{\xi\xi}^{-1} \Sigma_{\xi\theta} b,$$

alors (3.11) s'écrit comme

$$-b^T \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \Sigma_{\xi\theta} b + b^T \Sigma_{\theta\theta} b > 0,$$

pour tout $b \in \mathbb{R}^l$, d'où

$$\Sigma_{\theta\theta} - \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \Sigma_{\xi\theta} > 0.$$

(3) Le Théorème de corrélation normale admet l'interprétation géométrique suivante : soit $L_2^{\xi}(P, \mathbb{R}^l)$ le sous-espace linéaire de $L_2(P, \mathbb{R}^l)$ constitué de tous les vecteurs aléatoires dans $L_2(P, \mathbb{R}^l)$ mesurables par rapport à ξ. Supposons que $\mu = 0$. Alors $\Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \xi$ est la projection orthogonale de θ sur $L_2^{\xi}(P, \mathbb{R}^l)$ et le vecteur $\eta = \theta - \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \xi$ (le *résidu*) est orthogonal à $L_2^{\xi}(P, \mathbb{R}^l)$.

Preuve du Théorème de corrélation normale.

Etape 1. Calculons $E(\eta)$ et $V(\eta)$. On a :

$$E(\eta) = E(\theta - \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \xi) = \mu_\theta - \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \mu_\xi.$$

En utilisant les propriétés (C9) et (C8) des matrices de covariance on trouve

$$\begin{aligned} V(\eta) &= V(\theta) - C(\theta, \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \xi) - C(\Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \xi, \theta) + V(\Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \xi) \\ &= \Sigma_{\theta\theta} - \Sigma_{\theta\xi} \left(\Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \right)^T - \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \Sigma_{\xi\theta} + \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} V(\xi) \left(\Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \right)^T \\ &= \Sigma_{\theta\theta} - \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \Sigma_{\xi\theta}. \end{aligned}$$

Etape 2. Montrons que η est orthogonal à ξ. En effet,

$$C(\eta, \xi) = C(\theta, \xi) - \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} C(\xi, \xi) = \Sigma_{\theta\xi} - \Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} \Sigma_{\xi\xi} = \mathbf{0}.$$

Etape 3. Notons que la loi jointe du couple (ξ, η) est normale. En effet,

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = A\mathbf{x} = A \begin{pmatrix} \xi \\ \theta \end{pmatrix},$$

où

$$A = \begin{pmatrix} I_k & \mathbf{0} \\ -\Sigma_{\theta\xi} \Sigma_{\xi\xi}^{-1} & I_l \end{pmatrix},$$

où I_k et I_l sont les matrices unité $k \times k$ et $l \times l$. Vu la propriété (N3) du Paragraphe 3.2.7, $\begin{pmatrix} \xi \\ \eta \end{pmatrix}$ est un vecteur normal dans $\mathbb{R}^{k+l}$.

Etape 4. Le résultat de l'Etape 3 et la propriété (N5) impliquent que η est un vecteur normal. En utilisant les expressions pour $E(\eta)$ et $V(\eta)$ de l'Etape 1, on obtient

$$\eta \sim \mathcal{N}_l \left(\mu_\theta - \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}\mu_\xi, V^* \right).$$

Etape 5. On conclut. La propriété (N6) et les résultats des Etapes 2 et 3 impliquent que η et ξ sont indépendants, ce qui démontre la partie (iii) du Théorème. Par ailleurs, notons que

$$\theta = \eta + \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}\xi, \tag{3.12}$$

où η est indépendant de ξ. Il s'ensuit que

$$\begin{aligned} E(\theta|\xi) &= E(\eta|\xi) + \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}\xi = E(\eta) + \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}\xi, \\ V(\theta|\xi) &= V(\eta|\xi) = V(\eta), \end{aligned}$$

et en utilisant le résultat de l'Etape 1, on obtient la partie (i) du Théorème. La partie (ii) est une conséquence directe de (3.12), de l'indépendance $\eta \perp\!\!\!\perp \xi$ et de la normalité de η. En effet, la loi conditionnelle de θ sachant que $\xi = a$ est la loi conditionnelle de $\eta + \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}a$ sachant que $\xi = a$. Comme $\eta \perp\!\!\!\perp \xi$, c'est la loi (non-conditionnelle) de $\eta + \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1}a$. Or, la loi de η est trouvée dans l'Etape 4. ∎

Remarque. Le Théorème de corrélation normale s'étend au cas où la matrice Σ est dégénérée mais $\Sigma_{\xi\xi} > 0$. Il vient, de la démonstration donnée ci-dessus, que la partie (iii) du théorème est valable dans ce cas. Il est facile de voir que la partie (i) l'est aussi, si l'on définit l'espérance conditionnelle $E(\theta|\xi)$ pour un vecteur θ de loi normale dégénérée comme la meilleure prévision de θ étant donné ξ. Pour obtenir la partie (ii) au cas dégénéré, il suffit d'utiliser une modification convenable de la définition de la loi conditionnelle. En outre, on peut s'affranchir même de l'hypothèse $\Sigma_{\xi\xi} > 0$ en faisant recours à la notion de matrice pseudo-inverse (voir l'Exercice 3.17 ci-après).

Exemple 3.1. Supposons que le couple (X, Y) suit une loi normale dans $\mathbb{R}^2$ avec les moyennes $\mu_X = E(X)$, $\mu_Y = E(Y)$, les variances $\sigma_X^2 = \mathrm{Var}(X) > 0$, $\sigma_Y^2 = \mathrm{Var}(Y) > 0$ et la corrélation $\rho = \rho_{XY}$, $|\rho| < 1$. Notons $\mathbf{x} = \begin{pmatrix} X \\ Y \end{pmatrix}$, $\Sigma = V(\mathbf{x})$, alors

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

et $\mathrm{Det}(\Sigma) = \sigma_X^2\sigma_Y^2(1 - \rho^2) > 0$. Vu le Corollaire 3.2, la densité jointe de (X, Y) vaut

$$f_{X,Y}(x, y) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right] \right)}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}.$$

Si l'on pose $\xi = X$ et $\theta = Y$ dans le Théorème 3.4, alors

$$\Sigma_{\theta\xi} = \Sigma_{\xi\theta} = \rho\sigma_X\sigma_Y, \quad \Sigma_{\theta\xi}\Sigma_{\xi\xi}^{-1} = \rho\sigma_Y/\sigma_X.$$

Par conséquent, la fonction de régression et la variance conditionnelle sont données par

$$\begin{aligned} g(x) &= E(Y|X=x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \\ \gamma^2(x) &= V(Y|X=x) = \sigma_Y^2(1-\rho^2). \end{aligned}$$

La densité conditionnelle de Y sachant X est normale :

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_Y} \exp\left(-\frac{(y-g(x))^2}{2\gamma^2(x)}.\right)$$

C'est une densité de la loi $\mathcal{N}(g(x),\gamma^2(x))$. *La régression $g(x)$ est linéaire.*

Considérons le cas particulier où $\mu_X = \mu_Y = 0$ et $\sigma_X = \sigma_Y = 1$. Alors

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \Sigma^{-1} = (1-\rho^2)^{-1}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

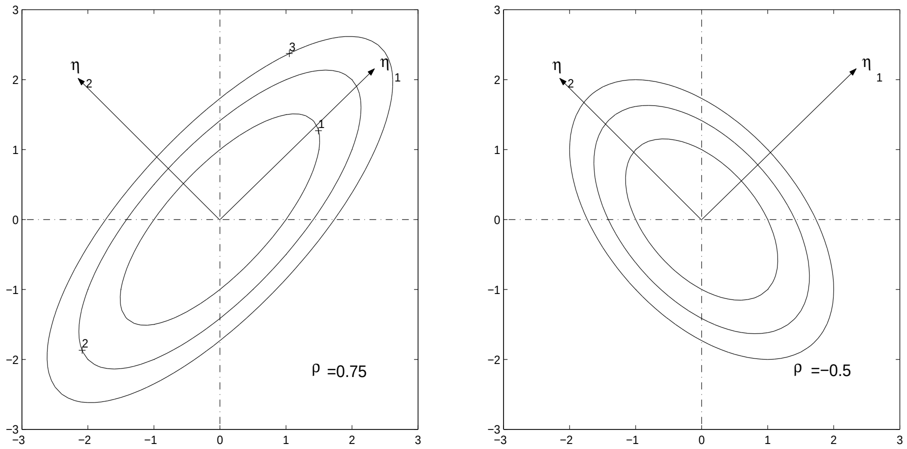

Figure 3.2. Ellipses de concentration : $\mathbf{x} = (\xi_1, \xi_2)$, $\mathbf{y} = (\eta_1, \eta_2)$, où $\mathbf{y} = \Sigma^{-1/2}\mathbf{x}$.

Les vecteurs propres de Σ sont $(1,1)^T$ et $(1,-1)^T$ correspondant aux valeurs propres, respectivement, $\lambda_1 = 1+\rho$ et $\lambda_2 = 1-\rho$. Les vecteurs propres orthonormés sont $\boldsymbol{\gamma}_{(1)} = 2^{-1/2}(1,1)^T$ et $\boldsymbol{\gamma}_{(2)} = 2^{-1/2}(1,-1)^T$. Si l'on note $\Gamma = (\boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(2)})$, la décomposition spectrale de Σ s'écrit sous la forme :

$$\Sigma = \Gamma\Lambda\Gamma^T = \Gamma\begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}\Gamma^T.$$

On peut considérer les ellipses de concentration de la densité jointe de (X,Y). Soit, pour $C > 0$,

$$E_C = \{x \in \mathbb{R}^2 : \ x^T\Sigma^{-1}x \le C\} = \{x \in \mathbb{R}^2 : \ |y|^2 \le C\},$$

où $y = \Sigma^{-1/2}x$. Si l'on note

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \ \ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

alors

$$y_1 = \frac{1}{\sqrt{2(1+\rho)}}\,(x_1 + x_2),$$
$$y_2 = \frac{1}{\sqrt{2(1-\rho)}}\,(x_1 - x_2),$$

et l'ellipse de concentration se présente sous la forme

$$E_C = \{x^T\Sigma^{-1}x \le C\} = \left\{\left(\frac{1}{\sqrt{2(1+\rho)}}\,(x_1+x_2)\right)^2 + \left(\frac{1}{\sqrt{2(1-\rho)}}\,(x_1-x_2)\right)^2 \le C\right\}.$$

3.5. Lois dérivées de la loi normale

3.5.1. Loi χ^2 de Pearson. C'est la loi de la somme

$$Y = \eta_1^2 + ... + \eta_p^2,$$

où $\eta_1, ..., \eta_p$ sont des variables aléatoires i.i.d. de loi $\mathcal{N}(0,1)$. On écrit alors $Y \sim \chi_p^2$ et on dit que Y suit **la loi chi-deux à** p **degrés de liberté**.

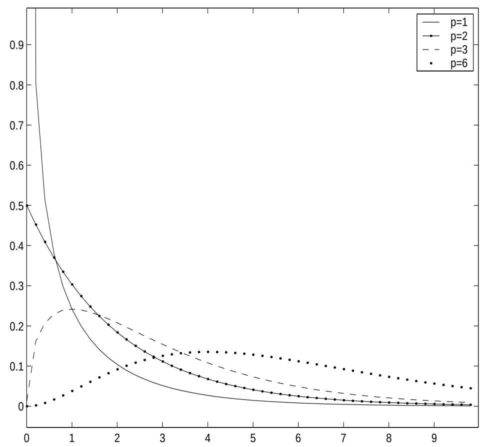

Figure 3.3. Densité de la loi de chi-deux pour différentes valeurs de p.

La densité de la loi χ_p^2 est

$$f_{\chi_p^2}(y) = C(p)y^{p/2-1}e^{-y/2}I\{y > 0\}, \tag{3.13}$$

où $C(p) = \left(2^{p/2}\Gamma(p/2)\right)^{-1}$ et $\Gamma(\cdot)$ est la fonction gamma $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u/2}du, \quad x > 0$. On a $E(Y) = p$, $\mathrm{Var}(Y) = 2p$ si $Y \sim \chi_p^2$.

EXERCICE 3.2. Montrer que la densité de la loi χ_p^2 est de la forme (3.13).

Proposition 3.6. *Soit* $\mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma)$, $\Sigma > 0$. *Alors la variable aléatoire*

$$\eta = \|\Sigma^{-1/2}(\mathbf{x} - \mu)\|^2 = (\mathbf{x} - \mu)^T\Sigma^{-1}(\mathbf{x} - \mu)$$

suit la loi χ_p^2.

Preuve. On utilise la propriété (N1) de la loi normale multivariée. ∎

3.5.2. Loi de Fisher-Snedecor. Soit $U \sim \chi^2_p$, $V \sim \chi^2_q$, deux v.a. indépendantes. La **loi de Fisher-Snedecor à degrés de liberté** p **et** q est la loi de la variable aléatoire

$$Y = \frac{U/p}{V/q}.$$

On écrit alors $Y \sim F_{p,q}$. La densité de $F_{p,q}$ est

$$f_{F_{p,q}}(y) = C(p,q) \frac{y^{p/2-1}}{(q+py)^{\frac{p+q}{2}}} I\{y > 0\}, \tag{3.14}$$

où

$$C(p,q) = \frac{p^{p/2} q^{q/2}}{B(p/2, q/2)} \quad \text{avec} \quad B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

On peut montrer que cette densité converge vers une densité de type $f_{\chi^2_p}$ quand $q \to \infty$.

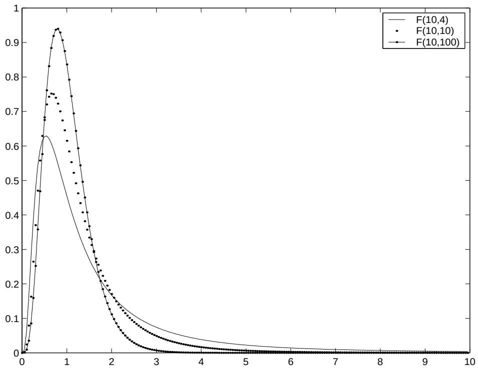

Figure 3.4. Densité de la loi de Fisher-Snedecor.

EXERCICE 3.3. Montrer que la densité de la loi de Fisher-Snedecor est de la forme (3.14).

3.5.3. Loi t de Student. Soit $\eta \sim \mathcal{N}(0,1)$, $\xi \sim \chi^2_q$ deux v.a. indépendantes. La **loi de Student à** q **degrés de liberté** est celle de la variable aléatoire

$$Y = \frac{\eta}{\sqrt{\xi/q}}.$$

On écrit alors $Y \sim t_q$. La densité de t_q est

$$f_{t_q}(y) = C(q)(1 + y^2/q)^{-(q+1)/2}, \quad y \in \mathbb{R}, \tag{3.15}$$

où

$$C(q) = \frac{1}{\sqrt{q} B(1/2, q/2)}.$$

Notons que
- la loi t_q est symétrique,
- t_1 est la loi de Cauchy,
- le carré de la variable t_q suit la loi $F_{1,q}$ ($t_q^2 = F_{1,q}$),
- la densité de t_q tend vers la densité de $\mathcal{N}(0,1)$ quand $q \to \infty$.

Les queues de t_q sont plus lourdes que celles de la loi normale standard.

Exercice 3.4. Montrer que la densité de la loi de Student est de la forme (3.15).

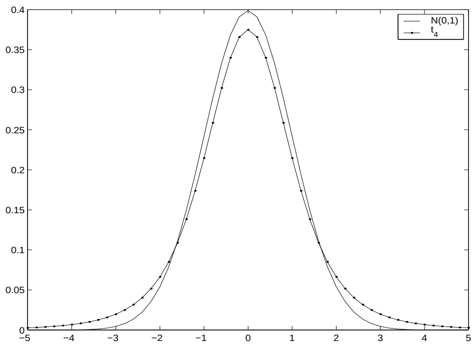

Figure 3.5. Densité de la loi de Student.

3.6. Théorème de Cochran

Théorème 3.5. *Soit* $\mathbf{x} \sim \mathcal{N}_p(0, I)$ *et soient* $A_1, ..., A_J$, $J \le p$, *des matrices* $p \times p$ *telles que*

(1) $A_j^2 = A_j$,

(2) A_j *est symétrique,* $\text{Rang}(A_j) = N_j$,

(3) $A_j A_k = \mathbf{0}$ *pour* $j \neq k$ *et* $\sum_{j=1}^J N_j \le p$.[3)]

Alors,

(i) les vecteurs aléatoires $A_j\mathbf{x}$, $j = 1, ..., J$, *sont mutuellement indépendants de lois* $\mathcal{N}_p(0, A_j)$, $j = 1, ..., J$, *respectivement;*

(ii) les variables aléatoires $\|A_j\mathbf{x}\|^2$, $j = 1, ..., J$, *sont mutuellement indépendantes de lois* $\chi^2_{N_j}$, $j = 1, ..., J$, *respectivement.*

Preuve. (i) Notons d'abord que $E(A_j\mathbf{x}) = 0$ et, d'après (C4),

$$V(A_j\mathbf{x}) = A_j V(\mathbf{x}) A_j^T = A_j A_j^T = A_j^2 = A_j.$$

Par ailleurs, la loi jointe de $A_1\mathbf{x}, \dots, A_J\mathbf{x}$ est normale (vérifiez ceci). De plus,

$$C(A_k\mathbf{x}, A_j\mathbf{x}) = E(A_k\mathbf{x}\mathbf{x}^T A_j^T) = A_k V(\mathbf{x}) A_j^T = A_k A_j^T = A_k A_j = \mathbf{0}$$

pour $j \neq k$. D'après la Proposition 3.5, on obtient alors que $A_1\mathbf{x}, \dots, A_J\mathbf{x}$ sont mutuellement indépendants.

(ii) Comme A_j est symétrique, il existe une matrice Γ orthogonale telle que $A_j = \Gamma\Lambda\Gamma^T$, où $\Lambda = \text{Diag}(\lambda_1, \dots, \lambda_p)$ est la matrice diagonale des valeurs propres de A_j. Alors,

$$\|A_j\mathbf{x}\|^2 = \mathbf{x}^T A_j^T A_j \mathbf{x} = \mathbf{x}^T A_j \mathbf{x} = (\mathbf{x}^T\Gamma)\Lambda(\Gamma^T\mathbf{x}) = \mathbf{y}^T\Lambda\mathbf{y} = \sum_{i=1}^p \lambda_i \eta_i^2,$$

où $\mathbf{y} = \Gamma^T\mathbf{x} = (\eta_1, ..., \eta_p)^T$ est un vecteur normal de loi $\mathcal{N}_p(0, I)$ (vu la propriété (N4)). Mais A_j est un projecteur, donc $\lambda_j \in \{0, 1\}$ et $\text{Card}(j : \lambda_j = 1) = \text{Rang}(A_j) = N_j$, d'où découle que $\|A_j\mathbf{x}\|^2 \sim \chi^2_{N_j}$. Finalement, la partie (i) du théorème et le fait que les transformations mesurables préservent l'indépendance impliquent que les variables aléatoires $\|A_1\mathbf{x}\|^2, \dots, \|A_J\mathbf{x}\|^2$ sont mutuellement indépendantes. ∎

[3)] Certaines versions de ce résultat supposent aussi que $A_1 + \cdots + A_J = I$.

3.7. Exercices

EXERCICE 3.5. Soit Q une matrice $q \times p$ (avec $q > p$) de rang p.
1^o. Montrer que la matrice $P = Q(Q^TQ)^{-1}Q^T$ est un projecteur.
2^o. Trouver le sous-espace $\mathcal{L}$ sur lequel projette P.

EXERCICE 3.6. Soit la matrice de covariance

$$\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Trouver $\Sigma^{1/2}$. Vérifier que $\Sigma = UU^T$ où $U \neq \Sigma^{1/2}$ est la matrice triangulaire donnée par

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{3} \end{pmatrix}.$$

Remarque. Ceci est un cas particulier de la *décomposition de Holesky* : pour toute matrice $p \times p$ symétrique positive Σ il existe une matrice $p \times p$ triangulaire U telle que $\Sigma = UU^T$.

EXERCICE 3.7. Soit (X, Y) un vecteur aléatoire de densité

$$f(x, y) = C \exp(-x^2 + xy - y^2/2).$$

1^o. Montrer que (X, Y) est un vecteur aléatoire normal. Calculer l'espérance, la matrice de covariance et la fonction caractéristique de (X, Y). Déterminer le coefficient de corrélation ρ_{XY} entre X et Y.
2^o. Déterminer la loi de X, de Y, de $2X - Y$.
3^o. Monter que X et $Y - X$ sont des variables aléatoires indépendantes et de même loi.

EXERCICE 3.8. Soit X une v.a. de loi $\mathcal{N}(0, 1)$ et Z une v.a. prenant les valeurs -1 ou 1 avec la probabilité $\frac{1}{2}$. On suppose X et Z indépendantes. On pose $Y = ZX$.
1^o. Montrer que Y suit la loi $\mathcal{N}(0, 1)$.
2^o. Calculer la covariance et la corrélation entre X et Y.
3^o. Calculer $P(X + Y = 0)$.
4^o. Le vecteur (X, Y) est-il un vecteur aléatoire normal ?

EXERCICE 3.9. Soient ξ et η deux v.a. indépendantes de loi $U[0, 1]$. Prouver que les v.a.

$$X = \sqrt{-2 \ln \xi}\, \cos(2\pi\eta), \quad Y = \sqrt{-2 \ln \xi}\, \sin(2\pi\eta)$$

sont telles que $Z = (X, Y)^T \sim \mathcal{N}_2(0, I)$. *Indication* : soit $(X, Y)^T \sim \mathcal{N}_2(0, I)$. Passer en coordonnées polaires.

EXERCICE 3.10. Soit $Z = (Z_1, Z_2, Z_3)^T$ un vecteur aléatoire normal, admettant une densité

$$f(z_1, z_2, z_3) = \frac{1}{4(2\pi)^{3/2}} \exp\left(-\frac{6z_1^2 + 6z_2^2 + 8z_3^2 + 4z_1z_2}{32} \right).$$

1^o. Déterminer la loi de (Z_2, Z_3) sachant que $Z_1 = z_1$.

Soient X et Y deux vecteurs aléatoires définis par :

$$X = \begin{pmatrix} 2 & 2 & 2 \\ 0 & 2 & 5 \\ 0 & 4 & 10 \\ 1 & 2 & 4 \end{pmatrix} Z \quad \text{et} \quad Y = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} Z.$$

2^o. Le vecteur (X, Y) de dimension 6 est-il normal ? Le vecteur X a-t-il une densité ? Le vecteur Y a-t-il une densité ?
3^o. Les vecteurs X et Y sont-ils indépendants ?
4^o. Déterminer les lois des coordonnées de Z.

Exercice 3.11. Soit $(X, Y, Z)^T$ un vecteur aléatoire normal de moyenne nulle et dont la matrice de covariance est

$$\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

1^o. On pose $U = -X + Y + Z$, $V = X - Y + Z$, $W = X + Y - Z$. Déterminer la loi du vecteur aléatoire $(U, V, W)^T$.
2^o. Déterminer la densité de la variable $T = U^2 + V^2 + W^2$.

Exercice 3.12. Parmi les matrices suivantes, lesquelles peuvent être la matrice de covariance d'un vecteur aléatoire $\mathbf{x} \in I\!R^2$:

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \begin{pmatrix} -1 & -1/2 \\ -1/2 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1/2 \\ 1/3 & 1 \end{pmatrix} ?$$

Dans la suite, on notera Σ les matrices répondant à la question et on supposera que $\mathbf{x}$ est de loi $\mathcal{N}_2(0, \Sigma)$.
1^o. Calculer, pour chaque matrice Σ, les valeurs propres (λ_1, λ_2) et les vecteurs propres associés (v_1, v_2).
2^o. Donner la loi jointe de $v_1^T \mathbf{x}$ et $v_2^T \mathbf{x}$.

Exercice 3.13. Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de loi $\mathcal{N}(0, 1)$ et $a_1, \ldots, a_n, b_1, \ldots, b_n$ des réels. Montrer que les v.a. $Y = \sum_{i=1}^n a_i X_i$ et $Z = \sum_{i=1}^n b_i X_i$ sont indépendantes si et seulement si $\sum_{i=1}^n a_i b_i = 0$.

Exercice 3.14. Soit X une variable aléatoire normale standard. Pour tout $c > 0$, on pose

$$X_c = X \left(I\{|X| < c\} - I\{|X| \geq c\} \right).$$

1^o. Déterminer la loi de X_c.
2^o. Calculer $\mathrm{Cov}(X, X_c)$ et montrer qu'il existe c_0 tel que $\mathrm{Cov}(X, X_{c_0}) = 0$.
3^o. Montrer que X et X_{c_0} ne sont pas indépendantes. Le vecteur (X, X_{c_0}) est-il normal ?

Exercice 3.15. Soit $(\varepsilon_Y, \varepsilon_Z, X)$ un vecteur aléatoire normal tel que $\varepsilon_Y, \varepsilon_Z, X$ sont indépendantes de lois $\mathcal{N}(0, 1)$, $\mathcal{N}(0, 1)$ et $\mathcal{N}(0, 2)$. On pose :

$$Z = 2Y - 3X + \varepsilon_Z,$$
$$Y = X + \varepsilon_Y.$$

Déterminer la loi du triplet (X, Y, Z). On notera Σ la matrice de covariance de ce vecteur. Calculer $E(Z|Y, X)$.

Exercice 3.16. Soit (X, Y, Z) un vecteur aléatoire normal tel que :
(i) la loi conditionnelle de (X, Y) sachant que $Z = z$ est

$$\mathcal{N}_2\left(\begin{pmatrix} -z \\ z-1 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}\right),$$

pour tout $z \in I\!\!R$,
(ii) la loi de Z sachant que $Y = y$ est $\mathcal{N}(y/4 + 1, 3/4)$ pour tout $y \in I\!\!R$,
(iii) $\mathrm{Var}(Z) = 1$.
Trouver la loi de (X, Y, Z) et celle de Z sachant (X, Y).

Exercice 3.17. *Matrice pseudo-inverse.* Soit A une matrice $p \times p$ symétrique de rang $k < p$ et soit $A = \Gamma\Lambda\Gamma^T$ sa représentation spectrale, où $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_k, 0, \ldots 0)$.

1^o. Vérifier que si $\Lambda_k = \mathrm{Diag}(\lambda_1, \ldots, \lambda_k)$ et $\Gamma_k = (\gamma_{(1)}, ..., \gamma_{(k)})$ est la matrice $p \times k$ de k premiers vecteurs propres orthonormés de A (qui correspondent aux valeurs propres non-nulles), alors

$$A = \Gamma_k \Lambda_k \Gamma_k^T.$$

2^o. Définissons la matrice

$$A^+ = \Gamma_k \Lambda_k^{-1} \Gamma_k^T$$

appelée *matrice pseudo-inverse de* A. Montrer que $AA^+A = A$. Vérifier que A^+A est le projecteur sur le sous-espace $\mathrm{Im}(A)$ et $I - A^+A$ est le projecteur sur $\mathrm{Ker}(A)$.

3^o. Montrer que les formules (3.10) du Théorème de corrélation normale restent valides si la matrice $\Sigma_{\xi\xi}$ est dégénérée et si au lieu de $\Sigma_{\xi\xi}^{-1}$ on considère $\Sigma_{\xi\xi}^+$.

Partie 2

Notions fondamentales de la Statistique

4

Échantillonnage et méthodes empiriques

4.1. Échantillon

Le matériel de départ de la démarche statistique sont *les données.*

Du point de vue d'applications, les données représentent une suite finie de nombres observés au cours d'une expérience, d'un essai. On désigne ces nombres par $X_1, ..., X_n$. Plus généralement, les X_i peuvent être des vecteurs, dans ce cas on parle de données multidimensionnelles ou multivariées.

Du point de vue mathématique, les données $X_1, ..., X_n$ sont considérées comme des variables aléatoires. C'est une hypothèse fondamentale de la Statistique. Théoriquement, on suppose qu'il existe une loi de probabilité inconnue (loi jointe de $X_1, ..., X_n$) qui "explique" le comportement des données. Dans le modèle le plus simple, les variables $X_1, ..., X_n$ sont i.i.d., de même loi F inconnue. Il est donc désirable de reconstituer F pour "expliquer" les données.

Dans cette optique, on peut voir la Statistique comme une matière dont l'objectif est d'estimer une loi inconnue (ou d'inférer au sujet d'une loi inconnue) à partir de variables aléatoires $X_1, ..., X_n$ qui suivent cette loi.

La suite de données $\mathcal{X}_n = (X_1, ..., X_n)$ s'appelle **l'échantillon**. Le nombre n est appelé **taille d'échantillon**. Au lieu du mot *données* on dit parfois *observations* ou *points d'échantillon.*

Dans ce chapitre, on suppose que l'échantillon vérifie l'hypothèse suivante.

Hypothèse (E0). *Soit X une variable aléatoire réelle, définie sur l'espace de probabilité $(\Omega, \mathcal{A}, P)$, de fonction de répartition F (on écrit $X \sim F$). L'échantillon $\mathcal{X}_n = (X_1, \ldots, X_n)$ est une* **réalisation** *de la variable X, c'est-à-dire les observations $X_1, ..., X_n$ sont des variables aléatoires i.i.d. de la même loi F que X ($X_i \sim F$).*

Notons qu'en général un échantillon peut contenir des données X_i dépendantes et/ou non-identiquement distribuées. Le fait que les X_i soient i.i.d. est formulé comme l'hypothèse supplémentaire (l'Hypothèse (E0)). Elle sera généralement imposée par la suite. Il est utile de noter que souvent dans la littérature statistique l'hypothèse de la structure i.i.d. des données est présupposée, de sorte que "l'échantillon" signifie "l'échantillon i.i.d.", sans explication particulière.

REMARQUES.

(1) Il est important de noter que d'habitude l'inférence statistique est de nature **asymptotique** : les conclusions sont valables si la taille n de l'échantillon est assez grande. Ceci est une conséquence du fait qu'elles sont, généralement, basées sur les résultats asymptotiques de la Théorie des probabilités, tels que la loi des grands nombres et le théorème central limite. La notion de "n assez grand" varie d'un exemple à l'autre et ne peut pas être précisée une fois pour toutes. Néanmoins, n de l'ordre de quelques centaines est souvent considérée comme une taille d'échantillon "confortable". Pour un n "petit" (par exemple, $n < 20$) l'approximation limite est typiquement en défaut, et on utilise, si possible, des méthodes non-asymptotiques dont l'arsenal est assez restreint.

(2) L'objectif de la Statistique est inverse de celui de la Théorie des probabilités. La Théorie des probabilités a pour but d'étudier, étant donnée une loi de probabilité, le comportement de ses réalisations aléatoires. La Statistique va dans le sens contraire : étant données des réalisations de la variable aléatoire, elle essaye de se renseigner sur sa loi de probabilité.

EXEMPLE 4.1. *Données de survie.* Supposons qu'on a mesuré les durées de vie (en mois depuis le début d'utilisation) de 10 ampoules électriques :

$$X_1 = 4.4,\ X_2 = 2.6,\ X_3 = 5.4,\ X_4 = 7.8,\ X_5 = 0.9,$$
$$X_6 = 0.5,\ X_7 = 2.7,\ X_8 = 9.1,\ X_9 = 2.9,\ X_{10} = 1.2.$$

Adoptons l'hypothèse suivante souvent utilisée pour les données de survie, à savoir que la loi de probabilité des X_i appartient à la famille des lois exponentielles $\mathcal{E}(\theta)$ de densité

$$f(x, \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} I\{x \geq 0\}, \tag{4.1}$$

où $\theta > 0$ est un paramètre inconnu. La f.d.r. F de X_i appartient donc à la famille $\mathcal{F} = \{F_\theta : \theta > 0\}$, où F_θ est la f.d.r. de la loi exponentielle de densité (4.1). Autrement dit, la fonction de répartition F est donnée par $F = F_{\theta^*}$ où $\theta^* > 0$ est la *vraie valeur du paramètre* θ inconnue. Pour reconstituer F il suffit d'estimer le paramètre θ^*.

L'échantillon $(X_1, ..., X_{10})$ peut être considéré comme une **réalisation** de la variable aléatoire X de densité $f(\cdot, \theta^*)$. La variable X dans cet exemple est continue (la loi de X admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$).

EXEMPLE 4.2. *Notes d'examen.* C'est un exemple de données discrètes. Trente étudiants ont reçu les notes suivantes à l'examen de statistique :

Note (j)	3	5	8	9	10	11	12	14	15	16
Nombre d'étudiants (n_j)	2	1	1	5	4	8	2	4	2	1

Notons d'abord que cette table présente les données "réduites". Les données de départ ne sont pas les n_i, mais les notes de $n = 30$ étudiants, de sorte que $X_i \in \{1, ..., 20\}$ est la note d'étudiant numéro i. Les X_i sont des variables aléatoires discrètes. Il est naturel d'attribuer aux vingt notes les probabilités p_j ($j = 1, ..., 20$), telles que $p_1 + ... + p_{20} = 1$. Les variables aléatoires n_j sont alors

$$n_j = \sum_{i=1}^{30} I\{X_i = j\}.$$

Les valeurs j non-présentes dans la table correspondent à $n_j = 0$.

On voit donc que dans cet exemple l'échantillon $X_1, ..., X_{30}$ peut être considéré comme une **réalisation** de la v.a. discrète X dont la loi inconnue est définie par $P(X = j) = p_j$, $j = 1, ..., 20$. Pour reconstituer cette loi, il suffit d'estimer $N = 20$ paramètres $p_1, ..., p_{20}$. (Comme $p_1 + ... + p_{20} = 1$, en effet seuls $N - 1$ paramètres $p_1, ..., p_{19}$ sont à estimer.) Notons que

$$\begin{aligned} P(X = x) &= \prod_{j=1}^{N} p_j^{I\{x=j\}} = p_N \prod_{j=1}^{N-1} (p_j/p_N)^{I\{x=j\}} \\ &= \exp\left(\sum_{j=1}^{N-1} I\{x = j\} \ln \frac{p_j}{p_N} + \ln p_N \right), \quad x = 1, \ldots, N. \end{aligned} \tag{4.2}$$

Ceci définit une loi discrète que l'on notera $\mathcal{D}(\{1, \ldots, N\}, (p_1, \ldots, p_N))$. Comme dans l'Exemple 4.1, on peut donc définir une famille $\mathcal{F}$ à laquelle appartient F :

$$\mathcal{F} = \{\text{toutes les lois } \mathcal{D}(\{1, \ldots, 20\}, (p_1, \ldots, p_{20}))\}.$$

Le paramètre inconnu θ^* ici est vectoriel : $\theta^* = (p_1, \ldots, p_{20})$.

Si $X_1, \ldots, X_n$ est un échantillon i.i.d. de loi $\mathcal{D}(\{1, \ldots, N\}, (p_1, \ldots, p_N))$, alors le vecteur aléatoire $\nu = (n_1, \ldots, n_N)$, où $n_j = \sum_{i=1}^{n} I\{X_i = j\}$, suit la loi

$$P(\nu = (k_1, \ldots, k_N)) = \frac{n!}{k_1! \cdots k_N!} p_1^{k_1} \cdots p_N^{k_N},$$

dite *loi multinomiale de degré* N.

4.2. Représentation graphique de l'échantillon

4.2.1. Fonction de répartition empirique. La **fonction de répartition empirique** $\widehat{F}_n$ associée à l'échantillon $X_1, ..., X_n$ est définie par

$$\boxed{\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \le x\}, \quad x \in \mathbb{R}.}$$

Pour tout x fixé, $\widehat{F}_n(x)$ est une variable aléatoire. Pour tout échantillon $X_1, ..., X_n$ fixé, $\widehat{F}_n$ est une fonction de x en escaliers, continue à droite, de sauts égaux à $1/n$ (si tous les X_i sont différents, comme dans le cas où X est une variable continue). De plus, $\lim_{x\to-\infty} \widehat{F}_n(x) = 0$, $\lim_{x\to+\infty} \widehat{F}_n(x) = 1$. Donc, pour tout échantillon $X_1, ..., X_n$ fixé, $\widehat{F}_n$ est une fonction de répartition de la *loi discrète uniforme sur* $\{X_1, ..., X_n\}$, i.e. de la loi qui attribue la masse $1/n$ à chaque X_i.

La f.d.r. empirique $\widehat{F}_n$ joue un rôle fondamental dans la Statistique, car elle fournit une bonne approximation de la vraie fonction de répartition F qui est inconnue. Un résultat important est la convergence de $\widehat{F}_n$ vers F.

Soit x fixé. Alors $\widehat{F}_n(x)$ est la moyenne arithmétique de n variables indépendantes de loi de Bernoulli de paramètre $F(x)$ et $E(\widehat{F}_n(x)) = F(x)$. D'après la loi forte des grands nombres,

$$\widehat{F}_n(x) \to F(x) \quad (p.s.), \quad \forall x \in \mathbb{R}, \tag{4.3}$$

quand $n \to \infty$. De plus, la convergence est uniforme :

Théorème 4.1. (Glivenko – Cantelli) *Si $X_1, ..., X_n$ sont i.i.d., $X_i \sim F$, alors*

$$\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \to 0 \quad (p.s.) \quad quand \quad n \to \infty.$$

Preuve. On va démontrer ce résultat seulement dans le cas où F est continue. Par continuité, il existe des points $x_1 < \cdots < x_{k-1}$ tels que $F(x_i) = i/k$. On pose $x_0 = -\infty$, $x_k = +\infty$. Grâce à la monotonie de F et de $\widehat{F}_n$ on obtient, pour tout $x \in [x_{i-1}, x_i]$,

$$\widehat{F}_n(x) - F(x) \leq \widehat{F}_n(x_i) - F(x_{i-1}) = \widehat{F}_n(x_i) - F(x_i) + 1/k,$$

et

$$\widehat{F}_n(x) - F(x) \geq \widehat{F}_n(x_{i-1}) - F(x_i) = \widehat{F}_n(x_{i-1}) - F(x_{i-1}) - 1/k.$$

Donc

$$|\widehat{F}_n(x) - F(x)| \leq \max_{i=1,...,k-1} |\widehat{F}_n(x_i) - F(x_i)| + 1/k, \quad \forall x \in \mathbb{R}.$$

Vu (4.3) ceci implique

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \leq 1/k \quad (p.s.).$$

On conclut en faisant k tendre vers l'infini. ∎

Notons que la f.d.r. empirique $\widehat{F}_n$ ne convient pas pour analyser visuellement le comportement d'une loi de probabilité. Par exemple, il n'est pas facile de comparer, en regardant le graphique de $\widehat{F}_n$, les zones de plus forte ou de moins forte concentration des points de l'échantillon. Il est plus pratique d'utiliser des analogues empiriques de la densité de probabilité que nous allons décrire maintenant.

4.2.2. Densités empiriques*. Soit X une variable continue, c'est-à-dire que F, la f.d.r. de X, admet une densité de probabilité f par rapport à la mesure de Lebesgue. A partir d'un échantillon $X_1, ..., X_n$, on cherche à construire une courbe $\widehat{f}_n(x)$ qui donnerait une bonne approximation de $f(x)$. Une telle courbe est appelée *densité empirique* ou *estimateur de densité*. Il existe plusieurs méthodes de construction de densités empiriques dont nous allons décrire ici quelques unes de plus élémentaires.

Histogramme et polygone des fréquences. Soit A un intervalle qui contient toutes les données $X_1, ..., X_n$ et soit $A_1, \ldots, A_m$ une partition de A en m sous-intervalles de longueur h

chacun. Soit $N_j = \sum_{i=1}^n I(X_i \in A_j)$ le nombre des points X_i dans l'intervalle A_j. **L'histogramme** est une fonction constante par morceaux définie par

$$\widehat{f}_n^H(x) = \frac{N_j}{nh}, \quad \text{si } x \in A_j,\ j = 1, \ldots, m.$$

Pour tout échantillon $X_1, ..., X_n$ fixé, $\widehat{f}_n^H$ est une densité de probabilité, car

$$\widehat{f}_n^H \geq 0, \qquad \int \widehat{f}_n^H = h \sum_j \frac{N_j}{nh} = 1.$$

L'histogramme est une fonction discontinue, non-régulière. Pour obtenir un estimateur plus lisse de la densité f on utilise une approximation linéaire : on construit un graphique linéaire par morceaux qui passe par les centres des “plateaux” de l'histogramme. Ce graphique porte le nom de **polygone des fréquences**.

Estimateurs à fenêtre mobile et à noyau. La densité f étant la dérivée de la fonction de répartition F, on peut écrire l'approximation

$$f(x) = F'(x) \approx \frac{F(x + h/2) - F(x - h/2)}{h} \ ,$$

si h est assez petit. Puisque la f.d.r. F est inconnue, remplaçons-la dans cette formule par la fonction de répartition empirique $\widehat{F}_n$ qui est réalisable à partir de l'échantillon et proche de F pour n assez grand (vu le Théorème de Glivenko – Cantelli). Ceci fournit l'approximation de $f(x)$ de la forme :

$$\widehat{f}_n(x) = \frac{\widehat{F}_n(x + h/2) - \widehat{F}_n(x - h/2)}{h} \tag{4.4}$$

que l'on appelle **estimateur à fenêtre mobile**. Ce nom est motivé par le fait que $\widehat{f}_n$ fait le comptage du nombre des points de l'échantillon X_i qui tombent dans la fenêtre $U_x = [x - h/2, x + h/2[$ autour du point x :

$$\frac{\widehat{F}_n(x + h/2) - \widehat{F}_n(x - h/2)}{h} = \frac{1}{nh} \sum_{i=1}^n I(X_i \in U_x) = \frac{1}{nh} \sum_{i=1}^n K_0\left(\frac{x - X_i}{h}\right) \tag{4.5}$$

où $K_0(u) = I(-1/2 < u \leq 1/2)$. Comme l'histogramme, l'estimateur à fenêtre mobile est une densité de probabilité pour $X_1, ..., X_n$ fixés. Notons aussi que $x \mapsto \widehat{f}_n(x)$ est une fonction constante par morceaux (pourquoi ?).

Une version plus régulière de l'estimateur à fenêtre mobile est l'estimateur à noyau. Il est obtenu quand on prend dans (4.5) au lieu de la fonction K_0 indicatrice une fonction K assez régulière que l'on appelle **noyau**. La définition de **l'estimateur à noyau** est donnée par

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

où K est une densité de probabilité symétrique sur $\mathbb{R}$. On utilise souvent le noyau gaussien $K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$. L'estimateur à noyau $\widehat{f}_n^N(x)$ est donc la moyenne arithmétique de n “fonctions-cloches”

$$\frac{1}{h} K\left(\frac{\cdot - X_i}{h}\right).$$

Chaque "cloche" est une densité de probabilité centrée en X_i et d'échelle h. Pour $X_1, ..., X_n$ fixés, la fonction $x \mapsto \widehat{f}_n^N(x)$ est une densité de probabilité, car

$$\widehat{f}_n^N \geq 0, \quad \int \widehat{f}_n^N = \int K = 1.$$

4.3. Caractéristiques de l'échantillon. Méthode de substitution

Dans les Exemples 4.1, 4.2, l'estimation de la loi de probabilité inconnue se réduit à l'estimation des paramètres θ^* (Exemple 4.1) ou $p_1, ..., p_{20}$ (Exemple 4.2). Comment les estimer ? Nous disposons seulement d'un échantillon, et la seule liberté que nous pouvons nous permettre pour estimer ces paramètres est de composer des fonctions appropriées des observations $X_1, ..., X_n$. Nous arrivons donc à la notion fondamentale suivante.

Définition 4.1. *On appelle* **statistique** *toute fonction borélienne des observations* $S = S(X_1, ..., X_n)$ *à valeurs dans un espace* $\mathbb{R}^l$.

Une statistique S est donc une variable aléatoire ou un vecteur aléatoire qui ne dépend que de l'échantillon.

Une statistique est aussi appelée **estimateur** si elle est utilisée pour estimer des paramètres (ou d'autres caractéristiques) d'une loi de probabilité.

La Définition 4.1 est très générale : par exemple, l'échantillon $(X_1, \ldots, X_n)$ est une statistique, la fonction $S(X_1, ..., X_n) \equiv 0$ l'est aussi, mais ces deux statistiques sont sans intérêt, car elles ne nous approchent pas de la connaissance de caractéristiques de la loi F sous-jacente.

Comment trouver des statistiques qui donnent une approximation convenable des paramètres d'une loi de probabilité ? On peut considérer la démarche suivante. Souvent les paramètres θ^* d'une loi F inconnue peuvent être présentés comme **fonctionnelles** de cette loi :

$$\theta^* = T(F). \tag{4.6}$$

En particulier, dans l'Exemple 4.1, où l'on suppose que la loi F est exponentielle de densité $f(x) = (\theta^*)^{-1} e^{-x/\theta^*} I\{x \geq 0\}$, il est facile de voir que

$$\theta^* = \int_0^\infty x f(x) dx = \int_0^\infty x dF(x).$$

Donc, dans ce cas particulier, (4.6) est vérifié avec la fonctionnelle

$$T(F) = \int_0^\infty x dF(x). \tag{4.7}$$

Puisque la f.d.r. F peut être approchée par la fonction de répartition empirique $\widehat{F}_n$, on peut prendre comme estimateur de $T(F)$ la statistique

$$S(X_1, \ldots, X_n) = T(\widehat{F}_n).$$

Dans notre exemple, la fonctionnelle $T(\cdot)$ est définie par (4.7), donc

$$T(\widehat{F}_n) = \int_0^\infty x d\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

(En effet, si l'on fixe $X_1, ..., X_n$, la f.d.r. empirique $\widehat{F}_n$ est une fonction de répartition d'une v.a. discrète qui prend les valeurs $X_i,\ i = 1, \dots, n$, avec les probabilités $1/n$.) L'estimateur ainsi obtenu est donc la moyenne arithmétique des X_i.

L'idée de construction de l'estimateur dans cet exemple peut être appelée méthode de substitution. On substitue $\widehat{F}_n$ à F. Plus généralement, on peut l'exprimer comme suit :

Méthode de substitution. *Soit $T(F)$ une fonctionnelle de fonction de répartition F inconnue. On prend comme estimateur de $T(F)$ la statistique $T(\widehat{F}_n)$ (la même fonctionnelle de la fonction de répartition empirique $\widehat{F}_n$).*

Sous des hypothèses assez générales,

$$T(\widehat{F}_n) \to T(F) \ \ \text{(p.s.)} \ \ \text{quand} \ \ n \to \infty, \tag{4.8}$$

ce qui justifie l'application de la méthode de substitution. Dans la suite, nous allons montrer (4.8) pour quelques exemples.

Pour éviter toute confusion, la vraie fonction de répartition F sera appelée fonction de répartition **théorique** et ses caractéristiques (fonctionnelles) seront appelées **caractéristiques théoriques**. Les fonctionnelles respectives de $\widehat{F}_n$ seront appelées **caractéristiques empiriques**.

Considérons quelques exemples de statistiques obtenues par la méthode de substitution.

4.3.1. Statistiques $\bar{X}$ et s^2. La **moyenne empirique** est la statistique

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Comme on l'a déjà vu, c'est un estimateur par la méthode de substitution de la fonctionnelle

$$T(F) = E(X) = \int_{-\infty}^{\infty} x dF(x),$$

i.e. de la moyenne théorique.

La **variance empirique** s^2 est définie par

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

Evidemment, s^2 est la variance de la f.d.r. empirique $\widehat{F}_n$:

$$s^2 = \int \left(x - \int x d\widehat{F}_n(x) \right)^2 d\widehat{F}_n(x) = \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x) \right)^2 = T(\widehat{F}_n)$$

où la fonctionnelle T est définie par

$$T(F) = \int x^2 dF(x) - \left(\int x dF(x) \right)^2 = \mathrm{Var}(X).$$

La caractéristique théorique correspondante à s^2 est la variance théorique $\sigma^2 = \mathrm{Var}(X)$. On appelle s, la racine carrée positive de la variance, **écart-type empirique**.

4.3.2. Estimateurs basés sur les statistiques d'ordre. Rangeons les observations $X_1, ..., X_n$ par ordre croissant :

$$X_{(1)} \leq ... \leq X_{(j)} \leq ... \leq X_{(n)},$$

La variable aléatoire $X_{(j)}$ (le j-ème plus petit élément de l'échantillon) s'appelle **la j-ème statistique d'ordre**. Le vecteur aléatoire $(X_{(1)}, \ldots, X_{(n)})$ s'appelle **la statistique d'ordre** associée à l'échantillon $X_1, ..., X_n$.

Le quantile q_p d'ordre $p \in]0,1[$ de la loi F est la fonctionnelle suivante (cf. Chapitre 1) :

$$q_p = T(F) = \frac{1}{2}\left(\inf\{q : F(q) > p\} + \sup\{q : F(q) < p\}\right).$$

D'après la méthode de substitution, la caractéristique empirique respective est donnée par

$$Q_{n,p} = T(\widehat{F}_n) = \frac{1}{2}\Big(\inf\{q : \widehat{F}_n(q) > p\} + \sup\{q : \widehat{F}_n(q) < p\}\Big).$$

On appelle $Q_{n,p}$ **quantile empirique d'ordre** p.

Notons que la fonction $\widehat{F}_n$ représente un cas difficile pour la définition des quantiles : son graphique est composé de sauts et de plateaux, donc la solution q de l'équation

$$\widehat{F}_n(q) = p$$

n'est pas unique ou n'existe pas. Par contre, si $\widehat{F}_n$ est considérée comme une multi-application, les quantiles empiriques vérifient

$$\widehat{F}_n(Q_{n,p}) = p.$$

Il est possible d'expliciter $Q_{n,p}$ à partir des statistiques d'ordre :

$$Q_{n,p} = \begin{cases} X_{(k)} & \text{si } p \in](k-1)/n, k/n[, \\ (X_{(k)} + X_{(k+1)})/2 & \text{si } p = k/n,\ k = 1, \ldots, n. \end{cases} \tag{4.9}$$

Exercice 4.1. Démontrer (4.9).

La **médiane empirique** (ou médiane de l'échantillon) notée M_n est définie comme le quantile empirique d'ordre 1/2. En utilisant (4.9) on obtient alors :

$$M_n = \begin{cases} X_{(\frac{n+1}{2})} & \text{pour } n \text{ impair}, \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{pour } n \text{ pair}. \end{cases}$$

Autrement dit, la médiane est une solution de l'équation

$$\widehat{F}_n(M_n) = \frac{1}{2}, \tag{4.10}$$

où $\widehat{F}_n$ est considérée comme une multi-application. Si la solution de (4.10) est unique, elle est prise pour médiane. Dans le cas contraire, s'il y a un intervalle de solutions, la médiane est définie comme le centre de l'intervalle. La caractéristique théorique correspondante est la médiane de la loi F. On définit la fonctionnelle $M = T(F)$ comme solution de

$$F(M) = \frac{1}{2}$$

si une telle M existe. Alors $M_n = T(\widehat{F}_n)$ pour ce choix de T.

Remarque. Si la loi F est symétrique et $E(|X|) < \infty$, la médiane théorique est égale à la moyenne théorique (Exercice 1.1). Mais cela n'implique pas l'égalité de la médiane et de la moyenne empiriques.

Intervalle interquartile empirique. C'est une mesure de dispersion des données basée sur les statistiques d'ordre et définie par

$$\mathcal{I}_n = Q_{n,3/4} - Q_{n,1/4}$$

où $Q_{n,1/4}$ et $Q_{n,3/4}$ sont les quartiles empiriques. Par exemple, pour la taille d'échantillon $n = 5$, $\mathcal{I}_n = X_{(4)} - X_{(2)}$. La caractéristique théorique correspondante est l'intervalle interquartile $\mathcal{I} = q_{3/4} - q_{1/4}$.

Remarque. Les statistiques $\bar{X}$ et M_n sont des caractéristiques de la tendance centrale, elles définissent une valeur autour de laquelle se groupent les observations. Par contre, l'écart-type s et l'intervalle interquartile $\mathcal{I}_n$ sont des caractéristiques empiriques de la dispersion des données.

Souvent on utilise le résumé graphique d'un échantillon basé sur les statistiques d'ordre et appelé **boxplot**. Il permet de repérer le centre des données (représenté par la médiane M_n), la dispersion (intervalle interquartile $\mathcal{I}_n$), la symétrie ou dissymétrie de la loi des données (localisation de la médiane par rapport aux quartiles), la présence des observations aberrantes.

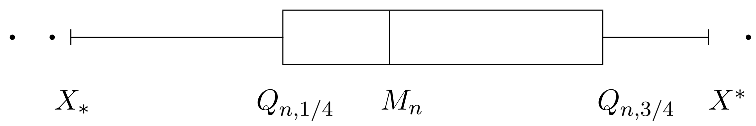

Figure 4.1. Le boxplot.

Les paramètres définissant le boxplot sont les statistiques $M_n, Q_{n,1/4}, Q_{n,3/4}$ et

$$X_* = \min\{X_i : |X_i - Q_{n,1/4}| \le \frac{3}{2}\mathcal{I}_n\}, \quad X^* = \max\{X_i : |X_i - Q_{n,3/4}| \le \frac{3}{2}\mathcal{I}_n\}.$$

Les observations aberrantes $X_i < X_*$ et $X_i > X^*$ sont représentées par les points isolés aux extrémités du graphique.

4.4. Statistiques exhaustives*

Le notion d'exhaustivité est introduite pour caractériser les statistiques

$$S = S(X_1, \dots, X_n)$$

qui résument *toute* l'information sur F contenue dans l'échantillon $X_1, \dots, X_n$. Il est clair qu'une statistique (la moins économique) qui contient toute cette information est l'échantillon $(X_1, \dots, X_n)$. Pourtant, peut-on trouver une statistique S beaucoup plus simple (par exemple, comme $\bar{X}$, s^2, M_n ou d'autres définies ci-dessus), telle qu'il suffise de connaître uniquement S, et qu'on puisse oublier l'échantillon initial sans aucun regret ? Généralement, la réponse à cette question est négative, mais il y a des cas remarquables où une telle statistique S existe. Tout dépend des hypothèses sur la f.d.r. F des X_i. On peut structurer ces hypothèses sous la forme : $F \in \mathcal{F}$ où $\mathcal{F}$ est une famille connue de fonctions de répartition (comme les familles $\mathcal{F}$ dans les Exemples 4.1 et 4.2).

Définition 4.2. *Une statistique* $S(X_1, \ldots, X_n)$ *est dite* **exhaustive pour la famille** $\mathcal{F}$ *si la loi conditionnelle de* $(X_1, \ldots, X_n)$ *sachant que* $S = s$ *ne dépend pas de* F *quand* $F \in \mathcal{F}$.

Interprétation : la Définition 4.2 dit que si l'on fixe la valeur de la statistique exhaustive S, on ne peut extraire aucune information supplémentaire sur F de l'échantillon $(X_1, \ldots, X_n)$. Autrement dit, toute l'information sur F est contenue dans S.

Notons quelques conséquences de la Définition 4.2 :

(1) Le concept d'exhaustivité dépend de la famille $\mathcal{F}$. Si S est exhaustive pour $\mathcal{F}$, alors S est exhaustive pour toute sous-famille $\mathcal{F}' \subset \mathcal{F}$.

(2) *Non-unicité* : si S est une statistique exhaustive pour $\mathcal{F}$ et l'application $s \mapsto g(s)$ est une bijection, alors $S' = g(S)$ est aussi une statistique exhaustive pour $\mathcal{F}$. Dans ce cas on dit que S' est *équivalente* à S.

(3) L'échantillon $S(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$ est une statistique exhaustive pour toute famille $\mathcal{F}$(dite *statistique exhaustive triviale*). Toute statistique équivalente à $(X_1, \ldots, X_n)$ est appelée triviale aussi.

La **statistique exhaustive minimale pour** $\mathcal{F}$ est définie comme une statistique S telle que toute statistique exhaustive pour $\mathcal{F}$ est fonction de S. Evidemment, la statistique exhaustive minimale n'est pas unique non plus.

Vérifions que pour les familles $\mathcal{F}$ relatives aux Exemples 4.1 et 4.2 il existe des statistiques exhaustives non-triviales.

Statistique exhaustive pour l'Exemple 4.1. Ici la famille $\mathcal{F} = \{F_\theta,\ \theta > 0\}$ où F_θ est la f.d.r. dont la densité est exponentielle de la forme (4.1). Si $X_i \sim F_\theta$, la densité jointe de $\mathbf{x} = (X_1, \ldots, X_n)$ est

$$f_{\mathbf{x}}(x_1, \ldots, x_n) = \theta^{-n} \exp(-\sum_{i=1}^{n} x_i/\theta) I\{x_1 > 0, \ldots, x_n > 0\} = \psi_\theta(S(x))h(x)$$

où $\psi_\theta(u) = \theta^{-n} \exp(-u/\theta)$, $S(x) = \sum_{i=1}^n x_i$ et $h(x) = I\{x_1 > 0, \ldots, x_n > 0\}$, $x = (x_1, ..., x_n)$. Nous allons montrer que la statistique $S = S(\mathbf{x}) = \sum_{i=1}^n X_i$ est exhaustive pour cette famille de lois. Considérons l'application linéaire $\mathbf{x} \mapsto \mathbf{y}$ (avec le Jacobien 1) où le vecteur aléatoire $\mathbf{y} = (Y_1, \ldots, Y_n)$ est défini par

$$Y_1 = \sum_{i=1}^{n} X_i = S,\ Y_2 = X_2,\ \ldots,\ Y_n = X_n.$$

Utilisant le Corollaire 3.1 on trouve la densité de $\mathbf{y}$:

$$f_{\mathbf{y}}(y_1, \ldots, y_n) = \theta^{-n} \exp(-y_1/\theta) I\{y_1 > y_2 + \ldots + y_n, y_2 > 0, \ldots, y_n > 0\}$$

d'où on obtient la densité marginale de $Y_1 = S$:

$$f_{Y_1}(y_1) = \int f_{\mathbf{y}}(y_1, \ldots, y_n) dy_2 \ldots dy_n = c(n)\theta^{-n} y_1^n \exp(-y_1/\theta) I\{y_1 > 0\}$$

(ici $c(n) > 0$ est une constante absolue). On en déduit que la densité conditionnelle $f_{\mathbf{y}|Y_1=s}(y_2,\ldots,y_n)$ ne dépend pas de θ :

$$f_{\mathbf{y}|Y_1=s}(y_2,\ldots,y_n) = \frac{f_{\mathbf{y}}(s,y_2,\ldots,y_n)}{f_{Y_1}(s)} = \frac{1}{c(n)s^n} I\{s > y_2+\ldots+y_n, y_2>0,\ldots,y_n>0\}.$$

Alors, la probabilité $P(\mathbf{y} \in B|Y_1 = s)$ n'est fonction que de s pour tout B borélien. Or, l'application $\mathbf{x} \mapsto \mathbf{y}$ est borélienne, donc aussi la probabilité $P(\mathbf{x} \in A|Y_1 = s)$ n'est fonction que de s pour tout A borélien : elle ne dépend pas de θ (et donc de F quand $F \in \mathcal{F} = \{F_\theta,\ \theta > 0\}$). Il s'ensuit que la statistique $S = Y_1$ est exhaustive pour $\mathcal{F}$.

Statistique exhaustive pour l'Exemple 4.2. Ici la famille $\mathcal{F}$ est l'ensemble de toutes les lois $\mathcal{D}(\{1,\ldots,N\},\theta)$ avec les paramètres $\theta = (p_1,\ldots,p_N)$, où $N = 20$.

Pour tout vecteur $x = (x_1,\ldots,x_n)$ avec les x_i appartenant à l'ensemble $\{1,\ldots,20\}$, définissons $S(x) = (n_1(x),\ldots,n_{N-1}(x))$, où

$$n_j(x) = \sum_{i=1}^n I\{x_i = j\}.$$

Soit $\mathbf{x} = (X_1,\ldots,X_n)$. Vérifions que la statistique $S = S(\mathbf{x})$ est exhaustive. Vu (4.2), la loi de $\mathbf{x}$ est donnée par

$$\begin{aligned} P(\mathbf{x} = (x_1,\ldots,x_n)) &= \prod_{i=1}^n \exp\left(\sum_{j=1}^{N-1} I\{x_i = j\}\ln\frac{p_j}{p_N} + \ln p_N\right) \\ &= \exp\left(\sum_{j=1}^{N-1} n_j(x)\ln\frac{p_j}{p_N} + n\ln p_N\right) \stackrel{\text{déf}}{=} \psi_\theta(S(x)) \end{aligned}$$

où $x_i \in \{1,\ldots,20\}$. On fixe maintenant le vecteur $s = (s_1,\ldots,s_{N-1})$ appartenant à l'ensemble des valeurs possibles de $S(\mathbf{x})$. Alors

$$P(\mathbf{x} = (x_1,\ldots,x_n), S(\mathbf{x}) = s) = \begin{cases} \psi_\theta(s) & \text{si } n_j(x) = s_j,\ j = 1,\ldots,N-1, \\ 0 & \text{sinon.} \end{cases}$$

Par conséquent,

$$P(\mathbf{x} = (x_1,\ldots,x_n)|S(\mathbf{x}) = s) = \begin{cases} 1/M(s) & \text{si } n_j(x) = s_j,\ j = 1,\ldots,N-1, \\ 0 & \text{sinon,} \end{cases}$$

où $M(s)$ est le nombre de tous les vecteurs $(x_1,\ldots,x_n)$ avec $x_i \in \{1,\ldots,20\}$ tels que $n_j(x) = s_j,\ j = 1,\ldots,N-1$. Evidemment, $M(s)$ ne dépend pas de $\theta = (p_1,\ldots,p_N)$ (et donc de $F \in \mathcal{F}$), ce qui implique l'exhaustivité de la statistique S. En utilisant la notation de l'Exemple 4.2, on peut écrire $S = (n_1,\ldots,n_{N-1})$, où $n_j = \sum_{i=1}^n I\{X_i = j\}$. L'exhaustivité de S explique pourquoi dans l'Exemple 4.2 il suffisait de considérer les données réduites $(n_1,\ldots,n_{20})$ au lieu des données initiales $(X_1,\ldots,X_n)$.

Les deux exemples ci-dessus sont des cas particuliers du résultat général de Théorie de la mesure connu sous le nom de Théorème de factorisation.

Théorème 4.2. (Théorème de factorisation.) *Soit $\mathcal{P}$ une famille de mesures de probabilité définies sur $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ telles que toute mesure $P \in \mathcal{P}$ est absolument continue par* *rapport à une mesure σ-finie μ_0 sur $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Soit $S : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ une fonction borélienne et soit $\mathbf{x}$ un vecteur aléatoire de loi P.*

Alors, la loi conditionnelle de $\mathbf{x}$ sachant que $S(\mathbf{x}) = s$ ne dépend pas de P pour tout $P \in \mathcal{P}$ si et seulement si il existe deux fonctions boréliennes positives $h(\cdot)$ (indépendante de P) et $\psi_P(\cdot)$ (dépendante de P) telles que

$$\frac{dP}{d\mu_0}(x) = \psi_P(S(x))h(x), \quad (\mu_0 - p.s.) \quad \forall P \in \mathcal{P}.$$

Dans les deux exemples ci-dessus, P est la mesure–produit qui correspond à l'échantillon $X_1, \ldots, X_n$, $\mathcal{P}$ est un ensemble de mesures–produits paramétrées par θ. La mesure dominante μ_0 est la mesure de Lebesgue dans l'Exemple 4.1 et la mesure de comptage dans l'Exemple 4.2.

Corollaire 4.1. *Soit l'Hypothèse (E0) vérifiée et soit $F \in \mathcal{F}$, où $\mathcal{F}$ est une famille de fonctions de répartition sur $\mathbb{R}$ absolument continues par rapport à une mesure σ-finie μ_0 sur $\mathbb{R}$. Soit f une densité de F par rapport à μ_0.*

Alors la statistique $S(X_1, \ldots, X_n)$ est exhaustive pour $\mathcal{F}$ si et seulement si il existe deux fonctions boréliennes positives $h(\cdot)$ (indépendante de F) et $\psi_F(\cdot)$ (dépendante de F) telles que

$$\prod_{i=1}^{n} f(x_i) = \psi_F(S(x))h(x), \quad (\mu_0 - p.s.) \quad \forall F \in \mathcal{F}, \tag{4.11}$$

où $x = (x_1, \ldots, x_n)$.

Remarque. Si $\mathcal{F}$ est une famille paramétrée par $\theta \in \Theta \subseteq \mathbb{R}^k$: $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ et si $f(\cdot, \theta)$ est la densité qui correspond à F_θ, la condition de factorisation (4.11) se traduit par

$$\prod_{i=1}^{n} f(x_i, \theta) = \psi_\theta(S(x))h(x), \quad (\mu_0 - p.s.) \quad \forall \theta \in \Theta.$$

Exercice 4.2. Montrer que le couple $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ est une statistique exhaustive pour la famille des lois normales $\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$ (par conséquent, le couple $(\bar{X}, s^2)$ est aussi une statistique exhaustive pour cette famille).

Exemple 4.3. Soit $\mathcal{F}$ la famille de tous les lois admettant une densité f par rapport à la mesure de Lebesgue. Alors

$$\prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_{(i)})$$

où $x_{(1)} \leq \ldots \leq x_{(n)}$ sont les valeurs $(x_1, \ldots, x_n)$ rangées par ordre croissant. Vu le Corollaire 4.1, on en déduit que la statistique d'ordre $(X_{(1)}, \ldots, X_{(n)})$ est exhaustive pour $\mathcal{F}$ (et donc pour toute sous-famille de $\mathcal{F}$).

Exemple 4.4. Soit $\mathcal{F}$ l'ensemble de tous les lois admettant une densité symétrique f par rapport à la mesure de Lebesgue. Alors $f(t) = f(|t|)$, et le Corollaire 4.1 permet de déduire que $(|X_1|, \ldots, |X_n|)$ est une statistique exhaustive. De plus, vu l'Exemple 4.3, $(|X|_{(1)}, \ldots, |X|_{(n)})$

est aussi exhaustive. Ici $|X|_{(1)} \leq \ldots \leq |X|_{(n)}$ sont les valeurs $|X_1|, \ldots, |X_n|$ rangées par ordre croissant.

Dans les Exemples 4.3 et 4.4, les statistiques exhaustives ne sont pas très différentes da la statistique exhaustive triviale. L'existence des statistiques exhaustives non-triviales pour une famille $\mathcal{F}$ n'est pas toujours garantie.

EXEMPLE 4.5. Soit $\mathcal{F}$ l'ensemble des lois de Cauchy sur $\mathbb{R}$ avec les densités

$$f(t, \theta) = \frac{1}{\pi(1 + (t - \theta)^2)}, \quad \theta \in \mathbb{R}.$$

Alors, la factorisation de type (4.11) de la densité-produit $\prod_{i=1}^n f(x_i, \theta)$ avec une statistique S à valeurs dans un espace de dimension $< n$ n'est pas possible. On peut montrer que la statistique exhaustive minimale dans cet exemple est la statistique d'ordre $(X_{(1)}, \ldots, X_{(n)})$. Le concept d'exhaustivité est donc sans intérêt pour cette famille des lois.

REMARQUE. Bien que la notion d'exhaustivité soit célèbre dans la littérature statistique, son rôle réel est modeste pour les raisons suivantes :

– on peut expliciter des statistiques exhaustives non-triviales seulement dans des cas exceptionnels, pour quelques familles $\mathcal{F}$ remarquables,
– dans la pratique, la famille $\mathcal{F}$ n'est pas donnée. Le statisticien peut se tromper du choix de $\mathcal{F}$ de façon qu'en vérité la loi sous-jacente F peut appartenir à une famille $\mathcal{F}_1$ inconnue et différente de $\mathcal{F}$. Une statistique exhaustive pour $\mathcal{F}$ n'est pas, en général, exhaustive pour $\mathcal{F}_1$. Le principe : "oublier l'échantillon initial et ne garder que la statistique exhaustive" n'est pas bien fondé dans ce contexte.

4.5. Propriétés des statistiques $\bar{X}$ et s^2

Proposition 4.1. *Pour tout c réel,*

$$\frac{1}{n}\sum_{i=1}^n (X_i - c)^2 = (\bar{X} - c)^2 + \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = (\bar{X} - c)^2 + s^2.$$

Preuve. On utilise la Proposition 1.1 pour la variable aléatoire ξ de loi discrète uniforme sur $\{X_1, \ldots, X_n\}$.

Proposition 4.2. *Si $E(X^2) < \infty$, $E(X) = \mu$, alors*

$$E(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\sigma^2}{n}, \quad E(s^2) = \frac{n-1}{n}\sigma^2.$$

Preuve. On utilise la Proposition 1.7 et on note que, d'après la Proposition 4.1 (avec $c = \mu$),

$$E(s^2) = \frac{1}{n}\sum_{i=1}^n E((X_i - \mu)^2) - E((\bar{X} - \mu)^2) = \mathrm{Var}(X) - \mathrm{Var}(\bar{X}).$$

∎

La proposition suivante est une conséquence immédiate de la loi forte des grands nombres.

Proposition 4.3. *Si $E(X^2) < \infty$, alors $\bar{X} \to \mu$ (p.s.) et $s^2 \to \sigma^2$ (p.s.) quand $n \to \infty$.*

Si les X_i sont des v.a. normales, on peut expliciter la loi jointe des statistiques $\bar{X}$ et s^2 pour tout n :

Proposition 4.4. *Soient $X_1, ..., X_n$ des variables aléatoires i.i.d. de loi normale, $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Alors,*

(i) $\bar{X} \perp\!\!\!\perp s^2$.

(ii) $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

(iii) $\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$.

Preuve. Introduisons le vecteur aléatoire normal

$$\xi = (X_1, ..., X_n)^T, \quad \xi \sim \mathcal{N}_n(m, \sigma^2 I),$$

avec $m = (\mu, ..., \mu)^T$. Soit $\eta = (\xi - E(\xi))/\sigma = (\xi - m)/\sigma$. Evidemment, $\eta \sim \mathcal{N}_n(0, I)$. Introduisons aussi la matrice $n \times n$ suivante :

$$A = \frac{1}{n}\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}.$$

Cette matrice est symétrique et idempotente :

$$A^2 = \frac{1}{n^2}\begin{pmatrix} n & \dots & n \\ \vdots & \ddots & \vdots \\ n & \dots & n \end{pmatrix} = \frac{1}{n}\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix} = A,$$

donc un projecteur. Posons

$$\eta_1 = A\eta = \frac{1}{\sigma}A(\xi - m) = \frac{1}{n\sigma}\begin{pmatrix} n(\bar{X} - \mu) \\ \vdots \\ n(\bar{X} - \mu) \end{pmatrix} = \frac{1}{\sigma}\begin{pmatrix} \bar{X} - \mu \\ \vdots \\ \bar{X} - \mu \end{pmatrix}$$

et

$$\eta_2 = (I - A)\eta = \frac{1}{\sigma}(I - A)(\xi - m) = \frac{1}{\sigma}\begin{pmatrix} X_1 - \mu \\ \vdots \\ X_n - \mu \end{pmatrix} - \frac{1}{\sigma}\begin{pmatrix} \bar{X} - \mu \\ \vdots \\ \bar{X} - \mu \end{pmatrix} = \frac{1}{\sigma}\begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix}.$$

Notons que $\mathrm{Rang}(A) = 1$ et $\mathrm{Rang}(I - A) = n - 1$. Les matrices $A_1 = A$ et $A_2 = I - A$ vérifient les hypothèses du Théorème de Cochran. Il s'ensuit que η_1 et η_2 sont indépendants et $\|\eta_2\|^2 \sim \chi^2_{n-1}$. Or,

$$\|\eta_2\|^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{ns^2}{\sigma^2},$$

d'où découle la partie (iii) de la proposition. Puisque $\eta_1 \perp\!\!\!\perp \eta_2$ et vu le fait que les transformations mesurables préservent l'indépendance, on obtient

$$\frac{\bar{X} - \mu}{\sigma} \perp\!\!\!\perp \frac{ns^2}{\sigma^2} \qquad \text{et} \qquad \bar{X} \perp\!\!\!\perp s^2,$$

ce qui démontre la partie (i) de la proposition. La partie (ii) est évidente. ∎

Corollaire 4.2. *Si $X_1, ..., X_n$ sont des variables aléatoires i.i.d $\mathcal{N}(\mu, \sigma^2)$, alors la variable aléatoire*

$$t = \sqrt{n-1}(\bar{X} - \mu)/s$$

suit la loi de Student t_{n-1} à $n-1$ degrés de liberté.

Preuve. Vu la Proposition 4.4 (ii), $\sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0,1)$, alors

$$\sqrt{n-1}\frac{\bar{X} - \mu}{s} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}\sqrt{\frac{(n-1)\sigma^2}{ns^2}} = \frac{\eta}{\sqrt{\chi/(n-1)}} ,$$

où $\eta \sim \mathcal{N}(0,1)$ et $\chi = ns^2/\sigma^2 \sim \chi^2_{n-1}$. De plus, les v.a. η et χ sont indépendantes d'après la Proposition 4.4 (i). ∎

4.6. Covariance et corrélation empiriques

Considérons maintenant un couple de variables aléatoires (X, Y) et l'échantillon de couples $(X_1, Y_1), \dots, (X_n, Y_n)$, où chaque (X_i, Y_i) suit la même loi que (X, Y). Introduisons les caractéristiques empiriques correspondant à la covariance $\mathrm{Cov}(X, Y)$ et à la corrélation $\mathrm{Corr}(X, Y) = \rho_{XY}$.

La **covariance empirique** entre X et Y est définie par :

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}.$$

Le **coefficient de corrélation empirique** (ou la **corrélation empirique**) entre X et Y est défini par :

$$r_{XY} = \frac{s_{XY}}{s_X s_Y},$$

où $s_X = \sqrt{n^{-1}\sum_{i=1}^{n} X_i^2 - \bar{X}^2}$ est l'écart-type de l'échantillon $(X_1, ..., X_n)$, s_Y est l'écart-type de l'échantillon $(Y_1, ..., Y_n)$ et l'on suppose que $s_X > 0, s_Y > 0$.

Proposition 4.5. *Soient (X, Y) deux v.a. telles que $E(X^2) < \infty$, $E(Y^2) < \infty$ et soient n couples indépendants de v.a. $(X_1, Y_1), \dots, (X_n, Y_n)$, tels que chaque (X_i, Y_i) suit la même loi que (X, Y). Alors les covariances empiriques convergent presque sûrement vers les covariances théoriques :*

$$s_{XY} \to \mathrm{Cov}(X, Y) \quad (p.s.) \quad \text{quand} \quad n \to \infty.$$

Si, de plus, $\mathrm{Var}(X) > 0$ et $\mathrm{Var}(Y) > 0$, alors les corrélations empiriques convergent presque sûrement vers les corrélations théoriques :

$$r_{XY} \to \rho_{XY} \quad (p.s.) \quad \text{quand} \quad n \to \infty.$$

Preuve. Elle est immédiate d'après la loi forte de grands nombres et le Premier théorème de continuité (cf. partie (i) de la Proposition 1.10). ∎

Propriétés des corrélations empiriques.

1°. $|r_{XY}| \leq 1$.

2°. $|r_{XY}| = 1$ si et seulement si il existe un lien linéaire entre (X_i) et (Y_i), i.e. il existe $a \neq 0$, $b \in \mathbb{R}$, tels que

$$Y_i = aX_i + b, \quad i = 1, ..., n.$$

On a l'interprétation géométrique suivante de r_{XY} : r_{XY} est le cosinus de l'angle φ entre les vecteurs $(X_1 - \bar{X}, ..., X_n - \bar{X})^T$ et $(Y_1 - \bar{Y}, ..., Y_n - \bar{Y})^T$. Alors $|r_{XY}| = 1$ implique que $\varphi = 0$ ou $\varphi = \pi$, i.e. que les deux vecteurs sont colinéaires.

3°. Si $r_{XY} = 0$, alors $\varphi = \pi/2$ et les deux vecteurs sont orthogonaux.

4°. La corrélation empirique est invariante par rapport aux transformations affines : pour tout $a \neq 0$, $b, d \in \mathbb{R}$,

$$r_{aX+b,aY+d} = r_{XY}.$$

De plus, si $c \neq 0$,

$$|r_{aX+b,cY+d}| = |r_{XY}|.$$

5°. La corrélation empirique n'est pas stable par rapport aux observations aberrantes, comme le montre la figure suivante.

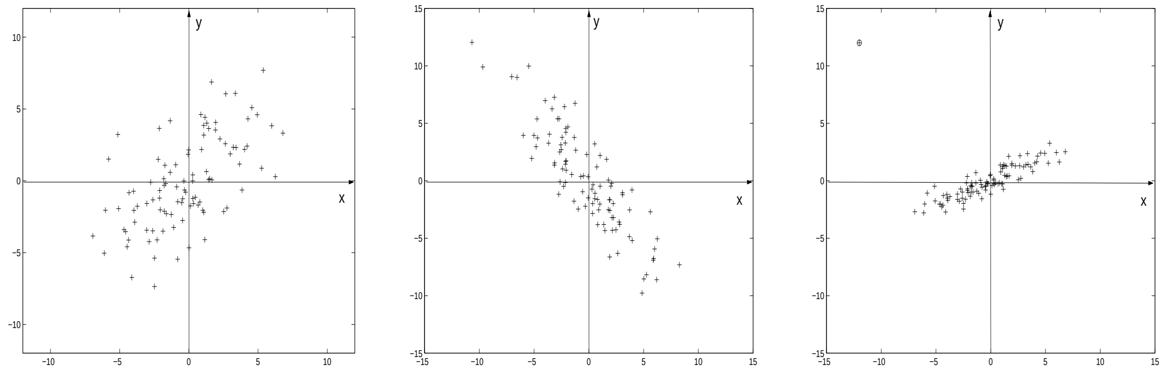

Figure 4.2. De gauche à droite : les "nuages" des points (X_i, Y_i) avec $r_{XY} > 0$, avec $r_{XY} < 0$ et le "nuage" perturbé par une observation aberrante tel que $r_{XY} < 0$ au lieu de $r_{XY} > 0$.

REMARQUES.

(1) La relation $|r_{XY}| = 1$ *n'implique pas* que les variables aléatoires théoriques X et Y soient liées d'un lien linéaire. Elle signifie seulement que les vecteurs de données (X_i) et (Y_i) sont liés linéairement. Il ne s'agit donc qu'une approximation, obtenue à partir de données, de la situation théorique sous-jacente.

(2) C'est rare, voire impossible, d'avoir $|r_{XY}| = 1$ ou $r_{XY} = 0$ pour les données réelles. Dans la pratique, il s'agit plutôt d'égalités approximatives $|r_{XY}| \approx 1$ ou $r_{XY} \approx 0$.

4.7. Construction d'un échantillon pseudo-aléatoire par simulation*

Dans les applications, on a souvent besoin de générer de façon artificielle (à l'aide d'un ordinateur) une suite $X_1, ..., X_n$ de nombres aléatoires i.i.d. suivant la loi donnée F. Les méthodes de simulation permettent d'obtenir seulement une valeur **pseudo-aléatoire** X_i, au lieu d'une valeur aléatoire. Cela signifie que les nombres $X_1, ..., X_n$ simulés sont **déterministes** – ils sont obtenus par un algorithme déterministe – mais les propriétés de la suite $X_1, ..., X_n$ sont proches de celles d'une suite aléatoire i.i.d. de loi donnée. Par exemple, pour les X_i pseudo-aléatoires on a la propriété de Glivenko-Cantelli :

$$\sup_x |\widehat{F}_n(x) - F(x)| \to 0 \text{ quand } n \to \infty,$$

mais il s'agit ici de la convergence au sens déterministe.

4.7.1. Simulation des variables uniformément distribuées. La f.d.r. $F^U(\cdot)$ de la loi uniforme $U[0,1]$ s'écrit sous la forme

$$F^U(x) = \begin{cases} 0, & x < 0, \\ x, & x \in [0,1], \\ 1, & x > 1. \end{cases}$$

Le programme-générateur d'un échantillon pseudo-aléatoire $U_1, \ldots, U_n$ de cette loi est disponible dans les nombreux logiciels. Le principe de son fonctionnement est le suivant. On se donne un réel $a > 1$ et un entier m (d'habitude a et m sont de très grands nombres). On commence par une valeur z_0 fixe. Pour tout $1 \le i \le n$ on définit

$$\begin{aligned} z_i &= \text{ le reste de division de } az_{i-1} \text{ par } m \\ &= az_{i-1} - \left[\frac{az_{i-1}}{m}\right] m, \end{aligned}$$

où $[\cdot]$ désigne la partie entière. Nous avons toujours $0 \le z_i < m$. On définit

$$U_i = \frac{z_i}{m} = \frac{az_{i-1}}{m} - \left[\frac{az_{i-1}}{m}\right].$$

Alors, $0 \le U_i < 1$. La suite $U_1, ..., U_n$ est considérée comme un échantillon de la loi uniforme $U[0,1]$. Bien que ce n'est pas une suite aléatoire, on peut montrer que la f.d.r. empirique

$$\widehat{F}_n^U(x) = \frac{1}{n} \sum_{i=1}^n I\{U_i \le x\}$$

est telle que $\sup_x |\widehat{F}_n^U - F^U(x)| = \sup_{0 \le x \le 1} |\widehat{F}_n^U - x| \le \epsilon(n,m)$ avec $\epsilon(n,m)$ qui converge très vite vers 0 quand $m \to \infty$ et $n \to \infty$. Autrement dit, on a la propriété de Glivenko – Cantelli au sens déterministe. Divers résultats mathématiques permettent de justifier de bons choix de z_0, a et m. Les valeurs suivantes sont souvent utilisées et donnent, en général, satisfaction : $a = 16807 (= 7^5)$, $m = 2147483647 (= 2^{31} - 1)$.

4.7.2. Simulation des variables d'une loi générale. Étant donné un échantillon i.i.d. $U_1, ..., U_n$ d'une loi uniforme, on peut obtenir un échantillon d'une loi générale $F(\cdot)$ par la **méthode d'inversion**. Elle est opérationnelle si F^{-1} est disponible sous la forme explicite. Cette méthode est basée sur la proposition suivante.

Proposition 4.6. *Soit F une f.d.r. continue et strictement croissante et soit U une variable aléatoire uniformément distribuée sur $[0,1]$. Alors la v.a.*

$$X = F^{-1}(U)$$

suit la loi F.

Preuve. On note que

$$F(x) = P(U \le F(x)) = P(F^{-1}(U) \le x) = P(X \le x).$$

∎

Il en découle l'algorithme de simulation suivant : si F est continue et strictement croissante, posons

$$X_i = F^{-1}(U_i),$$

où les U_i sont des nombres pseudo-aléatoires uniformément distribués sur $[0,1]$ générés comme expliqué précédemment. On obtient ainsi un échantillon simulé $(X_1, ..., X_n)$.

Si F n'est pas continue ou strictement croissante, il faut modifier la définition de F^{-1}. On pose

$$F^{-1}(y) \overset{\text{déf}}{=} \sup\{t : \ F(t) < y\}, \quad y \in [0,1].$$

Alors,

$$P(X_i \le x) = P(\sup\{t : \ F(t) < U_i\} \le x) = P(U_i \le F(x)) = F(x).$$

Exemple 4.6. *Simulation d'un échantillon de loi exponentielle* $\mathcal{E}(1)$. On a :

$$f(x) = e^{-x} I\{x > 0\}, \quad F(x) = (1 - e^{-x}) I\{x > 0\}.$$

Alors, $F^{-1}(y) = -\ln(1-y)$ pour $y \in (0,1)$. Posons $X_i = -\ln(1-U_i)$, où les U_i sont des nombres pseudo-aléatoires uniformément distribués sur $[0,1]$.

Exemple 4.7. *Simulation d'un échantillon de loi de Bernoulli.* Soit

$$P(X=1) = p, \quad P(X=0) = 1-p, \quad 0 < p < 1.$$

On utilise la méthode modifiée :

$$F^{-1}(y) = \sup\{t : \ F(t) < y\} = \begin{cases} 0, & y \in [0, 1-p], \\ 1, & y \in]1-p, 1]. \end{cases}$$

Si U_i est une v.a. de loi uniforme, alors $X_i = F^{-1}(U_i)$ suit la loi de Bernoulli. On pose alors

$$X_i = \begin{cases} 0, & U_i \in [0, 1-p], \\ 1, & U_i \in]1-p, 1]. \end{cases}$$

4.7.3. Simulation des variables transformées. Pour simuler un échantillon $Y_1, ..., Y_n$ de loi $F((\cdot - \mu)/\sigma)$, où $\sigma > 0$ et $\mu \in \mathbb{R}$, étant donné l'échantillon $X_1, ..., X_n$ de loi $F(\cdot)$, il suffit de prendre $Y_i = \sigma X_i + \mu$, $i = 1, ..., n$.

4.7.4. Simulation de la loi normale standard. La f.d.r. F de loi normale $\mathcal{N}(0,1)$ est continue et strictement croissante, mais F^{-1} n'est pas disponible sous la forme explicite. Alors, il est difficile d'appliquer la méthode d'inversion. Il existe néanmoins d'autres méthodes de simulation très performantes du point de vue du coût de calcul.

Utilisation du Théorème central limite. Pour $U \sim U[0,1]$ nous avons $E(U) = 1/2$ et $\mathrm{Var}(U) = 1/12$. Vu le Théorème central limite,

$$\frac{U_1 + \cdots + U_N - N/2}{\sqrt{N/12}} \xrightarrow{D} \mathcal{N}(0,1) \ \text{ quand } \ N \to \infty,$$

pour un échantillon i.i.d. $U_1, \ldots, U_N$ de loi uniforme sur $[0,1]$. La valeur $N = 12$ est déjà suffisante pour obtenir ainsi une bonne approximation de la loi normale. On en déduit la méthode de simulation suivante : on génère $U_1, U_2, \ldots, U_{nN}$, une suite de variables pseudo-aléatoires de loi $U[0,1]$ et on pose ensuite

$$X_i = \frac{U_{(i-1)N+1} + \cdots + U_{iN} - N/2}{\sqrt{N/12}}, \quad i = 1, ..., n.$$

On obtient ainsi un échantillon simulé $(X_1, \ldots, X_n)$ de la loi $\mathcal{N}(0,1)$.

Méthode de Box et Müller. Elle découle du résultat suivant (Exercice 3.9).

Proposition 4.7. *Soient ξ et η deux variables aléatoires indépendantes de loi $U[0,1]$. Alors les v.a.*

$$X = \sqrt{-2\ln\xi}\cos(2\pi\eta) \ \ et \ \ Y = \sqrt{-2\ln\xi}\sin(2\pi\eta)$$

sont normales et indépendantes avec $E(X) = E(Y) = 0$, $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$.

Ce résultat nous donne la méthode de simulation de $(X_1, ..., X_n)$ suivante : on génère des variables pseudo-aléatoires $U_1, \ldots, U_{2n}$ de loi $U[0,1]$ et on pose ensuite

$$\begin{aligned} X_{2i-1} &= \sqrt{-2\ln U_{2i}}\sin(2\pi U_{2i-1}), \\ X_{2i} &= \sqrt{-2\ln U_{2i}}\cos(2\pi U_{2i-1}), \end{aligned}$$

pour $i = 1, \ldots, n$.

4.8. Exercices

Exercice 4.3. Soit $X_1, \ldots, X_n$ un échantillon i.i.d., $X_i \sim F$. On considère la valeur de la fonction de répartition empirique $\widehat{F}_n(t)$ au point fixé t.

1^o. Quelle est la loi de $n\widehat{F}_n(t)$?
2^o. Calculer $E\left([\widehat{F}_n(t) - F(t)]^2\right)$ et en déduire que $\widehat{F}_n(t)$ converge en moyenne quadratique vers $F(t)$ lorsque $n \to \infty$.
3^o. Chercher la loi limite de $\sqrt{n}(\widehat{F}_n(t) - F(t))$ lorsque $n \to \infty$.

Exercice 4.4. Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et de même loi exponentielle, ayant comme densité $f(x) = \lambda\exp(-\lambda x)I(x > 0)$.
1^o. Donner la loi de $\bar{X}$. Calculer $E(1/\bar{X})$ et $\mathrm{Var}(1/\bar{X})$. Montrer que $E(1/\bar{X})$ tend vers λ quand n tend vers l'infini. Établir la relation

$$E\left((1/\bar{X} - \lambda)^2\right) = \mathrm{Var}(1/\bar{X}) + \left(E(1/\bar{X}) - \lambda\right)^2,$$

puis en déduire que $E\left((1/\bar{X} - \lambda)^2\right) \to 0$ quand n tend vers l'infini.
2^o. Montrer que $1/\bar{X}$ tend en probabilité vers λ. Donner la loi limite de $\sqrt{n}(\bar{X} - \frac{1}{\lambda})$, puis celle de

$$\sqrt{n}(1/\bar{X} - \lambda).$$

La variance de cette loi est–elle égale à $\lim_{n\to\infty} n\mathrm{Var}(1/\bar{X})$?

EXERCICE 4.5. Soit $X_1, ..., X_n$ un échantillon i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Considérons l'estimateur de σ de la forme :

$$\widehat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}.$$

Utilisez les théorèmes de continuité (Propositions 1.10 et 1.11) pour montrer la convergence $\widehat{\sigma}_n \to \sigma$ (p.s.) et établir la loi limite de $\sqrt{n}(\widehat{\sigma}_n - \sigma)$.

EXERCICE 4.6. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de fonction de répartition F. On suppose que F admet une densité f par rapport à la mesure de Lebesgue. On considère la statistique d'ordre $(X_{(1)}, \ldots, X_{(n)})$.

1^o. Déterminer la densité $f_k(x)$ de $X_{(k)}$. Calculer la fonction de répartition, notée $G_k(x)$, de $X_{(k)}$.
2^o. Donner la loi du couple $(X_{(1)}, X_{(n)})$ et la loi de la statistique $W = X_{(n)} - X_{(1)}$ (on appelle W *étendue*). Les variables $X_{(1)}$ et $X_{(n)}$ sont–elles indépendantes ?
3^o. Soient les variables aléatoires :

$$Y_k = F(X_{(k)}) \text{ et } Z_k = G_k(X_{(k)}).$$

Quelles lois suivent–elles ?

EXERCICE 4.7. Montrer que $X_{(n)}$ est une statistique exhaustive pour la famille des lois uniformes $\{U[0, \theta], \theta > 0\}$. Peut-on en déduire l'exhaustivité des statistiques $8X_{(n)}$, $X_{(n)} + \bar{X}$, $X_{(n)}^2 + 5$?

EXERCICE 4.8. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. ayant le moment d'ordre 4 fini. Le but de cet exercice est de calculer $\mathrm{Var}(s^2)$, où s^2 est la variance empirique associée à l'échantillon $(X_1, \ldots, X_n)$. On rappelle que $\sum_{i=1}^{n-1} i = \frac{n(n-1)}{2}$.
1^o. Montrer que l'on peut supposer sans perte de généralité que les X_i sont centrées : $E(X_i) = 0$. On fera cette hypothèse dans la suite.
2^o. Démontrer que :

$$s^2 = \frac{n-1}{n^2}\sum_{i=1}^{n} X_i^2 - \frac{2}{n^2}\sum_{k<l} X_k X_l\ .$$

3^o. Montrer que

$$\mathrm{Cov}(\sum_{i=1}^{n} X_i^2, \sum_{k<l} X_k X_l) = 0, \quad \mathrm{Var}(\sum_{k<l} X_k X_l) = n(n-1)\sigma^4/2.$$

En déduire que :

$$\mathrm{Var}(s^2) = \frac{n-1}{n^3}\left((n-1)E(X_1^4) - (n-3)(E(X_1^2))^2\right)\ .$$

4^o. Expliciter $\mathrm{Var}(s^2)$ quand $X_1 \sim \mathcal{N}(0, \sigma^2)$.

EXERCICE 4.9. Soient (X_i, ε_i), $i = 1, \ldots, n$, des couples de variables de même loi et indépendantes entre elles. On suppose que X_i et ε_i admettent des moments d'ordre 2 finis et que $E(\varepsilon_1) = 0$, $E(X_1^2) > 0$. Pour un réel b, on pose $Y_i = bX_i + \varepsilon_i$ et on note

$$\hat{b}_n = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2} .$$

1^o. En observant que

$$\hat{b}_n = b + \frac{\sum_{i=1}^n \varepsilon_i X_i/n}{\sum_{i=1}^n X_i^2/n},$$

déduire que $\hat{b}_n$ converge presque sûrement vers b. On pourra utiliser pour cela la loi forte des grands nombres.
2^o. Trouver la loi limite de $\sqrt{n}(\hat{b}_n - b)$ quand $n \to \infty$.

EXERCICE 4.10. *Méthode de Monte-Carlo.* On cherche à calculer l'intégrale $I = \int_0^1 f(x)dx$. Soit X une variable aléatoire de loi uniforme $U[0,1]$, alors

$$E(f(X)) = \int_0^1 f(x)dx = I.$$

Soient $X_1, ..., X_n$ des v.a. i.i.d de loi $U[0,1]$. Considérons l'estimateur de I de la forme :

$$I_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$$

et supposons que $\sigma^2 = \mathrm{Var}(f(X)) < \infty$. Montrer que $E(I_n) \to I$ et $I_n \to I$ (p.s.) quand $n \to \infty$.

EXERCICE 4.11. Décrire un algorithme de simulation d'une loi de Poisson par inversion. *Indication* : il n'y a pas d'expression simple pour la fonction de répartition F, et l'ensemble des valeurs possibles de F est dénombrable. On peut calculer les valeurs $F(k)$ au fur et à mesure. En effet, si X suit la loi de Poisson $\mathcal{P}(\lambda)$,

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} = \frac{\lambda}{k}P(X = k-1).$$

Il en découle que les valeurs $F(k)$ peuvent être calculées de façon récursive :

$$F(0) = P(0) = e^{-\lambda}, \quad P(X = k) = \frac{\lambda}{k}P(X = k-1), \quad F(k) = F(k-1) + P(X = k).$$

Voici les 6 premières valeurs de $F(k)$ pour $\lambda = 1$:

k	0	1	2	3	4	5
$F(k)$	0.3679	0.7358	0.9193	0.9810	0.9963	0.9994

Notons que dans 9994 cas sur 10000, les 6 valeurs précalculées suffiront.

5
Estimation des paramètres

Dans ce chapitre, nous supposerons que la loi de l'échantillon aléatoire est connue à un paramètre près. Le problème de reconstitution de cette loi se réduit alors à celui de l'estimation du paramètre. Nous allons étudier ici deux méthodes classiques de l'estimation : la méthode des moments et celle du maximum de vraisemblance. Tout d'abord, nous préciserons le modèle mathématique et introduirons quelques notions permettant de qualifier des estimateurs de bons ou de mauvais.

5.1. Modèle statistique. Problème d'estimation des paramètres

Comme dans le Chapitre 4, il s'agit ici d'un échantillon i.i.d. $\mathcal{X}_n = (X_1, ..., X_n)$. Cependant, les X_i peuvent être vectoriels. *L'hypothèse d'échantillonnage* suivante sera postulée tout au long de ce chapitre.

Hypothèse (E). *Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^m$ défini sur l'espace de probabilité $(\Omega, \mathcal{A}, P)$, de fonction de répartition F. L'échantillon $\mathcal{X}_n = (X_1, \ldots, X_n)$ est une* **réalisation** *de X, c'est-à-dire les observations $X_1, ..., X_n$ sont des vecteurs aléatoires i.i.d. de la même loi F que X ($X_i \sim F$).*

L'autre hypothèse fondamentale de ce chapitre est que la forme *paramétrique* de F est connue.

Hypothèse (P). *La fonction de répartition F des X_i appartient à une famille $\mathcal{F}$ paramétrique de fonctions de répartition :*

$$\mathcal{F} \stackrel{\text{déf}}{=} \{F_\theta,\ \theta \in \Theta\},$$

où $\Theta \subseteq \mathbb{R}^k$ est un ensemble connu et $F_\theta(x)$ est connue comme fonction de x et θ.

On appelle Θ *ensemble des paramèteres*. Sous l'Hypothèse (P), $F = F_{\theta^*}$, où $\theta^* \in \Theta$ est appelé *la vraie valeur du paramètre*. Le seul inconnu dans cette construction est θ^*. Pour identifier F, il suffit donc de trouver la vraie valeur θ^* du paramètre θ.

Le problème de l'estimation statistique consiste à construire une statistique (un estimateur) $\widehat{\theta}_n = \widehat{\theta}_n(X_1, ..., X_n)$ qui soit proche de θ^* en un sens probabiliste.

Le mot *estimateur* désignera aussi, pour abréger, une suite d'estimateurs $(\widehat{\theta}_n(X_1, ..., X_n))_{n\geq 1}$ ou bien la règle selon laquelle est définie la statistique $\widehat{\theta}_n(X_1, ..., X_n)$ pour tout n donné. Autrement dit, lorsque nous écrirons "l'estimateur $\widehat{\theta}_n(X_1, ..., X_n)$", nous entendrons par là "la *suite* d'estimateurs $(\widehat{\theta}_n(X_1, ..., X_n))_{n\geq 1}$". Cette précision sera utile à noter quand il s'agira des propriétés asymptotiques d'un estimateur $\widehat{\theta}_n$ pour $n \to \infty$.

Pour que θ^* soit défini de façon unique, il faut imposer la condition suivante (*Hypothèse d'identifiabilité*) sur la famille $\mathcal{F}$.

Hypothèse (Id). *Pour* $\theta, \theta' \in \Theta$,

$$F_\theta(\cdot) = F_{\theta'}(\cdot) \implies \theta = \theta'.$$

Si l'Hypothèse (Id) n'est pas vérifiée, deux valeurs différentes de θ peuvent donner des f.d.r. identiques, auquel cas l'unicité de la vraie valeur du paramètre θ^* est compromise.

5.1.1. Modèle statistique dans le cadre i.i.d. Dans ce chapitre, on supposera que les Hypothèses (E) et (P) énoncées ci-dessus sont vérifiées. On adopte la définition suivante.

Définition 5.1. *Soient les Hypothèses (E) et (P) vérifiées. Alors, la famille* $\{F_\theta,\ \theta \in \Theta\}$ *est appelée* **modèle statistique** *(ou* **modèle statistique paramétrique***).*

On dit qu'un modèle statistique est *identifiable* s'il vérifie l'Hypothèse (Id).

EXEMPLE 5.1. *Modèle normal à moyenne inconnue et variance connue* $\{\mathcal{N}(\theta, \sigma^2),\ \theta \in \mathbb{R}\}$. Soit $\sigma^2 > 0$ une valeur connue, θ un paramètre inconnu à estimer. Il s'agit du modèle statistique $\{F_\theta(x),\ \theta \in \Theta\}$, où $\Theta = \mathbb{R}$ et $F_\theta(\cdot)$ est la loi admettant la densité suivante par rapport à la mesure de Lebesgue sur $\mathbb{R}$:

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x-\theta)^2}{2\sigma^2}\Big).$$

EXEMPLE 5.2. *Modèle normal à moyenne et variance inconnues* $\{\mathcal{N}(\theta_1, \theta_2^2),\ \theta_1 \in \mathbb{R},\ \theta_2 > 0\}$. Il s'agit du modèle statistique avec le paramètre vectoriel inconnu $\theta = (\theta_1, \theta_2)$, l'ensemble des paramètres $\Theta = \mathbb{R}\times]0, \infty[$ et la loi F_θ de densité

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}\theta_2} \exp\Big(-\frac{(x-\theta_1)^2}{2\theta_2^2}\Big).$$

EXEMPLE 5.3. *Modèle de Poisson* $\{\mathcal{P}(\theta), \theta > 0\}$.
Pour ce modèle, F_θ est la loi de Poisson de paramètre $\theta > 0$, i.e. la loi de la v.a. discrète X à valeurs dans l'ensemble des entiers positifs définie par la fonction de probabilité

$$P(X = x) = f(x, \theta) = \frac{\theta^x}{x!}\, e^{-\theta}, \quad x = 0, 1, 2, \dots\ ,$$

L'ensembles des paramètres est $\Theta =]0, \infty[$.

EXEMPLE 5.4. *Modèle de Bernoulli* $\{\mathcal{B}e(\theta), 0 < \theta < 1\}$.
Pour ce modèle, F_θ est la loi de la v.a. X prenant les valeurs 0 et 1 avec les probabilités $P(X = 1) = \theta$ et $P(X = 0) = 1 - \theta$, où θ appartient à l'ensemble des paramètres $\Theta =]0, 1[$.

EXEMPLE 5.5. *Un modèle non-identifiable.*
Soit F_θ la fonction de répartition correspondant à la densité

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{(x - \theta^2)^2}{2} \Big), \quad x \in \mathbb{R},$$

et soit $\Theta = \mathbb{R}$. Alors le modèle $\{F_\theta,\ \theta \in \Theta\}$ n'est pas identifiable. Néanmoins, si l'on prend $\Theta = \{\theta \geq 0\}$, le modèle devient identifiable. Cet exemple montre que le choix correct de l'ensemble des paramètres Θ est important.

5.1.2. Modèles dominés. Modèles discrets et continus. Dans les Exemples 5.1-5.5, la loi F_θ admet une densité $f(x, \theta)$ soit par rapport à la mesure de Lebesgue (cas continu), soit par rapport à la mesure de comptage (cas discret). Plus généralement, nous supposerons partout dans ce chapitre que l'hypothèse suivante (*Hypothèse de dominance*) est vérifiée.

Hypothèse (D). *Il existe une mesure σ-finie μ sur $\mathcal{B}(\mathbb{R}^m)$ telle que, pour tout $\theta \in \Theta$, F_θ admet une densité $f(x, \theta)$ par rapport à μ.*

Si l'Hypothèse (D) est vérifiée, on dit que $\{F_\theta,\ \theta \in \Theta\}$ est un *modèle statistique dominé.*

Par la suite, nous considérerons principalement les deux cas suivants :

$$\mu = \begin{cases} \text{mesure de Lebesgue sur } \mathbb{R}^m \ (\textit{cas continu}), \\ \text{mesure de comptage } (\textit{cas discret}). \end{cases}$$

On parlera respectivement de *modèles statistiques discrets* et de *modèles statistiques continus.* Ces modèles sont entièrement définis par la donnée de familles de densités correspondantes $\{f(x, \theta), \theta \in \Theta\}$. Par la suite, P_θ désignera la loi jointe de $(X_1, ..., X_n)$ quand les X_i sont i.i.d. de loi F_θ :

$$P_\theta(dx_1, ..., dx_n) = \prod_{i=1}^{n} [f(x_i, \theta)\mu(dx_i)]. \tag{5.1}$$

On notera $E_\theta(\cdot)$ l'espérance par rapport à P_θ. En utilisant ces notations, l'espérance d'une statistique $\theta_n(X_1, ..., X_n)$ s'écrit sous la forme

$$E_\theta(\theta_n) = \int \theta_n dP_\theta = \begin{cases} \int \theta_n(x_1, ..., x_n) \prod_{i=1}^n [f(x_i, \theta) dx_i] & \text{(pour un modèle continu)}, \\ \sum_{(x_1, ..., x_n)} \theta_n(x_1, ..., x_n) \prod_{i=1}^n f(x_i, \theta) & \text{(pour un modèle discret)}. \end{cases}$$

EXERCICE 5.1. Considérons *le problème de régression* : estimer le paramètre inconnu $\theta \in \mathbb{R}$ à partir des observations aléatoires $X_1, \dots, X_n$ où $X_i = (Y_i, Z_i)$,

$$Y_i = \theta Z_i + \xi_i, \quad i = 1, ..., n,$$

les ξ_i sont des variables aléatoires i.i.d. de loi $\mathcal{N}(0,1)$, les Z_i sont des variables aléatoires i.i.d. de densité $p(\cdot)$ sur $\mathbb{R}$ et les vecteurs aléatores $(\xi_1, \dots, \xi_n)$ et $(Z_1, \dots, Z_n)$ sont indépendants. On remarque que les X_i sont des vecteurs dans $\mathbb{R}^2$. Quel est le modèle statistique correspondant ?

5.1.3. Modèles statistiques dans le cadre non-i.i.d.*. L'hypothèse que les X_i sont i.i.d. n'est pas toujours vérifiée. On peut définir des modèles statistiques sans cette hypothèse. Dans ce cadre plus général, un modèle statistique est défini par $\{P_\theta, \theta \in \Theta\}$, où P_θ est *la loi jointe* de $X_1, ..., X_n$ quand la vraie valeur du paramètre est θ. Cette loi peut ne pas être de la forme (5.1). Une généralisation de l'hypothèse d'identifiabilité est donnée par la condition suivante :

$$P_\theta = P_{\theta'} \implies \theta = \theta'.$$

EXEMPLE 5.6. *Modèle d'autorégression.*
Soient les observations $X_1, ..., X_n$ telles que

$$X_i = \theta X_{i-1} + \xi_i, \quad i = 1, ..., n, \quad X_0 = 0,$$

où $\theta \in \mathbb{R}$ est le paramètre inconnu et les v.a. ξ_i sont indépendantes de même loi $\mathcal{N}(0,1)$. Bien évidemment, X_i dépend de X_{i-1} et les X_i ne sont pas i.i.d. La loi jointe P_θ de $X_1, ..., X_n$ est de la forme

$$P_\theta(dx_1, ..., dx_n) = \Big[\varphi(x_1)\prod_{i=2}^n \varphi(x_i - \theta x_{i-1})\Big] dx_1 \dots dx_n,$$

où φ désigne la densité de $\mathcal{N}(0,1)$. Le modèle statistique est $\{P_\theta, \ \theta \in \mathbb{R}\}$.

5.2. Comparaison d'estimateurs

Dans ce paragraphe, l'ensemble Θ des paramètres sera un sous-ensemble de $\mathbb{R}$. On s'intéressera aux critères de sélection de bons estimateurs. Intuitivement, un estimateur $\widehat{\theta}_n$ est bon, s'il est proche de la vraie valeur du paramètre. Mais $\widehat{\theta}_n$ est une variable aléatoire, donc la notion de proximité peut avoir plusieurs interprétations. On peut l'interpréter, par exemple, comme convergence en probabilité de $\widehat{\theta}_n$ vers la vraie valeur du paramètre.

Définition 5.2. *Un estimateur $\widehat{\theta}_n(X_1, ..., X_n)$ est dit convergent (ou* **consistant***) si*

$$\widehat{\theta}_n \xrightarrow{P} \theta \quad \textit{(converge vers } \theta \textit{ en probabilité } P_\theta\textit{) pour tout } \theta \in \Theta,$$

i.e.

$$\lim_{n\to\infty} P_\theta(|\widehat{\theta}_n - \theta| \geq \epsilon) = 0 \;\; \textit{pour tout } \epsilon > 0, \, \theta \in \Theta.$$

REMARQUES.

(1) Dans cette définition :
   - la convergence doit avoir lieu *pour tout* $\theta \in \Theta$, ce qui garantit qu'elle a lieu pour la vraie valeur inconnue θ^*,

– la consistance est une propriété liée au modèle statistique : un estimateur $\widehat{\theta}_n$ peut être consistant pour un modèle et non-consistant pour un autre.

(2) Si l'on a la convergence presque sûre : $\widehat{\theta}_n \to \theta \quad$ (p.s.) au lieu de la convergence en probabilité, on dit que l'estimateur $\widehat{\theta}_n$ est *fortement consistant*.

EXEMPLE 5.7. Soit $\widehat{\theta}_n = \bar{X}$. On sait que $\bar{X} \to E_\theta(X_1)$ (p.s.), d'après la loi forte des grands nombres, pourvu que l'espérance soit finie. Par conséquent, si le modèle statistique est tel que $\theta = E_\theta(X_1)$, alors $\bar{X}$ est un estimateur (fortement) consistant de θ. Par exemple, $\bar{X}$ est un estimateur fortement consistant de θ dans le modèle $\{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$.

La consistance est une propriété assez faible. Il existe un nombre infini d'estimateurs consistants, s'il en existe au moins un. En effet, si $\widehat{\theta}_n \xrightarrow{P} \theta$ pour tout $\theta \in \Theta$ et (a_n) est une suite déterministe telle que $a_n \to 1$, alors

$$a_n \widehat{\theta}_n \xrightarrow{P} \theta$$

quand $n \to \infty$ pour tout $\theta \in \Theta$. La suite a_n peut être choisie de façon assez arbitraire. Par exemple, $\widehat{\theta}'_n = (1+10^6[\ln(\max\{2, \ln n\})]^{-1})\widehat{\theta}_n$ est un estimateur consistant si $\widehat{\theta}_n$ est consistant. Or, $|\widehat{\theta}'_n| \gg |\widehat{\theta}_n|$ pour toute valeur raisonnable de n. La différence entre deux estimateurs consistants peut donc être énorme pour n fini, et si l'un de ces deux estimateurs est bon, l'autre peut être très mauvais. On voit donc que la consistance d'un estimateur n'est pas du tout synonyme de bonne performance. La même remarque s'applique à la consistance forte.

En conclusion, la notion de consistance n'est pas assez informative pour nous guider dans le choix d'estimateurs. Néanmoins, elle n'est pas complètement inutile, car elle permet de rétrécir l'ensemble des estimateurs que l'on doit étudier. En effet, les estimateurs *non-consistants* doivent être avec certitude exclus de toute considération.

5.2.1. Risque quadratique d'un estimateur. Afin de comparer les estimateurs dans un modèle statistique pour une taille d'échantillon n finie, on utilise souvent le risque quadratique.

On appelle *risque quadratique* (erreur moyenne quadratique) de l'estimateur $\widehat{\theta}_n$ au point $\theta \in \Theta$ la quantité

$$R_n(\theta, \widehat{\theta}_n) = E_\theta[(\widehat{\theta}_n - \theta)^2].$$

Le risque quadratique est bien défini pour tout estimateur $\widehat{\theta}_n$. Il peut, en particulier, prendre la valeur $R_n(\theta, \widehat{\theta}_n) = +\infty$. Le risque permet de mesurer la distance entre l'estimateur $\widehat{\theta}_n$ et la valeur θ.

La proposition suivante découle directement de l'inégalité de Tchebychev.

Proposition 5.1. *Si $R_n(\theta, \widehat{\theta}_n) \to 0$ pour tout $\theta \in \Theta$, alors $\widehat{\theta}_n$ est consistant.*

Plus la valeur du risque est petite, plus l'estimateur $\widehat{\theta}_n$ est performant. La question qui se pose alors est : existe-t-il un estimateur θ_n^* qui soit meilleur que tous les autres estimateurs au sens du risque quadratique ? Autrement dit, est-il possible de concevoir un estimateur θ_n^* tel que

$$R_n(\theta, \theta_n^*) \le R_n(\theta, \widehat{\theta}_n) \text{ pour tout } \theta \in \Theta \text{ et tout estimateur } \widehat{\theta}_n \,?$$

La réponse à cette question est négative. Pour fixer les idées, considérons l'exemple suivant.

Exemple 5.8. Soit le modèle normal $\{\mathcal{N}(\theta,1),\ \theta \in \mathbb{R}\}$. Introduisons les deux estimateurs suivants : $\theta_n^{(1)} = \bar{X}$ (estimateur consistant et tout à fait sympathique), et $\theta_n^{(2)} \equiv 0$ (estimateur absurde, car il prend toujours la même valeur, indépendamment de l'échantillon). Les risques quadratiques de $\theta_n^{(1)}$ et de $\theta_n^{(2)}$ valent

$$\begin{aligned} R_n(\theta,\theta_n^{(1)}) &= E_\theta\left((\theta_n^{(1)}-\theta)^2\right) = \mathrm{Var}(\bar{X}) = \frac{1}{n}, \\ R_n(\theta,\theta_n^{(2)}) &= E_\theta(\theta^2) = \theta^2. \end{aligned}$$

Si $|\theta| < 1/\sqrt{n}$, le risque de $\theta_n^{(2)}$ est inférieur à celui de $\theta_n^{(1)}$. Donc, pour un intervalle de valeurs de θ, l'estimateur absurde $\theta_n^{(2)}$ est meilleur que l'estimateur raisonnable $\theta_n^{(1)}$. Cet intervalle $\{\theta : |\theta| < 1/\sqrt{n}\}$ devient de plus en plus petit quand $n \to \infty$, et pour tous les autres θ le meilleur estimateur est $\theta_n^{(1)}$.

De façon générale, supposons que Θ contienne au moins 2 points distincts $\theta_1 \neq \theta_2$ et les mesures $\{P_\theta,\ \theta \in \Theta\}$ sont deux à deux mutuellement absolument continues. Alors, pour tout estimateur θ_n^*, l'un des risques $R_n(\theta_1,\theta_n^*)$ ou $R_n(\theta_2,\theta_n^*)$ est non-nul. En effet, si $R_n(\theta_1,\theta_n^*) = R_n(\theta_2,\theta_n^*) = 0$, alors $\theta_n^* = \theta_1$ (P_{θ_1}-p. s.) et $\theta_n^* = \theta_2$ (P_{θ_2}-p. s.), ce qui contredit au fait que P_{θ_1} est absolument continue par rapport à P_{θ_2}. Supposons, sans perte de généralité, que c'est le risque $R_n(\theta_1,\theta_n^*)$ qui est non-nul et posons $\widehat{\theta}_n \equiv \theta_1$. Alors,

$$R_n(\theta_1,\widehat{\theta}_n) = 0 < R_n(\theta_1,\theta_n^*).$$

Il n'existe donc pas d'estimateur θ_n^* tel que $R_n(\theta,\theta_n^*) \le R_n(\theta,\widehat{\theta}_n)$ pour tout $\theta \in \Theta$ et *tout estimateur* $\widehat{\theta}_n$. Une conséquence de cette observation est qu'il n'existe pas d'échelle de comparaison *absolue* des estimateurs basée sur les risques. Néanmoins, une comparaison *relative* est toujours possible : on peut utiliser le risque quadratique pour comparer les estimateurs deux à deux.

Définition 5.3. *Soient $\widehat{\theta}_n^{(1)}$ et $\widehat{\theta}_n^{(2)}$ deux estimateurs dans le modèle statistique $\{F_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$. Si*

$$R_n(\theta,\widehat{\theta}_n^{(1)}) \le R_n(\theta,\widehat{\theta}_n^{(2)}) \quad \text{pour tout } \theta \in \Theta,$$

et si, de plus, il existe $\theta' \in \Theta$ tel que

$$R_n(\theta',\widehat{\theta}_n^{(1)}) < R_n(\theta',\widehat{\theta}_n^{(2)}),$$

on dit que $\widehat{\theta}_n^{(1)}$ est plus efficace que $\widehat{\theta}_n^{(2)}$ (ou meilleur que $\widehat{\theta}_n^{(2)}$) et que $\widehat{\theta}_n^{(2)}$ est **inadmissible**.

Un estimateur $\widehat{\theta}_n$ est dit **admissible** *s'il n'existe pas d'estimateurs plus efficaces que $\widehat{\theta}_n$.*

5.2.2. Structure du risque : biais et variance.

Proposition 5.2. *Le risque $R_n(\theta,\widehat{\theta}_n)$ admet la décomposition*

$$R_n(\theta,\widehat{\theta}_n) = b_n^2(\theta,\widehat{\theta}_n) + \sigma_n^2(\theta,\widehat{\theta}_n),$$

où

$$\begin{aligned} b_n(\theta,\widehat{\theta}_n) &\overset{\text{déf}}{=} E_\theta(\widehat{\theta}_n) - \theta, \\ \sigma_n^2(\theta,\widehat{\theta}_n) &\overset{\text{déf}}{=} E_\theta\left((\widehat{\theta}_n - E_\theta(\widehat{\theta}_n))^2\right). \end{aligned}$$

Preuve. Il suffit d'utiliser la Proposition 1.1 avec $\xi = \widehat{\theta}_n$ et $c = \theta$.

Définition 5.4. *On appelle* $b_n(\theta, \widehat{\theta}_n)$ **biais** *de l'estimateur* $\widehat{\theta}_n$ *et* $\sigma_n^2(\theta, \widehat{\theta}_n)$ **variance** *de* $\widehat{\theta}_n$.

On note aussi $\sigma_n^2(\theta, \widehat{\theta}_n) \stackrel{\text{déf}}{=} \mathrm{Var}_\theta(\widehat{\theta}_n)$. Le carré du biais $b_n^2(\theta, \widehat{\theta}_n)$ représente la partie déterministe de l'erreur d'estimation, alors que $\sigma_n^2(\theta, \widehat{\theta}_n)$ mesure la contribution de sa partie stochastique.

Définition 5.5. *On dit qu'un estimateur* $\widehat{\theta}_n$ *est* **sans biais** *si* $E_\theta(\widehat{\theta}_n) = \theta$ *(i.e.* $b_n(\theta, \widehat{\theta}_n) = 0$*) pour tout* $\theta \in \Theta$ *et tout* n. *Dans le cas contraire, on dit que* $\widehat{\theta}_n$ *est biaisé.*

Une approche dépassée de la détermination d'un estimateur optimal consiste à chercher un estimateur sans biais de variance minimale. Elle est motivée par le fait que la minimisation du risque quadratique sur l'ensemble de tous les estimateurs sans biais se réduit à la minimisation de la variance. Bien que cette approche soit souvent évoquée dans la littérature statistique, elle ne sera pas considérée ici, car son domaine de validité est très limité et elle ne présente pas d'intérêt pour les applications. En effet,

- les estimateurs sans biais n'existent que dans des cas exceptionnels, pour quelques modèles statistiques remarquables,
- si le statisticien se trompe légèrement de modèle (ce qui se passe souvent dans la pratique), l'estimateur n'est plus sans biais et l'approche n'est plus valide,
- même pour les modèles statistiques admettant des estimateurs sans biais, on peut souvent proposer des estimateurs biaisés ayant le risque quadratique plus petit. On le voit dans l'exemple suivant.

EXEMPLE 5.9. Soit le modèle normal $\{\mathcal{N}(0, \sigma^2),\ \sigma^2 \in]0, \infty[\}$. Considérons deux estimateurs $\widehat{\theta}_n^{(1)}$ et $\widehat{\theta}_n^{(2)}$ du paramètre $\theta = \sigma^2$:

$$\begin{aligned}\widehat{\theta}_n^{(1)} &= s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2,\\ \widehat{\theta}_n^{(2)} &= s_*^2 = \frac{n}{n-1}s^2.\end{aligned}$$

D'après la Proposition 4.2,

$$E_\theta(\widehat{\theta}_n^{(1)}) = E_\theta(s^2) = \frac{n-1}{n}\sigma^2,$$

donc

$$b_n(\sigma^2, \widehat{\theta}_n^{(1)}) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

On en déduit que l'estimateur $\widehat{\theta}_n^{(1)}$ est biaisé. Par contre, $\widehat{\theta}_n^{(2)}$ est sans biais : $b_n(\sigma^2, \widehat{\theta}_n^{(2)}) = 0$ pour tout $\sigma^2 > 0$. Calculons les variances de ces estimateurs (cf. Exercice 4.8) :

$$\sigma_n^2(\sigma^2, \widehat{\theta}_n^{(1)}) = \frac{2(n-1)\sigma^4}{n^2},$$

et, comme pour tout $a \in \mathbb{R}$ et toute variable aléatoire X, $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$,

$$\sigma_n^2(\sigma^2, \widehat{\theta}_n^{(2)}) = \left(\frac{n}{n-1}\right)^2 \sigma_n^2(\sigma^2, \widehat{\theta}_n^{(1)}) = \frac{2\sigma^4}{n-1}\ .$$

Ceci nous permet de comparer les risques quadratiques :

$$
\begin{aligned}
R_n(\sigma^2, \widehat{\theta}_n^{(1)}) &= \left(\frac{\sigma^2}{n}\right)^2 + \frac{2(n-1)}{n^2}\,\sigma^4 = \frac{2n-1}{n^2}\,\sigma^4,\\
R_n(\sigma^2, \widehat{\theta}_n^{(2)}) &= \frac{2\sigma^4}{n-1}.
\end{aligned}
$$

Pour tout $\sigma^2 > 0$ on a $R_n(\sigma^2, \widehat{\theta}_n^{(2)}) > R_n(\sigma^2, \widehat{\theta}_n^{(1)})$, i.e. l'estimateur $\widehat{\theta}_n^{(1)} = s^2$ est plus efficace que $\widehat{\theta}_n^{(2)} = s_*^2$. L'estimateur sans biais $\widehat{\theta}_n^{(2)} = s_*^2$ est donc inadmissible. On voit qu'un estimateur biaisé peut être plus efficace qu'un estimateur sans biais.

Cet exemple révèle aussi un défaut du concept de l'admissibilité. En effet, la différence entre les estimateurs s^2 et s_*^2 est négligeable pour n assez grand et elle disparaît quand $n \to \infty$. L'estimateur s_*^2 est tout a fait honorable, mais il est inadmissible pour tout n. D'autre part, pour tout n fini, l'estimateur constant $\theta_n^{(2)}$ de l'Exemple 5.8 (un estimateur absurde) est admissible : on ne peut pas l'améliorer au point $\theta = 0$. Par conséquent, la propriété d'admissibilité ne peut pas servir pour sélectionner de bons estimateurs.

Conclusions.

(1) La propriété de consistance n'est pas suffisamment informative pour nous guider dans le choix d'estimateurs.

(2) On peut comparer les estimateurs deux à deux et chercher des estimateurs admissibles, mais ceci ne donne pas satisfaction, car il existe des estimateurs absurdes qui sont admissibles.

(3) La recherche des estimateurs sans biais de variance minimale n'est pas une solution non plus : un estimateur sans biais peut être moins efficace qu'un estimateur biaisé.

Autrement dit, les propriétés d'être consistant, sans biais ou admissible ne sont pas suffisantes pour caractériser un bon estimateur.

On note aussi que quelques-uns des problèmes présentés ci-dessus disparaissent si, au lieu de comparer les risques pour n fixé, on les considère *asymptotiquement* quand $n \to \infty$. Par exemple, le rapport des risques de s^2 et de s_*^2 tend vers 1 quand $n \to \infty$, donc s_*^2 n'est pas "inadmissible dans l'asymptotique".

De la même façon, presque tous les estimateurs raisonnables sont *asymptotiquement* sans biais (i.e. leur biais tend vers 0 pour tout θ, lorsque $n \to \infty$). Cette propriété est proche de la consistance (et moins forte que la convergence du risque vers 0), donc elle est plus ou moins indispensable, à la différence de la propriété, très contraignante, que le biais soit nul pour tout n.

Ces remarques nous amènent à privilégier la comparaison *asymptotique* d'estimateurs. On en reviendra plus loin dans ce chapitre. Avant de le faire, définissons quelques méthodes générales de construction d'estimateurs.

5.3. Méthode des moments

La méthode des moments a été proposée par Karl Pearson en 1894.

Soit $X_1, ..., X_n$ un échantillon i.i.d., $X_i \sim F_{\theta^*}$, et soit $\{F_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$, le modèle statistique sous-jacent. Dans ce paragraphe, nous supposons que les X_i sont à valeurs dans $\mathbb{R}$ et que les moments d'ordre $\leq k$ de X_1 existent pour tout $\theta \in \Theta$. Notons

$$\mu_r(\theta) = E_\theta(X_1^r) = \int x^r dF_\theta(x), \quad r = 1, ..., k. \tag{5.2}$$

Comme la forme de F_θ est connue, $\theta \mapsto \mu_r(\theta)$ sont des fonctions connues sur Θ pour $r = 1, ..., k$. Si les vraies valeurs $\mu_r^* = \mu_r(\theta^*)$, $r = 1, ..., k$, étaient disponibles, on pourrait résoudre le système de k équations

$$\mu_r(\theta) = \mu_r^*, \quad r = 1, ..., k$$

pour trouver le vecteur θ^* (en supposant qu'une solution existe). Or, ces valeurs sont inconnues et nous disposons seulement d'un échantillon $X_1, ..., X_n$, $X_i \sim F_{\theta^*}$. Le principe de substitution nous suggère d'utiliser

$$m_r = \frac{1}{n}\sum_{i=1}^n X_i^r$$

comme un estimateur de $\mu_r^* = \mu_r(\theta^*)$. Puisque $m_r \xrightarrow{P} \mu_r(\theta^*)$ quand $n \to \infty$, on peut espérer, qu'au moins pour n assez grand, une solution par rapport à θ du système d'équations

$$\mu_r(\theta) = m_r, \quad r = 1, ..., k, \tag{5.3}$$

soit proche de θ^*.

Définition 5.6. *On appelle* **estimateur par la méthode des moments (EMM)** *du paramètre θ dans le modèle $\{F_\theta,\, \theta \in \Theta\}$ avec $\Theta \subseteq \mathbb{R}^k$ toute statistique $\widehat{\theta}_n^{MM}$ à valeurs dans Θ étant une solution du système de k équations (5.3). Autrement dit,*

$$\boxed{\mu_r(\widehat{\theta}_n^{MM}) = m_r, \quad r = 1, ..., k.}$$

Il est clair que l'EMM peut ne pas exister et, s'il existe, il n'est pas toujours unique.

Remarque. Au lieu d'utiliser les k premiers moments pour construire l'estimateur de $\theta \in \mathbb{R}^k$, on peut utiliser k moments quelconques $\mu_{r_1}, ..., \mu_{r_k}$ (pourvu qu'ils soient finis).

Exemple 5.10. *EMM pour le modèle normal à moyenne et variance inconnues* $\{\mathcal{N}(\mu, \sigma^2),\, \mu \in \mathbb{R},\, \sigma^2 > 0\}$.
On montre facilement que $\bar{X}$ et s^2 sont les estimateurs de μ et σ^2 par la méthode des moments.

Exemple 5.11. *EMM pour le modèle exponentiel* $\{\mathcal{E}(\theta),\, \theta > 0\}$. La densité de F_θ est

$$f(x, \theta) = \theta^{-1} e^{-x/\theta} \mathbb{1}_{\{x>0\}}.$$

Alors,

$$\mu_1(\theta) = E_\theta(X_1) = \frac{1}{\theta}\int_0^\infty x e^{-\frac{x}{\theta}} dx = \theta,$$

et la solution $\widehat{\theta}_n^{(1)} = \bar{X}$ de l'équation

$$\theta = \bar{X}, \quad (\text{ou } \mu_1(\theta) = m_1 = \bar{X})$$

est un estimateur par la méthode des moments du paramètre θ. On remarque aussi que

$$\mu_2(\theta) = \frac{1}{\theta}\int_0^\infty x^2 e^{-\frac{x}{\theta}}dx = 2\theta^2.$$

Par conséquent, la solution $\widehat{\theta}_n^{(2)} = \sqrt{\frac{\bar{X}^2+s^2}{2}}$ de l'équation $2\theta^2 = m_2$, où

$$m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2 = \bar{X}^2 + s^2$$

est un autre estimateur par la méthode des moments de θ. De plus, comme $\bar{X} \to \mu_1(\theta) = \theta$ (p.s.) et $s^2 \to \sigma^2 = E_\theta[(X_1 - \mu_1)^2] = \theta^2$ (p.s.), d'après le Premier théorème de continuité (Proposition 1.10),

$$\sqrt{\frac{\bar{X}^2 + s^2}{2}} \to \sqrt{\frac{\mu_2(\theta)}{2}} = \theta \quad (\text{p.s.}).$$

Exercice 5.2. En utilisant les théorèmes de continuité, chercher la loi limite de $\sqrt{n}(\widehat{\theta}_n^{(2)} - \theta)$ sous les hypothèses de l'Exemple 5.11.

5.3.1. Méthode des moments généralisée. Une démarche similaire à la méthode des moments peut être effectuée avec des fonctions générales $\varphi_r(x)$ au lieu de x^r dans (5.2). On raisonne de la même façon que précédemment, sauf que l'on définit

$$\mu_r(\theta) = E_\theta(\varphi_r(X_1)), \quad r = 1, \ldots, k,$$

et on pose $\frac{1}{n}\sum_{i=1}^n \varphi_r(X_i)$ à la place de m_r dans (5.3). La méthode des moments généralisée consiste à définir l'estimateur de θ^* comme une solution $\theta = \widehat{\theta}_n^{MG}$ du système

$$\mu_r(\theta) = \frac{1}{n}\sum_{i=1}^n \varphi_r(X_i), \quad r = 1, ..., k. \tag{5.4}$$

Exemple 5.12. *Estimation dans le modèle de Cauchy.*
On considère le modèle $\{F_\theta, \theta \in \mathbb{R}\}$ où la densité de F_θ est donnée par

$$f(x,\theta) = \frac{1}{\pi(1 + (x-\theta)^2)}, \quad x \in \mathbb{R}.$$

Pour cette densité, les moments n'existent pas et la méthode des moments n'est pas utilisable. Mais il est possible d'appliquer la méthode des moments généralisée avec, par exemple, $k = 1$, $\varphi_1(x) = \operatorname{sgn}(x)$ où

$$\operatorname{sgn}(x) = \begin{cases} -1, & \text{si } x \leq 0, \\ 1, & \text{si } x > 0, \end{cases}$$

Grâce à la symétrie de la loi de Cauchy,

$$\mu_1(\theta) = \int f(x,\theta)\operatorname{sgn}(x)dx = 1 - 2F_0(-\theta) = 2F_0(\theta) - 1$$

où

$$F_0(t) = \frac{1}{\pi}\int_{-\infty}^t \frac{du}{1+u^2} = \frac{1}{\pi}\arctan t + \frac{1}{2},$$

et alors l'équation (5.4) s'écrit sous la forme

$$\frac{2}{\pi}\arctan\theta = \frac{1}{n}\sum_{i=1}^{n}\mathrm{sgn}(X_i).$$

L'estimateur par la méthode des moments généralisée est donc

$$\widehat{\theta}_n^{MG} = \tan\left(\frac{\pi}{2n}\sum_{i=1}^{n}\mathrm{sgn}(X_i)\right).$$

En utilisant la loi des grands nombres, le théorème central limite et les théorèmes de continuité (Propositions 1.10, 1.11), on prouve que c'est un estimateur consistant et asymptotiquement normal :

$$\sqrt{n}\,(\widehat{\theta}_n^{MG} - \theta^*) \xrightarrow{D} \mathcal{N}(0, v(\theta^*)), \tag{5.5}$$

lorsque $n \to \infty$, où $v(\theta^*) > 0$ est sa variance asymptotique.

Exercice 5.3. Explicitez la variance asymptotique $v(\theta^*)$ dans (5.5).

Remarque. La méthode des moments est un cas particulier de la méthode de substitution. En effet, supposons que l'on peut écrire $\theta = T(F_\theta)$ pour $\theta \in \mathbb{R}^k$, où

$$T(F) = \tilde{T}(\mu_1(F), ..., \mu_k(F))$$

avec $\mu_j(F) = \int x^j dF(x)$ et $\tilde{T} : \mathbb{R}^k \to \mathbb{R}^k$. L'estimateur de $\theta = T(F)$ par la méthode de substitution est donc

$$T_n = \tilde{T}(\mu_1(\widehat{F}_n), ..., \mu_k(\widehat{F}_n)) = \tilde{T}(m_1, ..., m_k) = \widehat{\theta}_n^{MM}.$$

5.4. Méthode du maximum de vraisemblance

Quelques cas particuliers de la méthode du maximum de vraisemblance ont été connus depuis le XVIIIème siècle, mais sa définition générale et l'argumentation de son rôle fondamental en Statistique sont dues à Fisher (1922).

Définition 5.7. *Considérons un modèle statistique $\{F_\theta, \theta \in \Theta\}$, où $\Theta \subseteq \mathbb{R}^k$, vérifiant l'Hypothèse (D). On appelle* **fonction de vraisemblance** *l'application $\theta \mapsto L(\mathcal{X}_n, \theta)$, où*

$$L(\mathcal{X}_n, \theta) = \prod_{i=1}^{n} f(X_i, \theta), \quad \theta \in \Theta,$$

avec $\mathcal{X}_n = (X_1, ..., X_n)$.

Exemple 5.13. Soient $X_1, ..., X_n$ des variables aléatoires discrètes, réalisations i.i.d. d'une v.a. X à valeurs dans un ensemble fini A. Alors, pour toute valeur fixée $a \in A$,

$$f(a, \theta) = P_\theta(X_1 = a).$$

Si l'on fixe l'échantillon $\mathcal{X}_n = (x_1, ..., x_n)$, on peut écrire

$$L(\mathcal{X}_n, \theta) = L\left((x_1, ..., x_n), \theta\right) = \prod_{i=1}^{n} P_\theta(X_1 = x_i) = \prod_{i=1}^{n} P_\theta(X_i = x_i). \tag{5.6}$$

On voit donc que $L(\mathcal{X}_n, \theta)$ est la probabilité d'obtenir la réalisation $(x_1, ..., x_n)$ quand la vraie valeur du paramètre est égale à θ. Supposons que, pour deux valeurs θ_1 et θ_2,

$$L\left((x_1, ..., x_n), \theta_1\right) > L\left((x_1, ..., x_n), \theta_2\right).$$

Alors (cf. (5.6)) la probabilité d'obtenir la réalisation $\mathcal{X}_n = (x_1, ..., x_n)$ est plus grande si la vraie valeur du paramètre était $\theta^* = \theta_1$ que si elle était $\theta^* = \theta_2$. Autrement dit, la valeur θ_1 est "plus vraisemblable" que θ_2 étant donnée la réalisation $(x_1, ..., x_n)$. En général, la valeur

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L\big((x_1, ..., x_n), \theta\big)$$

est "la plus vraisemblable". Ceci nous conduit à la définition suivante.

Définition 5.8. *Toute statistique* $\widehat{\theta}_n^{MV} = \widehat{\theta}_n^{MV}(\mathcal{X}_n)$ *telle que*

$$L(\mathcal{X}_n, \widehat{\theta}_n^{MV}) = \max_{\theta \in \Theta} L(\mathcal{X}_n, \theta)$$

est appelée **estimateur du maximum de vraisemblance (EMV)** *du paramètre* θ *dans le modèle statistique* $\{F_\theta,\, \theta \in \Theta\}$. *Autrement dit,*

$$\boxed{\widehat{\theta}_n^{MV} = \arg\max_{\theta \in \Theta} L(\mathcal{X}_n, \theta).}$$

Remarques.

(1) L'EMV peut ne pas exister (voir l'Exemple 5.18 ci-après).

(2) Si un EMV existe, il n'est pas toujours unique (voir les Exemples 5.15 – 5.17 ci-après).

(3) La fonction

$$l_n(\theta) = -\frac{1}{n} \ln L(\mathcal{X}_n, \theta) = -\frac{1}{n} \sum_{i=1}^{n} \ln f(X_i, \theta),$$

bien définie si $f(x, \theta) > 0$, est appelée **fonction de log-vraisemblance**. Alors

$$\widehat{\theta}_n^{MV} = \arg\min_{\theta \in \Theta} l_n(\theta).$$

(4) Si le maximum de $L(\mathcal{X}_n, \cdot)$ (respectivement, le minimum de $l_n(\cdot)$) n'est pas atteint sur la frontière de Θ et si l'application $\theta \mapsto \nabla_\theta L(\mathcal{X}_n, \theta)$ est continue, *condition nécessaire de maximum* est l'annulation du gradient :

$$\nabla_\theta L(\mathcal{X}_n, \theta)\big|_{\theta = \widehat{\theta}_n^{MV}} = 0,$$

ce qui représente un système de k équations, car $\theta \in \mathbb{R}^k$. De façon similaire, condition nécessaire de minimum de la fonction de log-vraisemblance est

$$\nabla l_n(\theta) = 0. \tag{5.7}$$

On appelle (5.7) **équation de vraisemblance** si $\theta \in \mathbb{R}$ et **système des équations de vraisemblance** si $\theta \in \mathbb{R}^k, k > 1$.

Définition 5.9. *On appelle* **racine de l'équation de vraisemblance (REV)** *dans le modèle* $\{F_\theta, \theta \in \Theta\}$ *avec* $\Theta \subseteq \mathbb{R}^k$ *toute statistique* $\widehat{\theta}_n^{RV}$ *à valeurs dans* Θ *étant une solution du système de* k *équations (5.7). Autrement dit,*

$$\boxed{\nabla l_n(\widehat{\theta}_n^{RV}) = 0.}$$

Notons qu'en résolvant le système (5.7) on obtient tous les maxima et tous les minima locaux de $l_n(\cdot)$, ainsi que ses points d'inflexion. Il est clair que la REV peut ne pas exister et, si elle existe, elle n'est pas toujours unique.

Pour que les Définitions 5.8 et 5.9 donnaient les mêmes estimateurs, i.e. pour que tous EMV soient des REV et vice versa, il faut que les conditions suivantes soient réunies.

(E1) $f(x, \theta) > 0$ pour $\theta \in \Theta$ et pour tout x.
(E2) La fonction $\theta \mapsto f(x, \theta)$ est différentiable sur l'ensemble Θ pour tout x.
(E3) La fonction l_n atteint son minimum global pour tous les θ tels que $\nabla l_n(\theta) = 0$.

La condition (E3) est très restrictive : on ne peut effectivement la vérifier que si la fonction l_n est convexe et son minimum global n'est pas atteint sur la frontière de Θ. L'équivalence des deux définitions n'a donc pas lieu que dans une situation très particulière. Il s'agit essentiellement de deux estimateurs différents, sauf cas exceptionnel.

Exemple 5.14. *EMV pour le modèle normal à moyenne et variance inconnues* $\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$. La densité de la loi $\mathcal{N}(\mu, \sigma^2)$ est

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \quad \theta = (\mu, \sigma),$$

donc les fonctions de vraisemblance et de log-vraisemblance correspondantes valent

$$L(\mathcal{X}_n, \theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\Big),$$

$$l_n(\theta) = \frac{1}{2}\ln(2\pi) + \ln\sigma + \frac{1}{2n\sigma^2}\sum_{i=1}^n (X_i - \mu)^2.$$

Les équations de vraisemblance ont la forme

$$\begin{cases} \frac{\partial l_n(\theta)}{\partial \mu} = 0, \\ \\ \frac{\partial l_n(\theta)}{\partial \sigma} = 0. \end{cases} \qquad \begin{cases} \frac{1}{\sigma^2 n}\sum_{i=1}^n (X_i - \mu) = 0, \\ \\ \frac{1}{\sigma} - \frac{1}{\sigma^3 n}\sum_{i=1}^n (X_i - \mu)^2 = 0. \end{cases}$$

Pour $n \geq 2$, ce système admet une seule solution (μ, σ) donnée par :

$$\mu = \bar{X},$$
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} = s.$$

La seule REV est donc $(\bar{X}, s)$. C'est aussi l'EMV au sens de la Définition 5.8. Pour $n = 1$ il n'existe pas d'EMV, mais ce cas n'est pas intéressant, car dans la pratique il s'agit toujours d'un échantillon de taille $n > 1$.

Conclusion : l'estimateur du maximum de vraisemblance de (μ, σ) est $\widehat{\theta}_n^{MV} = (\bar{X}, s) = \widehat{\theta}_n^{RV}$.

EXEMPLE 5.15. *EMV pour le modèle de Laplace.*
Considérons le modèle statistique $\{F_\theta,\, \theta \in \Theta\}$, où F_θ admet la densité de Laplace

$$f(x, \theta) = \frac{1}{2\sigma} \exp\Big(- \frac{|x-\theta|}{\sigma}\Big), \quad x \in \mathbb{R},$$

où $\sigma > 0$ est connu, le paramètre $\theta \in \mathbb{R}$ est inconnu et $\Theta = \mathbb{R}$. Les fonctions de vraisemblance et de log-vraisemblance pour ce modèle sont, respectivement,

$$\begin{aligned} L(\mathcal{X}_n, \theta) &= \left(\frac{1}{2\sigma}\right)^n \exp\left(-\frac{1}{\sigma}\sum_{i=1}^n |X_i - \theta|\right), \\ l_n(\theta) &= \ln(2\sigma) + \frac{1}{n\sigma}\sum_{i=1}^n |X_i - \theta|. \end{aligned}$$

Si l'on cherche à minimiser $l_n(\theta)$, cela revient à minimiser $\sum_{i=1}^n |X_i - \theta|$. Cette fonction est différentiable presque partout et sa dérivée admet la version suivante :

$$\frac{d}{d\theta}\Big(\sum_{i=1}^n |X_i - \theta|\Big) = -\sum_{i=1}^n \operatorname{sgn}(X_i - \theta) \stackrel{\text{déf}}{=} h(\theta).$$

Si n est pair, l'EMV n'est pas unique : en effet, tout point de l'intervalle $[X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)}]$ est un EMV et tout point de l'intervalle $]X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)}[$ est une REV. Si n est impair, l'EMV est unique : $\widehat{\theta}_n^{MV} = X_{(\frac{n+1}{2})}$, mais il n'existe pas de REV.

Rappelons-nous que la médiane empirique est définie par

$$M_n = \begin{cases} X_{(\frac{n+1}{2})} & \text{pour } n \text{ impair}, \\ \frac{X_{(n/2)} + X_{(n/2+1)}}{2} & \text{pour } n \text{ pair}. \end{cases}$$

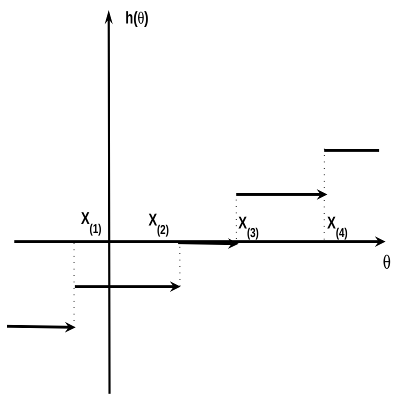

Figure 5.1. La fonction $h(\theta)$ pour $n = 4$ (le même type de graphique pour tout n pair). L'estimateur du maximum de vraisemblance n'est pas unique dans ce cas : tout point de $[X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)}]$ est un EMV.

Conclusion : dans le modèle de Laplace, la médiane empirique est *l'un des estimateurs* du maximum de vraisemblance, au sens de la Définition 5.8. C'est aussi une REV si n est pair. Par contre, si n est impair, il n'existe pas de REV.

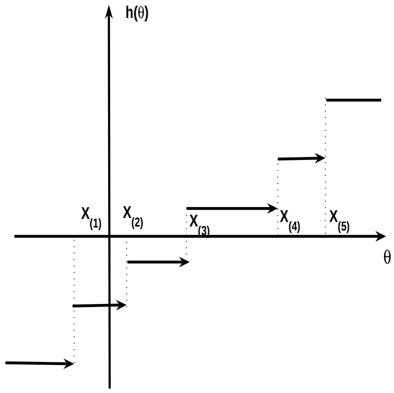

Figure 5.2. La fonction $h(\theta)$ pour $n = 5$ (le même type de graphique pour tout n impair). L'estimateur du maximum de vraisemblance est unique : $\widehat{\theta}_n^{MV} = X_{(\frac{n+1}{2})}$.

Exemple 5.16. *EMV pour le modèle uniforme* $\{U(0, \theta), \theta > 0\}$.
La densité de F_θ vaut

$$f(x, \theta) = \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x).$$

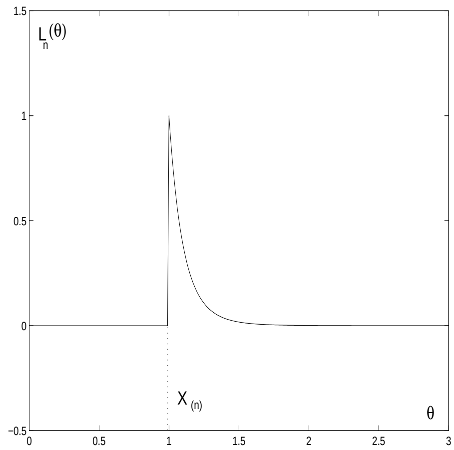

Figure 5.3. Modèle uniforme : $\widehat{\theta}_n^{MV} = X_{(n)}$.

La fonction de vraisemblance pour ce modèle s'écrit sous la forme

$$L(\mathcal{X}_n, \theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I\{0 \leq X_i \leq \theta\}.$$

Notons que

$$\frac{1}{\theta^n}\prod_{i=1}^n I\{0 \le X_i \le \theta\} = \begin{cases} 0 & \text{si } \theta < X_{(n)} = \max(X_1, ..., X_n), \\ \frac{1}{\theta^n} & \text{sinon.} \end{cases}$$

On voit que $\widehat{\theta}_n^{MV} = X_{(n)}$ est l'unique EMV. Il n'existe pas de REV, car la fonction de log-vraisemblance n'est pas dérivable.

EXEMPLE 5.17. *EMV pour le modèle de Cauchy :* $f(x,\theta) = \dfrac{1}{\pi(1+(x-\theta)^2)}$, $\theta \in \mathbb{R},\ x \in \mathbb{R}$.
La fonction de vraisemblance pour ce modèle est

$$L(\mathcal{X}_n, \theta) = \pi^{-n}\prod_{i=1}^n \frac{1}{1+(X_i-\theta)^2}\ .$$

La fonction de log-vraisemblance correspondante est de la forme

$$l_n(\theta) = \ln\pi + \frac{1}{n}\sum_{i=1}^n \ln(1+(X_i-\theta)^2)),$$

et l'équation de vraisemblance $l_n'(\theta) = 0$ équivaut à

$$\sum_{i=1}^n \frac{X_i-\theta}{1+(X_i-\theta)^2} = 0.$$

Généralement, cette équation admet plusieurs solutions que l'on ne peut pas trouver sous la forme explicite. Il y a, en général, plusieurs EMV et plusieurs REV.

EXEMPLE 5.18. *Modèle statistique pour lequel il n'existe pas d'EMV.*
Considérons le modèle $\{F_\theta, \theta \in \mathbb{R}\}$ tel que F_θ admet la densité $f_0(\cdot - \theta)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$ avec

$$f_0(x) = \frac{e^{-|x|/2}}{2\sqrt{2\pi|x|}} = f_{\chi_1^2}(|x|)/2,$$

où $f_{\chi_1^2}$ est la densité de la loi chi-deux à 1 degré de liberté. Alors, la fonction de vraisemblance

$$L(\mathcal{X}_n, \theta) = \prod_{i=1}^n f_0(X_i - \theta)$$

vérifie $\lim_{\theta\to X_i} L(\mathcal{X}_n, \theta) = +\infty$ pour tout $i = 1, \ldots, n$, ce qui implique que la borne supérieure de la fonction de vraisemblance n'est pas atteinte et alors il n'existe pas d'EMV.

La fonction de log-vraisemblance et ses fonctionnelles jouent un rôle important en Statistique. Pour élucider les propriétés de l'EMV, il est utile de comprendre comment se comporte la fonction de log-vraisemblance quand $n \to \infty$.

5.5. Comportement asymptotique de la fonction de log-vraisemblance

Soit $\theta^* \in \Theta$ la vraie valeur du paramètre, c'est-à-dire la valeur θ^* telle que $X_i \sim F_{\theta^*}$ pour $i = 1, \ldots, n$, et soit l'hypothèse suivante vérifiée :

$$\int |\ln f(x,\theta)| f(x,\theta^*) d\mu(x) < \infty, \quad \forall \theta \in \Theta. \tag{5.8}$$

Pour tout θ fixé, les variables aléatoires $Z_i = -\ln f(X_i, \theta)$ sont i.i.d. de moyenne

$$E(Z_i) = -E_{\theta^*}(\ln f(X_i, \theta)) = -\int f(x, \theta^*) \ln f(x, \theta) d\mu(x).$$

D'après la loi des grands nombres, on obtient la convergence en P_{θ^*}-probabilité :

$$l_n(\theta) \xrightarrow{P} J(\theta) \ \text{ quand } n \to \infty,$$

pour tout $\theta \in \Theta$, où

$$J(\theta) \stackrel{\text{déf}}{=} -\int f(x, \theta^*) \ln f(x, \theta) d\mu(x)$$

est appelée **fonction de contraste** associée à l'estimation du maximum de vraisemblance dans le modèle statistique $\{F_\theta, \theta \in \Theta\}$. On voit donc que, pour n assez grand, la fonction de log-vraisemblance $l_n(\cdot)$ est suffisamment bien approchée par la fonction de contraste $J(\cdot)$. Ceci nous incite à étudier plus en détail les propriétés de la fonction de contraste.

Lemme 5.1. *Supposons que la condition (5.8) est vérifiée. Alors*

$$J(\theta) \geq J(\theta^*), \quad \forall \ \theta \in \Theta.$$

Si, de plus, l'Hypothèse (Id) est vérifiée, alors

$$J(\theta) > J(\theta^*), \quad \forall \ \theta \neq \theta^*.$$

Preuve. Notons que, pour tout $t \geq -1$, on a : $\ln(1+t) - t \leq 0$, avec $\ln(1+t) - t = 0$ si et seulement si $t = 0$. On pose, pour abréger, $f(x) = f(x, \theta)$, $f^*(x) = f(x, \theta^*)$. Comme $f/f^* \geq 0$, on a :

$$\ln \frac{f}{f^*} - \Big(\frac{f}{f^*} - 1\Big) = \ln\Big(1 + \Big(\frac{f}{f^*} - 1\Big)\Big) - \Big(\frac{f}{f^*} - 1\Big) \leq 0, \tag{5.9}$$

où l'égalité est atteinte si et seulement si $f/f^* = 1$ (dans ces calculs on suppose que $f^* > 0$). Évidemment,

$$\int \Big(\frac{f}{f^*} - 1\Big) f^* d\mu = 0. \tag{5.10}$$

En utilisant (5.9) et (5.10), on obtient

$$J(\theta) - J(\theta^*) = -\int f^* \ln \frac{f}{f^*} d\mu = -\int f^* \Big[\ln \frac{f}{f^*} - \Big(\frac{f}{f^*} - 1\Big)\Big] d\mu \geq 0.$$

Le membre dans les crochets est négatif, donc la dernière inégalité se transforme en égalité si et seulement si ce membre est nul pour μ-presque tous x tels que $f^*(x) > 0$, i.e. l'égalité dans (5.9) est atteinte pour tout x tel que $f^*(x) > 0$. Mais ce n'est possible que si

$$f(x)/f^*(x) = 1 \quad \text{pour } \mu\text{-presque tous } x \text{ tels que } \ f^*(x) > 0. \tag{5.11}$$

Notons que (5.11) est vrai si et seulement si

$$f(x) = f^*(x) \quad \text{ pour } \mu\text{-presque tous } x. \tag{5.12}$$

En effet, comme f et f^* sont deux densités de probabilité, la condition (5.11) implique :

$$\int f I(f^* = 0) d\mu = 1 - \int f I(f^* > 0) d\mu = 1 - \int f^* I(f^* > 0) d\mu = 0,$$

d'où on obtient que $f = 0$ pour μ-presque tous x tels que $f^*(x) = 0$. Donc, la condition (5.11) implique (5.12). L'implication réciproque est évidente.

On a alors,

$$J(\theta) = J(\theta^*)$$

si et seulement si

$$f(x,\theta) = f(x,\theta^*) \quad \text{pour } \mu\text{-presque tous } x. \tag{5.13}$$

D'après l'Hypothèse (Id), (5.13) implique que $\theta = \theta^*$. On voit donc que (5.13) n'est pas possible pour $\theta \neq \theta^*$. Par conséquent, $J(\theta) = J(\theta^*)$ n'est pas possible pour $\theta \neq \theta^*$. ∎

Le résultat du Lemme 5.1 peut être considéré comme une justification de l'estimateur du maximum de vraisemblance : puisque $l_n(\theta)$ converge en probabilité vers $J(\theta)$ pour tout $\theta \in \Theta$, on peut espérer que l'EMV

$$\widehat{\theta}_n^{MV} = \operatorname*{argmin}_{\theta\in\Theta} l_n(\theta)$$

ne soit pas loin de

$$\theta^* = \operatorname*{argmin}_{\theta\in\Theta} J(\theta).$$

Dans la suite, on donnera un sens mathématique précis à cette idée heuristique.

5.6. Consistance de l'estimateur du maximum de vraisemblance

Quelquefois on peut montrer la consistance de l'EMV directement, sans utiliser des théorèmes généraux. Par exemple, pour le modèle normal $\{\mathcal{N}(\theta,1),\ \theta \in \mathbb{R}\}$ ou pour celui de Bernoulli $\{\mathcal{B}e(\theta),\ 0 < \theta < 1\}$, l'EMV est la moyenne empirique $\bar{X}$, alors que la vraie valeur du paramètre est la moyenne théorique, $\theta^* = E_{\theta^*}(X)$. La consistance de l'EMV découle donc directement de la loi des grands nombres. La liste des exemples de ce genre peut être encore prolongée, mais ce ne sont que des cas exceptionnels. Un résultat plus général est donné dans le théorème suivant.

Théorème 5.1. *Supposons que Θ est un ouvert de $\mathbb{R}$ et*

(i) la densité $f(x,\theta)$ est continue comme fonction de θ, pour tout x,

(ii) l'Hypothèse (Id) d'identifiabilité est satisfaite,

(iii) la condition (5.8) est vérifiée pour tout $\theta^* \in \Theta$,

(iv) pour tout n, l'EMV $\widehat{\theta}_n^{MV}$ existe et l'ensemble des minima locaux de la fonction de log-vraisemblance $l_n(\theta)$ est un intervalle fermé et borné à l'intérieur de Θ.

Alors $\widehat{\theta}_n^{MV}$ est un estimateur consistant.

Preuve. Fixons $\theta^* \in \Theta$. Il faut démontrer que

$$\lim_{n\to\infty} P_{\theta^*}(|\widehat{\theta}_n^{MV} - \theta^*| \geq \varepsilon) = 0, \quad \forall \varepsilon > 0. \tag{5.14}$$

Il suffit ici de considérer seulement les valeurs $\varepsilon > 0$ telles que $\theta^* + \varepsilon \in \Theta$, $\theta^* - \varepsilon \in \Theta$. Fixons alors $\varepsilon > 0$ qui vérifie ces hypothèses. Notons

$$J_n(\theta) = l_n(\theta) - l_n(\theta^*).$$

Évidemment,

$$J_n(\theta^*) = 0.$$

D'après la loi des grands nombres, pour tout $\theta \in \Theta$,

$$J_n(\theta) \stackrel{P}{\to} J(\theta) - J(\theta^*) \text{ quand } n \to \infty.$$

(On note ici $\stackrel{P}{\to}$ la convergence en P_{θ^*} -probabilité.) En particulier, pour tout $\varepsilon > 0$,

$$J_n(\theta^* - \varepsilon) \stackrel{P}{\to} E_{\theta^*}(J_n(\theta^* - \varepsilon)) = J(\theta^* - \varepsilon) - J(\theta^*) = \delta_- > 0, \tag{5.15}$$

$$J_n(\theta^* + \varepsilon) \stackrel{P}{\to} E_{\theta^*}(J_n(\theta^* + \varepsilon)) = J(\theta^* + \varepsilon) - J(\theta^*) = \delta_+ > 0, \tag{5.16}$$

lorsque $n \to \infty$. La positivité des constantes δ_- et δ_+ découle de l'hypothèse (ii) du théorème et du Lemme 5.1.

Soit l'événement aléatoire $A \stackrel{\text{déf}}{=} \{|\widehat{\theta}_n^{MV} - \theta^*| < \varepsilon\}$. Montrons que

$$A \supseteq \{J_n(\theta^* - \varepsilon) > 0\} \cap \{J_n(\theta^* + \varepsilon) > 0\}. \tag{5.17}$$

En effet, la continuité de $J_n(\theta)$ (d'après l'hypothèse (i)) et les inégalités

$$J_n(\theta^* - \varepsilon) > 0,$$
$$J_n(\theta^*) = 0,$$
$$J_n(\theta^* + \varepsilon) > 0$$

impliquent que $J_n(\theta)$ atteint au moins un minimum local sur l'intervalle $]\theta^* - \varepsilon, \theta^* + \varepsilon[$. La condition (iv) du théorème implique que tous les minima locaux de $J_n(\theta)$ sont ses minima globaux et ils forment un intervalle fermé et borné. Par conséquent, cet intervalle est contenu dans $]\theta^* - \varepsilon, \theta^* + \varepsilon[$. Donc, $\widehat{\theta}_n^{MV} \in]\theta^* - \varepsilon, \theta^* + \varepsilon[$, et (5.17) est démontré.

D'après (5.17), (5.15) et (5.16), on a

$$P_{\theta^*}(|\widehat{\theta}_n^{MV} - \theta^*| \geq \varepsilon) = P_{\theta^*}(\bar{A})$$

$$\leq P_{\theta^*}(J_n(\theta^* - \varepsilon) \leq 0) + P_{\theta^*}(J_n(\theta^* + \varepsilon) \leq 0)$$

$$= P_{\theta^*}\left(J_n(\theta^* - \varepsilon) - E_{\theta^*}(J_n(\theta^* - \varepsilon) \leq -\delta_-\right)$$

$$+P_{\theta^*}\left(J_n(\theta^* + \varepsilon) - E_{\theta^*}(J_n(\theta^* + \varepsilon) \leq -\delta_+\right) \to 0 \text{ quand } n \to \infty.$$

∎

Remarques.

(1) Les hypothèses (i) - (iii) sont peu restrictives. Évidemment, l'hypothèse (i) est satisfaite dans plusieurs situations. L'hypothèse (ii) est nécessaire pour l'unicité de θ^*, la vraie valeur du paramètre. La condition (iii) est nécessaire pour l'existence d'une limite finie $J(\theta)$ de la fonction de log-vraisemblance.

(2) Seule l'hypothèse (iv) du Théorème 5.1 est restrictive. On peut montrer qu'elle n'est pas nécessaire. Wald (1949) a démontré que, sous les hypothèses un peu différentes de (i) - (iii), mais aussi très générales, et sans aucune hypothèse sur le comportement de l'ensemble des minima locaux de l_n, l'EMV est consistant. Un résultat similaire est vrai pour toute suite de racines de l'équation de vraisemblance (voir, par exemple, A.A.Borovkov *Statistique mathématique*, 1984).

EXEMPLE 5.19. *Consistance de l'EMV pour le modèle de Laplace.*
Soit $\{F_\theta, \theta \in \mathbb{R}\}$ le modèle de Laplace défini dans l'Exemple 5.15. Il est évident que les conditions (i) et (ii) du Théorème 5.1 sont vérifiées. La condition (iv) aussi est vraie, vu l'Exemple 5.15. Pour vérifier l'hypothèse (iii), il suffit de noter que, pour le modèle de Laplace, $E_{\theta^*}(|\ln f(X,\theta)|) < \infty \iff E_{\theta^*}(|X-\theta)|) < \infty$ et que la dernière inégalité est évidemment satisfaite. Le Théorème 5.1 permet donc de déduire que l'EMV dans le modèle de Laplace est consistant.

EXEMPLE 5.20. *Consistance de l'EMV pour le modèle de Weibull.*
Soit le modèle statistique $\{F_\theta, \theta > 1\}$, où F_θ admet la densité suivante par rapport à la mesure de Lebesgue :

$$f(x,\theta) = \theta x^{\theta-1}\exp(-x^\theta)I\{x>0\}.$$

La fonction de vraisemblance correspondante est donnée par

$$L(\mathcal{X}_n,\theta) = \prod_{i=1}^n f(X_i,\theta) = \theta^n \exp(-\sum_{i=1}^n X_i^\theta)\prod_{i=1}^n X_i^{\theta-1}I\{X_i>0\}.$$

La fonction de log-vraisemblance et ses dérivées sont données par (on ne regarde que l'ensemble où tous les X_i sont strictement positifs) :

$$\begin{aligned}
l_n(\theta) &= -\ln\theta - (\theta-1)\frac{1}{n}\sum_{i=1}^n \ln X_i + \frac{1}{n}\sum_{i=1}^n X_i^\theta,\\
l_n'(\theta) &= -\frac{1}{\theta} - \frac{1}{n}\sum_{i=1}^n \ln X_i + \frac{1}{n}\sum_{i=1}^n X_i^\theta \ln X_i,\\
l_n''(\theta) &= \frac{1}{\theta^2} + \frac{1}{n}\sum_{i=1}^n X_i^\theta(\ln X_i)^2 > 0, \ \text{ pour tout } \ \theta > 0.
\end{aligned}$$

L'existence d'une racine de l'équation de vraisemblance est claire vu les convergences $l_n'(\theta) \to +\infty$ quand $\theta \to +\infty$, et $l_n'(\theta) \to -\infty$ quand $\theta \to +0$. En outre, cette racine est unique car $l_n''(\theta) > 0$ pour tout θ et n. La condition (iv) du Théorème 5.1 est donc satisfaite. On note que, pour le modèle de Weibull, $\widehat{\theta}_n^{MV} = \widehat{\theta}_n^{RV}$.

Finalement, la condition (iii) est satisfaite, car

$$E_{\theta^*}(|\ln f(X,\theta)|) \le C_1 E_{\theta^*}\left(|X| + |X|^\theta\right) + C_2$$

avec des constantes C_1, C_2 positives (dépendantes de θ) et la dernière expression est finie pour tout θ, ce qui implique (5.8). On peut donc appliquer le Théorème 5.1 pour en déduire que l'EMV dans le modèle de Weibull est consistant.

EXEMPLE 5.21. *Un EMV non-consistant.*
Considérons le modèle statistique où X est une v.a. discrète à valeurs dans $\{0, 1\}$, de fonction

de probabilité $P_\theta(X = x) = f(x, \theta)$ donnée par

$$f(x,\theta) = \begin{cases} \theta^x(1-\theta)^{(1-x)}, & \text{si } \theta \text{ est rationnel,} \\ \theta^{(1-x)}(1-\theta)^x, & \text{si } \theta \text{ est irrationnel,} \end{cases}$$

où $x \in \{0, 1\}$ et $0 < \theta < 1$. L'EMV de θ basé sur l'échantillon $X_1, \ldots, X_n$ est $\widehat{\theta}_n^{MV} = \bar{X}$. D'après la loi des grands nombres,

$$\bar{X} \xrightarrow{P} \begin{cases} \theta, & \text{si } \theta \text{ est rationnel,} \\ 1-\theta, & \text{si } \theta \text{ est irrationnel,} \end{cases}$$

(il s'agit ici de la convergence en P_θ-probabilité) et donc l'EMV $\widehat{\theta}_n^{MV} = \bar{X}$ n'est pas consistant. Ce contre-exemple est assez artificiel, mais il montre que, pour avoir la consistance de l'EMV, il faut que l'application $\theta \mapsto f(x, \theta)$ ne soit pas trop oscillante sur des petits ensembles (cf. hypothèse (i) du Théorème 5.1).

5.7. Modèles statistiques réguliers

Pour le reste de ce chapitre, on supposera que $\Theta \subseteq \mathbb{R}$. Le cas multi-dimensionnel peut être traité de la même manière.

Notre but local est d'introduire des hypothèses sur le modèle statistique qui permettent la différentiation de $J(\cdot)$ deux fois sous le signe intégrale. On verra plus loin que, sous ces hypothèses, l'EMV jouit de bonnes propriétés, telles que la normalité asymptotique.

Notons pour abréger

$$l(x,\theta) \stackrel{\text{déf}}{=} \ln f(x,\theta), \quad l'(x,\theta) \stackrel{\text{déf}}{=} \frac{\partial}{\partial\theta} \ln f(x,\theta), \quad l''(x,\theta) \stackrel{\text{déf}}{=} \frac{\partial^2}{\partial\theta^2} \ln f(x,\theta),$$

$$f'(x,\theta) \stackrel{\text{déf}}{=} \frac{\partial}{\partial\theta} f(x,\theta),$$

où les notations $l'(x, \theta)$, $l''(x, \theta)$ et $f'(x, \theta)$ sont valables seulement si les dérivées en question existent.

Définition 5.10. *Soit la fonction $\theta \mapsto l(x, \theta)$ différentiable pour presque tout x par rapport à la mesure μ. La fonction $I(\cdot)$ sur Θ à valeurs positives définie par*

$$\boxed{I(\theta) = \int (l'(x,\theta))^2 f(x,\theta) d\mu(x) = E_\theta\left([l'(X,\theta)]^2\right)}$$

est appelée **information de Fisher** *associée à une seule observation dans le modèle statistique $\{F_\theta,\ \theta \in \Theta\}$.*

Sous les hypothèses de la Définition 5.10, on peut aussi écrire

$$I(\theta) = \int_{\{x: f(x,\theta)>0\}} \frac{(f'(x,\theta))^2}{f(x,\theta)} d\mu(x).$$

La Définition 5.10 n'exclut pas la possibilité $I(\theta) = +\infty$. Cependant, par la suite, on s'intéressera seulement aux modèles statistiques ayant une information de Fisher finie. Introduisons les hypothèses suivantes.

Hypothèses de régularité.

(H1) L'ensemble des paramètres Θ est un intervalle ouvert de $\mathbb{R}$ et

$$f(x,\theta) > 0 \iff f(x,\theta') > 0, \quad \forall \theta, \theta' \in \Theta.$$

(H2) Pour μ presque tout x, les fonctions $\theta \mapsto f(x,\theta)$ et $\theta \mapsto l(x,\theta)$ sont deux fois continûment dérivables sur Θ.

(H3) Pour tout $\theta^* \in \Theta$ il existe un intervalle ouvert $U_{\theta^*} \subseteq \Theta$ contenant θ^* et une fonction borélienne $\Lambda(x)$ tels que $|l''(x,\theta)| \leq \Lambda(x)$, $|l'(x,\theta)| \leq \Lambda(x)$, $|l'(x,\theta)|^2 \leq \Lambda(x)$, pour tout $\theta \in U_{\theta^*}$ et μ presque tout x, et

$$\int \Lambda(x) \sup_{\theta \in U_{\theta^*}} f(x,\theta) d\mu(x) < \infty.$$

(H4) L'information de Fisher $I(\theta)$ vérifie

$$I(\theta) > 0, \quad \forall \theta \in \Theta.$$

Remarque. Comme le voisinage U_{θ^*} peut être choisi aussi petit que l'on veut, l'hypothèse **(H3)** n'est pas beaucoup plus forte que la condition que les intégrales

$$\int |l''(x,\theta^*)| f(x,\theta^*) d\mu(x), \quad \int (l'(x,\theta^*))^2 f(x,\theta^*) d\mu(x) = I(\theta^*)$$

soient finies pour tout $\theta^* \in \Theta$.

Définition 5.11. *Un modèle statistique $\{F_\theta, \theta \in \Theta\}$ est appelé* **modèle régulier** *s'il vérifie les Hypothèses (D), (H1) – (H4).*

Exercice 5.4. Montrer que les modèles normal, de Bernoulli et de Cauchy définis dans les Exemples 5.1, 5.4, 5.17 respectivement sont réguliers.

Par la suite, l'appellation *hypothèses de régularité* sera réservée aux hypothèses **(H1) - (H4)**.

Notons que l'hypothèse **(H3)** implique que l'information de Fisher $I(\theta)$ est finie pour tout $\theta \in \Theta$ et

$$\int \sup_{\theta \in U_{\theta^*}} |l'(x,\theta)| f(x,\theta^*) d\mu(x) < \infty, \tag{5.18}$$

$$\int \sup_{\theta \in U_{\theta^*}} \Big(|l'(x,\theta)| f(x,\theta) \Big) d\mu(x) < \infty. \tag{5.19}$$

5.7.1. Régularité de la fonction de contraste. Calculons les dérivées d'ordre 1 et 2 de la fonction de contraste $J(\cdot)$. On aura besoin du lemme suivant que l'on utilisera par la suite pour justifier la dérivation sous le signe intégrale.

Lemme 5.2. *Soit $(\mathcal{X}, \mathcal{A})$ un espace mesurable et $g : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ une fonction mesurable de $(\mathcal{X} \times \mathbb{R}, \mathcal{A} \otimes \mathcal{B}(\mathbb{R}))$ vers $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, telle que $g(x,\theta)$ est continûment différentiable en θ pour*

ν presque tout tout $x \in \mathcal{X}$. Soit, de plus,

$$\int \sup_{\theta \in U} \Big| \frac{\partial}{\partial \theta} g(x,\theta) \Big| d\nu(x) < \infty, \tag{5.20}$$

$$\int |g(x,\theta)| d\nu(x) < \infty, \quad \forall\, \theta \in U, \tag{5.21}$$

où U est un intervalle ouvert de $\mathbb{R}$ et ν est une mesure σ-finie sur $\mathcal{A}$. Alors la fonction

$$G(\theta) = \int g(x,\theta)\, d\nu(x)$$

est continûment différentiable sur U et on peut dériver sous le signe intégrale :

$$\frac{d}{d\theta} \int g(x,\theta)\, d\nu(x) = \int \frac{\partial}{\partial \theta}\, g(x,\theta)\, d\nu(x), \qquad \theta \in U.$$

Preuve. En utilisant (5.20) et le théorème de Fubini, on obtient, pour tout $\theta_1 \in U, \theta_2 \in U$,

$$\begin{aligned}\int_{\theta_1}^{\theta_2} \Big[\int \frac{\partial}{\partial \theta} g(x,\theta) d\nu(x) \Big] d\theta &= \int \Big[\int_{\theta_1}^{\theta_2} \frac{\partial}{\partial \theta} g(x,\theta) d\theta \Big] d\nu(x) \\ &= \int \big(g(x,\theta_2) - g(x,\theta_1) \big) d\nu(x) = G(\theta_2) - G(\theta_1).\end{aligned}$$

On a donc

$$G(\theta_2) - G(\theta_1) = \int_{\theta_1}^{\theta_2} G'(\theta) d\theta.$$

avec

$$G'(\theta) = \int \frac{\partial}{\partial \theta} g(x,\theta) d\nu(x).$$

Il ne reste qu'à vérifier que la fonction G' est continue. Il vient, pour tout $\theta \in U, \theta' \in U$,

$$|G'(\theta) - G'(\theta')| \le \int \Big| \frac{\partial}{\partial \theta} g(x,\theta) - \frac{\partial}{\partial \theta'} g(x,\theta') \Big| d\nu(x)$$

L'expression sous l'intégrale dans le membre de droite converge vers 0 quand $\theta' \to \theta$ pour ν presque tout x et elle est uniformément bornée par la fonction $2 \sup_{\theta \in U} \big| \frac{\partial}{\partial \theta} g(x,\theta) \big|$ qui est intégrable vu (5.20). L'application du théorème de convergence dominée permet donc de conclure. ∎

Le Lemme 5.2 entraîne le résultat suivant.

Lemme 5.3. *Soit $\{F_\theta, \theta \in \Theta\}$ un modèle régulier. Supposons que la condition (5.8) soit vérifiée. Alors, la fonction de contraste J est deux fois continûment différentiable sur un voisinage de θ^* et*

$$J'(\theta^*) = 0, \tag{5.22}$$

$$J''(\theta^*) = -E_{\theta^*}(l''(X,\theta^*)) = -\int l''(x,\theta^*) f(x,\theta^*) d\mu(x). \tag{5.23}$$

Preuve. Montrons d'abord que la fonction J est différentiable. Utilisons le Lemme 5.2 avec

$$\nu = \mu, \quad g(x,\theta) = f(x,\theta^*) \ln f(x,\theta).$$

Les conditions (5.20) et (5.21) sont vérifiées vu (5.18) et (5.8) respectivement. La fonction J est donc continûment différentiable sur un voisinage U_{θ^*} de θ^* et, pour tout $\theta \in U_{\theta^*}$,

$$J'(\theta) = -\int l'(x,\theta) f(x,\theta^*) d\mu(x). \tag{5.24}$$

D'après le Lemme 5.1, $J(\theta)$ atteint son minimum pour $\theta = \theta^*$. Par conséquent, $J'(\theta^*) = 0$.

Montrons maintenant qu'il est possible de dériver sous le signe intégrale dans (5.24) pour $\theta \in U_{\theta^*}$, ce qui entraîne (5.23). Il suffit d'appliquer le Lemme 5.2 à la fonction $G(\theta) = J'(\theta)$. Posons, dans le Lemme 5.2,

$$g(x,\theta) = -l'(x,\theta) f(x,\theta^*).$$

Il est facile de voir que dans ce cas les hypothèses (5.20) et (5.21) du Lemme 5.2 découlent de **(H3)** et de (5.18) respectivement. ∎

5.7.2. Propriétés de l'information de Fisher. Donnons d'abord deux exemples de modèles non-réguliers pour lesquels l'information de Fisher n'est pas définie ou n'est pas strictement positive.

Exemple 5.22. L'information de Fisher n'est pas définie pour le modèle uniforme

$$\{U[0,\theta],\ \theta > 0\},$$

car $l(x,\theta)$ n'est pas différentiable. Ce modèle n'est pas régulier.

Exemple 5.23. Soit le modèle statistique de densité

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(x-\theta^2)^2}{2}\Big), \quad \theta \in \Theta,$$

où l'ensemble des paramètres $\Theta = \mathbb{R}$. Alors $l'(x,\theta) = -2\theta(x-\theta^2)$ et l'information de Fisher vaut $I(\theta) = 4\theta^2$. En particulier, $I(0) = 0$, donc le modèle n'est pas régulier. Rappelons que ce modèle n'est pas indentifiable (cf. Exemple 5.5). Par contre, le modèle devient régulier si l'on prend $\Theta = \{\theta : \theta > 0\}$.

Lemme 5.4. *Soit $\{F_\theta, \theta \in \Theta\}$ un modèle régulier. Alors*

$$\boxed{I(\theta) = -\int l''(x,\theta) f(x,\theta) d\mu(x),}$$

pour tout $\theta \in \Theta$.

Preuve. Comme $f(x,\theta)$ est une densité de probabilité par rapport à la mesure μ,

$$\int f(x,\theta) d\mu(x) = 1. \tag{5.25}$$

Pour obtenir le lemme, on va dériver cette égalité 2 fois sous le signe intégrale. La démonstration utilise la double application du Lemme 5.2, avec $\nu = \mu$. D'abord, on pose dans le Lemme 5.2

$$g(x,\theta) = f(x,\theta).$$

Dans ce cas, la condition (5.20) du Lemme 5.2 est vérifiée vu (5.19). La condition (5.21) du Lemme 5.2 est évidente. On peut donc dériver (5.25) sous le signe intégrale et, pour tout $\theta \in \Theta$,

$$\int f'(x,\theta)d\mu(x) = 0. \tag{5.26}$$

Ceci équivaut à

$$\int l'(x,\theta)f(x,\theta)d\mu(x) = 0, \quad \forall\, \theta \in \Theta. \tag{5.27}$$

Utilisons le Lemme 5.2 encore une fois pour justifier la différentiation sous le signe intégrale dans (5.27). Posons, dans le Lemme 5.2,

$$g(x,\theta) = f'(x,\theta) = l'(x,\theta)f(x,\theta).$$

Alors,

$$\frac{\partial}{\partial\theta} g(x,\theta) = l''(x,\theta)f(x,\theta) + [l'(x,\theta)]^2 f(x,\theta). \tag{5.28}$$

Vu (5.28) et l'hypothèse (**H3**), on obtient la condition (5.20) du Lemme 5.2. La condition (5.21) du Lemme 5.2 découle de (5.19). On peut donc dériver sous le signe intégrale dans (5.27) et on obtient, en utilisant (5.28),

$$0 = \int \frac{\partial}{\partial\theta} g(x,\theta)d\mu(x) = \int l''(x,\theta)f(x,\theta)d\mu(x) + I(\theta),$$

ce qui démontre le lemme. ∎

EXEMPLE 5.24. *Information de Fisher pour le modèle normal* $\{\mathcal{N}(\theta,\sigma^2),\ \theta \in \mathbb{R}\}$ *avec* $\sigma^2 > 0$ *connu* :

$$I(\theta) \equiv \frac{1}{\sigma^2}, \ \forall\, \theta \in \mathbb{R}.$$

L'information de Fisher ne dépend donc pas de θ et la fonction de contraste correspondante vaut

$$J(\theta) = J(\theta^*) + \frac{1}{2\sigma^2}(\theta - \theta^*)^2.$$

5.7.3. Interprétation graphique de l'information de Fisher. Les Lemmes 5.3 et 5.4 impliquent que, sous les hypothèses de régularité,

$$\boxed{I(\theta^*) = J''(\theta^*).}$$

D'après le Lemme 5.1, la fonction J atteint son minimum au point θ^*. Si la valeur $I(\theta^*)$ est petite, le rayon de courbure du graphique $\theta \mapsto J(\theta)$ sur un voisinage du minimum de J est grand, donc la fonction J est "plate" sur ce voisinage. Si l'information $I(\theta^*)$ est grande, la situation est différente : J est "pointue" sur un voisinage de θ^*. Mais la fonction de contraste J est la limite en probabilité de la fonction de log-vraisemblance l_n. Autrement dit, l_n, avec une probabilité proche de 1 pour n assez grand, oscille dans un petit "tube" autour de J. Si l'information $I(\theta^*)$ est petite (J est "plate"), ses oscillations peuvent amener l'estimateur du maximum de vraisemblance $\widehat{\theta}_n^{MV}$ loin de θ^*. Par contre, si $I(\theta^*)$ est grande (J est "pointue"), l'estimateur du maximum de vraisemblance $\widehat{\theta}_n^{MV}$ est très proche de θ^*, la vraie valeur du paramètre.

Ceci nous conduit à la conclusion suivante : plus grande est l'information de Fisher, plus proche est l'EMV de la vraie valeur du paramètre. Le sens précis mathématique de cette remarque sera élucidé dans le résultat sur la normalité asymptotique de l'EMV (voir le Théorème 5.2 plus loin).

Il est utile de noter que l'information de Fisher est une caractéristique *locale* en ce sens qu'elle décrit le comportement de la fonction de contraste J seulement sur un voisinage de son minimum θ^*. Dans le même esprit, la notion de modèle régulier est locale. Le fait qu'un modèle statistique soit régulier ne signifie pas que $J(\theta) > J(\theta^*)$ globalement pour tout $\theta \in \Theta$, ni même qu'il existe des estimateurs consistants de θ : un modèle régulier peut ne pas être identifiable. En effet, le modèle de l'Exemple 5.23 avec l'ensemble des paramètres $\Theta = \{\theta : |\theta| < 1, \theta \neq 0\}$ est régulier, mais il ne vérifie évidemment pas l'Hypothèse (Id).

REMARQUE. Les quantités $J(\theta)-J(\theta^*)$ et $J(\theta^*)$ apparaissent souvent dans l'usage statistique. On appelle

$$K(\theta,\theta^*) = J(\theta) - J(\theta^*) = \int \ln \frac{f(x,\theta^*)}{f(x,\theta)} f(x,\theta^*)d\mu(x)$$

information de Kullback (ou *divergence de Kullback*) de $f(x,\theta)$ par rapport à $f(x,\theta^*)$. C'est une mesure de divergence (dissymétrique) entre $f(x,\theta)$ et $f(x,\theta^*)$. Son interprétation est similaire à celle de l'information de Fisher : elle permet de juger si la fonction $J(\theta)$ est "plate" où "pointue". Mais, à la différence de l'information de Fisher, la valeur $K(\theta,\theta^*)$ est une caractéristique globale (non restreinte à un voisinage de θ^*) et elle est bien définie pour certains modèles non-réguliers.

La valeur

$$J(\theta^*) = -\int f(x,\theta^*) \ln f(x,\theta^*)d\mu(x)$$

est appelée *entropie de Shannon* associée à la densité $f(x,\theta^*)$. Elle est parfois utilisée comme une mesure de dispersion, par exemple, quand la variance

$$\int x^2 f(x,\theta^*)d\mu(x) - \left(\int x f(x,\theta^*)d\mu(x)\right)^2$$

n'est pas finie, comme pour la loi de Cauchy. L'entropie de Shannon joue un rôle important dans la Théorie de l'Information.

5.7.4. Information de Fisher du produit des densités. L'information de Fisher *associée à l'échantillon i.i.d.* $X_1, ..., X_n$ dans un modèle statistique régulier $\{F_\theta,\ \theta \in \Theta\}$ est définie par :

$$I_n(\theta) = E_\theta\Big(\Big[\frac{L'_\theta(\mathcal{X}_n,\theta)}{L(\mathcal{X}_n,\theta)}\Big]^2\Big).$$

(on remplace $f(X,\theta)$ par la densité produit $L(\mathcal{X}_n,\theta)$ dans la Définition 5.10). Il est facile de voir que

$$I_n(\theta) = E_\theta([l'_n(\theta)]^2) = nI(\theta), \tag{5.29}$$

où $I(\theta)$ est l'information de Fisher associée à une seule observation.

5.8. Normalité asymptotique de l'estimateur du maximum de vraisemblance

Théorème 5.2. *Soit $\{F_\theta, \theta \in \Theta\}$ un modèle régulier et soit $(\widehat{\theta}_n^{RV})_{n\geq 1}$ une suite consistante de racines de l'équation de vraisemblance. Alors, pour tout $\theta^* \in \Theta$,*

$$\boxed{\sqrt{n}(\widehat{\theta}_n^{RV} - \theta^*) \overset{D}{\to} \mathcal{N}(0, 1/I(\theta^*)).} \tag{5.30}$$

Remarque. Dans ce théorème, comme partout dans ce chapitre, l'estimateur $\widehat{\theta}_n^{RV} = \widehat{\theta}_n^{RV}(X_1, \ldots, X_n)$ est basé sur l'échantillon $(X_1, \ldots, X_n)$, où $X_i \sim F_{\theta^*}$, i.e. θ^* est la vraie valeur du paramètre. Le résultat (5.30) se traduit donc par la convergence

$$P_{\theta^*}\big(\sqrt{nI(\theta^*)}\,(\widehat{\theta}_n^{RV} - \theta^*) \leq t\big) \to \Phi(t) \quad \text{quand} \quad n \to \infty, \quad \forall\, t \in \mathbb{R}, \quad \forall\, \theta^* \in \Theta,$$

où Φ est la fonction de répartition de $\mathcal{N}(0,1)$.

Preuve.
Étape 1. Comme $\widehat{\theta}_n^{RV}$ est une racine de l'équation de vraisemblance, $l_n'(\widehat{\theta}_n^{RV}) = 0$. Il s'ensuit que

$$-l_n'(\theta^*) = l_n'(\widehat{\theta}_n^{RV}) - l_n'(\theta^*) = \int_0^1 l_n''\big(t\widehat{\theta}_n^{RV} + (1-t)\theta^*\big)dt\; (\widehat{\theta}_n^{RV} - \theta^*),$$

donc

$$-\sqrt{n}\, l_n'(\theta^*) = A_n \sqrt{n}\,(\widehat{\theta}_n^{RV} - \theta^*) \tag{5.31}$$

où

$$A_n \overset{\text{déf}}{=} \int_0^1 l_n''\big(t\widehat{\theta}_n^{RV} + (1-t)\theta^*\big)\, dt.$$

Étape 2. On montre que

$$-\sqrt{n}\, l_n'(\theta^*) \overset{D}{\to} \mathcal{N}(0, I(\theta^*)) \quad \text{quand} \quad n \to \infty. \tag{5.32}$$

En effet,

$$-\sqrt{n}\, l_n'(\theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n l'(X_i, \theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i,$$

où les variables aléatoires i.i.d. $Z_i = l'(X_i, \theta^*)$ sont de moyenne

$$E_{\theta^*}(Z_i) = \int f'(x, \theta^*) d\mu(x) = 0$$

et de variance

$$E_{\theta^*}(Z_i^2) = \int (l'(x, \theta^*))^2 f(x, \theta^*) d\mu(x) = I(\theta^*).$$

D'après le Théorème central limite, on obtient donc (5.32).

Étape 3. On montre que

$$A_n \overset{P}{\to} I(\theta^*) \quad \text{quand} \quad n \to \infty, \tag{5.33}$$

en P_{θ^*}-probabilité. (La démonstration de cette étape sera donnée ci-après.)

Étape 4. On conclut : (5.30) découle de (5.31) – (5.33) et du résultat 1^o de l'Exercice 1.5. ∎

Preuve de l'Étape 3. On fixe $\varepsilon > 0$. Alors

$$P_{\theta^*}\big(|A_n - I(\theta^*)| > \varepsilon\big) \le P_{\theta^*}\big(|A_n - l''_n(\theta^*)| > \varepsilon/2\big) + P_{\theta^*}\big(|l''_n(\theta^*) - I(\theta^*)| > \varepsilon/2\big). \quad (5.34)$$

Or,

$$l''_n(\theta^*) = -\frac{1}{n}\sum_{i=1}^n l''(X_i,\theta^*) \xrightarrow{P} I(\theta^*) \quad \text{quand} \quad n \to \infty,$$

d'après la loi des grands nombres. En effet, les v.a. $l''(X_i,\theta^*)$ sont i.i.d. et $E_{\theta^*}(l''(X_i,\theta^*)) = I(\theta^*)$. Il s'ensuit que la dernière probabilité dans (5.34) tend vers 0 lorsque $n \to \infty$ et

$$\limsup_{n\to\infty} P_{\theta^*}(|A_n - I(\theta^*)| > \varepsilon) \le \limsup_{n\to\infty} P_{\theta^*}(|A_n - l''_n(\theta^*)| > \varepsilon/2). \quad (5.35)$$

On fixe maintenant $\delta > 0$. Évidemment,

$$\begin{aligned} P_{\theta^*}(|A_n - l''_n(\theta^*)| > \varepsilon/2) \;&\le\; P_{\theta^*}(|A_n - l''_n(\theta^*)| > \varepsilon/2,\ |\widehat{\theta}_n^{RV} - \theta^*| \le \delta) \\ &\quad + \; P_{\theta^*}(|\widehat{\theta}_n^{RV} - \theta^*| > \delta). \end{aligned} \quad (5.36)$$

Comme $\widehat{\theta}_n^{RV}$ est un estimateur consistant, la dernière probabilité dans (5.36) tend vers 0 lorsque $n \to \infty$. De (5.35) et (5.36) on obtient

$$\limsup_{n\to\infty} P_{\theta^*}(|A_n - I(\theta^*)| > \varepsilon) \le \limsup_{n\to\infty} P_{\theta^*}(|A_n - l''_n(\theta^*)| > \varepsilon/2,\ |\widehat{\theta}_n^{RV} - \theta^*| \le \delta). \quad (5.37)$$

Si $|\widehat{\theta}_n^{RV} - \theta^*| \le \delta$, on a :

$$\begin{aligned} |A_n - l''_n(\theta^*)| \;&=\; \Big| \int_0^1 \Big(l''_n(t\widehat{\theta}_n^{RV} + (1-t)\theta^*) - l''_n(\theta^*)\Big)dt\Big| \\ &\le \sup_{\theta:|\theta-\theta^*|\le\delta} |l''_n(\theta) - l''_n(\theta^*)| \le \frac{1}{n}\sum_{i=1}^n \Delta(X_i,\delta) \end{aligned}$$

où

$$\Delta(x,\delta) \overset{\text{déf}}{=} \sup_{\theta:|\theta-\theta^*|\le\delta} |l''(x,\theta) - l''(x,\theta^*)|.$$

Par conséquent,

$$\begin{aligned} P_{\theta^*}(|A_n - l''_n(\theta^*)| > \varepsilon/2,\ |\widehat{\theta}_n^{RV} - \theta^*| \le \delta) \;&\le\; P_{\theta^*}\Big(\frac{1}{n}\sum_{i=1}^n \Delta(X_i,\delta) > \varepsilon/2\Big) \\ &\le (2/\varepsilon)E_{\theta^*}(\Delta(X_1,\delta)), \end{aligned} \quad (5.38)$$

où on a utilisé l'inégalité de Markov. Or, pour tout x, $\Delta(x,\delta)$ décroît de façon monotone vers 0 quand $\delta \to 0$, et

$$0 \le \Delta(x,\delta) \le 2 \sup_{\theta:|\theta-\theta^*|\le\delta} |l''(x,\theta)|.$$

D'après l'Hypothèse **(H3)**, il existe $\delta_0 > 0$ assez petit, tel que

$$E_{\theta^*}\bigg(\sup_{\theta:|\theta-\theta^*|\le\delta_0} |l''(X_1,\theta)|\bigg) < \infty.$$

On peut donc utiliser le Théorème de convergence dominée, ce qui implique

$$\lim_{\delta\to 0} E_{\theta^*}\big[\Delta(X_1,\delta)\big] = 0. \quad (5.39)$$

On conclut en notant que (5.33) découle de (5.37) – (5.39). ∎

Corollaire 5.1. *Soit $\{F_\theta, \theta \in \Theta\}$ un modèle régulier et soit $(\widehat{\theta}_n^{MV})_{n\geq 1}$ une suite consistante des estimateurs du maximum de vraisemblance. Alors, pour tout $\theta^* \in \Theta$,*

$$\boxed{\sqrt{n}(\widehat{\theta}_n^{MV} - \theta^*) \xrightarrow{D} \mathcal{N}(0, 1/I(\theta^*)).} \tag{5.40}$$

Preuve. Elle est immédiate d'après le Théorème 5.2, compte tenu du fait que, sous les hypothèses de régularité, tout EMV est une racine de l'équation de vraisemblance. ∎

5.9. Comparaison asymptotique d'estimateurs

On peut proposer la démarche asymptotique suivante pour définir un estimateur optimal. Tout d'abord, on considère uniquement les estimateurs asymptotiquement normaux, i.e. $\widehat{\theta}_n$ tels que

$$\sqrt{n}\,(\widehat{\theta}_n - \theta) \xrightarrow{D} \mathcal{N}(0, v(\theta)) \quad \text{quand} \quad n \to \infty, \quad \forall\, \theta \in \Theta \subseteq \mathbb{R}, \tag{5.41}$$

où $v(\theta) > 0$ est la variance asymptotique de $\widehat{\theta}_n$ (précisons que la convergence dans (5.41) est en loi P_θ). On désigne par $\widehat{\Theta}_{AN}$ la classe de tous les estimateurs asymptotiquement normaux, i.e. vérifiant (5.41), tels que la variance $v(\cdot)$ est une fonction continue et strictement positive sur Θ. Cette classe est assez large. Sous des hypothèses appropriées, elle contient, par exemple, les estimateurs par la méthode des moments et ceux du maximum de vraisemblance. En particulier, le Théorème 5.2 montre que pour l'EMV la variance asymptotique est $v(\theta) = 1/I(\theta)$ et les Exercices 5.2, 5.3 montrent (5.41) pour quelques estimateurs par la méthode des moments.

Plus petite est la variance asymptotique $v(\theta^*)$, plus proche est $\widehat{\theta}_n$ de la vraie valeur du paramètre θ^* quand n est assez grand. Ceci nous conduit à la méthode de comparaison asymptotique d'estimateurs suivante.

Définition 5.12. *Soient $\widehat{\theta}_n^{(1)}$ et $\widehat{\theta}_n^{(2)}$ deux estimateurs de classe $\widehat{\Theta}_{AN}$ dans le modèle statistique $\{F_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$. Notons $v_1(\cdot)$ et $v_2(\cdot)$ les variances asymptotiques de $\widehat{\theta}_n^{(1)}$ et $\widehat{\theta}_n^{(2)}$ respectivement. On dit que l'estimateur $\widehat{\theta}_n^{(1)}$ est* **asymptotiquement plus efficace** *que $\widehat{\theta}_n^{(2)}$ si*

$$v_1(\theta) \leq v_2(\theta) \quad \textit{pour tout } \theta \in \Theta$$

et si, de plus, il existe $\theta' \in \Theta$ tel que

$$v_1(\theta') < v_2(\theta').$$

Un estimateur $\widehat{\theta}_n$ est appelé **asymptotiquement efficace** *s'il n'existe pas d'estimateurs asymptotiquement plus efficaces que $\widehat{\theta}_n$.*

Cette définition ressemble à la Définition 5.3 de l'estimateur admissible, mais il s'agit ici de la propriété asymptotique. De plus, on considère ici la classe restreinte d'estimateurs asymptotiquement normaux $\widehat{\Theta}_{AN}$. Ceci permet, en particulier, d'éliminer les estimateurs absurdes, comme $\widehat{\theta}_n \equiv c$, où c est une constante. On peut montrer que, sous les hypothèses de régularité, un estimateur asymptotiquement efficace a toujours la variance asymptotique

$v(\theta) = 1/I(\theta)$. Par conséquent, l'EMV pour les modèles statistiques vérifiant les hypothèses du Théorème 5.2 est asymptotiquement efficace. Bien sûr, il s'agit ici de l'optimalité de l'EMV par rapport à une classe restreinte d'estimateurs $\widehat{\Theta}_{AN}$. Une approche plus fine de l'optimalité due à Le Cam permet de montrer que, sous des hypothèses assez générales, l'EMV est aussi asymptotiquement optimal parmi *tous* les estimateurs.

5.10. Exercices

EXERCICE 5.5. Soit $(X_1, \dots, X_n)$ un échantillon i.i.d. de loi de Bernoulli de paramètre θ $(0 < \theta < 1)$.
1°. Estimer θ par la méthode des moments et du maximum de vraisemblance.
2°. Montrer que l'estimateur du maximum de vraisemblance de θ est sans biais.
3°. On cherche à estimer la variance $\theta(1-\theta)$; $\bar{X}$ étant la moyenne empirique, on propose l'estimateur $T = \bar{X}(1-\bar{X})$. Vérifier qu'il n'est pas sans biais et donner un estimateur sans biais de $\theta(1-\theta)$.

EXERCICE 5.6. Supposons que l'on observe n variables aléatoires i.i.d. $X_1, \dots, X_n$. Calculer l'estimateur du maximum de vraisemblance lorsque la loi des variables X_i est :
1°. Une loi de Poisson $\mathcal{P}(\theta)$ de paramètre $\theta > 0$.
2°. Une loi exponentielle $\mathcal{E}(\theta)$ de paramètre $\theta > 0$.
3°. Une loi admettant la densité $\exp\{-(x-\theta)\}\mathbb{1}_{\{x \geq \theta\}}$, $\theta \in \mathbb{R}$.
On vérifiera dans chaque cas que l'on obtient bien le maximum global de la fonction de vraisemblance. Dans quels cas EMV = REV ?

EXERCICE 5.7. Soient n variables aléatoires i.i.d. $X_1, \dots, X_n$, de densité uniforme $U[\theta, \theta+1]$, $\theta \in \mathbb{R}$. Montrer que tout point de l'intervalle $[X_{(n)} - 1, X_{(1)}]$ est un estimateur du maximum de vraisemblance de θ.

EXERCICE 5.8. Soient n variables aléatoires i.i.d. $X_1, \dots, X_n$, de loi normale $\mathcal{N}(\theta, 2\theta)$, $\theta > 0$. Calculer l'estimateur du maximum de vraisemblance de θ et montrer qu'il est consistant.

EXERCICE 5.9. Soient n variables aléatoires i.i.d. $X_1, \dots, X_n$, de densité de Pareto

$$\frac{\theta}{x^{\theta+1}} \mathbb{1}_{\{x \geq 1\}},$$

où $\theta > 0$ est un paramètre inconnu que l'on souhaite estimer.
1°. On suppose d'abord que l'ensemble des paramètres est $\Theta = \{\theta > 1\}$. Estimer θ par la méthode des moments.
2°. On suppose maintenant que l'ensemble des paramètres est $\Theta = \{\theta > 0\}$. Montrer que la méthode des moments n'est pas applicable. Estimer θ par la méthode des moments généralisée et par celle du maximum de vraisemblance.
3°. Étudier la loi limite de l'estimateur du maximum de vraisemblance et calculer l'information de Fisher $I(\theta)$. Comparer $(nI(\theta))^{-1}$ avec la variance asymptotique de l'estimateur du maximum de vraisemblance.
4°. Le modèle statistique en question est-il régulier ?

EXERCICE 5.10. Une chaîne de montage produit des objets, dont on veut estimer la durée moyenne de fabrication. On suppose que les temps de fabrication T_i sont indépendants et de loi exponentielle de paramètre θ. Le n-ième objet est donc fabriqué à la date $T_1 + \ldots + T_n$, et on observe le nombre d'objets N_t fabriqués à la date t.
1°. Montrer que $P(N_t \leq n) = P(T_1 + \ldots + T_{n+1} \geq t)$.
2° Quelle est la loi de $T_1 + \ldots + T_n$? On pourra utiliser les propriétés des lois Gamma. Montrer, par intégration par parties, que N_t suit une loi de Poisson dont on donnera le paramètre.
3°. Construire un estimateur de θ par la méthode des moments et par celle du maximum de vraisemblance. Étudier le comportement des risques quadratiques respectifs lorsque t tend vers l'infini.

EXERCICE 5.11. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d., dont la densité est

$$\theta^2 x \exp(-\theta x) \mathbb{1}_{\{x \geq 0\}},$$

où $\theta > 0$.
1°. Le modèle statistique est-il régulier ?
2°. Chercher l'estimateur $\widehat{\theta}_n^{MM}$ de θ par la méthode des moments.
3°. Chercher l'estimateur du maximum de vraisemblance $\widehat{\theta}_n^{MV}$ et donner son risque quadratique. Proposer un estimateur sans biais et comparer le à $\widehat{\theta}_n^{MV}$.
4°. Quelle est la loi limite de $\sqrt{n}(\widehat{\theta}_n^{MV} - \theta)$?

EXERCICE 5.12. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. pouvant prendre les valeurs 0, 1, 2 avec probabilités $p/2$, $p/2$, $1-p$. Dans cet exercice, on note n_0, n_1 et n_2 le nombre de 0, de 1 et de 2 dans l'échantillon.
1°. Dans quel intervalle de $\mathbb{R}$ varie p ? Proposer un estimateur $\hat{p}_1$ de p par la méthode des moments et calculer son risque quadratique. Calculer $\hat{p}_1$ en fonction de n_0, n_1, n_2 et n.
2°. Calculer en fonction de n_0, n_1, n_2 l'estimateur $\hat{p}_2$ obtenu par la méthode du maximum de vraisemblance. En remarquant que $n_k = \sum_{i=1}^n \mathbb{1}_{\{X_i = k\}}$, $k = 0, 1, 2$, calculer son risque quadratique et comparer le à celui de $\hat{p}_1$.

EXERCICE 5.13. *Modèle de mélange.* Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de densité f qui est un mélange de deux densités gaussiennes $\mathcal{N}(0,1)$ et $\mathcal{N}(0,4)$:

$$f(x) = p \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) + (1-p) \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{x^2}{8}\right),$$

où $0 < p < 1$ est un paramètre inconnu que l'on souhaite estimer. Quelle difficulté rencontre-t-on pour traiter l'estimateur du maximum de vraisemblance ? Expliciter $\hat{p}_n$, l'estimateur de p obtenu à l'aide de la méthode des moments (on utilisera le 2ème moment). Montrer que l'estimateur $\hat{p}_n$ est consistant et déterminer la loi limite de $\sqrt{n}(\hat{p}_n - p)$ lorsque $n \to \infty$.

EXERCICE 5.14. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de densité

$$f(x, \theta) = (1+\theta) \mathbb{1}_{\{0 \leq x \leq 1/2\}} + (1-\theta) \mathbb{1}_{\{1/2 < x \leq 1\}},$$

où $\theta \in]-1, 1[$ est un paramètre inconnu que l'on souhaite estimer. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{MV}$ de θ. Est-il consistant ? sans bias ? Déterminer la loi limite de $\sqrt{n}(\hat{\theta}_n^{MV} - \theta)$ quand $n \to \infty$.

EXERCICE 5.15. Soit la densité de probabilité $f(x) = 2(1-x)\mathbb{1}_{\{0\le x\le 1\}}$. On dispose d'un échantillon i.i.d. $(X_1, \dots, X_n)$ de densité $f(x-\theta)$, où $\theta \in \mathbb{R}$ est un paramètre inconnu.
1°. Le modèle statistique est-il régulier ?
2°. Chercher $\widehat{\theta}_n^{MM}$, l'estimateur de θ par la méthode des moments (en utilisant seulement le premier moment).
3°. L'estimateur $\widehat{\theta}_n^{MM}$ est-il consistant ? sans biais ? Quelle est la loi asymptotique de $\widehat{\theta}_n^{MM}$?
4°. Montrer que l'estimateur du maximum de vraisemblance de θ est $\widehat{\theta}_n^{MV} = X_{(1)}$.

EXERCICE 5.16. Le but de cet exercice est de montrer qu'il existe des modèles statistiques *non-réguliers* tels que :

– l'EMV pour ces modèles converge à la vitesse plus rapide que pour les modèles réguliers,
– l'EMV pour ces modèles est inadmissible.

Considérons le modèle uniforme $\{U[0,\theta],\ \theta > 0\}$. Soient $X_1, \dots, X_n$ des variables aléatoires i.i.d. de loi $U[0,\theta]$.

1°. Calculer le biais, la variance et le risque quadratique de l'estimateur du maximum de vraisemblance $\widehat{\theta}_n^{MV}$. Noter que le risque quadratique converge vers 0 à la vitesse $1/n^2$.
2°. Parmi les estimateurs de la forme $c\widehat{\theta}_n^{MV}$, $c \in \mathbb{R}$, déterminer tel que son risque quadratique soit minimal. On note cet estimateur $\tilde{\theta}_n$. Déduire que l'estimateur du maximum de vraisemblance $\widehat{\theta}_n^{MV}$ est inadmissible.
3°. Chercher les lois limites de $n(\widehat{\theta}_n^{MV} - \theta)$ et de $n(\tilde{\theta}_n - \theta)$.

6

Tests d'hypothèses et régions de confiance

6.1. Le problème de test d'hypothèse

Dans ce chapitre, nous considérerons des hypothèses sur la valeur du paramètre inconnu d'une loi de probabilité et nous proposerons des méthodes permettant de décider si celles-ci sont ou non correctes. Commençons par l'exemple suivant.

Exemple 6.1. *Détection de missile.* Une des premières applications de la théorie des tests statistiques était liée au problème militaire de détection de la présence d'un missile à l'aide de radar. L'écho de radar est "grand" si un missile est présent et il est "petit" dans le cas contraire. Supposons que l'on observe une suite de valeurs $X_1, \ldots, X_n$ de l'echo de radar aux instants $1, \ldots, n$. On peut supposer que les X_i sont des variables aléatoires (effet de bruit de propagation d'ondes, erreurs de mesures, etc.), qu'elles sont i.i.d. et, plus particulièrement, se placer dans le cadre d'un modèle paramétrique, de même qu'au Chapitre 5. Notamment, supposons que l'on connaît la famille paramétrique de fonctions de répartition $\mathcal{F} = \{F_\theta,\ \theta \in \Theta\}$ telle que la fonction de répartition F des X_i appartient à $\mathcal{F}$, i.e. $F = F_{\theta^*}$ pour une valeur inconnue $\theta^* \in \Theta$ (θ^* est la vraie valeur du paramètre). Supposons aussi que l'ensemble Θ peut être décomposé en deux sous-ensembles disjoints Θ_0 et Θ_1 :

$$\Theta = \Theta_0 \cup \Theta_1, \qquad \Theta_0 \cap \Theta_1 = \varnothing,$$

de sorte que

$$\theta^* \in \Theta_0 \ \text{ si et seulement si un missile est présent,}$$

alors que

$$\theta^* \in \Theta_1 \ \text{ si et seulement si il n'y a pas de missile.}$$

Notre objectif est le suivant : à partir des observations $X_1, \ldots, X_n$, décider si le missile est présent (i.e. $\theta^* \in \Theta_0$) ou non (i.e. $\theta^* \in \Theta_1$).

On appelle Θ_0 **hypothèse** ou *hypothèse nulle* et Θ_1 **alternative** ou *hypothèse alternative* et on utilise l'écriture symbolique suivante pour définir le problème de test :

$$H_0 : \ \theta \in \Theta_0, \qquad H_1 : \ \theta \in \Theta_1. \tag{6.1}$$

REMARQUES.

(1) Par la suite, nous écrirons "l'hypothèse H_0" et "l'alternative H_1" aussi bien que "l'hypothèse Θ_0" et "l'alternative Θ_1".

(2) L'écriture "$\theta \in \Theta_0$" et "$\theta \in \Theta_1$" dans la définition symbolique (6.1) de problème de test est standard dans la littérature statistique, mais elle n'est pas très précise. Il serait plus précis d'écrire que $\theta^* \in \Theta_0$ ou $\theta^* \in \Theta_1$, où θ^* est la vraie valeur du paramètre.

L'hypothèse (ou l'alternative) est dite **simple** si Θ_0 (ou Θ_1) ne contient qu'un seul élément. Dans le cas contraire, on l'appelle hypothèse (ou alternative) **composite** (ou *multiple*).

Définition 6.1. *Un test d'hypothèse H_0 est une règle qui, pour tout échantillon donné $\mathcal{X}_n = (X_1, ..., X_n)$, dit si l'on accepte ou rejette H_0.*

Un test est donc identique à une décomposition de l'ensemble de tous les échantillons possibles $\mathcal{X}_n$ en deux parties disjointes : la partie R où l'on rejette H_0 et son complément R^c où l'on ne rejette pas H_0. On appelle R **région critique** du test (ou **région de rejet**) :

$$\text{si } \mathcal{X}_n \in R \text{ on rejette } H_0,$$
$$\text{si } \mathcal{X}_n \notin R \text{ on accepte } H_0.$$

REMARQUE. Comme un test est entièrement défini par la donnée de sa région critique R, on écrira souvent dans la suite, pour abréger,"test R" au lieu de "test à région critique R".

EXEMPLE 6.2. *Un test "naïf"*. Considérons le modèle statistique $\{\mathcal{N}(\theta, 1), \theta \in \{0, 1\}\}$ avec $\Theta = \{0, 1\}$, $\Theta_0 = \{0\}$ et $\Theta_1 = \{1\}$. Étant donné l'échantillon $\mathcal{X}_n = (X_1, ..., X_n)$, on souhaite choisir entre les hypothèses $H_0 : \theta = 0$ et $H_1 : \theta = 1$. Notre hypothèse de préférence est $\theta = 0$, on cherche à accepter ou à rejeter cette hypothèse. Notons que $\bar{X}$ est un bon estimateur de θ dans le modèle normal, i.e. il est proche de la vraie valeur du paramètre pour n assez grand. Ceci nous incite de construire un test qui semblerait, à la première vue, intuitif : on rejette l'hypothèse H_0 si $\bar{X} > 1/2$, i.e. si $\bar{X}$ est plus proche de 1 que de 0. La région de rejet de ce test est

$$R = \{\mathcal{X}_n : \ \bar{X} > 1/2\}.$$

On verra dans la suite qu'un tel test n'est pas toujours très adéquat : il traite l'hypothèse et l'alternative de façon "égalitaire", tandis qu'il est souvent utile de tenir compte d'une certaine dyssymétrie entre l'hypothèse et l'alternative. En effet, l'hypothèse peut s'avérer plus "dangereuse" que l'alternative, comme dans l'Exemple 6.1.

Voici quelques exemples d'hypothèses H_0 et H_1 sur le paramètre $\theta \in \mathbb{R}$ (avec une valeur donnée $\theta_0 \in \mathbb{R}$) :

– $H_0 : \theta = \theta_0, \ H_1 : \theta \neq \theta_0$ – test d'une hypothèse simple contre une alternative composite ;
– $H_0 : \theta > \theta_0, \ H_1 : \theta \leq \theta_0$ – test d'une hypothèse composite contre une alternative composite ;

– $H_0 : \theta > \theta_0$, $H_1 : \theta = \theta_0$ – test d'une hypothèse composite contre une alternative simple.

Tout au long de ce chapitre on supposera que l'Hypothèse (E) (hypothèse d'échantillonnage), l'Hypothèse (P) (hypothèse de paramétrisation) et l'Hypothèse (D) (hypothèse de dominance) du Chapitre 5 soient vérifiées.

6.2. Test d'hypothèse simple contre l'alternative simple

Dans ce paragraphe, nous étudierons le cas basique où l'hypothèse et l'alternative sont simples :

$$\begin{aligned} H_0 &: \theta = \theta_0, \\ H_1 &: \theta = \theta_1. \end{aligned} \tag{6.2}$$

Ici θ_0 et θ_1 sont deux valeurs données. Le modèle statistique est $\{F_\theta, \theta \in \Theta\}$ où $\Theta = \{\theta_0, \theta_1\}$. Par la suite, P_θ désignera la loi jointe de $(X_1, \ldots, X_n)$ quand les X_i sont i.i.d. de loi F_θ (cf. Paragraphe 5.1.2).

En choisissant entre l'hypothèse et l'alternative on s'expose à deux types d'erreurs :

Erreur de $\mathbf{1^{ère}}$ espèce : rejeter l'hypothèse H_0, alors qu'elle est vraie.

Erreur de $\mathbf{2^{ème}}$ espèce : accepter l'hypothèse H_0, alors qu'elle est fausse.

On associe à ces erreurs les deux risques suivants.

$$\boxed{\textbf{Risque de } \mathbf{1^{ère}} \textbf{ espèce} = P_{\theta_0}(\mathcal{X}_n \in R).}$$

C'est la probabilité de rejeter l'hypothèse H_0, alors qu'elle est vraie (l'indice θ_0 de la probabilité signale qu'elle est calculée sous l'hypothèse que la vraie valeur du paramètre θ est égale à θ_0).

$$\boxed{\textbf{Risque de } \mathbf{2^{ème}} \textbf{ espèce} = P_{\theta_1}(\mathcal{X}_n \notin R).}$$

C'est la probabilité d'accepter l'hypothèse H_0, alors qu'elle est fausse (l'indice θ_1 de la probabilité signale qu'elle est calculée sous l'hypothèse que la vraie valeur du paramètre θ est égale à θ_1).

Comment choisir la région critique R de façon optimale ? Il est clair que plus petits sont les deux risques, mieux est il. Cependant, on ne peut pas minimiser en R les deux risques simultanément. En effet, pour minimiser le risque de $1^{ère}$ espèce il faut choisir R aussi petit que possible. Pour minimiser le risque de $2^{ème}$ espèce, au contraire, il faut choisir R aussi grand que possible.

Donc, il faut chercher une méthode de choix de R permettant d'établir un compromis entre les deux risques. L'approche la plus courante est celle de Neyman – Pearson. Elle est fondée sur l'idée de *dissymétrie entre H_0 et H_1*. Afin de comprendre son origine, revenons à l'Exemple 6.1 (détection de missile). Si l'on commet l'erreur de $1^{ère}$ espèce (i.e. on rejette sans raison l'hypothèse qu'un *missile est présent*), cela peut nous coûter beaucoup plus cher et les conséquences peuvent être beaucoup plus dangereuses que si l'on commet l'erreur de $2^{ème}$

espèce, i.e. l'erreur de *fausse alerte*. Ceci explique le fait que d'habitude on fixe une borne pour le risque de $1^{\text{ère}}$ espèce : on veut que

$$P_{\theta_0}(\mathcal{X}_n \in R) \leq \alpha, \quad \text{où } \alpha \in]0,1[\text{ est “petit”}. \tag{6.3}$$

Les valeurs couramment utilisées de α sont $\alpha = 0.01$, $\alpha = 0.05$, $\alpha = 0.1$. Ayant borné le risque de $1^{\text{ère}}$ espèce par (6.3), il est naturel de chercher à minimiser le risque de $2^{\text{ème}}$ espèce

$$\beta = \beta(R) = P_{\theta_1}(\mathcal{X}_n \notin R),$$

i.e. de chercher un test R tel que son risque de $2^{\text{ème}}$ espèce β soit minimal parmi tous les tests qui vérifient la contrainte (6.3).

Définition 6.2. *Soit* $0 < \alpha < 1$. *Un test* R *de l'hypothèse simple* $H_0 : \theta = \theta_0$ *est dit* **test de niveau** α *si*

$$P_{\theta_0}(\mathcal{X}_n \in R) \leq \alpha$$

et **test de taille** α *si*

$$P_{\theta_0}(\mathcal{X}_n \in R) = \alpha.$$

La valeur α *est dite* **niveau** *(ou* **niveau de signification***) du test.*

Nous pouvons donc formuler une approche suivante du choix optimal de test.

Paradigme de Neyman – Pearson. Soit $0 < \alpha < 1$ un niveau donné. On déclare optimal tout test R^* de niveau α qui atteint le minimum du risque de $2^{\text{ème}}$ espèce parmi tous les tests de niveau α.

Définissons la **puissance** du test R par

$$\pi(R) = 1 - \beta(R) = P_{\theta_1}(\mathcal{X}_n \in R).$$

Définition 6.3. *Un test* R^* *test de niveau* α *de l'hypothèse simple* $H_0 : \theta = \theta_0$ *contre l'alternative simple* $H_1 : \theta = \theta_1$ *est appelé* **test le plus puissant de niveau** α *(en abrégé* **test PP de niveau** α*) si*

$$\pi(R^*) \geq \pi(R)$$

pour tout test R *de niveau* α.

Vu cette définition, une façon équivalente de formuler le Paradigme de Neyman – Pearson est la suivante : **on déclare optimal tout test PP de niveau** α.

Il est remarquable qu'un test PP de niveau α existe dans plusieurs situations et qu'on peut le trouver de façon explicite. Il appartient à la famille de tests ayant les régions critiques de la forme :

$$R^*(c) = \{\mathcal{X}_n : \, L(\mathcal{X}_n, \theta_1) > cL(\mathcal{X}_n, \theta_0)\},$$

où $L(\mathcal{X}_n, \theta) = \prod_{i=1}^n f(X_i, \theta)$ est la fonction de vraisemblance et $c > 0$ est une constante à préciser. Tout test qui correspond à une région critique de cette forme s'appelle **test du rapport de vraisemblance**. Si, de plus, on a partout $L(\mathcal{X}_n, \theta_0) > 0$, on peut définir la variable aléatoire $\dfrac{L(\mathcal{X}_n, \theta_1)}{L(\mathcal{X}_n, \theta_0)}$ appelée **rapport de vraisemblance** et écrire

$$R^*(c) = \left\{ \mathcal{X}_n : \, \frac{L(\mathcal{X}_n, \theta_1)}{L(\mathcal{X}_n, \theta_0)} > c \right\}.$$

Le résultat suivant est fondamental dans la théorie des tests.

Théorème 6.1. (Lemme fondamental de Neyman – Pearson.) *S'il existe une valeur* $c_\alpha > 0$ *telle que*

$$P_{\theta_0}\big(L(\mathcal{X}_n,\theta_1) > c_\alpha L(\mathcal{X}_n,\theta_0)\big) = \alpha, \tag{6.4}$$

alors le test du rapport de vraisemblance à région critique $R^*(c_\alpha)$ *fournit le minimum du risque de* $2^{\text{ème}}$ *espèce parmi tous les tests de niveau* α. *Autrement dit, ce test est PP de niveau* α.

Preuve. Notons pour abréger $L_i = L(\mathcal{X}_n,\theta_i)$, $i = 0,1$. Il faut montrer que pour tout R vérifiant l'inégalité

$$P_{\theta_0}(\mathcal{X}_n \in R) \le \alpha \tag{6.5}$$

on a :

$$P_{\theta_1}(\mathcal{X}_n \notin R) \ge P_{\theta_1}(\mathcal{X}_n \notin R^*) \tag{6.6}$$

où $R^* = R^*(c_\alpha)$. Or, (6.6) équivaut à $P_{\theta_1}(\mathcal{X}_n \in R) \le P_{\theta_1}(\mathcal{X}_n \in R^*)$ et on a :

$$\begin{aligned} P_{\theta_1}(\mathcal{X}_n \in R^*) - P_{\theta_1}(\mathcal{X}_n \in R) &= \int_{R^*} L_1 d\mu - \int_R L_1 d\mu \\ &= \int_{R^*\setminus R} L_1 d\mu - \int_{R\setminus R^*} L_1 d\mu, \end{aligned}$$

où μ est la mesure dominante (voir l'Hypothèse (D), Chapitre 5). Comme $R^*\setminus R \subset R^*$, on obtient : $L_1 > c_\alpha L_0$ sur $R^*\setminus R$. De la même façon, $L_1 \le c_\alpha L_0$ sur $R\setminus R^*$. Alors,

$$\begin{aligned} \int_{R^*\setminus R} L_1 d\mu - \int_{R\setminus R^*} L_1 d\mu &\ge c_\alpha \left[\int_{R^*\setminus R} L_0 d\mu - \int_{R\setminus R^*} L_0 d\mu\right] \\ &= c_\alpha \left[\int_{R^*} L_0 d\mu - \int_R L_0 d\mu\right] \\ &= c_\alpha \left[P_{\theta_0}(\mathcal{X}_n \in R^*) - P_{\theta_0}(\mathcal{X}_n \in R)\right] \ge 0, \end{aligned}$$

vu (6.5) et le fait que $P_{\theta_0}(\mathcal{X}_n \in R^*) = \alpha$ d'après (6.4). ∎

Exemple 6.3. Considérons le modèle statistique $\{\mathcal{N}(\theta,\sigma^2), \theta \in \mathbb{R}\}$ avec σ^2 connu. Supposons que l'on souhaite tester l'hypothèse $H_0 : \theta = 0$, contre l'alternative $H_1 : \theta = 1$ (i.e. $\theta_0 = 0$, $\theta_1 = 1$). Dans ce cas la fonction de vraisemblance vaut

$$L(\mathcal{X}_n,\theta) = \prod_{i=1}^n f(X_i,\theta) \quad \text{avec} \quad f(x,\theta) = (2\pi)^{-1/2}\sigma^{-1}\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

et le rapport de vraisemblance est

$$\frac{L(\mathcal{X}_n,1)}{L(\mathcal{X}_n,0)} = \prod_{i=1}^n \exp\left\{\frac{1}{2\sigma^2}[2X_i - 1]\right\} = \exp\left\{\frac{n}{2\sigma^2}[2\bar{X} - 1]\right\}.$$

Le test du rapport de vraisemblance a pour région critique

$$R^* = \left\{\exp\left\{\frac{n}{2\sigma^2}[2\bar{X} - 1]\right\} \ge c'\right\}$$

avec une constante $c' > 0$ à préciser. On peut l'écrire sous la forme équivalente :

$$R^* = \{\bar{X} \geq c\} \quad \text{avec} \quad c = \frac{\sigma^2}{n} \ln c' + \frac{1}{2}.$$

Choisissons la constante c de façon à obtenir $P_{\theta_0}(\mathcal{X}_n \in R^*) = \alpha$, i.e. $P_0(\bar{X} \geq c) = \alpha$. Notons que sous P_0 (i.e. sous l'hypothèse H_0) les X_i suivent la loi $\mathcal{N}(0, \sigma^2)$ correspondant à la valeur du paramètre $\theta = 0$. On a donc $\bar{X} \sim \mathcal{N}(0, \sigma^2/n)$ sous P_0. Alors,

$$P_0(\bar{X} \geq c) = P_0\left(\frac{\sqrt{n}\bar{X}}{\sigma} \geq \frac{\sqrt{n}c}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right),$$

où $\Phi(\cdot)$ est la f.d.r. de la loi normale standard. Pour que R^* soit un test de taille α, il faut prendre c comme solution de

$$1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right) = \alpha,$$

ce qui équivaut à

$$c = c_\alpha = \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha},$$

où $q^N_{1-\alpha}$ désigne le quantile d'ordre $1 - \alpha$ de la loi $\mathcal{N}(0, 1)$. La région critique du test PP de niveau α est donc

$$R^* = \left\{\bar{X} \geq \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha}\right\}.$$

Considérons l'exemple numérique : $\sigma = 2$, $\alpha = 0,05$ et $n = 25$. Dans ce cas $q^N_{0,95} \approx 1,64$, $c_{0,05} \approx \frac{2}{5} \cdot 1,64 = 0,656$. On rejette donc l'hypothèse $H_0 :\ \theta = 0$ au niveau $0,05$ si $\bar{X} \geq 0,656$ et on ne rejette pas H_0 au niveau $0,05$ si $\bar{X} < 0,656$.

Pour calculer la puissance de ce test, on remarque que sous P_1, la variable $\sqrt{n}(\bar{X} - 1)/\sigma$ suit la loi normale $\mathcal{N}(0, 1)$, donc

$$\begin{aligned} \pi &= P_{\theta_1}(\mathcal{X}_n \in R^*) = P_1(\bar{X} \geq c) \\ &= P_1\left(\frac{\sqrt{n}(\bar{X} - 1)}{\sigma} \geq \frac{\sqrt{n}(c - 1)}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(c - 1)}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}(1 - c)}{\sigma}\right). \end{aligned}$$

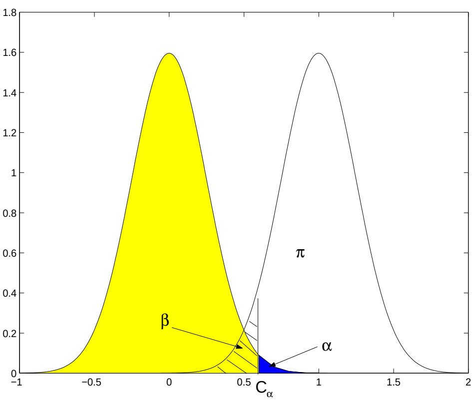

Fig. 6.1. Densité de la statistique $\bar{X}$ sous P_0 et sous P_1.

Remarques.

(1) On ne peut pas simultanément diminuer le risque de $1^{\text{ère}}$ espèce α et augmenter la puissance π (cf. Fig. 6.1).

(2) Quand $n \to \infty$ le test devient de plus en plus puissant : $\pi \to 1$.

(3) Dans les applications, on évite la formulation "accepter H_0". On dit plutôt "ne pas rejeter H_0". Ceci s'explique par le fait que, dans la pratique, le statisticien n'est pas toujours sûr de la définition de l'hypothèse H_0 qu'il choisit pour tester. S'il ne rejette pas H_0, il y a beaucoup d'autres hypothèses H_0' qu'il ne rejette pas non plus. Par contre, si le résultat du test est la rejection de H_0 au niveau α (où α est très petit), ceci signifie que H_0 est vraiment très peu probable, autrement dit, la rejection est sûre.

Dans les applications, une pratique couramment répandue consiste à se référer au *seuil critique* α^* *(p-value)* du test. Il s'agit de donner, pour un échantillon fixé et un test fixé, la plus grande valeur de niveau α, pour laquelle l'hypothèse H_0 n'est pas rejetée par le test. La donnée du seuil critique permet de trouver l'ensemble de tous les α tels que l'hypothèse H_0 est rejetée (ou ne pas rejetée) au niveau α, sans refaire les calculs pour chaque α particulier.

Définition 6.4. *Supposons que l'échantillon $\mathcal{X}_n$ est fixé et on utilise un test fixé. La valeur $\alpha^* = \alpha^*(\mathcal{X}_n)$ est dite* **seuil critique (p-value)** *du test si l'on rejette H_0 pour tout $\alpha > \alpha^*$ et on ne rejette pas H_0 pour tout $\alpha < \alpha^*$.*

Dans l'Exemple 6.3, le test PP de niveau α est $R^* = \{\bar{X} \geq c_\alpha\}$ avec $c_\alpha = \sigma q^N_{1-\alpha}/\sqrt{n}$ et

$$1 - \alpha = \Phi\left(\frac{\sqrt{n}c_\alpha}{\sigma}\right).$$

Pour un $\bar{X}$ donné, on passe de l'acceptation au rejet de H_0 à partir de $\alpha = \alpha^*(\bar{X})$ tel que $c_{\alpha^*} = \bar{X}$. Ce choix correspond à la valeur α^* vérifiant

$$1 - \alpha^* = \Phi\left(\frac{\sqrt{n}\bar{X}}{\sigma}\right).$$

Alors, le seuil critique (p-value) de ce test est donné par

$$\alpha^* = 1 - \Phi\left(\frac{\sqrt{n}\bar{X}}{\sigma}\right).$$

On rejette l'hypothèse $H_0 : \theta = 0$ à tout niveau $\alpha > \alpha^*$ et on ne rejette pas H_0 pour $\alpha < \alpha^*$.

Si le seuil critique α^* est relativement grand ($\alpha^* > 0.1$), on peut l'interpréter comme une indication en faveur de l'hypothèse H_0 : par exemple, on ne peut pas rejeter H_0 aux niveaux habituels $\alpha = 0.01$, $\alpha = 0.05$, $\alpha = 0.1$. Le fait que α^* soit petit ($\alpha < 0.1$) s'interpète comme une indication contre l'hypothèse H_0.

Remarque. Dans la pratique, une question importante est de bien poser le problème de test, i.e. de choisir laquelle des deux hypothèses en question doit être nommée hypothèse nulle H_0. Il y a plusieurs règles heuristiques de le faire, dont les suivantes.

– Choisir comme H_0 l'hypothèse que l'on cherche à rejeter : e.g., si l'on teste un médicament, on prend comme H_0 l'hypothèse que le médicament n'est pas efficace (ceci doit être traduit, bien évidemment, en termes de paramètres des lois statistiques).

– Si l'une de deux hypothèses est plus simple ou "de dimension plus petite" que l'autre, c'est elle qui est généralement nommée H_0 (exemple : $H_0 : \theta = 0$, $H_1 : \theta \neq 0$).
– Très souvent H_0 est plus "importante" ou plus "dangereuse" que H_1 (cf. Exemple 6.1 lié à la détection de missile).

6.3. Tests des hypothèses composites

Considérons maintenant les tests de deux hypothèses composites, i.e. tels que les ensembles Θ_0 et Θ_1 peuvent contenir plus d'un élément.

$$H_0 : \ \theta \in \Theta_0, \quad H_1 : \ \theta \in \Theta_1.$$

On peut alors formuler une généralisation du paradigme de Neyman – Pearson. Pour ce faire, définissons d'abord le risque de $1^{\text{ère}}$ espèce d'un test R de l'hypothèse composite H_0 :

$$\boxed{\textbf{Risque de } \mathbf{1^{\text{ère}}} \textbf{ espèce} = \sup_{\theta \in \Theta_0} P_\theta(\mathcal{X}_n \in R).}$$

Si Θ_0 ne contient qu'un seul élément : $\Theta_0 = \{\theta_0\}$, on retrouve la définition du risque de $1^{\text{ère}}$ espèce pour l'hypothèse simple donnée au paragraphe précédent.

Définition 6.5. *Soit* $0 < \alpha < 1$. *Un test* R *de l'hypothèse composite* $H_0 : \theta \in \Theta_0$ *est dit* **test de niveau** α *si*

$$\sup_{\theta \in \Theta_0} P_\theta(\mathcal{X}_n \in R) \leq \alpha$$

et **test de taille** α *si*

$$\sup_{\theta \in \Theta_0} P_\theta(\mathcal{X}_n \in R) = \alpha.$$

Autrement dit, pour un test de niveau α de l'hypothèse composite H_0, le maximum des risques de $1^{\text{ère}}$ espèce pour toutes les hypothèses simples $H_0 : \theta = \theta_0$ avec θ_0 appartenant à Θ_0 est borné par α.

Si l'alternative Θ_1 est composite, il n'y a pas de notion de risque de $2^{\text{ème}}$ espèce : on le remplace par la notion de *fonction puissance.*

Définition 6.6. *La fonction* $\pi : \Theta \to [0, 1]$ *définie par*

$$\pi(\theta) = P_\theta(\mathcal{X}_n \in R)$$

est appelée **fonction puissance** *du test* R *(ou* **caractéristique opérationnelle** *du test* R*).*

Quand il s'agit une alternative composite Θ_1, l'ensemble des valeurs $\{\pi(\theta), \theta \in \Theta_1\}$ joue un rôle analogue à celui du risque de $2^{\text{ème}}$ espèce pour le cas d'alternative simple. Soulignons que

$$0 \leq \pi(\theta) \leq 1.$$

Définition 6.7. *Un test R^* de niveau α est dit* **uniformément plus puissant (UPP) de niveau** α *contre l'alternative $H_1 : \theta \in \Theta_1$ si*

$$\pi(\theta) = P_\theta(\mathcal{X}_n \in R) \le P_\theta(\mathcal{X}_n \in R^*) = \pi^*(\theta)$$

pour tout $\theta \in \Theta_1$ et tout test R de niveau α.

Le paradigme de Neyman – Pearson pour des hypothèses composites se généralise de la façon suivante : **déclarer optimal tout test UPP de niveau** α.

Il est utile de noter que les tests UPP n'existent que dans quelques cas exceptionnels. Nous allons ici en décrire un : celui du modèle normal et d'alternative unilatérale.

Exemple 6.4. *Test UPP pour le modèle normal $\{\mathcal{N}(\theta, \sigma^2),\ \theta \in \mathbb{R}\}$ avec $\sigma > 0$ connu.* Considérons le problème de test de deux hypothèses composites suivantes :

$$\begin{aligned} H_0 &: \theta \le 0, \\ H_1 &: \theta > 0. \end{aligned} \tag{6.7}$$

Introduisons le test

$$\tilde{R} = \{\bar{X} > c_\alpha\} \quad \text{avec} \quad c_\alpha = \frac{\sigma}{\sqrt{n}}\, q^N_{1-\alpha} \tag{6.8}$$

et calculons sa fonction puissance $\pi(\cdot)$. Notons que, pour tout $\theta \in \mathbb{R}$, la variable aléatoire $\sqrt{n}(\bar{X} - \theta)/\sigma$ suit la loi normale standard sous P_θ. On a alors,

$$\begin{aligned} \pi(\theta) &= P_\theta(\bar{X} > c_\alpha) = P_\theta\left(\frac{\sqrt{n}(\bar{X} - \theta)}{\sigma} > \frac{\sqrt{n}(c_\alpha - \theta)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(c_\alpha - \theta)}{\sigma}\right) \\ &= \Phi\left(\frac{\sqrt{n}(\theta - c_\alpha)}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}\theta}{\sigma} - q^N_{1-\alpha}\right) \end{aligned} \tag{6.9}$$

où Φ est la fonction de répartition de la loi normale standard $\mathcal{N}(0, 1)$. En utilisant la symétrie de Φ, on obtient

$$\pi(0) = \Phi(-q^N_{1-\alpha}) = 1 - \Phi(q^N_{1-\alpha}) = 1 - (1 - \alpha) = \alpha.$$

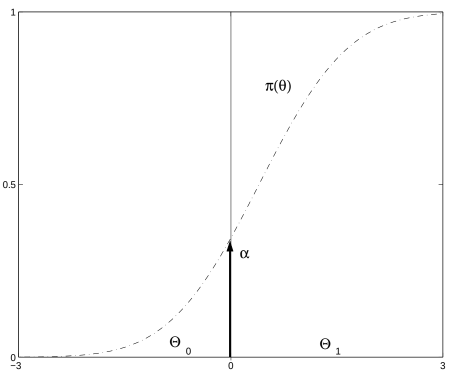

Fig. 6.2. Fonction puissance du test $\tilde{R}$.

Vu la monotonie de Φ, le test $\tilde{R}$ est de niveau (et de taille) α :

$$\sup_{\theta \in \Theta_0} P_\theta(\mathcal{X}_n \in \tilde{R}) = \sup_{\theta \in \Theta_0} \pi(\theta) = \pi(0) = \alpha,$$

où $\Theta_0 = \{\theta \le 0\}$. Montrons que $\tilde{R}$ défini par (6.8) est un test uniformément plus puissant de niveau α. Fixons une valeur $\theta' \in \Theta_1 = \{\theta > 0\}$. Considérons les hypothèses simples

$$\begin{aligned} H_0 &: \ \theta = 0, \\ H_1 &: \ \theta = \theta'. \end{aligned}$$

D'après le lemme de Neyman – Pearson et l'Exemple 6.3, le test $\tilde{R}$ donné par (6.8) de l'hypothèse simple $H_0 : \theta = 0$ contre l'alternative simple $H_1 : \theta = \theta'$ satisfait :

$$P_{\theta'}(\mathcal{X}_n \in \tilde{R}) \ge P_{\theta'}(\mathcal{X}_n \in R) \tag{6.10}$$

pour tout test R de niveau α, i.e. tel que

$$P_0(\mathcal{X}_n \in R) \le \alpha. \tag{6.11}$$

Mais si un test R vérifie $\sup_{\theta \in \Theta_0} P_\theta(\mathcal{X}_n \in R) \le \alpha$, il vérifie aussi (6.11), car $0 \in \Theta_0$.

Par conséquent, pour tout test R de niveau α de l'hypothèse composite $H_0 : \theta \le 0$ contre l'alternative composite $H_1 : \theta > 0$ et pour tout $\theta' > 0$, on a (6.10). Ceci équivaut à dire que $\tilde{R}$ est un test uniformément plus puissant de niveau α pour le problème de test (6.7).

Il est facile de voir que le test $\tilde{R}$ est aussi uniformément plus puissant de niveau α pour le problème de test de l'hypothèse simple $H_0 : \theta = 0$ contre l'alternative composite $H_1 : \theta > 0$.

La fonction puissance du test $\tilde{R}$ défini dans (6.8) est donnée par (cf. (6.9)) :

$$\pi(\theta) = \Phi\Big(\frac{\sqrt{n}}{\sigma}\theta - q^N_{1-\alpha}\Big).$$

Sa dérivée vaut

$$\pi'(\theta) = \frac{\sqrt{n}}{\sigma}\varphi\Big(\frac{\sqrt{n}}{\sigma}\theta - q^N_{1-\alpha}\Big),$$

où $\varphi(x) = \Phi'(x)$ est la densité de la loi normale $\mathcal{N}(0,1)$. En utilisant ces formules on peut analyser le comportement asymptotique quand $n \to \infty$ de la fonction puissance $\pi(\cdot)$ (voir les graphiques suivants).

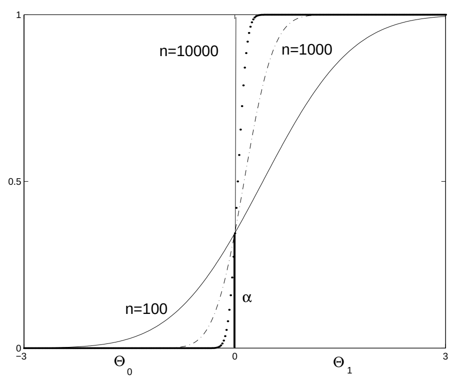

Fig. 6.3. Pour tout $\theta \in \Theta_1$, $\pi(\theta) \to 1$ lorsque $n \to \infty$.

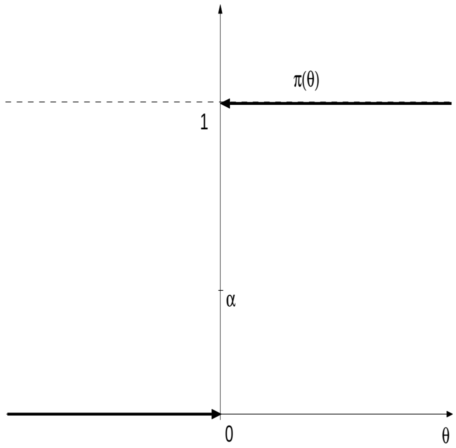

Fig. 6.4. On obtient asymptotiquement, quand $n \to \infty$, une fonction puissance “idéale”.

Définition 6.8. *Un test R s'appelle* **consistant** *si* $\pi(\theta) \to 1$ *lorsque* $n \to \infty$ *pour tout* $\theta \in \Theta_1$.

Définition 6.9. *Un test R est dit* **sans biais** *si*

$$\sup_{\theta \in \Theta_0} \pi(\theta) \le \inf_{\theta \in \Theta_1} \pi(\theta).$$

EXERCICE 6.1. Montrer que le test $\tilde{R}$ défini par (6.8) est consistant et sans biais.

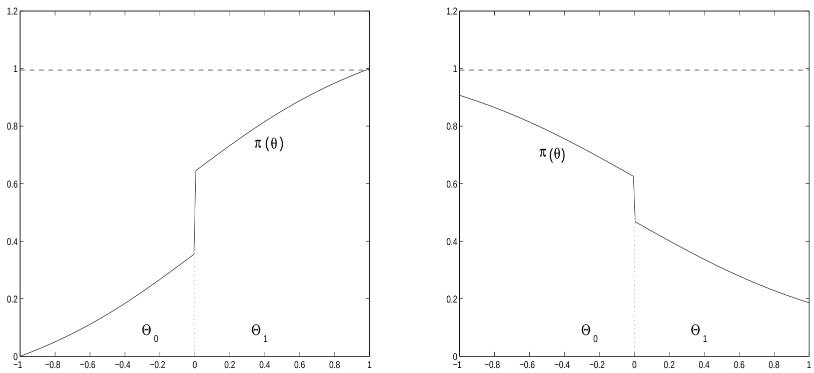

Fig. 6.5. Fonctions puissance d'un test sans biais et d'un test biaisé.

6.4. Tests dans le modèle normal

Le modèle statistique que nous considérerons dans ce paragraphe est $\{\mathcal{N}(\theta, \sigma^2),\ \theta \in \mathbb{R}, \sigma > 0\}$. Soit θ_0 une valeur donnée. On examinera d'abord les tests d'hypothèse sur le paramètre θ quand σ est connu, en étudiant séparement le cas *d'alternative unilatérale* $H_1 : \theta > \theta_0$ $(\theta \ge \theta_0)$ ou $H_1 : \theta < \theta_0$ $(\theta \le \theta_0)$ et celui *d'alternative bilatérale* $H_1 : \theta \ne \theta_0$.

6.4.1. Alternative unilatérale, σ connu. *Cas $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$ avec $\sigma > 0$ connu.* Notons $X_i' = X_i - \theta_0$, $\theta' = \theta - \theta_0$, alors le problème de test se réécrit comme

$$H_0 : \theta' = 0, \quad H_1 : \theta' > 0.$$

Pour ce dernier, comme on l'a déjà vu, le test $R = \{\bar{X}' > \frac{\sigma}{\sqrt{n}} q_{1-\alpha}^N\}$, où $\bar{X}' = n^{-1}\sum_{i=1}^n X_i'$, est uniformément plus puissant de niveau α. Alors le test

$$R = \left\{ \bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} q_{1-\alpha}^N \right\}$$

est uniformément plus puissant de niveau α pour le problème initial.

Étudions la fonction puissance de ce test. Rappelons-nous que, sous P_{θ_0}, la v.a. $\frac{\sqrt{n}(\bar{X}-\theta_0)}{\sigma}$ suit la loi $\mathcal{N}(0,1)$. On a alors

$$\begin{aligned} \pi(\theta) &= P_\theta\Big(\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} q_{1-\alpha}^N\Big) = P_\theta\left(\frac{\sqrt{n}(\bar{X}-\theta)}{\sigma} > \frac{\sqrt{n}(\theta_0-\theta)}{\sigma} + q_{1-\alpha}^N\right) \\ &= 1 - \Phi\Big(q_{1-\alpha}^N + \frac{\sqrt{n}(\theta_0-\theta)}{\sigma}\Big) = \Phi\Big(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma} - q_{1-\alpha}^N\Big). \end{aligned}$$

Comme

$$\pi(\theta_0) = \Phi(-q_{1-\alpha}^N) = 1 - \Phi(q_{1-\alpha}^N) = \alpha,$$

c'est un test de niveau α. Notons que $\pi(\theta)$ représente une translation par θ_0 de la fonction puissance (6.9) du test (6.8) de l'hypothèse $H_0 : \theta = 0$ contre l'alternative $H_1 : \theta > 0$.

Cas $H_0 : \theta \le \theta_0$, $H_1 : \theta > \theta_0$, avec σ connu. Le même test

$$R = \left\{ \bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} q_{1-\alpha}^N \right\}$$

est UPP de niveau α (cf. Paragraphe 6.3).

Les cas ($H_0 : \theta = \theta_0$, $H_1 : \theta < \theta_0$) et ($H_0 : \theta \ge \theta_0$, $H_1 : \theta < \theta_0$) avec σ connu peuvent être traités de façon similaire aux cas précédents. Notamment, le test

$$R = \left\{ \bar{X} < \theta_0 - \frac{\sigma}{\sqrt{n}} q_{1-\alpha}^N \right\}$$

est uniformément plus puissant de niveau α pour ces problèmes (démontrez ceci à titre d'exercice). La fonction puissance du test R est

$$\begin{aligned} \pi(\theta) &= P_\theta\Big(\bar{X} < \theta_0 - \frac{\sigma}{\sqrt{n}} q_{1-\alpha}^N\Big) = P_\theta\left(\frac{\sqrt{n}(\bar{X}-\theta)}{\sigma} < \frac{\sqrt{n}(\theta_0-\theta)}{\sigma} - q_{1-\alpha}^N\right) \\ &= \Phi\Big(\frac{\sqrt{n}(\theta_0-\theta)}{\sigma} - q_{1-\alpha}^N\Big), \end{aligned}$$

et $\pi(\theta_0) = \Phi(-q^N_{1-\alpha}) = \alpha$.

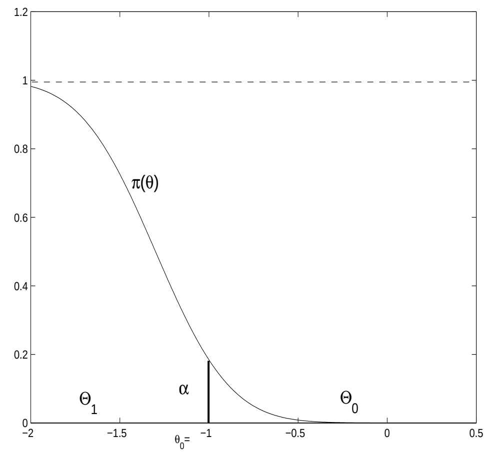

Fig. 6.6. Fonction puissance du test R.

REMARQUE. Notons que si l'alternative est unilatérale, le test du rapport de vraisemblance dans le cas gaussien s'écrit de manière très simple : l'alternative $\theta > \theta_0$ (ou $\theta \geq \theta_0$) est associée avec la région critique de la forme $\{\bar{X} > C\}$, alors que l'alternative $\theta < \theta_0$ (ou $\theta \leq \theta_0$) est associée avec $\{\bar{X} < C\}$ (*même sens des inégalités dans la définition de l'alternative et de la région critique*), pour une constante $C = C(\alpha, \theta_0)$ que l'on choisit de façon à s'assurer que le test soit de niveau α.

6.4.2. Alternative bilatérale, σ connu. Considérons l'hypothèse et l'alternative

$$H_0 : \ \theta = \theta_0, \quad H_1 : \ \theta \neq \theta_0.$$

Introduisons la région critique

$$R = \{|\bar{X} - \theta_0| > c\},$$

où $c > 0$ est choisi de façon à obtenir un test de taille α, i.e. $c = C_\alpha$ est tel que

$$P_{\theta_0}(|\bar{X} - \theta_0| > C_\alpha) = \alpha.$$

Sous P_{θ_0}, la v.a. $\frac{\sqrt{n}(\bar{X}-\theta_0)}{\sigma}$ suit la loi normale $\mathcal{N}(0,1)$. Donc,

$$P_{\theta_0}(|\bar{X} - \theta_0| > C_\alpha) = P_{\theta_0}\left(\left|\frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma}\right| > \frac{\sqrt{n}C_\alpha}{\sigma}\right) = P\left(|\eta| > \frac{\sqrt{n}C_\alpha}{\sigma}\right),$$

où $\eta \sim \mathcal{N}(0,1)$. On veut que cette dernière expression soit égale à α, ce qui équivaut à

$$\Phi\left(\frac{\sqrt{n}C_\alpha}{\sigma}\right) - \Phi\left(-\frac{\sqrt{n}C_\alpha}{\sigma}\right) = 1 - \alpha,$$

ou bien à

$$C_\alpha = \frac{\sigma}{\sqrt{n}}\, q^N_{1-\alpha/2}.$$

Il en découle que la région critique

$$R^* = \left\{|\bar{X} - \theta_0| > \frac{\sigma}{\sqrt{n}}\, q^n_{1-\alpha/2}\right\} \tag{6.12}$$

définit un test de niveau α de l'hypothèse $H_0 : \theta = \theta_0$ contre l'alternative bilatérale $H_1 : \theta \neq \theta_0$. La fonction puissance de ce test est donnée par

$$\begin{aligned}
\pi^*(\theta) &= P_\theta\Big(\mathcal{X}_n \in R^*\Big) = P_\theta\Big(|\bar{X} - \theta_0| > \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}\Big) \\
&= P_\theta\Big(\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}\Big) + P_\theta\Big(\bar{X} < \theta_0 - \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}\Big) \\
&= P\left(\eta > \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} + q^N_{1-\alpha/2}\right) + P\left(\eta < \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} - q^N_{1-\alpha/2}\right) \\
&= 1 - \Phi\Big(q^N_{1-\alpha/2} + \frac{\sqrt{n}(\theta_0 - \theta)}{\sigma}\Big) + \Phi\Big(\frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} - q^N_{1-\alpha/2}\Big) \\
&= \Phi\Big(\frac{\sqrt{n}(\theta - \theta_0)}{\sigma} - q^N_{1-\alpha/2}\Big) + \Phi\Big(\frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} - q^N_{1-\alpha/2}\Big), \qquad (6.13)
\end{aligned}$$

où $\eta = \sqrt{n}(\bar{X} - \theta)/\sigma \sim \mathcal{N}(0,1)$ sous P_θ.

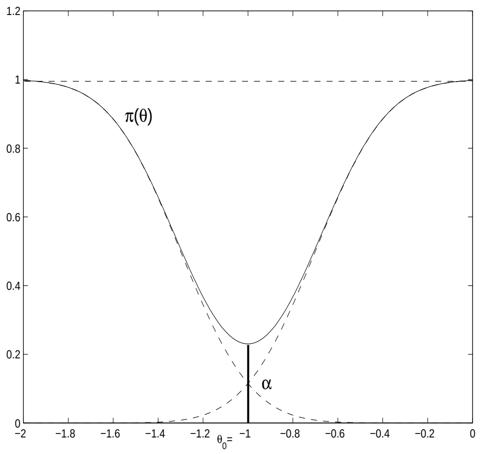

Fig. 6.7. La fonction puissance du test bilatéral.

Quand $n \to \infty$, la valeur $\pi^*(\theta_0)$ reste fixée : $\pi^*(\theta_0) = \alpha$, mais pour tout $\theta \neq \theta_0$, on a $\pi^*(\theta) \to 1$, i.e. le test (6.12) est consistant. C'est aussi un test sans biais.

Étant donnée la valeur de la statistique $\bar{X}$, on peut calculer la *p-value* $\alpha^* = \alpha^*(\bar{X})$ correspondant au test bilatéral (6.12). Elle est déterminée par l'équation

$$|\bar{X} - \theta_0| = \frac{\sigma}{\sqrt{n}} q^n_{1-\alpha^*/2}$$

qui a comme solution

$$\alpha^* = 2\left(1 - \Phi\left(\frac{\sqrt{n}}{\sigma}|\bar{X} - \theta_0|\right)\right).$$

Notons que *le test bilatéral défini par (6.12) n'est pas un test uniformément plus puissant de niveau* α. En effet, il existe au moins un test de niveau α qui est plus puissant que R^* sur un sous-ensemble de Θ_1. Ce test est défini par la région critique

$$R_1 = \Big\{\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha}\Big\}.$$

En effet, la fonction puissance correspondant à R^* (cf. (6.13)) est

$$\pi^*(\theta) = \Phi\Big(\frac{\sqrt{n}(\theta - \theta_0)}{\sigma} - q^N_{1-\alpha/2}\Big) + \Phi\Big(\frac{\sqrt{n}(\theta_0 - \theta)}{\sigma} - q^N_{1-\alpha/2}\Big),$$

alors que celle correspondant à R_1 est

$$\pi_1(\theta) = \Phi\Big(\frac{\sqrt{n}(\theta - \theta_0)}{\sigma} - q^N_{1-\alpha}\Big).$$

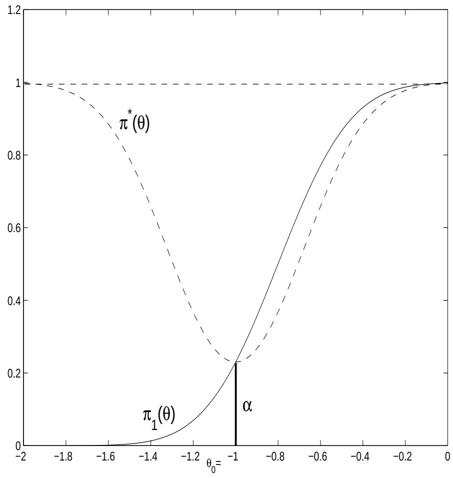

Fig. 6.8. Les fonctions puissance des tests R^* et R_1.

Les deux tests sont de niveau α : $\pi^*(\theta_0) = \alpha$ et $\pi_1(\theta_0) = \alpha$, mais $\pi_1(\theta) > \pi^*(\theta)$ pour un intervalle de valeurs $\theta > \theta_0$. En effet,

$$\begin{aligned} \frac{d}{d\theta}\pi^*(\theta)\Big|_{\theta=\theta_0} &= \frac{\sqrt{n}}{\sigma}\varphi(-q^N_{1-\alpha/2}) - \frac{\sqrt{n}}{\sigma}\varphi(-q^N_{1-\alpha/2}) = 0, \quad \text{où } \varphi(x) = \Phi'(x), \\ \frac{d}{d\theta}\pi_1(\theta)\Big|_{\theta=\theta_0} &= \frac{\sqrt{n}}{\sigma}\varphi(-q^N_{1-\alpha}) > 0. \end{aligned}$$

Notons que néanmoins le test R_1 n'est pas intéressant : ce n'est même pas un test consistant, car pour $\theta < \theta_0$, $\pi_1(\theta) \to 0$ quand $n \to \infty$.

6.4.3. Versions des tests avec σ inconnu. Si le paramètre σ est inconnu, on le remplace par un estimateur convenable. D'après la Proposition 4.2, la statistique $s^2/(n-1)$, où

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

est un estimateur sans biais de σ^2/n. On peut considérer une méthode de construction des tests avec σ inconnu qui consiste à remplacer $\sigma/\sqrt{n}$ par $s/\sqrt{n-1}$ et les quantiles $q^N_{1-\alpha}$ de la loi normale par ceux de la loi de Student, dans la défintion de la région critique. Par exemple, au lieu de la région critique

$$R^* = \Big\{\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}}\, q^N_{1-\alpha}\Big\}$$

du test de l'hypothèse $H_0 : \theta = \theta_0$ contre l'alternative $H_1 : \theta > \theta_0$, on prend

$$R' = \Big\{ \bar{X} > \theta_0 + \frac{s}{\sqrt{n-1}} q_{1-\alpha}(t_{n-1}) \Big\}, \tag{6.14}$$

où $q_{1-\alpha}(t_{n-1})$ est le quantile d'ordre $1-\alpha$ de la loi de Student à $n-1$ degrés de liberté. Ceci donne un test de niveau (et de taille) α. En effet, notons $P_{\theta,\sigma}$ la loi de probabilité (dépendant maintenant de σ qui est inconnu) de $(X_1, \dots, X_n)$, où les X_i sont i.i.d. de loi $\mathcal{N}(\theta, \sigma^2)$. Sous $P_{\theta,\sigma}$, pour tout $\theta \in \mathbb{R}, \sigma > 0$, la variable aléatoire $\sqrt{n-1}(\bar{X} - \theta)/s$ suit la loi de Student t_{n-1} (cf. Corollaire 4.2). On a alors

$$P_{\theta_0,\sigma}\Big(\bar{X} > \theta_0 + \frac{s}{\sqrt{n-1}} q_{1-\alpha}(t_{n-1}) \Big) = P_{\theta_0,\sigma}\left(\frac{\sqrt{n-1}(\bar{X} - \theta_0)}{s} > q_{1-\alpha}(t_{n-1}) \right) = \alpha.$$

Si l'on considère le problème de test bilatéral :

$$H_0 : \ \theta = \theta_0, \qquad H_1 : \ \theta \neq \theta_0$$

avec σ inconnu, la région critique d'un test de niveau α basé sur la même idée est de la forme

$$\tilde{R} = \Big\{ |\bar{X} - \theta_0| > \frac{s}{\sqrt{n-1}} q_{1-\alpha/2}(t_{n-1}) \Big\}.$$

6.4.4. Tests d'hypothèse sur la variance σ^2. Considérons un échantillon i.i.d. $\mathcal{X}_n = (X_1, \dots, X_n)$ d'une loi normale $\mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$. On souhaite tester au niveau α dans le problème suivant :

$$H_0 : \ \sigma^2 \le \sigma_0^2, \quad H_1 : \ \sigma^2 > \sigma_0^2.$$

Cas de μ connu. Pour un $\sigma > \sigma_0$ fixé, considérons le test du rapport de vraisemblance

$$R = \{ L(\mathcal{X}_n, \sigma) > C L(\mathcal{X}_n, \sigma_0) \},$$

où $C > 0$ est une constante à préciser. Clairement,

$$\frac{L(\mathcal{X}_n, \sigma)}{L(\mathcal{X}_n, \sigma_0)} = \frac{\sigma_0}{\sigma} \exp\left(\Big(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma^2} \Big) \sum_{i=1}^n (X_i - \mu)^2 \right),$$

et la région de rejet du test est donc de la forme

$$R = \Big\{ \sum_{i=1}^n (X_i - \mu)^2 > C' \Big\},$$

où $C' > 0$ est une constante. Choisissons $C' = C_\alpha$ de façon à obtenir

$$P_{\sigma_0}(\mathcal{X}_n \in R) = \alpha.$$

Sous P_{σ_0} la variable aléatoire $\sum_{i=1}^n (X_i - \mu)^2 / \sigma_0^2$ suit la loi χ^2_n, donc

$$\begin{aligned} P_{\sigma_0}(\mathcal{X}_n \in R) &= P_{\sigma_0}\left(\sum_{i=1}^n (X_i - \mu)^2 > C_\alpha \right) \\ &= P_{\sigma_0}\left(\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma_0^2} > \frac{C_\alpha}{\sigma_0^2} \right) = 1 - F_{\chi^2_n}\Big(\frac{C_\alpha}{\sigma_0^2} \Big), \end{aligned}$$

où $F_{\chi^2_n}(\cdot)$ est la f.d.r. de loi de χ^2_n. On obtient alors

$$C_\alpha = \sigma_0^2 q_{1-\alpha}(\chi^2_n),$$

ce qui nous amène au test de la forme

$$R = \left\{ \sum_{i=1}^{n} (X_i - \mu)^2 > \sigma_0^2 q_{1-\alpha}(\chi_n^2) \right\}, \tag{6.15}$$

où $q_{1-\alpha}(\chi_n^2)$ désigne le quantile d'ordre $1-\alpha$ de la loi χ_n^2. Pour calculer la fonction puissance de ce test on remarque que sous P_σ la variable aléatoire $\sum_{i=1}^{n}(X_i-\mu)^2/\sigma^2$ suit la loi χ_n^2 et

$$\begin{aligned}\pi(\sigma) &= P_\sigma(\mathcal{X}_n \in R) = P_\sigma\left(\sum_{i=1}^{n}(X_i-\mu)^2 > \sigma_0^2 q_{1-\alpha}(\chi_n^2)\right) \\ &= P_\sigma\left(\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2} q_{1-\alpha}(\chi_n^2)\right) = 1 - F_{\chi_n^2}\Big(\frac{\sigma_0^2}{\sigma^2} q_{1-\alpha}(\chi_n^2)\Big).\end{aligned}$$

Notons que (6.15) définit un test de niveau α. En effet, pour tout $0 < \sigma \le \sigma_0$,

$$1 - F_{\chi_{n-1}^2}\Big(\frac{\sigma_0^2}{\sigma^2} q_{1-\alpha}(\chi_{n-1}^2)\Big) \le 1 - F_{\chi_{n-1}^2}(q_{1-\alpha}(\chi_{n-1}^2)) = \alpha.$$

Application numérique : pour $n = 20$, on souhaite tester au niveau $\alpha = 0.05$ l'hypothèse $H_0 : \sigma \le 2$ contre l'alternative $H_1 : \sigma > 2$ avec $\mu = 0$. Le quantile $q_{0.95}(\chi_{20}^2)$ vaut 31.41 et la région critique du test est

$$R = \Big\{ \sum_{i=1}^{20} X_i^2 > 125.64 \Big\}.$$

La puissance π de ce test au point $\sigma = 4$ est donnée par :

$$\pi(4) = P(\chi_{20}^2 > 125.64/16) = P(\chi_{20}^2 > 7.85) \approx 0.9928.$$

Cas de moyenne μ inconnue. Considérons l'hypothèse et l'alternative

$$H_0 : \ \sigma^2 \le \sigma_0^2, \quad H_1 : \ \sigma^2 > \sigma_0^2$$

quand μ est inconnu. Le paramètre μ est une nuisance, sa valeur ne nous intéresse pas dans ce problème particulier. On peut utiliser le fait que, d'après la Proposition 4.4, sous $P_{\mu,\sigma}$ la statistique ns^2/σ^2 suit la loi χ_{n-1}^2. (Ici $P_{\mu,\sigma}$ désigne la loi jointe de $(X_1, \dots, X_n)$ quand les X_i sont i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$.) On peut donc modifier le test (6.15) de la façon suivante :

$$R = \{ns^2 > \sigma_0^2 q_{1-\alpha}(\chi_{n-1}^2)\}.$$

C'est bien un test de niveau α.

6.5. Tests asymptotiques

Dans la pratique, la loi des variables aléatoires X_i est souvent inconnue. Supposons que $E_\theta(X_i^2) < \infty$ et que $E_\theta(X_i) = \theta \in \mathbb{R}$. Alors, d'après le Théorème central limite, sous P_θ,

$$\sqrt{n}\frac{(\bar{X} - \theta)}{\sigma(\theta)} \xrightarrow{D} \mathcal{N}(0,1) \quad \text{quand } n \to \infty,$$

où $\sigma(\theta)$ est défini par

$$\sigma^2(\theta) = \mathrm{Var}_\theta(X_1) = E_\theta(X_1^2) - (E_\theta(X_1))^2$$

et on suppose que $\sigma(\theta) > 0$ pour tout θ. Si l'application $\theta \mapsto \sigma(\theta)$ est continue en θ, vu la convergence $\bar{X} \xrightarrow{P} \theta$ et le Théorème de Slutsky, on obtient

$$\sqrt{n}\frac{(\bar{X} - \theta)}{\sigma(\bar{X})} \xrightarrow{D} \mathcal{N}(0,1) \quad \text{quand } n \to \infty. \tag{6.16}$$

Basé sur ce résultat donnant une approximation asymptotique de la loi de la statistique $\sqrt{n}(\bar{X} - \theta)/\sigma(\bar{X})$, on peut proposer le test d'hypothèse $H_0 : \theta = \theta_0$ contre l'alternative $H_1 : \theta > \theta_0$ défini par

$$R = \Big\{ \bar{X} > \theta_0 + \frac{\sigma(\bar{X})}{\sqrt{n}}\, q^N_{1-\alpha} \Big\}.$$

Pour ce test

$$\lim_{n\to\infty} P_{\theta_0}(\mathcal{X}_n \in R) = \alpha.$$

En effet, compte tenu de (6.16),

$$\lim_{n\to\infty} P_{\theta_0}\left(\bar{X} > \theta_0 + \frac{\sigma(\bar{X})}{\sqrt{n}}\, q^N_{1-\alpha} \right) = \lim_{n\to\infty} P_{\theta_0}\left(\frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma(\bar{X})} > q^N_{1-\alpha} \right) = \alpha.$$

Définition 6.10. *Un test R de l'hypothèse $H_0 : \theta \in \Theta_0$ contre l'alternative $H_1 : \theta \in \Theta_1$ est dit* **test de niveau asymptotique** α *si*

$$\sup_{\theta\in\Theta_0} \lim_{n\to\infty} P_\theta(\mathcal{X}_n \in R) \le \alpha.$$

Soit $\widehat{\theta}_n^{MV}$ l'estimateur de maximum de vraisemblance de θ dans un modèle statistique $\{F_\theta, \theta \in \Theta \in \mathbb{R}^k\}$. Si le modèle statistique vérifie les hypothèses du Théorème 5.2, pour tout $\theta \in \Theta$,

$$\widehat{\theta}_n^{MV} \xrightarrow{P} \theta,$$
$$\sqrt{nI(\theta)}(\widehat{\theta}_n^{MV} - \theta) \xrightarrow{D} \mathcal{N}(0,1)$$

quand $n \to \infty$. Si l'information de Fisher $I(\theta)$ est continue sur Θ, par le Premier théorème de continuité (Proposition 1.9),

$$I(\widehat{\theta}_n^{MV}) \xrightarrow{P} I(\theta) \quad \text{quand} \quad n \to \infty,$$

et, vu le Théorème de Slutsky, on obtient

$$\sqrt{nI(\widehat{\theta}_n^{MV})}(\widehat{\theta}_n^{MV} - \theta) \xrightarrow{D} \mathcal{N}(0,1) \quad \text{quand} \quad n \to \infty. \tag{6.17}$$

On peut utiliser (6.17) pour construire un test de niveau asymptotique α des hypothèses classiques considérées précédemment. Par exemple, pour le problème bilatéral,

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \ne \theta_0,$$

on peut définir un test de niveau asymptotique α par

$$R = \left\{ |\widehat{\theta}_n^{MV} - \theta_0| > \frac{1}{\sqrt{nI(\widehat{\theta}_n^{MV})}}\, q^N_{1-\alpha/2} \right\}.$$

On remarque que c'est un test de type (6.12), où on substitue $\widehat{\theta}_n^{MV}$ à $\bar{X}$ et $I(\widehat{\theta}_n^{MV})^{-1}$ à σ^2.

6.6. Tests de comparaison de deux lois normales*

Souvent on cherche à comparer deux lois de probabilité à partir de deux échantillons différents. Supposons ici que ces échantillons sont i.i.d. et issus de deux lois normales : $(X_1^{(1)},\dots,X_{n_1}^{(1)})$ de loi $\mathcal{N}(\mu_1,\sigma_1^2)$, et $(X_1^{(2)},\dots,X_{n_2}^{(2)})$ de loi $\mathcal{N}(\mu_2,\sigma_2^2)$. Supposons aussi l'indépendance des échantillons : $(X_1^{(1)},\dots,X_{n_1}^{(1)}) \perp\!\!\!\perp (X_1^{(2)},\dots,X_{n_2}^{(2)})$.

6.6.1. Test d'égalité des variances. On se place dans le cadre général, où les espérances μ_1 et μ_2 sont inconnues. Considérons le problème de test :

$$H_0:\ \sigma_1^2=\sigma_2^2, \quad H_1:\ \sigma_1^2\neq\sigma_2^2.$$

D'après la Proposition 4.4,

$$\frac{n_1 s_1^2}{\sigma_1^2}\sim\chi^2_{n_1-1} \quad \text{et} \quad \frac{n_2 s_2^2}{\sigma_2^2}\sim\chi^2_{n_2-1}. \tag{6.18}$$

On en déduit que les statistiques

$$\begin{aligned} S_1^2 &= s_1^2\frac{n_1}{n_1-1}=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i^{(1)}-\bar X_1)^2,\\ S_2^2 &= s_1^2\frac{n_2}{n_2-1}=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_i^{(2)}-\bar X_2)^2, \end{aligned}$$

où $\bar X_1=n_1^{-1}\sum_{i=1}^{n_1}X_i^{(1)}$, $\bar X_2=n_2^{-1}\sum_{i=1}^{n_2}X_i^{(2)}$, vérifient

$$\frac{S_1^2}{S_2^2}\,\frac{\sigma_2^2}{\sigma_1^2}\sim F_{n_1-1,n_2-1},$$

où $F_{p,q}$ est la loi de Fisher-Snedecor à degrés de liberté p et q. En particulier, la statistique $U=\frac{S_1^2}{S_2^2}$ suit la loi de Fisher-Snedecor sous l'hypothèse $H_0:\sigma_1^2=\sigma_2^2$. Alors un test de niveau (et de taille) α peut être construit à l'aide de la région critique

$$R=\{U<C_1,\ \ U>C_2\}$$

avec C_1 et C_2 tels que $1-\alpha=P(C_1\le F_{n_1-1,n_2-1}\le C_2)$. Un choix de $\alpha=\alpha_1+\alpha_2$ avec $0<\alpha_1,\alpha_2<1$ étant fait (par exemple, $\alpha_1=\alpha_2=\alpha/2$), on peut prendre C_1 comme le quantile d'ordre α_1 de F_{n_1-1,n_2-1}, noté $q_{\alpha_1}(n_1-1,n_2-1)$ et C_2 comme le quantile d'ordre $1-\alpha_2$ de la même loi, noté $q_{1-\alpha_2}(n_1-1,n_2-1)$. On obtient alors la région critique

$$R=\{U<q_{\alpha_1}(n_1-1,n_2-1),\ \ U>q_{1-\alpha_2}(n_1-1,n_2-1)\}.$$

Si n_1 et n_2 sont grands (par exemple, $n_1,n_2\ge 20$), on utilise souvent un test asymptotique construit comme suit. Notons que l'indépendance de s_1^2 et s_2^2 et le Théorème central limite impliquent :

$$\frac{s_1^2-s_2^2-E(s_1^2-s_2^2)}{\sqrt{\mathrm{Var}(s_1^2-s_2^2)}}\xrightarrow{D}\mathcal{N}(0,1) \quad \text{quand } n_1,n_2\to\infty, \tag{6.19}$$

où $E(s_1^2-s_2^2)=\sigma_1^2-\sigma_2^2$ et $\mathrm{Var}(s_1^2-s_2^2)=\mathrm{Var}(s_1^2)+\mathrm{Var}(s_2^2)=\frac{2\sigma_1^4}{n_1}+\frac{2\sigma_2^4}{n_2}$ (cf. Exercice 4.8). Remplaçons ici $\frac{2\sigma_i^4}{n_i}$ par $\frac{2s_i^4}{n_i}$ car $s_i^2\xrightarrow{P}\sigma_i^2$, $i=1,2$. Vu le Théorème de Slutsky, il en découle la convergence en loi

$$\frac{s_1^2 - s_2^2 - (\sigma_1^2 - \sigma_2^2)}{\sqrt{\frac{2s_1^4}{n_1} + \frac{2s_2^4}{n_2}}} \xrightarrow{D} \mathcal{N}(0,1).$$

Puisque $\sigma_1^2 = \sigma_2^2$ sous H_0, on obtient un test de niveau asymptotique α :

$$R = \left\{ \frac{|s_1^2 - s_2^2|}{\sqrt{\frac{2s_1^4}{n_1} + \frac{2s_2^4}{n_2}}} \geq q^N_{1-\alpha/2} \right\}.$$

6.6.2. Test d'égalité des espérances. Testons maintenant l'hypothèse $H_0 : \mu_1 = \mu_2$ contre l'alternative $H_1 : \mu_1 \neq \mu_2$ au niveau α, en supposant que $\sigma_1 = \sigma_2 = \sigma$ avec $\sigma > 0$ inconnu.

Puisque $(X_1^{(1)}, \ldots, X_{n_1}^{(1)}) \perp\!\!\!\perp (X_1^{(2)}, \ldots, X_{n_2}^{(2)})$, les variables aléatoires $n_1 s_1^2/\sigma^2$ et $n_2 s_2^2/\sigma^2$ sont indépendantes. Soit $n = n_1 + n_2$, alors vu (6.18),

$$\frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2} \sim \chi^2_{n-2}. \tag{6.20}$$

Pour $i = 1, 2$, notons $P_{\mu_i,\sigma}$ les lois de $(X_1^{(i)}, \ldots, X_{n_i}^{(i)})$ quand $\sigma_1 = \sigma_2 = \sigma$. On a alors

$$\text{sous } P_{\mu_1,\sigma},\ \tfrac{\sqrt{n_1}}{\sigma}(\bar{X}_1 - \mu_1) \sim \mathcal{N}(0,1),$$
$$\text{sous } P_{\mu_2,\sigma},\ \tfrac{\sqrt{n_2}}{\sigma}(\bar{X}_2 - \mu_2) \sim \mathcal{N}(0,1),$$

et respectivement

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_1 - \mu_1) \sim \mathcal{N}(0, \frac{n}{n_1}), \quad \frac{\sqrt{n}}{\sigma}(\bar{X}_2 - \mu_2) \sim \mathcal{N}(0, \frac{n}{n_2}).$$

L'indépendance de $\bar{X}_1$ et $\bar{X}_2$ implique que si $\mu_1 = \mu_2$,

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_1 - \bar{X}_2) \sim \mathcal{N}\left(0, \frac{n}{n_1} + \frac{n}{n_2}\right).$$

Par conséquent, sous H_0,

$$\sqrt{\frac{n_1 n_2}{n}} \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma} \sim \mathcal{N}(0,1).$$

Vu (6.20) ceci permet de déduire que la variable aléatoire

$$Z = \sqrt{\frac{n_1 n_2 (n-2)}{n(n_1 s_1^2 + n_2 s_2^2)}} (\bar{X}_1 - \bar{X}_2)$$

suit la loi t_{n-2} (loi de Student à $n-2$ degrés de liberté). On peut donc considérer le test à région critique

$$R = \left\{ |\bar{X}_1 - \bar{X}_2| > C_\alpha \sqrt{\frac{n(n_1 s_1^2 + n_2 s_2^2)}{n_1 n_2 (n-2)}} \right\}.$$

Si l'on choisit ici $C_\alpha = q_{1-\alpha/2}(t_{n-2})$, le quantile d'ordre $1-\alpha$ de la loi t_{n-2}, on obtient un test de taille α.

Comme la f.d.r. de loi de Student t_{n-2} tend vers celle de $\mathcal{N}(0,1)$ quand $n \to \infty$, un test de niveau asymptotique α est donné par la région critique suivante :

$$R^a = \left\{ \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}} > q^N_{1-\alpha/2} \right\}, \tag{6.21}$$

où $q^N_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0,1)$.

6.7. Régions de confiance

Le modèle statistique ici est, comme précédemment, $\{F_\theta,\, \theta \in \Theta\}$, $\Theta \in \mathbb{R}^k$, et $\mathcal{X}_n = (X_1, ..., X_n)$ est l'échantillon observé.

Définition 6.11. *Soit* $0 < \alpha < 1$. *Une* **région de confiance de niveau** $1 - \alpha$ *pour* θ *est un ensemble aléatoire* $\mathcal{C}(\mathcal{X}_n) \subseteq \mathbb{R}^k$ *tel que, pour tout* $\theta \in \Theta$,

$$P_\theta(\theta \in \mathcal{C}(\mathcal{X}_n)) \geq 1 - \alpha.$$

On dit que $\mathcal{C}(\mathcal{X}_n)$ *est* **une région de confiance de taille** $1 - \alpha$ *pour* θ *si, pour tout* $\theta \in \Theta$,

$$P_\theta(\theta \in \mathcal{C}(\mathcal{X}_n)) = 1 - \alpha.$$

Dans le cas unidimensionnel ($k = 1$) on utilise le plus souvent les régions de confiance de forme particulière, notamment les *intervalles de confiance*. Un intervalle de confiance de niveau $1 - \alpha$ est un intervalle de la forme

$$\mathcal{C}(\mathcal{X}_n) = [a(\mathcal{X}_n), b(\mathcal{X}_n)],$$

où $a(\cdot)$ et $b(\cdot)$ sont des fonctions boréliennes à valeurs dans $\mathbb{R}$, telles que $a(\mathcal{X}_n) < b(\mathcal{X}_n)$ pour tout $\mathcal{X}_n$ et

$$P_\theta(a(\mathcal{X}_n) \leq \theta \leq b(\mathcal{X}_n)) \geq 1 - \alpha$$

pour tout $\theta \in \Theta$.

Exemple 6.5. *Intervalle de confiance de niveau* $1-\alpha$ *pour* θ *dans le modèle* $\{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$ *avec* $\sigma > 0$ *connu.* Considérons l'intervalle aléatoire

$$\mathcal{C}(\mathcal{X}_n) = \left[\bar{X} - \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2} \right], \tag{6.22}$$

i.e. posons $a(\mathcal{X}_n) = \bar{X} - \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}$ et $b(\mathcal{X}_n) = \bar{X} + \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2}$. C'est un intervalle de confiance de taille $1 - \alpha$ pour θ, car

$$P_\theta(\theta \in \mathcal{C}(\mathcal{X}_n)) = P_\theta \left(|\bar{X} - \theta| \leq \frac{\sigma}{\sqrt{n}} q^N_{1-\alpha/2} \right) = P(|\eta| \leq q^N_{1-\alpha/2}) = 1 - \alpha.$$

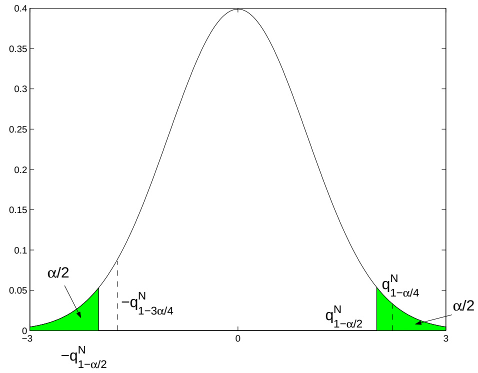

Fig. 6.9. Quantiles de la loi $\mathcal{N}(0,1)$ correspondant aux intervalles symétrique et non-symétique.

Un autre exemple d'intervalle de confiance de niveau $1-\alpha$ est un intervalle non-symétrique de type

$$\mathcal{C}(\mathcal{X}_n) = \Big[\bar{X} - \frac{\sigma}{\sqrt{n}}\, q^N_{1-3\alpha/4}, \bar{X} + \frac{\sigma}{\sqrt{n}}\, q^N_{1-\alpha/4}\Big].$$

On peut montrer que cet intervalle est de longueur plus grande que l'intervalle symétrique (6.22). Il permet donc de localiser la vraie valeur du paramètre θ avec moins de précision que l'intervalle (6.22). La même remarque reste vraie pour d'autres intervalles non-symétriques et ceci explique pourquoi ils ne sont pas intéressants dans cet exemple.

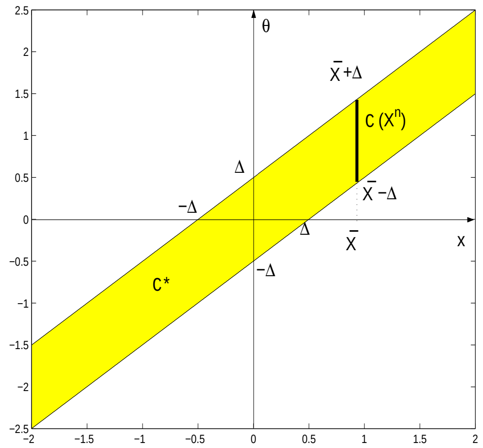

Fig. 6.10. Diagramme Tests/IC : $\mathcal{C}^* = \{(x,\theta) : |x-\theta| \le \Delta\}$ avec $\Delta = \frac{\sigma}{\sqrt{n}}\, q^N_{1-\alpha/2}$.

Considérons maintenant un outil graphique que l'on appelle *diagramme Tests/Intervalles de confiance* (en abrégé *diagramme Tests/IC*). Sur le plan (x,θ) dans $\mathbb{R}^2$ introduisons la région

$$\mathcal{C}^* = \{(x,\theta) : \ |x-\theta| \le \Delta\}$$

où $\Delta = \frac{\sigma}{\sqrt{n}}\, q^N_{1-\alpha/2}$. Les sections verticales de cette région par les droites $x = \bar{X}$ représentent les intervalles de confiance de niveau $1-\alpha$ obtenus pour différentes valeurs de $\bar{X}$. Les sections horizontales de $\mathcal{C}^*$ par les droites $\theta = \theta_0$ représentent les régions d'acceptation $A(\theta_0)$ des tests de niveau α des hypothèses de la forme $H_0 : \ \theta = \theta_0$ contre $H_1 : \ \theta \ne \theta_0$.

Remarque. Le diamètre $|\mathcal{C}(\mathcal{X}_n)| \to 0$ quand $n \to \infty$. Par contre, $|\mathcal{C}(\mathcal{X}_n)|$ grandit quand $\alpha \to 0$.

6.8. Méthodes de construction des régions de confiance

Nous examinerons ici trois méthodes différentes de construction de régions de confiance en dimension 1 (i.e. quand $\Theta \subseteq \mathbb{R}$).

6.8.1. Intervalles de confiance pour n fini : méthode des fonctions pivotales. Cete méthode ne s'applique que pour quelques modèles statistiques très particuliers. Supposons qu'il existe $S_n(X_1, \ldots, X_n, \theta) = S_n(\mathcal{X}_n, \theta)$ (une fonction borélienne de $X_1, \ldots, X_n, \theta$) telle que, pour tout $t \in \mathbb{R}$, la probabilité

$$P_\theta(S_n(\mathcal{X}_n, \theta) \leq t)$$

ne dépend pas de θ. Si cette condition est vérifiée, on appelle $\theta \mapsto S_n(\mathcal{X}_n, \theta)$ **fonction pivotale** (ou **pivot**) pour le modèle statistique $\{F_\theta, \theta \in \Theta\}$. Notons qu'une fonction pivotale n'est pas une statistique, car elle dépend du paramètre θ. Pour une fonction pivotale $S_n(\mathcal{X}_n, \theta)$, il existe $\Delta_1(\alpha)$, $\Delta_2(\alpha)$ ne dépendant pas de θ, tels que

$$P_\theta(\Delta_1(\alpha) \leq S_n(\mathcal{X}_n, \theta) \leq \Delta_2(\alpha)) \geq 1 - \alpha$$

pour tout $\theta \in \Theta$. Cette inégalité signifie que

$$\mathcal{C}(\mathcal{X}_n) = \{\theta : \Delta_1(\alpha) \leq S_n(\mathcal{X}_n, \theta) \leq \Delta_2(\alpha)\}$$

est une région de confiance de niveau $1 - \alpha$ pour θ. Dans l'Exemple 6.5, la fonction pivotale est donnée par $S_n(\mathcal{X}_n, \theta) = \sqrt{n}(\bar{X} - \theta)/\sigma$. Puisque la v.a. $\sqrt{n}(\bar{X} - \theta)/\sigma$ est distribuée selon la loi normale $\mathcal{N}(0, 1)$, sous P_θ, indépendamment de θ, les quantités $\Delta_1(\alpha)$ et $\Delta_2(\alpha)$ peuvent être choisies sous la forme : $\Delta_1(\alpha) = -q^N_{1-\alpha/2}$, $\Delta_2(\alpha) = q^N_{1-\alpha/2}$.

Plus généralement, les inégalités $\Delta_1(\alpha) \leq S_n(\mathcal{X}_n, \theta) \leq \Delta_2(\alpha)$ définissent une région de confiance pour θ de façon implicite : pour l'expliciter il faut résoudre le système de ces deux inégalités par rapport à θ, ce qui n'est pas toujours facile. Néanmoins, il existe quelques exemples remarquables pour lesquels cette démarche mène au succès, dont le suivant.

Exemple 6.6. *Intervalle de confiance pour θ dans le modèle exponentiel $\mathcal{E}(\theta)$.* Soit le modèle statistique $\{F_\theta,\ \theta > 0\}$, où F_θ est la f.d.r. de densité $f_\theta(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} I\{x > 0\}$. Notons que la variable aléatoire $Y = \frac{2}{\theta} X$ suit la loi χ^2_2. Ceci implique que la v.a.

$$Z = \frac{2}{\theta} \sum_{i=1}^n X_i = \frac{2n}{\theta} \bar{X}$$

est distribuée selon la loi χ^2_{2n} sous P_θ, indépendamment de θ. On voit donc que $S_n(\mathcal{X}_n, \theta) = \frac{2n}{\theta} \bar{X}$ est un pivot et qu'on peut trouver Δ_1 et Δ_2 tels que, pour tout $\theta > 0$,

$$P_\theta(\Delta_1 \leq S_n(\mathcal{X}_n, \theta) \leq \Delta_2) = 1 - \alpha.$$

En effet, on peut choisir, par exemple, $\Delta_1 = q_{\alpha/2}(\chi^2_{2n})$ et $\Delta_2 = q_{1-\alpha/2}(\chi^2_{2n})$, les quantiles d'ordre $\alpha/2$ et $1 - \alpha/2$ de la loi χ^2_{2n}. Alors, l'ensemble

$$\mathcal{C}^*(\mathcal{X}_n) = \left\{\theta : q_{\alpha/2}(\chi^2_{2n}) \leq \frac{2n}{\theta} \bar{X} \leq q_{1-\alpha/2}(\chi^2_{2n})\right\} = \left\{\theta : \frac{2n\bar{X}}{q_{1-\alpha/2}(\chi^2_{2n})} \leq \theta \leq \frac{2n\bar{X}}{q_{\alpha/2}(\chi^2_{2n})}\right\}$$

est un intervalle de confiance de niveau (et de taille) $1-\alpha$ pour θ. La figure ci-dessous présente la diagramme Tests/IC pour cet exemple permettant une construction graphique de $\mathcal{C}^*(\mathcal{X}_n)$.

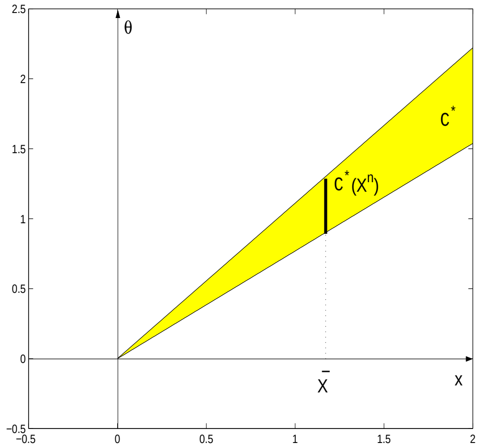

Fig. 6.11. Diagramme Tests/IC : $\mathcal{C}^* = \left\{(x,\theta) : \frac{2nx}{q_{1-\alpha/2}(\chi^2_{2n})} \le \theta \le \frac{2nx}{q_{\alpha/2}(\chi^2_{2n})}\right\}$.

Pour quelques exemples, on peut définir un pivot en utilisant la démarche suivante. Soit $\widehat{\theta}_n$ une statistique. Posons

$$G_\theta(x) = P_\theta(\widehat{\theta}_n \le x).$$

Supposons que les hypothèses suivantes soient vérifiées :
– la fonction $G_\theta(x)$ est monotone en θ pour tout x fixé ;
– la fonction $G_\theta(x)$ est continue en x pour tout θ fixé.

Définissons la pivot $S_n(\mathcal{X}_n,\theta) \stackrel{\text{déf}}{=} G_\theta(\widehat{\theta}_n)$. La loi de $S_n(\mathcal{X}_n,\theta)$ sous P_θ est alors uniforme. En particulier,

$$P_\theta\left(\frac{\alpha}{2} \le G_\theta(\widehat{\theta}_n) \le 1-\frac{\alpha}{2}\right) = 1-\alpha.$$

Si $G_\theta(x)$ est une fonction croissante de θ, et ceci pour tout x fixé, alors :

$$P_\theta\left(\frac{\alpha}{2} \le G_\theta(\widehat{\theta}_n) \le 1-\frac{\alpha}{2}\right) = P_\theta\left(b_{\frac{\alpha}{2}}(\widehat{\theta}_n) \le \theta \le b_{1-\frac{\alpha}{2}}(\widehat{\theta}_n)\right),$$

où $b_\alpha(x)$ est tel que $G_{b_\alpha(x)} = \alpha$.

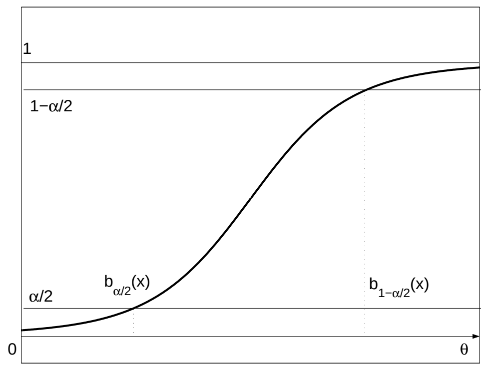

Fig. 6.12. La fonction $\theta \mapsto G_\theta(x)$ pour un x fixé.

On voit donc que $\mathcal{C}(\mathcal{X}_n) = [b_{\frac{\alpha}{2}}(\widehat{\theta}_n), b_{1-\frac{\alpha}{2}}(\widehat{\theta}_n)]$ est un intervalle de confiance de niveau (et de taille) $1-\alpha$ pour θ sous les hypothèses ci-dessus.

Par un raisonnement similaire on obtient que si $G_\theta(x)$ est une fonction monotone décroissante de θ pour tout x fixé, $\mathcal{C}(\mathcal{X}_n) = [b_{1-\frac{\alpha}{2}}(\widehat{\theta}_n), b_{\frac{\alpha}{2}}(\widehat{\theta}_n)]$ est un intervalle de confiance de taille $1-\alpha$ pour θ. Nous avons donc démontré la proposition suivante.

Proposition 6.1. *Soit $G_\theta(x)$ continue en x et monotone comme fonction de θ pour tout x. Alors l'intervalle*

$$\mathcal{C} = [\underline{\theta}, \overline{\theta}], \quad \textit{où} \quad \begin{cases} \underline{\theta} = \min(b_{1-\frac{\alpha}{2}}(\widehat{\theta}_n), b_{\frac{\alpha}{2}}(\widehat{\theta}_n)), \\ \overline{\theta} = \max(b_{1-\frac{\alpha}{2}}(\widehat{\theta}_n), b_{\frac{\alpha}{2}}(\widehat{\theta}_n)) \end{cases}$$

est un intervalle de confiance de taille $1-\alpha$ *pour* θ*.*

Exemple 6.7. *Intervalle de confiance pour θ dans le modèle de Bernoulli $\mathcal{B}e(\theta)$.* Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de loi de Bernoulli, $X_i \sim \mathcal{B}e(\theta)$ avec $0 < \theta < 1$. Posons $\widehat{\theta}_n = \bar{X}$ et

$$G_\theta(x) = P_\theta(\bar{X} \le x) = \sum_{k \le nx} C_n^k \theta^k (1-\theta)^{n-k}.$$

Il est facile de voir que $\theta \mapsto G_\theta(x)$ est une fonction décroissante pour tout x. Or, l'application $x \mapsto G_\theta(x)$ n'est pas continue. Néanmoins, on peut construire des intervalles de confiance de la même manière que dans la Proposition 6.1, en utilisant la monotonie. La seule différence est dans le fait que les intervalles ainsi obtenus ne sont plus de taille $1-\alpha$, mais seulement de niveau $1-\alpha$. Considérons l'application numérique avec $\alpha = 0.05$, $\bar{X} = 0.3$ et $n = 100$. L'utilisation des tables spéciales donnant les intervalles de confiance basés sur la loi exacte de $n\bar{X}$ (qui est la loi binomiale $\mathcal{B}(n, \theta)$) nous amène à l'intervalle de confiance de niveau $0,95$ suivant : $\mathcal{C}^0(\mathcal{X}_{100}) = [0.2124, 0.3998]$.

6.8.2. Intervalles de confiance pour n fini : utilisation des inégalités. Cette méthode de construction des intervalles de confiance est basée sur l'application des diverses inégalités probabilistes, par exemple, celle de Tchebychev.

Plaçons-nous dans le cadre de l'Exemple 6.7. On remarque que

$$P_\theta(|\bar{X} - \theta| \le \Delta) = 1 - P_\theta(|\bar{X} - \theta| > \Delta).$$

Cherchons un intervalle de confiance de la forme $\mathcal{C}(\mathcal{X}_n) = [\bar{X} - \Delta, \bar{X} + \Delta]$ avec $\Delta > 0$. Comme les X_i sont indépendants de moyenne θ et de variance $\theta(1-\theta)$ sous P_θ, l'inégalité de Tchebychev donne la borne

$$P_\theta(|\bar{X} - \theta| > \Delta) \le \frac{E_\theta(|\bar{X} - \theta|^2)}{\Delta^2} = \frac{1}{n\Delta^2} E_\theta((X_1 - \theta)^2) = \frac{\theta(1-\theta)}{n\Delta^2}.$$

On obtient, pour tout θ tel que $0 < \theta < 1$,

$$P_\theta\left(|\bar{X} - \theta| \le \Delta\right) \ge 1 - \frac{\theta(1-\theta)}{n\Delta^2} \ge 1 - \frac{1}{4n\Delta^2}.$$

Ceci nous permet de déterminer la valeur de Δ_α telle que $1 - \frac{1}{4n\Delta_\alpha^2} = 1 - \alpha$ et de construire finalement un intervalle de confiance de niveau $1-\alpha$:

$$\mathcal{C}(\mathcal{X}_n) = [\bar{X} - \Delta_\alpha, \bar{X} + \Delta_\alpha].$$

Pour la même application numérique que dans l'Exemple 6.7, on a $\Delta_\alpha = 0.2236$ et cet intervalle est de la forme :

$$\mathcal{C}^{Tcheb}(\mathcal{X}_{100}) = [\bar{X} - 0.2236, \bar{X} + 0.2236] = [0.0763, 0.5236].$$

L'intervalle de confiance $\mathcal{C}^{Tcheb}(\mathcal{X}_{100})$ est *plus conservatif* (i. e. moins précis) que l'intervalle de confiance $\mathcal{C}^0(\mathcal{X}_{100}) = [0.2124, 0.3998]$ obtenu dans l'Exemple 6.7 à l'aide de la méthode des fonctions pivotales.

6.8.3. Intervalles de confiance asymptotiques. Sous les hypothèses de l'Exemple 6.7 on obtient :

$$\frac{\sqrt{n}}{\sigma(\theta)}(\bar{X} - \theta) \xrightarrow{D} \mathcal{N}(0,1) \quad \text{quand } n \to \infty,$$

où $\sigma^2(\theta) = E_\theta((X_1-\theta)^2) = \theta(1-\theta)$. La fonction $\sqrt{x(1-x)}$ est continue, donc, par le Premier théorème de continuité, $\sigma(\bar{X}) = \sqrt{\bar{X}(1-\bar{X})} \xrightarrow{P} \sigma(\theta)$ quand $n \to \infty$. On en déduit que

$$\sqrt{\frac{n}{\bar{X}(1-\bar{X})}}(\bar{X} - \theta) \xrightarrow{D} \mathcal{N}(0,1) \quad \text{quand } n \to \infty,$$

et pour tout $\Delta > 0$

$$P_\theta\left(\left|\sqrt{\frac{n}{\bar{X}(1-\bar{X})}}(\bar{X} - \theta)\right| \le \Delta\right) \to P\left(|\eta| \le \Delta\right), \tag{6.23}$$

où $\eta \sim \mathcal{N}(0,1)$. Vu la forme du membre de droite dans (6.23), on peut choisir $\Delta = q^N_{1-\alpha/2}$, ce qui implique que l'ensemble de tous les θ tels que

$$\left|\sqrt{\frac{n}{\bar{X}(1-\bar{X})}}(\bar{X} - \theta)\right| \le q^N_{1-\alpha/2},$$

est un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour θ. On notera cet intervalle $\mathcal{C}^a(\mathcal{X}_n)$:

$$\mathcal{C}^a(\mathcal{X}_n) = \left[\bar{X} - \frac{\sqrt{\bar{X}(1-\bar{X})}}{\sqrt{n}}\, q^N_{1-\alpha/2}, \bar{X} + \frac{\sqrt{\bar{X}(1-\bar{X})}}{\sqrt{n}}\, q^N_{1-\alpha/2}\right].$$

L'intervalle de confiance asymptotique $\mathcal{C}^a(\mathcal{X}_n)$ vérifie

$$\lim_{n\to\infty} P_\theta(\theta \in \mathcal{C}^a(\mathcal{X}_n)) = P(|\eta| \le q^N_{1-\alpha/2}) = 1 - \alpha \quad \text{pour tout} \quad 0 < \theta < 1. \tag{6.24}$$

Si $\alpha = 0.05$, alors $q^N_{1-\alpha/2} = q^N_{0.975} \approx 1.96$ et, pour l'application numérique de l'Exemple 6.7,

$$\mathcal{C}^a(\mathcal{X}_{100}) = [0.2102, 0.3898].$$

On voit donc que l'intervalle asymptotique $\mathcal{C}^a(\mathcal{X}_{100})$ est essentiellement le même que l'intervalle $\mathcal{C}^0(\mathcal{X}_{100})$ basé sur la loi exacte. Ils sont plus courts que l'intervalle $\mathcal{C}^{Tcheb}(\mathcal{X}_{100})$. Néanmoins $\mathcal{C}^a(\mathcal{X}_n)$ n'est pas nécessairement un intervalle de confiance de niveau $1 - \alpha$ pour n fini, i.e. il peut ne pas vérifier la Définition 6.11. C'est un intervalle de confiance de niveau *asymptotique* $1 - \alpha$ au sens de (6.24). Pour avoir une idée, à partir de quel n les intervalles de confiance asymptotiques deviennent valables, considérons l'application numérique avec les mêmes valeurs $\alpha = 0.05$ et $\bar{X} = 0.3$ que précédemment, mais avec la taille d'échantillon n plus petite. Les résultats sont donnés dans le tableau suivant.

n	10	20	30
$\mathcal{C}^0(\mathcal{X}_n)$	[0.0667, 0.6525]	[0.1189, 0.5428]	[0.1473, 0.4940]
$\mathcal{C}^a(\mathcal{X}_n)$	[0.0159, 0.5840]	[0.0992, 0.5008]	[0.1360, 0.4640]

Sans surprise, les intervalles deviennent de plus en plus courts quand n croît. De plus, $\mathcal{C}^0$ et $\mathcal{C}^a$ se rapprochent. Cependant, les intervalles asymptotiques $\mathcal{C}^a$ sont toujours plus courts que les intervalles $\mathcal{C}^0$ basés sur la loi exacte et ils sont un peu biaisés vers la gauche par rapport à ces derniers. En conclusion, l'approximation asymptotique peut s'avérer trompeuse pour $n \leq 30$: il est plus prudent d'utiliser l'intervalle $\mathcal{C}^0$. Par contre, pour $n = 100$, comme on l'a déjà vu, la différence entre les deux intervalles devient négligeable, ce qui signifie que l'utilisation des intervalles asymptotiques est bien fondée.

L'approche asymptotique est la plus répandue dans la pratique, car elle permet, pour n assez grand, d'obtenir facilement de bons intervalles de confiance pour plusieurs modèles statistiques. Nous allons maintenant donner la définition générale sur laquelle est basée cette approche.

Définition 6.12. *Un ensemble $\mathcal{C}^a(\mathcal{X}_n)$ est dit* **région de confiance de niveau asymptotique** $1-\alpha$ *pour θ si, pour tout $\theta \in \Theta$,*

$$\liminf_{n\to\infty} P_\theta(\theta \in \mathcal{C}^a(\mathcal{X}_n)) \geq 1-\alpha.$$

Comme pour les tests asymptotiques, on peut utiliser la normalité asymptotique des statistiques classiques pour construire des intervalles de confiance de niveau asymptotique $1-\alpha$. Une approche possible est de fonder l'intervalle de confiance sur l'estimateur du maximum de vraisemblance $\widehat{\theta}_n^{MV}$ de θ qui, sous les hypothèses de régularité, satisfait

$$\sqrt{nI(\theta)}(\widehat{\theta}_n^{MV} - \theta) \xrightarrow{D} \mathcal{N}(0,1) \quad \text{quand} \quad n \to \infty,$$

pour tout $\theta \in \Theta$ (cf. Théorème 5.2). Sous les hypothèses de régularité, comme il a été expliqué au Paragraphe 6.5, nous avons aussi

$$\sqrt{nI(\widehat{\theta}_n^{MV})}(\widehat{\theta}_n^{MV} - \theta) \xrightarrow{D} \mathcal{N}(0,1) \quad \text{quand} \quad n \to \infty,$$

ce qui permet d'obtenir l'intervalle de confiance de niveau asymptotique $1-\alpha$ sous la forme

$$\mathcal{C}^a(\mathcal{X}_n) = \left[\widehat{\theta}_n^{MV} - \frac{q_{1-\alpha/2}^N}{\sqrt{nI(\widehat{\theta}_n^{MV})}},\ \widehat{\theta}_n^{MV} + \frac{q_{1-\alpha/2}^N}{\sqrt{nI(\widehat{\theta}_n^{MV})}}\right].$$

Pour l'Exemple 6.7 on a : $\widehat{\theta}_n^{MV} = \bar{X}$ et $I(\theta) = (\theta(1-\theta))^{-1}$, donc $I(\widehat{\theta}_n^{MV}) = (\bar{X}(1-\bar{X}))^{-1}$.

6.9. Dualité entre tests et régions de confiance

Considérons d'abord le modèle statistique $\{\mathcal{N}(\theta, \sigma^2),\, \theta \in \mathbb{R}\}$ avec $\sigma > 0$ connu et définissons les ensembles

$$A(\theta_0) = \Big\{\bar{X} :\ |\theta_0 - \bar{X}| \leq \frac{\sigma}{\sqrt{n}}\, q^N_{1-\alpha/2}\Big\},$$
$$R(\theta_0) = \Big\{\bar{X} :\ |\theta_0 - \bar{X}| > \frac{\sigma}{\sqrt{n}}\, q^N_{1-\alpha/2}\Big\} = A^c(\theta_0),$$

où A^c désigne le complémentaire de A.

L'ensemble $R(\theta_0)$ est la région critique d'un test de niveau α de l'hypothèse $H_0 : \theta = \theta_0$ contre l'alternative $H_1 : \theta \neq \theta_0$, $A(\theta_0)$ est donc la région d'acceptation associée à ce test. Comme il a été expliqué précédemment, $\mathcal{C}(X^n)$ et $A(\theta_0)$ peuvent être obtenus à l'aide de la diagramme Tests/IC.

Plus généralement, on a le résultat suivant qui explique les propriétés de la diagramme Tests/IC.

Théorème 6.2.

(i) Si pour tout $\theta_0 \in \Theta$ il existe un test $R(\theta_0)$ de niveau α de l'hypothèse simple $H_0 : \theta = \theta_0$ contre l'alternative $H_1 : \theta \neq \theta_0$, alors

$$\mathcal{C}(\mathcal{X}_n) = \{\theta :\ \mathcal{X}_n \in A(\theta)\}, \quad \textit{où}\ \ A(\theta) = R^c(\theta),$$

est une région de confiance de niveau $1-\alpha$ pour θ.

(ii) Soit $\mathcal{C}(\mathcal{X}_n)$ une région de confiance de niveau $1-\alpha$ pour θ. Alors pour tout $\theta_0 \in \Theta$, le test de l'hypothèse simple $H_0 : \theta = \theta_0$ contre l'alternative $H_1 : \theta \neq \theta_0$ ayant la région critique $R(\theta_0) = A^c(\theta_0)$, où

$$A(\theta_0) = \{\mathcal{X}_n :\ \theta_0 \in \mathcal{C}(\mathcal{X}_n)\}$$

est un test de niveau α.

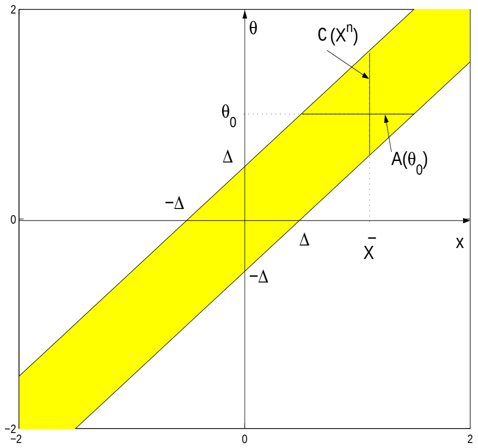

Fig. 6.13. Diagramme Tests/IC. L'intervalle de confiance $\mathcal{C}(X^n)$ pour θ et la région d'acceptation $A(\theta_0)$ du test.

Preuve. (i) On vérifie facilement que pour tout $\theta \in \Theta$,

$$P_\theta(\theta \in \mathcal{C}(\mathcal{X}_n)) = P_\theta(\mathcal{X}_n \in A(\theta)) = 1 - P_\theta(\mathcal{X}_n \in R(\theta)) \geq 1 - \alpha.$$

(ii) Pour montrer le réciproque, il suffit de noter que, pour tout $\theta_0 \in \Theta$,

$$P_{\theta_0}(\mathcal{X}_n \in R(\theta_0)) = 1 - P_{\theta_0}(\mathcal{X}_n \in A(\theta_0)) = 1 - P_{\theta_0}(\theta_0 \in \mathcal{C}(\mathcal{X}_n)) \leq \alpha,$$

donc le test $R(\theta_0)$ est effectivement de niveau α. ∎

6.10. Exercices

Exercice 6.2. On observe X_1, de loi $U[0,1]$ sous H_0, ou $U[2,3]$ sous H_1. Proposer un test de l'hypothèse H_0 contre l'alternative H_1 et calculer ses risques de première et seconde espèce.

Exercice 6.3. Soit $(X_1, \dots, X_n)$ un échantillon i.i.d. de la loi uniforme $U[0,\theta]$, $\theta > 0$. On souhaite tester l'hypothèse $H_0 : \theta = \theta_0$ contre l'alternative $H_1 : \theta < \theta_0$, où $\theta_0 > 0$. Montrer que le test à région critique

$$R = \{X_{(n)} \leq \theta_0 \alpha^{1/n}\}$$

est UPP de niveau α.

Exercice 6.4. La limite légale d'un polluant contenu dans les déchets d'une usine est de 6mg/kg. On effectue un dosage sur 12 prélèvements, pour lesquels on observe une moyenne de 7mg/kg avec un écart-type de 2.4mg/kg. On admet que la loi de dosage est gaussienne.
1°. Préciser le modèle statistique et poser le problème de test d'hypothèses.
2°. Quel test ferait le directeur de cette usine ? Quelle serait sa conclusion ?
3°. Sachant que si la moyenne est supérieure à 8 mg/kg, il y a danger, quel test ferait le député écologiste de la région ou se situe cette usine ? Quelle serait sa conclusion ?
4°. Commenter les résultats de 2° et 3° en utilisant la notion de *p-value*.

Exercice 6.5. Soient $X_1, \dots, X_n$ des variables aléatoires i.i.d. dont la loi admet la densité $f(x - \theta)$, où $f(x) = 2(1 - x)I\{0 \leq x \leq 1\}$. On veut tester l'hypothèse $H_0 : \theta \geq 1$ contre l'alternative $H_1 : \theta < 1$. Introduisons les régions critiques

$$R_c = \{X_{(1)} < c\}$$

et

$$\tilde{R}_c = \{X_{(n)} < c\}.$$

Le but de cet exercice est de comparer le test basé sur R_c avec celui basé sur $\tilde{R}_c$.
1°. Calculer la fonction puissance π associée à R_c et montrer que cette fonction est monotone.
2°. Quelle valeur critique c faut-il choisir pour que le test associé à R_c soit de niveau 5% ?
3° Calculer la fonction puissance $\tilde{\pi}$ associée à $\tilde{R}_c$, où c est choisi de telle façon que le test soit de niveau 5%.
4°. Comparer les fonctions puissance π et $\tilde{\pi}$ pour les tests de niveau 5%. Peut–on affirmer qu'un de ces tests est plus puissant que l'autre ?
5° Analyser l'asymptotique de π et $\tilde{\pi}$ quand $n \to \infty$ et c reste fixé.

Exercice 6.6. Un client de supermarché a pesé 16 paquets de café de même marque de poids nominal 500g. Les résultats des mesures sont les suivants :

$$487.5,\ 500.1,\ 480.3,\ 519.8,\ 470.3,\ 500.2,\ 485.2,\ 499.4,$$
$$499.7,\ 503.1,\ 504.9,\ 480.7,\ 505.1,\ 494.7,\ 488.3,\ 473.3$$

avec $\bar{X} = 493.29$, $s = 12.82$.

On admet que les poids des paquets forment un échantillon d'une loi normale de moyenne $\mu > 0$ et d'écart-type σ.

1°. Faire un test de contrôle de qualité sur la moyenne ($H_0 : \ \mu = 500$; $H_1 : \ \mu \neq 500$). Calculer le seuil critique (*p-value*) de ce test. Conclusion ?
2° Le client qui a pesé les paquets fait son propre test dont les hypothèses sont $H_0 : \ \mu \le 490$ et $H_1 : \ \mu > 490$. Calculer le seuil critique de ce test et commenter le résultat.
3°. On souhaite maintenant tester si l'écart-type σ dépasse le seuil autorisé de 20g. Effectuer un test de niveau 0.05.

Exercice 6.7. On admet que la durée de vie, exprimée avec une unité de temps convenablement choisie, d'un certain type de matériel est représentée par une variable aléatoire X suivant une loi de Weibull de paramètres θ, a et c strictement positifs. Cette loi, notée $W(\theta, a, c)$ a pour fonction de répartition

$$F(x) = \left[1 - \exp\left(-\frac{(x-a)^c}{\theta}\right)\right] I\{x > a\},$$

et donc elle admet la densité

$$f(x) = \frac{c}{\theta}(x-a)^{c-1} \exp\left(-\frac{(x-a)^c}{\theta}\right) I\{x > a\}.$$

1°. Montrer que la variable aléatoire $Y = \frac{2}{\theta}(X-a)^c$ suit la loi χ^2_2.
2°. Que représente le paramètre a ? Pour $x > a$, on appelle *taux de panne instantanné* à l'instant x la quantité

$$\tau(x) = \lim_{\Delta x \downarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x [1 - F(x)]} = \frac{f(x)}{1 - F(x)}.$$

Quelle interprétation peut–on en donner ? Déduire la valeur de $\tau(x)$. Pour quelles valeurs des paramètres θ et c, ce taux sera–t–il constant ? proportionnel à $(x-a)$?

Dans la suite on supposera connus les paramètres a et c de la loi $W(\theta, a, c)$, le paramètre θ étant inconnu. De plus, on disposera d'un échantillon $(X_1, \dots, X_n)$ des durées de vie observées sur n matériels du type considéré, les X_i étant des réalisations i.i.d. de la variable aléatoire X.

3°. Montrer que l'estimateur du maximum de vraisemblance de θ est

$$\hat{\theta}_n^{MV} = \frac{1}{n} \sum_{i=1}^{n} (X_i - a)^c.$$

Cet estimateur est–il consistant ?

4°. Construire un intervalle de confiance pour θ de niveau 90%, puis l'intervalle de confiance de niveau asymptotique 90%.

5°. Considérons le problème de test de l'hypothèse simple $H_0 : \ \theta = \theta_0$ contre l'alternative simple $H_1 : \ \theta = \theta_1$, vérifiant $\theta_1 > \theta_0 > 0$.
5.1°. Montrer que le lemme de Neyman–Pearson conduit à une région critique de la forme

$$R = \{(X_1, \ldots, X_n) : \sum_{i=1}^{n} (X_i - a)^c \geq k\},$$

où $k > 0$ est une constante.
5.2°. Soit $0 < \alpha < 1$ un niveau de signification donné et soit $q_\alpha(\chi^2_{2n})$ le quantile d'ordre α de la loi χ^2_{2n}. Déterminer, en fonction de θ_0 et de $q_\alpha(\chi^2_{2n})$, la région critique d'un test de niveau α.
5.3°. Exprimer, en fonction de θ_0, θ_1 et de $q_\alpha(\chi^2_{2n})$ et à l'aide de la fonction de répartition F de la loi χ^2_{2n}, le risque de seconde espèce et la puissance de ce test.
5.4°. Préciser comment varient la région critique, le risque de seconde espèce et la puissance de ce test en fonction de α, puis en fonction de θ_1.

6°. On considère maintenant le problème de test de l'hypothèse $H_0 : \ \theta \leq 1$ contre l'alternative $H_1 : \theta > 1$.
6.1°. Proposer un test uniformément plus puissant de niveau α.
6.2°. Soit $a = 0, \ c = 1/2, \ n = 15$ et $\sum_{i=1}^{n} X_i^{1/2} = 20,23$. Tester H_0 au niveau $\alpha = 0,05, \ \alpha = 0,1$. Calculer le seuil critique (*p-value*) de ce test.
6.3°. En utilisant la loi limite de $\hat{\theta}_n^{MV}$, proposer un test de niveau asymptotique α. Avec les mêmes valeurs numériques que dans la question 6.2°, tester H_0 au niveau $\alpha = 0,05$ et $\alpha = 0,1$. Comparer les résultats avec ceux des tests non–asymptotiques de 6.2°.

Exercice 6.8. Les résultats d'un examen noté sur 20 sont les suivants.

$$\text{Filles} : 2; 12,5; 4; 4; 2; 15,5; 10; 17,5; 11; 12,5; 2.$$

$$\text{Garçons} : 7; 6; 6; 8; 9; 10; 9; 11; 16,5; 16,5; 14; 12; 4; 11; 2.$$

On suppose que les notes sont des variables aléatoires indépendantes, identiquement distribuées selon la loi $\mathcal{N}(\mu_F, \sigma^2)$ pour les filles et selon la loi $\mathcal{N}(\mu_G, \sigma^2)$ pour les garçons. Les paramètres μ_F, μ_G et σ^2 sont inconnus. Dans la suite, on notera n_F et n_G le nombre de filles et de garçons.

Le but de l'exercice est de tester l'hypothèse que les notes des filles sont en moyenne de même niveau que celles des garçons, i.e.

$$H_0 \ : \ \mu_F = \mu_G,$$

contre l'alternative

$$H_1 : \ \mu_F \neq \mu_G.$$

1°. Proposer un estimateur de $\mu_F - \mu_G$. Proposer un estimateur $\hat{\sigma}^2$ convergeant vers σ^2 à la fois sous H_0 et sous H_1. Donner la loi limite de $\sqrt{n_F n_G/(n_F + n_G)}(\hat{\mu}_F - \hat{\mu}_G)/\hat{\sigma}$ lorsque n_F et n_G tendent simultanément vers $+\infty$.
2°. A partir des résultats de la question précédente, proposer un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour $\mu_F - \mu_G$.
3°. Construire un test de niveau asymptotique α. Quelle est la *p-value* de ce test ? Accepte–t–on H_0 au niveau 0,10 ; 0,05 ; 0,01 ? Commenter le résultat. Pourrait–on obtenir des résultats

non asymptotiques ?
4°. Est-il réaliste de supposer que les deux lois ont la même variance ?

EXERCICE 6.9. *Test du signe.* Soit F une fonction de répartition sur $\mathbb{R}$ et soit $\theta \in \mathbb{R}$ un paramètre inconnu. On dispose d'un échantillon i.i.d. $(X_1, \ldots, X_n)$ de $F(\cdot - \theta)$ et on considère la *statistique du signe*

$$W_n = \sum_{i=1}^{n} I\{X_i > 0\}.$$

On suppose d'abord que F est connue.

1°. Donner la loi exacte de W_n pour n fixé.
2°. Montrer que la loi limite quand $n \to \infty$ de $(W_n - nw)/\sqrt{n}$ est normale où w est une constante à préciser. Donner la moyenne et la variance de cette loi limite.

On suppose maintenant que F est une fonction de répartition symétrique inconnue et on souhaite tester l'hypothèse $H_0 :\ \theta = 0$ contre l'alternative $H_1 :\ \theta > 0$. Soit $0 < \alpha < 1$.

3°. Proposer un test de niveau exact α basé sur W_n.
4°. Proposer un test de niveau asymptotique α basé sur W_n.
5°. Quelle est la *p-value* du test asymptotique si $n = 16$ et $W_n = 2$?

EXERCICE 6.10. Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de densité

$$f(x, \theta) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right),$$

où $\sigma > 0$ et $\mu \in \mathbb{R}$ sont des paramètres, $\theta = (\mu, \sigma)$.

1°. Trouver $\hat{\sigma}_n$, l'estimateur du maximum de vraisemblance de σ, dans les deux cas suivants :
  (i) μ n'est pas connu,
  (ii) μ est connu.
Dans chacun de ces deux cas, l'estimateur du maximum de vraisemblance est-il unique ?

On supposera désormais que $\mu = 0$.

2°. Chercher la loi asymptotique de $\sqrt{n}(\hat{\sigma}_n - \sigma)$ quand $n \to \infty$.
3°. En utilisant $\hat{\sigma}_n$, construire un test de niveau asymptotique α de l'hypothèse $H_0 :\ \sigma = 1$ contre l'alternative $H_1 :\ 0 < \sigma < 1$.
4°. Donner un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour σ basé sur $\hat{\sigma}_n$.

EXERCICE 6.11. Soient $X_1, \ldots, X_n$ des variables aléatores i.i.d. de densité

$$f(x, \theta) = (2/\sqrt{\pi\theta}) \exp(-x^2/\theta) I\{x > 0\},$$

par rapport à la mesure de Lebesgue dans $\mathbb{R}$, où θ est un paramètre inconnu.
1°. Déterminer la loi de probabilité de la variable $X_1/\sqrt{\theta}$. Déduire de ce résultat que la loi de la variable $\zeta = m_2/\theta$ ne depend pas de θ (ici m_2 désigne le moment empirique d'ordre 2).
2°. Déterminer les réels a et b tels que $[m_2/a, m_2/b]$ soit un intervalle de confiance de niveau $1 - \alpha$ pour θ (pour un $0 < \alpha < 1$ donné).
3°. En utilisant l'approximation de la loi de ζ par une loi normale, chercher les réels a_1 et b_1 tels que $[m_2/a_1, m_2/b_1]$ soit un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour θ.

Exercice 6.12. On dispose d'un échantillon de taille $n = 400$ d'une loi de Poisson $\mathcal{P}(\theta)$ de paramètre θ inconnu. Proposer un intervalle de confiance au niveau asymptotique 0.99 pour θ basé sur l'estimateur du maximum de vraisemblance.

Exercice 6.13. On souhaite comparer les moyennes μ et μ' de deux échantillons de taille n gaussiens indépendants et de même variance connue. On utilise la démarche suivante : si deux intervalles de confiance de niveau a obtenus à partir des échantillons ont une intersection vide, on décide que $\mu \neq \mu'$. Etudier le test correspondant à cette procédure.

Exercice 6.14. Soit $(X_1, \ldots, X_n)$ un échantillon i.i.d. de la loi uniforme $U[0, \theta]$, $\theta > 0$.
1°. Montrer que $\theta / X_{(n)}$ est une fonction pivotale et donner sa densité de probabilité.
2°. Soit $0 < \alpha < 1$. Montrer que l'intervalle de confiance pour θ le plus court de niveau $1 - \alpha$ basé sur cette fonction pivotale est de la forme $[X_{(n)}, \alpha^{-1/n} X_{(n)}]$.
3°. Tracer la diagramme Tests/IC. L'utiliser pour construire un test de niveau α de l'hypothèse $H_0 : \theta = 1$ contre l'alternative $H_1 : \theta \neq 1$.

Exercice 6.15. *Problème de sondages.* Soit N le nombre d'habitants d'une commune. Il s'agit de faire un sondage de popularité de deux candidats (candidat A et candidat B) qui se présentent aux élections municipales. On choisit un échantillon de n habitants auxquels on pose la question : "Pour qui voteriez–vous aux élections ? " A l'issue de ce sondage, on obtient les données $X_1, \ldots, X_n$, où

$$X_i = \begin{cases} 1, \text{ si le } i\text{–ème habitant questionné préfère le candidat } A, \\ 0, \text{ si le } i\text{–ème habitant questionné préfère le candidat } B, \end{cases}$$

$i = 1, \ldots, n$. Pour des raisons évidentes, il est impossible de questionner tous les habitants. Donc $n < N$ (dans la pratique, on a toujours $n \ll N$). Notons μ la part d'habitants de la commune qui préfèrent le candidat A. Le but du sondage est d'estimer μ et de donner un intervalle confiance pour μ.

1°. Proposer un modèle statistique pour ce problème. Observer qu'il s'agit d'un tirage au hasard sans remise d'une population de taille N, car chaque habitant peut apparaître au maximum une fois dans l'échantillon. Définissons les valeurs déterministes $x_1, \ldots, x_N$ par

$$x_j = \begin{cases} 1, & \text{si le } j\text{–ème habitant préfère le candidat } A, \\ 0, & \text{si le } j\text{–ème habitant préfère le candidat } B, \end{cases}$$

$j = 1, \ldots, N$. On a alors

$$\mu = \frac{1}{N} \sum_{j=1}^{N} x_j.$$

Définissons aussi

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{N} (x_j - \mu)^2.$$

On appelle μ *moyenne de population* et σ^2 *variance de population.*

2°. Montrer que $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ est un estimateur sans biais de μ.

3°. Montrer que

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1} \quad \text{pour} \quad i \neq j$$

et
$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$

4°. Calculer $E(s^2)$, où $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, et proposer un estimateur sans biais $\hat{v}^2$ de la variance $\mathrm{Var}(\bar{X})$.
5°. On se place maintenant dans le cadre asymptotique où $N \to \infty$, $n = n(N) \to \infty$ et $n/N \to 0$.
5.1°. Montrer que $\bar{X}$ et $\hat{v}^2$ sont des estimateurs consistants de μ et σ^2.
5.2°. Démontrer la normalité asymptotique
$$\sqrt{n}\,\frac{\bar{X} - \mu}{\hat{v}} \xrightarrow{D} \mathcal{N}(0, 1).$$
En déduire l'intervalle de confiance de niveau asymptotique $1 - \alpha$ pour μ $(0 < \alpha < 1)$.
Application numérique : donner l'intervalle de confiance de niveau asymptotique 95% pour μ lorsque $N = 8000,\ n = 100,\ n_1 = \sum_{i=1}^{n} I\{X_i = 1\} = 65$.

Partie 3

Analyse statistique multivariée

7

Analyse en composantes principales

7.1. Données multivariées

Soit $\mathbf{x} \in \mathbb{R}^p$ un vecteur aléatoire : $\mathbf{x} = (\xi_1, \ldots, \xi_p)^T$, où $\mathbf{v}^T$ désigne le transposé du vecteur $\mathbf{v}$. Un échantillon multidimensionnel est une suite $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de réalisations aléatoires du vecteur $\mathbf{x}$, c'est-à-dire que chaque $\mathbf{x}_i$ est de même loi que $\mathbf{x}$ pour tout $i = 1, \ldots, n$.

Dans ce chapitre, X_{ij} désignera la $j^{\text{ème}}$ composante du vecteur $\mathbf{x}_i$, c'est-à-dire la $i^{\text{ème}}$ réalisation de la variable aléatoire ξ_j. Les X_{ij} forment la matrice aléatoire

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

que l'on appelle **matrice des données** ou **tableau des données**. A partir de la matrice des données $\mathbf{X}$, on peut calculer les statistiques suivantes :

a) *Les moyennes empiriques*

$$\bar{X}_k = \frac{1}{n}\sum_{i=1}^n X_{ik}, \qquad k = 1, \ldots, p,$$

qui forment le vecteur

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{pmatrix} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i = \frac{1}{n}\,\mathbf{X}^T\mathbf{1} \quad \text{avec} \quad \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n.$$

b) *Les covariances empiriques*

$$s_{jk} = \frac{1}{n}\sum_{i=1}^n X_{ij}X_{ik} - \bar{X}_j\bar{X}_k, \quad k, j = 1, \ldots, p,$$

qui forment la matrice

$$S = (s_{jk})_{k,j=1,\dots,p}$$

que l'on appelle **matrice de covariance empirique**.

c) *Les corrélations empiriques* définies, pour $s_{jj} > 0,\ j = 1,\dots,p$, par

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{kk}s_{jj}}}\ , \qquad k,j = 1,\dots,p$$

qui forment la matrice

$$R = (r_{jk})_{k,j=1,\dots,p}$$

que l'on appelle **matrice de corrélation empirique**.

Il est facile de voir que

$$S = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T - \bar{\mathbf{x}}\bar{\mathbf{x}}^T = \frac{1}{n}\,\mathbf{X}^T\mathbf{X} - \bar{\mathbf{x}}\bar{\mathbf{x}}^T = \frac{1}{n}\,\mathbf{X}^T\mathbf{X} - \frac{1}{n^2}\,\mathbf{X}^T\mathbf{1}\mathbf{1}^T\mathbf{X} = \frac{1}{n}\,\mathbf{X}^T H\mathbf{X}$$

où la matrice $H = I_n - n^{-1}\mathbf{1}\mathbf{1}^T$ est appelée *matrice centring.*

Exercice 7.1. Montrer que H est un projecteur, i.e. $H = H^2$ et $H^T = H$. Sur quel sous-espace vectoriel de $\mathbb{R}^n$ projette-t-il ?

Notons que la matrice de covariance empirique S est positive, en effet pour tout vecteur $a \in \mathbb{R}^p$ on a

$$a^T S a = \frac{1}{n}\,a^T\mathbf{X}^T H\mathbf{X}a = \frac{1}{n}\,a^T\mathbf{X}^T HH\mathbf{X}a = \frac{1}{n}\,\mathbf{y}^T\mathbf{y} \geq 0,$$

où $\mathbf{y} = H^T\mathbf{X}a$. De plus, si l'on note par D la matrice diagonale $\mathrm{diag}\{\sqrt{s_{11}},\dots,\sqrt{s_{pp}}\}$, on obtient $S = DRD$, donc la matrice de corrélation empirique R est aussi positive.

7.2. L'idée de l'Analyse en composantes principales (ACP)

L'Analyse en composantes principales (ACP) est une méthode de traitement des données multidimensionnelles qui poursuit les deux objectifs suivants :

– visualiser les données,
– réduire la dimension effective des données.

Géométriquement, les données multidimensionnelles constituent un *nuage de points* dans $\mathbb{R}^p$ (un point de ce nuage correspond à un $\mathbf{x}_i$). Si la dimension p est supérieure à 3, ce qui est le plus souvent le cas, on ne peut pas visualiser ce nuage. Le seul moyen de visualiser les données est alors de considérer leurs projections sur des droites, sur des plans ou éventuellement sur des espaces de dimension 3. Ainsi, si $a = (a_1,\dots,a_p) \in \mathbb{R}^p$ est une direction de projection (c'est-à-dire un vecteur de norme un : $\|a\|^2 = a_1^2 + \cdots + a_p^2 = 1$), les données projetées $(a^T\mathbf{x}_1,\dots,a^T\mathbf{x}_n)$ forment un échantillon de dimension 1 que l'on peut visualiser et qui est donc plus facile à interpréter que l'échantillon de départ $(\mathbf{x}_1,\dots,\mathbf{x}_n)$.

Si la dimension p est grande, elle est d'habitude redondante. En réalité la "vraie" dimension des données p^* est souvent beaucoup plus petite que p. L'ACP a pour objectif de trouver un sous-espace linéaire de $\mathbb{R}^p$ de dimension $p^* \ll p$ tel que la projection sur ce sous-espace "capte" presque toute la structure des données.

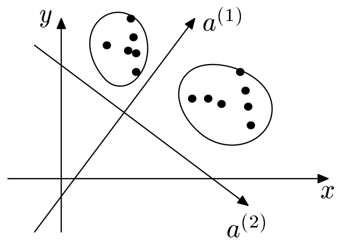

Fig. 7.1. Bonne et mauvaise directions de projection.

Dans l'exemple de la Figure 7.1, on voit que si l'on projette les données $\mathbf{x}_i$ (représentées par des points noirs) sur la direction $a^{(1)}$, certaines projections coïncideront. Par contre, la projection de ces données sur la direction $a^{(2)}$ donne des valeurs deux à deux distinctes. On voit que la projection sur cette dernière direction est plus informative que sur la première, donc plus intéressante.

L'idée de base de l'ACP est de chercher la direction $a \in \mathbb{R}^p$ "la plus intéressante", pour laquelle les données projetées seront le plus dispersées possibles, c'est-à-dire la direction qui maximise en a la variance empirique de l'échantillon unidimensionnel $(a^T\mathbf{x}_1, \ldots, a^T\mathbf{x}_n)$ (cf. définition de la variance empirique au Chapitre 4) :

$$\begin{aligned} s_a^2 &\stackrel{\text{déf}}{=} \frac{1}{n}\sum_{i=1}^n (a^T\mathbf{x}_i)^2 - \left(\frac{1}{n}\sum_{i=1}^n (a^T\mathbf{x}_i)\right)^2 \\ &= \frac{1}{n}\, a^T\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\right)a - \frac{1}{n^2}\, a^T\left(\sum_{i=1}^n \mathbf{x}_i \sum_{i=1}^n \mathbf{x}_i^T\right)a = a^T S a, \end{aligned}$$

où S désigne la matrice de covariance empirique introduite au paragraphe précédent. Par conséquent, la direction la plus intéressante $\hat{a}$ est une solution de

$$\max_{a\in\mathbb{R}^p:\|a\|=1} a^T S a = \hat{a}^T S \hat{a},$$

où $\|\cdot\|$ est la norme euclidienne de $\mathbb{R}^p$. On peut écrire cette égalité sous la forme équivalente

$$\hat{a} = \arg\max_{a\in\mathbb{R}^p:\|a\|=1} a^T S a. \tag{7.1}$$

Le vecteur $\hat{a}$ ainsi défini maximise la variance empirique unidimensionnelle s_a^2 en a tels que $\|a\| = 1$. De la même manière, on peut définir la direction "idéale" pour projeter les données, comme le vecteur a^* qui maximise la variance théorique :

$$a^* = \arg\max_{a\in\mathbb{R}^p:\|a\|=1} \mathrm{Var}[a^T\mathbf{x}]. \tag{7.2}$$

Pour que cette variance soit bien finie, on suppose que $E[\|\mathbf{x}\|^2] < \infty$. Dans ce qui suit, on utilisera les notations suivantes pour la moyenne et la matrice de covariance de $\mathbf{x}$:

$$E(\mathbf{x}) = \mu, \qquad V(\mathbf{x}) = \Sigma.$$

(ici μ est un vecteur de $\mathbb{R}^p$ et Σ est une matrice symétrique et positive de dimension $p \times p$).

7.3. ACP : cadre théorique

Nous nous intéresserons ici à la solution du problème de maximisation (7.2). Soit $\Sigma = \Gamma\Lambda\Gamma^T$ une décomposition spectrale de la matrice de covariance, où Γ est une matrice $p \times p$ orthogonale et Λ est une matrice $p \times p$ diagonale. On notera

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}, \qquad \Gamma = \big(\boldsymbol{\gamma}_{(1)}, \dots, \boldsymbol{\gamma}_{(p)}\big),$$

où les λ_i sont les valeurs propres de Σ rangées par ordre décroissant : $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, et les $\boldsymbol{\gamma}_{(i)}$ sont les vecteurs propres orthonormés de Σ correspondants,

$$\|\boldsymbol{\gamma}_{(i)}\| = 1, \qquad \boldsymbol{\gamma}_{(j)}^T \boldsymbol{\gamma}_{(k)} = 0, \quad j \neq k.$$

Définition 7.1. *La variable aléatoire* $\eta_j = \boldsymbol{\gamma}_{(j)}^T(\mathbf{x} - \mu)$ *est dite* **j^{ème} composante principale** *du vecteur aléatoire* $\mathbf{x} \in \mathbb{R}^p$.

EXEMPLE 7.1. Soit $\mathbf{x}$ un vecteur aléatoire de $\mathbb{R}^2$ de moyenne nulle et de matrice de covariance

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad 0 \leq \rho \leq 1.$$

Considérons les vecteurs propres orthonormés de cette matrice

$$\boldsymbol{\gamma}_{(1)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\gamma}_{(2)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Donc si les coordonnées de $\mathbf{x}$ sont ξ_1 et ξ_2, les composantes principales de x valent

$$\eta_1 = \frac{\xi_1 + \xi_2}{\sqrt{2}}, \qquad \eta_2 = \frac{\xi_1 - \xi_2}{\sqrt{2}}.$$

D'une part, on peut facilement vérifier que la variable aléatoire η_j est centrée, c'est-à-dire $E[\eta_j] = 0$. D'autre part, en utilisant le fait que les $\boldsymbol{\gamma}_{(j)}$ sont les vecteurs propres de la matrice de covariance Σ du vecteur aléatoire $\mathbf{x}$, on obtient

$$\mathrm{Var}[\eta_j] = E[\boldsymbol{\gamma}_{(j)}^T(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \boldsymbol{\gamma}_{(j)}] = \boldsymbol{\gamma}_{(j)}^T \Sigma \boldsymbol{\gamma}_{(j)} = \boldsymbol{\gamma}_{(j)}^T \lambda_j \boldsymbol{\gamma}_{(j)} = \lambda_j,$$

où λ_j désigne la valeur propre correspondant au vecteur propre $\boldsymbol{\gamma}_{(j)}$. De même, pour $j \neq k$,

$$\mathrm{Cov}(\eta_j, \eta_k) = E[\boldsymbol{\gamma}_{(j)}^T(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \boldsymbol{\gamma}_{(k)}] = \boldsymbol{\gamma}_{(j)}^T \Sigma \boldsymbol{\gamma}_{(k)} = \boldsymbol{\gamma}_{(j)}^T \lambda_k \boldsymbol{\gamma}_{(k)} = 0,$$

car les vecteurs $\boldsymbol{\gamma}_{(j)}$ sont orthonormés.

Théorème 7.1. *Soit* $\mathbf{x} \in \mathbb{R}^p$ *un vecteur aléatoire tel que* $E(\|\mathbf{x}\|^2) < \infty$. *Alors* $\hat{a} = \boldsymbol{\gamma}_{(1)}$ *est une solution du problème (7.2), c'est-à-dire :*

$$\mathrm{Var}[\hat{a}^T \mathbf{x}] = \max_{a \in \mathbb{R}^p:\, \|a\|=1} \mathrm{Var}[a^T \mathbf{x}] = \max_{a \in \mathbb{R}^p:\; \|a\|=1} \mathrm{Var}[a^T(\mathbf{x} - \mu)].$$

Preuve. La décomposition spectrale de la matrice Σ est de la forme

$$\Sigma = \Gamma\Lambda\Gamma^T = \sum_{j=1}^{p} \lambda_j \boldsymbol{\gamma}_{(j)} \boldsymbol{\gamma}_{(j)}^T.$$

On a donc

$$\mathrm{Var}[a^T \mathbf{x}] = \sum_{j=1}^{p} \lambda_j (a^T \boldsymbol{\gamma}_{(j)})(\boldsymbol{\gamma}_{(j)}^T a) = \sum_{j=1}^{p} \lambda_j c_j^2,$$

où $c_j = a^T \boldsymbol{\gamma}_{(j)}$ est la projection du vecteur a sur la direction $\boldsymbol{\gamma}_{(j)}$. Puisque les vecteurs $\boldsymbol{\gamma}_{(j)}$ forment une base orthonormée de $\mathbb{R}^p$, on a $c_1^2 + \cdots + c_p^2 = \|a\|^2$. Comme $\lambda_j \leq \lambda_1$, on en déduit que

$$\mathrm{Var}[a^T \mathbf{x}] = \sum_{j=1}^{p} \lambda_j c_j^2 \leq \lambda_1 \sum_{j=1}^{p} c_j^2 = \lambda_1 \|a\|^2 = \lambda_1.$$

Par ailleurs, si $a = \hat{a} = \gamma_{(1)}$, les coefficients c_j sont tous nuls sauf le premier $c_1 = 1$. On a donc $\mathrm{Var}[\hat{a}^T \mathbf{x}] = \lambda_1$. Par conséquent, $\hat{a}$ est une solution du problème de maximisation (7.2) et $\mathrm{Var}[\hat{a}^T \mathbf{x}] = \lambda_1 = \mathrm{Var}[\eta_1]$. ∎

Deuxième composante principale. De la même façon, on peut prouver que $\boldsymbol{\gamma}_{(2)}$ est l'un des vecteurs qui maximise la variance $\mathrm{Var}[a^T \mathbf{x}]$ sur l'ensemble $A_1 = \{a \in \mathbb{R}^p \ : \ \|a\| = 1 \text{ et } a \perp \boldsymbol{\gamma}_{(1)}\}$. En effet, comme a est orthogonal à $\boldsymbol{\gamma}_{(1)} = \hat{a}$, sa projection c_1 sur $\boldsymbol{\gamma}_{(1)}$ est nulle. Par conséquent, pour tout vecteur de A_1, on a

$$\mathrm{Var}[a^T \mathbf{x}] = \sum_{j=2}^{p} \lambda_j c_j^2 \leq \lambda_2 \sum_{j=2}^{p} c_j^2 = \lambda_2 \|a\|^2 = \lambda_2.$$

On voit donc que $\mathrm{Var}[\gamma_{(2)}^T \mathbf{x}] = \lambda_2 = \mathrm{Var}(\eta_2)$.

k-ème composante principale. On démontre de la même manière que $\boldsymbol{\gamma}_{(k)}$ est l'un des vecteurs $a \in \mathbb{R}^p$ qui maximise $\mathrm{Var}[a^T \mathbf{x}]$ sur l'ensemble A_{k-1} de tous les vecteurs de norme 1 orthogonaux aux $\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(k-1)}$. On trouve dans ce cas $\max_{a \in A_{k-1}} \mathrm{Var}[a^T \mathbf{x}] = \mathrm{Var}[\eta_k]$.

On voit donc que, du point de vue mathématique, l'ACP se réduit à la diagonalisation de la matrice de covariance de $\mathbf{x}$.

7.4. ACP : cadre empirique

Considérons maintenant le problème de maximisation (7.1). Nous pouvons obtenir une solution de ce problème par la même méthode qu'au paragraphe précédent, en remplaçant la matrice de covariance Σ par la matrice de covariance empirique S (il suffit de noter que dans (7.2) $\mathrm{Var}[a^T \mathbf{x}] = a^T \Sigma a$ et de comparer (7.1) et (7.2)).

Comme S est une matrice symétrique, il existe une matrice orthogonale G et une matrice diagonale L telles que $S = GLG^T$. Bien évidemment, ces matrices dépendent de l'échantillon $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Les éléments diagonaux $l_1, \ldots, l_p$, de la matrice L sont alors les valeurs propres de S. De plus, les l_j sont positifs, car S est une matrice positive. On suppose que les l_j sont numérotés par ordre décroissant :

$$l_1 \geq l_2 \geq \ldots \geq l_p \geq 0.$$

On note $g_{(j)}$ le vecteur propre de norme 1 associé à la valeur propre l_j.

Définition 7.2. *La* **$j^{\text{ème}}$ composante principale empirique** *associée à l'échantillon* $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *est la fonction* $y_j : \mathbb{R}^p \to \mathbb{R}$ *définie par*

$$y_j(\mathbf{z}) = g_{(j)}^T(\mathbf{z} - \bar{\mathbf{x}}) \quad \textit{pour} \quad \mathbf{z} \in \mathbb{R}^p.$$

Soit $y_{ij} = y_j(\mathbf{x}_i)$. Considérons la matrice $\mathbf{Y} = (y_{ij})_{i=1,\ldots,n, j=1,\ldots,p}$, de dimension $n \times p$. Elle remplace la matrice des données $\mathbf{X}$ initiale. Les vecteurs-lignes $\mathbf{y}_1, \ldots, \mathbf{y}_n$ de la matrice $\mathbf{Y}$ peuvent être considérés comme un nouvel échantillon de données transformées (il s'agit d'une transformation affine de l'échantillon initial $\mathbf{x}_1, \ldots, \mathbf{x}_n$). Dans la pratique, l'application de l'ACP est intéressante s'il s'avère que les $\mathbf{y}_i$ résident "essentiellement" dans un sous-espace affine de $\mathbb{R}^p$ de dimension beaucoup plus petite que p.

Remarques.

(1) Si les variables ξ_i sont de nature différente (par exemple, ξ_1 est le prix d'un produit en dollars et ξ_2 est son poids en kilogrammes), dans la pratique on utilise l'ACP sur la matrice de corrélation R plutôt que l'ACP sur la matrice de covariance S, i.e. on cherche à maximiser $a^T Ra$ au lieu de maximiser $a^T Sa$. Ceci est motivé par le fait que les éléments de R n'ont pas d'unité de mesure.

(2) Si tous les éléments de la matrice S sont strictement positifs, comme dans l'exemple numérique qui sera analysé à la fin de ce chapitre, toutes les coordonnées de $g_{(1)}$ ont le même signe (cf. Théorème de Perron – Frobenius ci-après). Dans ce cas, la première composante principale empirique $y_1(\cdot)$ s'appelle *facteur de taille*. La valeur $y_1(\mathbf{x}_i)$ est alors interprétée comme une caractéristique de "taille" ou d'importance de l'individu i. Ainsi, dans l'exemple numérique qui sera examiné à la fin de ce chapitre, $y_1(\mathbf{x}_i)$ peut être considérée comme une caractéristique du niveau général de l'étudiant numéro i calculée à partir de ses notes.

Proposition 7.1. (Théorème de Perron – Frobenius.) *Soit* $A = (a_{ij})_{i,j=1,\ldots,p}$ *une matrice* $p \times p$ *symétrique dont tous les éléments sont strictement positifs. Alors toutes les coordonnées du premier vecteur propre de* A *ont le même signe.*

Preuve. Soit $g = (g_1, \ldots, g_p)$ un vecteur propre orthonormé de A correspondant à sa plus grande valeur propre. Notons $\tilde{g} = (|g_1|, \ldots, |g_p|)$ le vecteur dont les coordonnées sont les valeurs absolues des coordonnées respectives de g. D'une part, il est évident que $\|g\| = \|\tilde{g}\| = 1$ et

$$g^T Ag = \max_{\|\bar{g}\|=1} \bar{g}^T A\bar{g},$$

ce qui implique que $g^T Ag \geq \tilde{g}^T A\tilde{g}$. D'autre part, comme tous les éléments a_{ij} de A sont positifs, on obtient

$$g^T Ag = \sum_{i,j=1}^{p} a_{ij} g_i g_j \leq \sum_{i,j=1}^{p} a_{ij} |g_i||g_j| = \tilde{g}^T A\tilde{g}.$$

On a alors $g^T Ag = \tilde{g}^T A\tilde{g}$. De plus, $\tilde{g}^T Ag = g^T A\tilde{g}$, car la matrice A est symétrique. Ces deux égalités impliquent que

$$(g - \tilde{g})^T A(g + \tilde{g}) = 0. \tag{7.3}$$

Soit maintenant $w = A(g + \tilde{g})$. Comme tous les éléments de A sont strictement positifs et $g_i + |g_i| \geq 0$, toutes les coordonnées du vecteur w sont positives.

On peut avoir les deux cas suivants.

Cas 1 : toutes les coordonnées $w_1, \dots, w_p$ de w sont strictement positives. Dans ce cas, les relations $(g-\tilde{g})w = 0$ et $\tilde{g}_i \geq g_i$ impliquent que $g_i = \tilde{g}_i$ pour tout $i = 1, \dots, p$. Par conséquent, tous les g_i sont positifs.

Cas 2 : il existe j_0 tel que $w_{j_0} = 0$. Comme $w = A(g + \tilde{g})$, la coordonnée w_{j_0} vaut

$$w_{j_0} = \sum_i a_{ij_0}(\tilde{g}_i + g_i).$$

D'après l'hypothèse de la proposition, tous les coefficients a_{ij_0} sont strictement positifs. Il en résulte que $\tilde{g}_i + g_i = 0$ pour tout i. On en déduit que toutes les coordonnées de g sont négatives. $\blacksquare$

7.5. Etude des corrélations : cadre théorique

Soit $\mathbf{x} \in \mathbb{R}^p$ un vecteur aléatoire de moyenne μ et de matrice de covariance Σ. On définit **la variance totale** de $\mathbf{x}$ par

$$E(\|\mathbf{x} - \mu\|^2) = E\big((\mathbf{x} - \mu)^T(\mathbf{x} - \mu)\big) = E\big((\mathbf{x} - \mu)^T\Gamma\Gamma^T(\mathbf{x} - \mu)\big).$$

où, d'après les définitions introduites au Paragraphe 7.3,

$$\Gamma^T(\mathbf{x} - \mu) = \begin{pmatrix} \gamma_{(1)}^T(\mathbf{x} - \mu) \\ \vdots \\ \gamma_{(p)}^T(\mathbf{x} - \mu) \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_p \end{pmatrix} \stackrel{\text{déf}}{=} \mathbf{y}.$$

Compte tenu de ces notations et de l'égalité $E(\eta_i^2) = \lambda_i$, où λ_i est la $i^{\text{ème}}$ valeur propre de Σ, on obtient l'expression suivante pour la variance totale :

$$E(\|\mathbf{x} - \mu\|^2) = E(\eta_1^2 + \cdots + \eta_p^2) = \lambda_1 + \cdots + \lambda_p = \mathrm{Tr}(\Sigma).$$

Rappelons que la trace $\mathrm{Tr}(\Sigma)$ est la somme de ses éléments diagonaux de la matrice Σ.

7.5.1. La part de variance expliqueé.

Définition 7.3. *On appelle* **part de la variance totale de x expliquée par les** k **premières composantes principales** $(\eta_1, \dots, \eta_k)$ *la quantité*

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} = \frac{\lambda_1 + \cdots + \lambda_k}{\mathrm{Tr}(\Sigma)}\ .$$

On appelle part de la variance totale de $\mathbf{x}$ *expliquée par la* $j^{\text{ème}}$ *composante principale* η_j *la quantité*

$$\frac{\lambda_j}{\lambda_1 + \cdots + \lambda_p}.$$

Si pour un $k < p$, la part de la variance totale expliquée par les k premières composantes principales est égale à 1, alors on dit que la variance totale est *entièrement expliquée* par les composantes $\eta_1, \dots, \eta_k$. Cela signifie que seules les k premières composantes principales contribuent à la variance totale du vecteur $\mathbf{x}$, les $(p-k)$ composantes restantes étant des valeurs déterministes.

Analysons maintenant l'influence de la composante principale η_j sur la variable ξ_i, la $i^{\text{ème}}$ coordonnée du vecteur aléatoire $\mathbf{x}$. Nous allons caractériser cette influence par la valeur du coefficient de corrélation $\mathrm{Corr}(\xi_i, \eta_j)$. Plus la valeur absolue de $\mathrm{Corr}(\xi_i, \eta_j)$ est proche de 1, mieux la composante principale η_j "explique" la variable ξ_i. Calculons d'abord la matrice de covariance des vecteurs aléatoires $\mathbf{x}$ et $\mathbf{y}$. On a

$$C(\mathbf{x}, \mathbf{y}) = E[(\mathbf{x}-\mu)\mathbf{y}^T] = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\Gamma] = \Sigma\Gamma = \Gamma\Lambda\Gamma^T\Gamma = \Gamma\Lambda.$$

Comme $\mathrm{Cov}(\xi_i, \eta_j)$ est le $(i,j)^{\text{ème}}$ élément de cette matrice, on obtient

$$\mathrm{Cov}(\xi_i, \eta_j) = \gamma_{ij}\lambda_j.$$

La corrélation $\tilde\rho_{ij} = \mathrm{Corr}(\xi_i, \eta_j)$ entre ξ_i et η_j vaut

$$\tilde\rho_{ij} = \frac{\mathrm{Cov}(\xi_i, \eta_j)}{\sqrt{\mathrm{Var}(\xi_i)\mathrm{Var}(\eta_j)}} = \gamma_{ij}\sqrt{\frac{\lambda_j}{\sigma_{ii}}}.$$

Proposition 7.2. *Soit $\mathbf{x} \in \mathbb{R}^p$ un vecteur aléatoire, tel que $E(\|\mathbf{x}\|^2) < \infty$ et $\sigma_{ii} > 0$ pour tout $i = 1, \dots, p$. Alors,*

$$\sum_{j=1}^p \tilde\rho_{ij}^2 = 1 \quad \textit{pour} \quad i = 1, \dots, p.$$

Preuve. Soit $\tilde P$ la matrice carrée dont les éléments sont les corrélations $\tilde\rho_{ij}$, $i = 1, \dots, p$, $j = 1, \dots, p$. Soit encore Δ une matrice diagonale dont les éléments diagonaux sont σ_{ii} :

$$\Delta = \mathrm{diag}(\sigma_{11}, \dots, \sigma_{pp}).$$

Il est facile alors de vérifier que $\tilde P = \Delta^{-1/2}\Gamma\Lambda^{1/2}$. Par conséquent,

$$\tilde P\tilde P^T = \Delta^{-1/2}\Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma^T\Delta^{-1/2} = \Delta^{-1/2}\Sigma\Delta^{-1/2} = P, \tag{7.4}$$

où P est la matrice formée par les corrélations $\rho_{ij} = \mathrm{Corr}(\xi_i, \xi_j)$ entre les coordonnées ξ_i et ξ_j de $\mathbf{x}$. Pour conclure, il suffit de remarquer que d'une part $\rho_{ii} = 1$ et d'autre part, d'après (7.4), $\rho_{ii} = \sum_{j=1}^p \tilde\rho_{ij}^2$. ∎

Définition 7.4. *On appelle* $\tilde\rho_{ij}^2$ **part de variance de la variable** ξ_i **expliquée par la** $j^{\text{ème}}$ **composante principale** η_j.

Proposition 7.3. *Supposons que les hypothèses de la Proposition 7.2 soient vérifiées. Alors, pour tout sous-ensemble J de $\{1, \dots, p\}$,*

$$\sum_{j\in J}\lambda_j = \sum_{i=1}^p \sigma_{ii}\tilde\rho_{iJ}^2,$$

où $\tilde\rho_{iJ}^2 = \sum_{j\in J}\tilde\rho_{ij}^2$.

Preuve.

$$\sum_{i=1}^{p} \sigma_{ii}\tilde{\rho}_{iJ}^{2} = \sum_{i=1}^{p} \sigma_{ii} \sum_{j\in J} \gamma_{ij}^{2} \frac{\lambda_j}{\sigma_{ii}} = \sum_{j\in J} \lambda_j \sum_{i=1}^{p} \gamma_{ij}^{2}.$$

Le résultat de la proposition découle du fait que la dernière somme vaut 1, car $\|\boldsymbol{\gamma}_{(j)}\|^2 = \sum_{i=1}^{p} \gamma_{ij}^2 = 1$. ∎

7.5.2. Disque des corrélations. D'après la Proposition 7.2, la somme des carrés des deux corrélations $\tilde{\rho}_{i1}^2 + \tilde{\rho}_{i2}^2$ est inférieure ou égale à 1, donc tous les points de $\mathbb{R}^2$ ayant les coordonnées $(\tilde{\rho}_{i1}, \tilde{\rho}_{i2})$ appartiennent au disque de rayon 1 centré en 0, que l'on appelle dans le contexte de l'ACP **disque des corrélations**. Sa frontière est appelée **cercle des corrélations**. Plus le point $(\tilde{\rho}_{i1}, \tilde{\rho}_{i2})$ est proche du cercle des corrélations, mieux la variable ξ_i est expliquée par les deux premières composantes principales. Considérons maintenant la situation idéale quand les points $(\tilde{\rho}_{i1}, \tilde{\rho}_{i2})$ et $(\tilde{\rho}_{k1}, \tilde{\rho}_{k2})$ se trouvent exactement sur le cercle, ce qui correspond au fait que les variables ξ_i et ξ_k sont entièrement expliquées par les deux premières composantes principales.

Proposition 7.4. *Soient ξ_i et ξ_k deux variables entièrement expliquées par les deux premières composantes principales, i.e.*

$$\tilde{\rho}_{i1}^2 + \tilde{\rho}_{i2}^2 = 1 \quad \textit{et} \qquad \tilde{\rho}_{k1}^2 + \tilde{\rho}_{k2}^2 = 1.$$

Alors, la corrélation de ξ_i et ξ_k est donnée par la formule

$$\rho_{ik} = \tilde{\rho}_{i1}\tilde{\rho}_{k1} + \tilde{\rho}_{i2}\tilde{\rho}_{k2} = \cos(\varphi),$$

où φ est l'angle formé par les vecteurs $(\tilde{\rho}_{i1}, \tilde{\rho}_{i2})$ et $(\tilde{\rho}_{k1}, \tilde{\rho}_{k2})$.

Preuve. Vu que la variable ξ_i est entièrement expliquée par η_1 et η_2, on a $\tilde{\rho}_{im} = 0$, quel que soit $m \geq 3$. De même, pour ξ_k, on a $\tilde{\rho}_{km} = 0$ pour tout $m \geq 3$. Comme $P = \tilde{P}\tilde{P}^T$, cela implique que

$$\rho_{ik} = \tilde{\rho}_{i1}\tilde{\rho}_{k1} + \tilde{\rho}_{i2}\tilde{\rho}_{k2}.$$

Soit φ_1 l'angle formé par les vecteurs $(\tilde{\rho}_{i1}, \tilde{\rho}_{i2})$ et $(1, 0)$, et φ_2 l'angle formé par les vecteurs $(\tilde{\rho}_{k1}, \tilde{\rho}_{k2})$ et $(1, 0)$. Il est évident que $\varphi = |\varphi_1 - \varphi_2|$ et

$$\tilde{\rho}_{i1}\tilde{\rho}_{k1} + \tilde{\rho}_{i2}\tilde{\rho}_{k2} = \cos(\varphi_1)\cos(\varphi_2) + \sin(\varphi_1)\sin(\varphi_2) = \cos(\varphi_1 - \varphi_2) = \cos(\varphi).$$

∎

D'après cette proposition, si les variables ξ_i et ξ_k sont entièrement expliquées par les deux premières composantes principales, l'angle formé par les vecteurs $(\tilde{\rho}_{i1}, \tilde{\rho}_{i2})$ et $(\tilde{\rho}_{k1}, \tilde{\rho}_{k2})$ décrit la dépendance mutuelle de ces variables. En effet, si l'angle φ est zéro, alors $\rho_{ik} = 1$, ce qui signifie qu'il y a un lien linéaire déterministe entre ces variables :

$$\exists\ a > 0,\ b \in \mathbb{R} \quad \text{tels que} \quad \xi_i = a\xi_k + b.$$

Si les deux points $(\tilde{\rho}_{i1}, \tilde{\rho}_{i2})$ et $(\tilde{\rho}_{k1}, \tilde{\rho}_{k2})$ de $\mathbb{R}^2$ sont diamétralement opposés, alors $\cos\varphi = \rho_{ik} = -1$ et

$$\exists\ a > 0,\ b \in \mathbb{R} \quad \text{tels que} \quad \xi_i = -a\xi_k + b.$$

Dans le contexte de l'ACP, on dit dans ce cas que *les variables ξ_i et ξ_k sont opposées.* Finalement, si l'angle φ est de 90°, alors $\rho_{ik} = 0$: les variables ξ_i et ξ_k sont non-corrélées.

7.6. Etude des corrélations : cadre empirique

Dans ce paragraphe, on se place dans le cadre, habituel pour une étude statistique, où la moyenne μ et de la matrice de covariance Σ ne sont pas connues. Comme cela a déjà été fait précédemment, on remplace dans toutes les définitions du Paragraphe 7.5 les paramètres inconnus par leurs estimateurs empiriques. Ainsi, μ est remplacé par $\bar{\mathbf{x}}$, Σ par S, $\boldsymbol{\gamma}_{(j)}$ par $g_{(j)}$, λ_j par l_j et η_j par y_j. On donne maintenant les versions empiriques des définitions principales du paragraphe précédent.

Définition 7.5. *On appelle* **part de la variance empirique expliquée par les k premières composantes principales** $(y_1, \ldots, y_k)$ *la quantité suivante :*

$$\frac{l_1 + \cdots + l_k}{l_1 + \cdots + l_p} = \frac{l_1 + \cdots + l_k}{\mathrm{Tr}(S)}.$$

On appelle la quantité $l_i/\mathrm{Tr}(S)$ part de la variance empirique expliquée par la $i^{\text{ème}}$ composante principale y_i.

Pour introduire la définition suivante, rappelons que les s_{ii} désignent les éléments diagonaux de la matrice de covariance empirique S et l_j est la $j^{\text{ème}}$ valeur propre de S. Notons g_{ij} la $i^{\text{ème}}$ coordonnée du vecteur propre $g_{(j)}$.

Définition 7.6. *On appelle* $\tilde{r}_{ij}^2 = g_{ij}^2 l_j / s_{ii}$ **part de la variance empirique de la $i^{\text{ème}}$ variable expliquée par la $j^{\text{ème}}$ composante principale**.

En utilisant le même raisonnement qu'au paragraphe précédent (cf. Propositions 7.2 et 7.3), on trouve que

$$\sum_{j=1}^{p} \tilde{r}_{ij}^2 = 1 \quad \text{pour tout} \quad i = 1, \ldots, p,$$

$$\sum_{j \in J} l_j = \sum_{i=1}^{p} s_{ii} \tilde{r}_{iJ}^2 \quad \text{avec} \quad \tilde{r}_{iJ}^2 = \sum_{j \in J} \tilde{r}_{ij}^2.$$

On introduit également le disque des corrélations auquel appartiennent les points $(\tilde{r}_{i1}, \tilde{r}_{i2})$ pour $i = 1, \ldots, p$. Les résultats de l'ACP sont facilement interpétables si ces points sont proches du cercle des corrélations. L'interprétation est basée sur la comparaison du graphique obtenu avec l'une des trois configurations idéales :

(1) L'angle φ formé par les vecteurs $(\tilde{r}_{i1}, \tilde{r}_{i2})$ et $(\tilde{r}_{k1}, \tilde{r}_{k2})$ est zéro : la $i^{\text{ème}}$ et la $k^{\text{ème}}$ variables sont liées par une relation linéaire déterministe avec la pente strictement positive.

(2) L'angle φ est de 180° : la $i^{\text{ème}}$ et la $k^{\text{ème}}$ variables sont liées par une relation linéaire déterministe avec la pente strictement négative.

(3) L'angle φ est de 90° : la $i^{\text{ème}}$ et la $k^{\text{ème}}$ variables sont non-corrélées.

Il est clair que, dans la pratique, ces trois possibilités peuvent se réaliser seulement de façon approximative, car il s'agit ici de corrélations empiriques $\tilde{r}_{ij}$ qui approchent les corrélations théoriques $\tilde{\rho}_{ij}$ seulement quand la taille d'échantillon n est assez grande (cf. Proposition 4.5).

7.7. Exemple d'application numérique de l'ACP

Analysons ici un exemple d'application de l'ACP emprunté du livre de K.V. Mardia, J.T. Kent et J.M. Bibby *Multivariate Analysis* (Academic Press, London, 1992). Le tableau suivant donne les notes (sur 100) de 88 étudiants obtenues à l'issue de différentes épreuves écrites (E) et orales (O). C'est un exemple de tableau des données $\mathbf{X}$. Les $n = 88$ lignes de ce tableau sont les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_{88}$. Il y a $p = 5$ variables : les notes des 5 examens.

No	Mécanique (O)	Algèbre lin. (O)	Algèbre (E)	Analyse (E)	Statistique (E)
1.	77	82	67	67	81
2.	63	78	80	70	81
3.	75	73	71	66	81
4.	55	72	63	70	68
5.	63	63	65	70	63
6.	53	61	72	64	73
7.	51	67	65	65	68
8.	59	70	68	62	56
9.	46	52	53	41	40
10.	62	60	58	62	70
11.	64	72	60	62	45
12.	52	64	60	63	54
13.	55	67	59	62	44
14.	50	50	64	55	63
15.	65	63	58	56	37
16.	31	55	60	57	73
17.	60	64	56	54	40
18.	44	69	53	53	53
19.	42	69	61	55	45
20.	62	46	61	57	45
21.	31	49	62	63	62
22.	44	61	52	62	46
23.	49	41	61	49	64
24.	12	58	61	63	67
25.	49	53	49	62	47
26.	54	49	56	47	53
27.	54	53	46	59	44
28.	44	56	55	61	36
29.	18	44	50	57	81
30.	46	52	65	50	35
31.	32	45	49	57	64
32.	30	69	50	52	45
33.	46	49	53	59	37
34.	40	27	54	61	61
35.	31	42	48	54	68

No	Mécanique (O)	Algèbre lin. (O)	Algèbre (E)	Analyse (E)	Statistique (E)
36.	36	59	51	45	51
37.	56	40	56	54	35
38.	46	56	57	49	32
39.	45	42	55	56	40
40.	42	60	54	49	33
41.	40	63	53	54	25
42.	23	55	59	53	44
43.	48	48	49	51	37
44.	41	63	49	46	34
45.	46	61	46	38	41
46.	40	57	51	52	31
47.	49	49	45	48	39
48.	22	58	53	56	41
49.	35	60	47	54	33
50.	48	56	49	42	32
51.	31	57	50	54	34
52.	17	53	57	43	51
53.	49	57	47	39	26
54.	59	50	47	15	46
55.	37	56	49	28	45
56.	40	43	48	21	61
57.	35	35	41	51	50
58.	38	44	54	47	24
59.	43	43	38	34	49
60.	39	46	46	32	43
61.	62	44	36	22	42
62.	48	38	41	44	33
63.	34	42	50	47	29
64.	18	51	40	56	30
65.	35	36	46	48	29
66.	59	53	37	22	19
67.	41	41	43	30	33
68.	31	52	37	27	40
69.	17	51	52	35	31
70.	34	30	50	47	36
71.	46	40	47	29	17
72.	10	46	36	47	39
73.	46	37	45	15	30
74.	30	34	43	46	18
75.	13	51	50	25	31
76.	49	50	38	23	9
77.	18	32	31	45	40
78.	8	42	48	26	40
79.	23	38	36	48	15
80.	30	24	43	33	25

No	Mécanique (O)	Algèbre lin. (O)	Algèbre (E)	Analyse (E)	Statistique (E)
81.	3	9	51	47	40
82.	7	51	43	17	22
83.	15	40	43	23	18
84.	15	38	39	28	17
85.	5	30	44	36	18
86.	12	30	32	35	21
87.	5	26	15	20	20
88.	0	40	21	9	14

La moyenne et la matrice de covariance empiriques associées à ce tableau des données sont

$$\bar{\mathbf{x}} = \begin{pmatrix} 38.95 \\ 50.59 \\ 50.60 \\ 46.68 \\ 42.31 \end{pmatrix}, \qquad S = \begin{pmatrix} 305.77 & 127.22 & 101.58 & 106.27 & 117.40 \\ 127.22 & 172.84 & 85.16 & 94.67 & 99.01 \\ 101.58 & 85.16 & 112.88 & 112.11 & 121.87 \\ 106.27 & 94.67 & 112.11 & 220.38 & 155.53 \\ 117.40 & 99.01 & 121.87 & 155.53 & 297.75 \end{pmatrix}.$$

En utilisant la décomposition spectrale de la matrice S, on trouve ses vecteurs propres orthonormés :

$$g_{(1)} = \begin{pmatrix} 0.50 \\ 0.37 \\ 0.35 \\ 0.45 \\ 0.53 \end{pmatrix}, \qquad g_{(2)} = \begin{pmatrix} -0.75 \\ -0.21 \\ 0.08 \\ 0.30 \\ 0.55 \end{pmatrix}, \qquad g_{(3)} = \begin{pmatrix} -0.30 \\ 0.42 \\ 0.14 \\ 0.60 \\ -0.60 \end{pmatrix},$$
$$g_{(4)} = \begin{pmatrix} 0.30 \\ -0.78 \\ 0.00 \\ 0.52 \\ -0.18 \end{pmatrix}, \qquad g_{(5)} = \begin{pmatrix} 0.08 \\ 0.19 \\ -0.92 \\ 0.28 \\ 0.15 \end{pmatrix},$$

et les valeurs propres correspondantes :

$$l_1 = 687.00, \qquad l_2 = 202.11, \qquad l_3 = 103.75, \qquad l_4 = 84.63, \qquad l_5 = 32.15.$$

En portant ces valeurs dans la définition

$$\tilde{r}_{ij} = g_{ij}\sqrt{\frac{l_j}{s_{ii}}},$$

on obtient le tableau des corrélations empiriques suivant :

$\tilde{r}_{ij}$	1	2	3	4	5
1	0.76	−0.61	−0.17	0.16	0.03
2	0.73	−0.22	0.32	−0.55	0.08
3	0.85	0.10	0.14	0.00	−0.49
4	0.80	0.29	0.41	0.32	0.11
5	0.81	0.45	−0.35	−0.09	0.05

Dans ce tableau, la $i^{\text{ème}}$ ligne correspond aux racines carrées des parts de la variance de la variable ξ_i (où, par exemple, ξ_2 est le vecteur des notes de l'épreuve d'algèbre linéaire) expliquées par les composantes principales.

7.8. Représentation graphique des résultats de l'ACP

1. Scree graph. Il s'agit de représenter dans un repère orthogonal l'interpolation linéaire des parts de la variance empirique expliquées par la première, deuxième, ... , $p^{\text{ème}}$ composantes principales. Pour l'exemple numérique du paragraphe précédent, $p = 5$ et

$$\begin{aligned} &\frac{l_1}{\sum_{j=1}^5 l_j} = 62\%, \quad \frac{l_2}{\sum_{j=1}^5 l_j} = 18\%, \quad \frac{l_3}{\sum_{j=1}^5 l_j} = 9\%, \\ &\frac{l_4}{\sum_{j=1}^5 l_j} = 8\%, \quad \frac{l_5}{\sum_{j=1}^5 l_j} = 3\%. \end{aligned} \tag{7.5}$$

Le scree graph est donc la courbe présentée dans la Figure 7.3. On utilise le scree graph pour choisir le nombre des composantes principales qu'il faut retenir. Plus précisément, on se donne un seuil α (par exemple, $\alpha = 0,05$) et on retient toutes les composantes principales pour lesquelles la part de la variance expliquée est supérieure à ce seuil.

2. Projection des individus. Dans le contexte de l'ACP, on appelle *individus* les n porteurs des données $\mathbf{x}_1, \dots, \mathbf{x}_n$. Ainsi, dans l'exemple numérique du paragraphe précédent, les individus sont les $n = 88$ étudiants. Le vecteur $\mathbf{x}_i$ représente l'ensemble des caractéristiques observées de l'individu numéro i. Si les $\mathbf{x}_i$ sont de dimension supérieure à deux, on ne peut pas représenter ces données de façon graphique sur le plan. Afin de visualiser les données statistiques multidimensionnelles, on les projette sur le plan engendré par les deux premiers vecteurs propres $g_{(1)}$ et $g_{(2)}$ de la matrice de covariance empirique S. On obtient ainsi la projection bidimensionnelle de l'échantillon initial :

$$\big(y_1(\mathbf{x}_1), y_2(\mathbf{x}_1)\big), \big(y_1(\mathbf{x}_2), y_2(\mathbf{x}_2)\big), \dots, \big(y_1(\mathbf{x}_n), y_2(\mathbf{x}_n)\big), \tag{7.6}$$

qui peut être visualisée à l'aide d'un nuage de points sur le plan. Ici $y_1(\cdot)$ et $y_2(\cdot)$ sont les deux premières composantes principales empiriques. Le graphique du nuage de points (7.6) sur $\mathbb{R}^2$ s'appelle **projection des individus**. Pour l'exemple numérique du paragraphe précédent, la projection des individus est présentée sur la Figure 7.2.

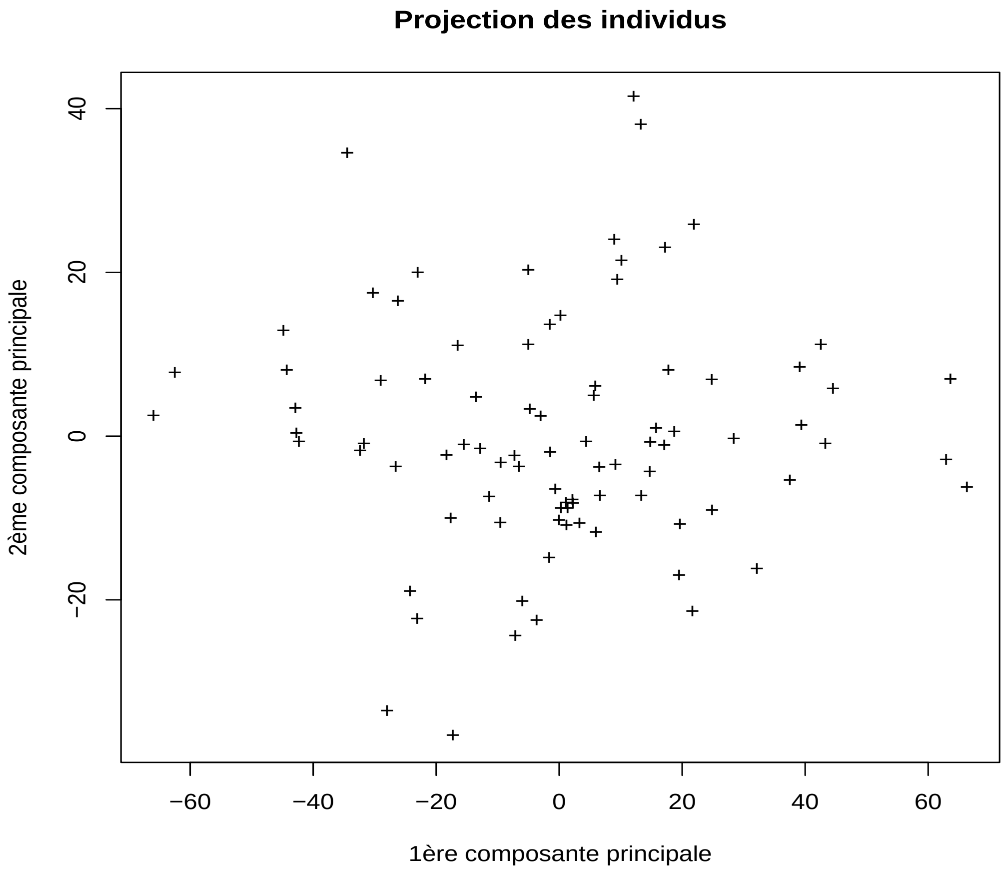

Fig. 7.2. Projection des individus.

3. Projection des variables. Les deux premières composantes principales sont souvent les plus importantes, en ce sens qu'elles expliquent la part dominante de la variance empirique. Ainsi, dans l'exemple numérique du paragraphe précédent, cette part est égale à 80% (cf. (7.5)). Dans ce cas, les corrélations empiriques $\tilde{r}_{i1}, \tilde{r}_{i2},\ i = 1, \dots, p$, entre les p variables et les deux premières composantes principales sont beaucoup plus informatives que les corrélations restantes $\tilde{r}_{ij}$ pour $j \geq 3$. Cette remarque justifie l'utilisation de l'outil graphique appelé **projection des variables sur le disque des corrélations** (ou, en abrégé, **projection des variables**). C'est un graphique sur lequel on trace le cercle des corrélations et les p points $(\tilde{r}_{i1}, \tilde{r}_{i2}),\ i = 1, \dots, p$, qui se trouvent dans le disque des corrélations. Si ces points sont proches du cercle, le graphique nous permet de juger de la dépendance linéaire ou de l'absence de corrélation entre la $i^{\text{ème}}$ et la $k^{\text{ème}}$ variables en utilisant les remarques faites à la fin du Paragraphe 7.5 (cf. Proposition 7.4) et du Paragraphe 7.6.

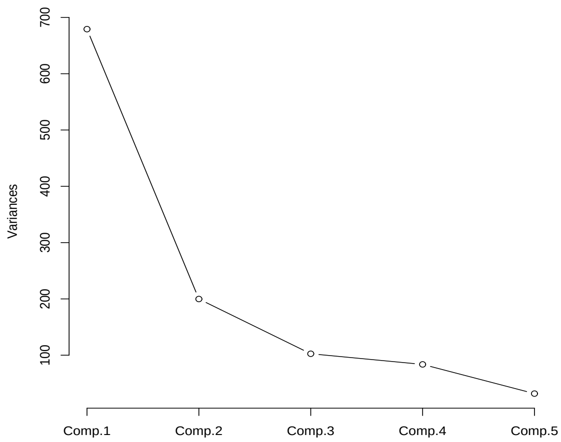

Fig. 7.3. Scree graph.

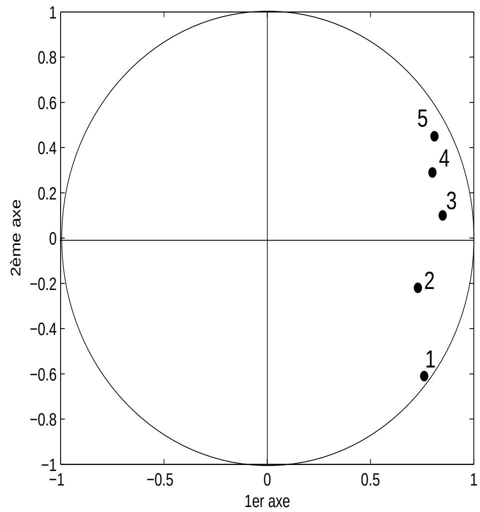

Fig. 7.4. Projection des variables.

7.9. Limites d'utilisation de l'ACP

Comme il a été expliqué au Chapitre 2, les coefficients de corrélation sont essentiellement adaptés pour décire un lien linéaire entre des variables aléatoires, si un tel lien existe. L'ACP est aussi un **outil linéaire**, en ce sens qu'elle est basée sur l'information contenue dans les corrélations. C'est pourquoi l'ACP est souvent sans intérêt si les données présentent des liens non-linéaires, tels que, par exemple, des liens quadratiques (cf. Exercice 7.9).

De manière schématique, on peut considérer que l'ACP fournit un bon résultat lorsque les données $\mathbf{x}_i$ forment un nuage de points dans $\mathbb{R}^p$ de structure ellipsoïdale, alors qu'elle donne un résultat peu satisfaisant si les données ont une structure très différente de l'ellipsoïdale, par exemple, celle de “banane” qui correspond plutôt à un lien quadratique (cf. Figure 7.5).

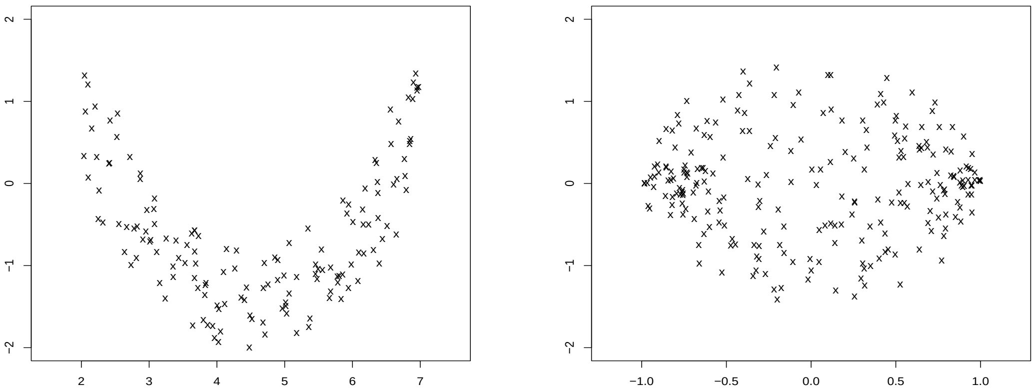

Fig. 7.5. Deux nuages de points : structure de “banane” et structure ellipsoïdale.

Finalement, il est utile de noter que, comme les corrélations empiriques ne sont pas stables par rapport aux observations aberrantes (cf. Paragraphe 4.6), les résultats de l'ACP ne le sont pas non plus. Cela signifie que la présence d'une seule observation aberrante (i.e. d'une observation $\mathbf{x}_j$ très éloignée de tous les autres $\mathbf{x}_i$) peut changer de façon radicale les résultats de l'ACP.

7.10. Exercices

Exercice 7.2. Soit (f, u_1, u_2) un vecteur aléatoire de loi $\mathcal{N}_3(0, I)$ et $\beta \in \mathbb{R}$, $\sigma \geq 0$. Posons

$$\begin{aligned} \xi_1 &= \beta f + \sigma u_1, \\ \xi_2 &= -\beta f + \sigma u_2 \end{aligned}$$

et notons $\mathbf{x} = (\xi_1, \xi_2)^T$.
1°. Donner la loi de $\mathbf{x}$. Calculer les vecteurs propres et les valeurs propres $\lambda_1 \geq \lambda_2$ de la matrice de covariance de $\mathbf{x}$.
2°. Calculer, en fonction de ξ_1 et ξ_2, puis en fonction de f, u_1 et u_2 les composantes principales η_1 et η_2 associées à $\mathbf{x}$. Montrer que $\mathrm{Var}(\eta_i) = \lambda_i$, $\mathrm{Cov}(\eta_1, \eta_2) = 0$.
3°. Calculer les corrélations $\tilde{\rho}_{ij}$ entre ξ_i et η_j. Montrer que $\tilde{\rho}_{i1}^2 + \tilde{\rho}_{i2}^2 = 1,\ i = 1, 2$.
4°. Donner le scree–graph dans les cas limites $\sigma = 0,\ \sigma = +\infty$.
5°. Tracer la projection des variables sur le disque des corrélations lorsque σ est proche de 0 ou de $+\infty$.

Exercice 7.3. Supposons qu'on ait un échantillon de n individus caractérisés par quatre variables $\xi_1, \xi_2, \xi_3, \xi_4$ dont les moyennes et les variances sont finies. On se propose d'effectuer l'ACP sur la matrice de covariance Σ du vecteur aléatoire $\mathbf{x} = (\xi_1, \xi_2, \xi_3, \xi_4)^T$. Supposons que cette matrice se met sous la forme :

$$\Sigma = \begin{pmatrix} 1 & a & b & c \\ a & 1 & c & b \\ b & c & 1 & a \\ c & b & a & 1 \end{pmatrix}$$

où $a,\ b$ et c sont des réels.
1°. Quelle est la signification des coefficients $a,\ b,\ c$ et entre quelles valeurs varient–ils ?
2°. Trouver tous les vecteurs propres de Σ, ainsi que les valeurs propres associées. Quelles inégalités doivent vérifier $a,\ b,\ c$ pour que Σ soit une matrice de corrélation ?
3°. On suppose dans toute la suite du problème que $0 \le a \le b \le c$. Quelles relations doivent satisfaire $a,\ b,\ c$ pour que le support de $\mathbf{x}$ se réduise à une droite ? à un plan ? à un espace de dimension 3 ?
4°. Soit η_j la $j^{\text{ème}}$ composante principale pour l'ACP sur la matrice de covariance Σ. Calculer la corrélation $\tilde{\rho}_{ij}$ entre η_j et ξ_i pour $i, j = 1, \ldots, 4$. On disposera ces corrélations dans un tableau carré.
5°. Que peut–on dire de la projection des variables sur le disque des corrélations lorsque $a = b = c$? $a = b$? $b = c$?
6°. Application numérique : soit $a = 0.1,\ b = 0.4,\ c = 0.6$. Préciser les valeurs propres de Σ, les composantes principales et les parts de variance expliquées. Tracer le scree–graph et la projection des variables sur le disque des corrélations.

EXERCICE 7.4. Pendant 28 ans, un laboratoire a observé des réalisations de 4 variables météorologiques suivantes :

$$\begin{aligned}
\xi_1 &= \text{précipitations en juillet (en mm)},\\
\xi_2 &= \text{température moyenne en juillet (en degrés Celsius)},\\
\xi_3 &= \text{vitesse moyenne du vent en juillet (en km/h)},\\
\xi_4 &= \text{précipitations en septembre (en mm)}
\end{aligned}$$

La matrice de covariance empirique obtenue à partir de ces observations est la suivante :

$$S = \begin{pmatrix} 140,017 & 107,881 & 139,068 & 109,095 \\ & 106,038 & 110,0439 & 82,627 \\ & & 168,752 & 125,136 \\ & & & 108,960 \end{pmatrix},$$

alors que les corrélations empiriques $\tilde{r}_{ij}$ entre les variables et les composantes principales valent :

$$(\tilde{r}_{ij})_{i,j=1,\ldots,4} = \begin{pmatrix} 0.969 & -0.103 & 0.191 & 0.119 \\ 0.906 & -0.394 & -0.105 & -0.111 \\ 0.970 & 0.160 & -0.156 & 0.090 \\ 0.943 & 0.249 & 0.096 & -0.197 \end{pmatrix}.$$

1°. Calculer les variances empiriques l_i des composantes principales et tracer le scree–graph.
2°. Calculer la part de variance de la première variable expliquée par les deux dernières composantes principales, et la part de variance de la deuxième variable expliquée par les deux premières composantes principales.
3°. Faire la projection des variables sur le disque des corrélations et commenter le résultat.

EXERCICE 7.5. Soit $\mathbf{x} \in \mathbb{R}^4$ un vecteur aléatoire de moyenne μ et de matrice de covariance Σ. On suppose que les éléments diagonaux de Σ sont $\sigma_{ii} = 1$. On souhaite effectuer l'analyse en composantes principales basé sur la matrice de covariance Σ.
1°. Soit $0 < \rho < 1$. L'un des deux graphiques ci–contre présente la projection des variables

sur le disque des corrélations. Lequel ?

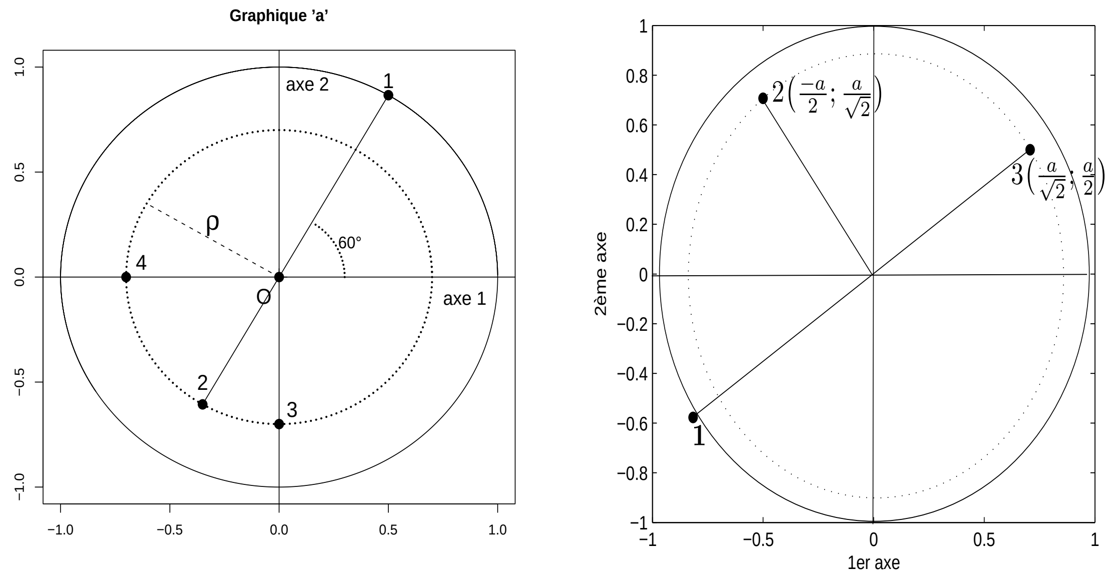

Fig. 7.6.

Les deux questions suivantes utilisent la projection des variables choisie en 1°.
2°. Sans effectuer les calculs donner l'interprétation la plus complète possible de
- corrélations entre les variables,
- corrélations entre les variables et les composantes principales.

Que se passe-t-il si $\rho = 1$?
3°. Calculer la part de la variance totale expliquée par les deux premières composantes principales.

Exercice 7.6. Soit un vecteur aléatoire $\mathbf{x} = (\xi_1, \xi_2, \xi_3)^T$ de moyenne 0 et de matrice de covariance

$$\Sigma = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$$

où $\rho \geq 0$ est une valeur donnée.
1°. Chercher la plus grande valeur ρ_1 telle que Σ soit bien une matrice de covariance quand $\rho \in P = [0, \rho_1]$. On suppose dans la suite que $\rho \in P$.
2°. Déterminer les composantes principales η_j de $\mathbf{x}$, ainsi que leurs variances.
3°. Calculer les parts de variance de chacune des variables ξ_1, ξ_2, ξ_3 expliquées par η_1, puis par η_2. Quelle est la valeur minimale, pour $\rho \in P$, de la part de variance de ξ_1 expliquée par le couple (η_1, η_2) ?
4°. Faire la projection des variables sur le disque des corrélations. Commenter le graphique obtenu dans les deux cas limites : $\rho = 0$ et $\rho = \rho_1$.

Exercice 7.7. Soit $\mathbf{x} \in \mathbb{R}^\mathbf{3}$ un vecteur aléatoire de moyenne μ et de matrice de covariance Σ. On souhaite effectuer l'analyse en composantes principales basé sur la matrice de covariance Σ. Soit $0 < a < 1$. Le graphique ci-contre présente la projection des variables sur le disque des corrélations.

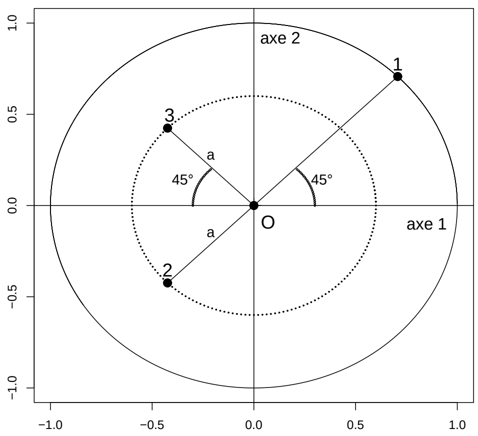

Fig. 7.7.

1°. Commenter le graphique.
2°. Calculer la corrélation entre la 2$^{\text{ème}}$ variable et la 2$^{\text{ème}}$ composante principale.
3°. Démontrer que la corrélation entre la 1$^{\text{ère}}$ et la 2$^{\text{ème}}$ variables est négative.

EXERCICE 7.8. Soit $\mathbf{x} \in \mathbb{R}^{\mathbf{3}}$ un vecteur aléatoire de moyenne μ et de matrice de covariance Σ. On souhaite effectuer l'analyse en composantes principales basé sur la matrice de covariance Σ. Soit $0 < a < 1$. Le graphique ci-dessous présente la projection des variables sur le disque des corrélations.

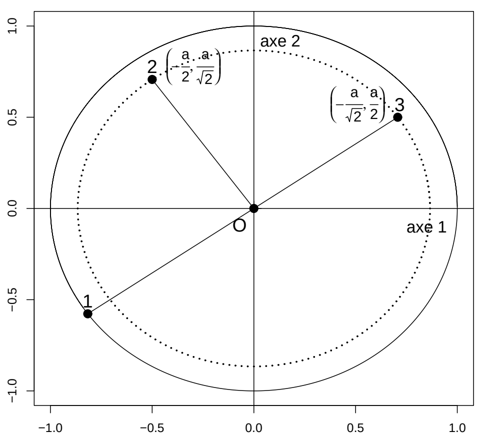

Fig. 7.8.

1°. Calculer :
– la part de variance de la 2$^{\text{ème}}$ variable expliquée par la 3$^{\text{ème}}$ composante principale,
– la corrélation entre la 1$^{\text{ère}}$ variable et la 2$^{\text{ème}}$ composante principale.

2°. Déterminer la corrélation entre la 1$^{\text{ère}}$ et la 2$^{\text{ème}}$ variable, puis la corrélation entre la 1$^{\text{ère}}$ et la 3$^{\text{ème}}$ variable. Commenter le résultat.

3^o. On suppose maintenant que la matrice Σ se met sous la forme

$$\Sigma = \begin{pmatrix} b & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & b & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & b \end{pmatrix}$$

où les σ_{ij} sont des constantes réelles inconnues et $b > 0$. En utilisant les valeurs données sur le graphique, déterminer les variances de deux premières composantes principales.

Exercice 7.9. Soit $\mathbf{x} = (\xi_1, \xi_2)^T$, où $\xi_2 = \xi_1^2$ et $\xi_1 \sim \mathcal{N}(0, 1)$. Effectuer l'ACP sur la matrice de covariance de $\mathbf{x}$ et remarquer que la part de variance de ξ_2 expliquée par la deuxième composante principale $\eta_2 = \xi_1$ vaut 0, alors que ξ_1 et ξ_2 sont fonctionnellement liés.

8

Régression linéaire multivariée

8.1. Le problème d'estimation de régression multivariée

Soient $\mathbf{x}$ un vecteur aléatoire p-dimensionnel et Y une variable aléatoire réelle, tels que $E(\|\mathbf{x}\|^2) < \infty$ et $E(Y^2) < \infty$, où $\|\cdot\|$ désigne la norme Euclidienne. La *fonction de régression* de Y sur $\mathbf{x}$ est une fonction $g : \mathbb{R}^p \to \mathbb{R}$ définie par :

$$g(z) = E(Y|\,\mathbf{x} = z), \quad z \in \mathbb{R}^p.$$

Cette fonction, comme dans le cas unidimensionnel, jouit de la propriété de *meilleure prévision*, i.e.

$$E\big[\big(Y - g(\mathbf{x})\big)^2\big] = \min_{h(\cdot)} E\big[\big(Y - h(\mathbf{x})\big)^2\big],$$

où le minimum est cherché dans l'ensemble de toutes les fonctions boréliennes $h(\cdot)$ (cf. Paragraphe 3.3). On peut alors écrire

$$Y = g(\mathbf{x}) + \xi, \quad \text{où} \quad E(\xi\,|\,\mathbf{x}) = 0$$

(cf. Chapitres 2 et 3).

Dans ce chapitre, nous supposerons que l'on dispose d'un échantillon $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$ tel que

$$Y_i = g(\mathbf{x}_i) + \xi_i, \quad i = 1, \ldots, n,$$

où les ξ_i sont des variables aléatoires centrées et mutuellement indépendantes. Nous considérerons le problème statistique de l'estimation de la fonction de régression g à partir de cet échantillon. Plus particulièrement, nous nous intéresserons seulement à la situation quand la régression est *linéaire* :

$$g(\mathbf{x}) = \theta^T \mathbf{x},$$

où $\theta \in \mathbb{R}^p$ est un paramètre vectoriel : $\theta = (\theta_1, \ldots, \theta_p)^T$. Les observations Y_i sont alors de la forme

$$\boxed{Y_i = \theta^T \mathbf{x}_i + \xi_i, \quad i = 1, \ldots, n,} \tag{8.1}$$

et l'estimation de la fonction g se réduit à l'estimation du paramètre inconnu θ. Le modèle statistique défini par (8.1) s'appelle **modèle de régression linéaire multidimensionnelle** (ou multivariée). L'importance de ce modèle pour les appications statistiques s'explique d'une part par sa relative simplicité et d'autre part par le fait qu'il permet d'inclure comme des cas particuliers un certain nombre de modèles qui semblent, à la première vue, non-linéaires.

EXEMPLE 8.1. *Régression linéaire simple.* Posons $\theta = (a, b)^T$ et $\mathbf{x} = (1, Z)^T$ avec $a, b \in \mathbb{R}$, où Z une variable aléatoire réelle. Notons que dans ce cas la première composante du vecteur aléatoire $\mathbf{x}$ est déterministe (non aléatoire). Les observations Y_i sont alors de la forme

$$Y_i = a + bZ_i + \xi_i, \quad i = 1, \ldots, n,$$

où les Z_i sont des réalisations de la variable Z.

EXEMPLE 8.2. *Régression polynomiale.* Soit Z une variable aléatoire réelle. Puisque toute fonction suffisamment régulière peut être décomposée selon la formule de Taylor, il est naturel de chercher la dépendance entre Y et Z sous une forme polynomiale :

$$Z \mapsto \theta_1 + \theta_2 Z + \cdots + \theta_p Z^{p-1},$$

où $p \geq 1$ est un entier et $\theta_1, \ldots, \theta_p$ sont des coefficients inconnus. Si l'on définit les vecteurs $\mathbf{x} = (1, Z, \ldots, Z^{p-1})^T$ et $\theta = (\theta_1, \ldots, \theta_p)^T$, on obtient

$$g(\mathbf{x}) = \theta^T \mathbf{x}.$$

On voit donc que la régression polynomiale est un cas particulier de la régression linéaire multidimensionnelle. Dans ce cas aussi, comme pour la régression linéaire simple, la première composante du vecteur aléatoire $\mathbf{x}$ est déterministe.

EXEMPLE 8.3. *Régression non-linéaire transformée.* Il existe des modèles non-linéaires de régression qui peuvent être réduits aux modèles linéaires par une transformation. Par exemple, supposons que la fonction de régression $g(\cdot)$ est de la forme

$$g(\mathbf{x}) = A\, e^{v^T \mathbf{x}} \quad \text{avec} \quad \mathbf{x}, v \in \mathbb{R}^k,$$

où v est un vecteur des paramètres inconnus et $A > 0$ est une constante inconnue. Des fonctions de régression de ce type sont utilisés, par exemple, dans les applications en économie, pour modéliser la productivité des entreprises. En prenant les logarithmes, on obtient

$$\ln g(\mathbf{x}) = \ln A + v^T \mathbf{x}.$$

Afin de se ramener à une régression linéaire, on pose $\theta = (\ln A, v^T)^T$, $\mathbf{x}' = (1, \mathbf{x}^T)^T$ et on obtient

$$Y_i' = \ln Y_i = \theta^T \mathbf{x}_i' + \xi_i', \qquad i = 1, \ldots, n. \tag{8.2}$$

C'est un modèle de régression linéaire par rapport à l'échantillon transformé

$$(\mathbf{x}_1', Y_1'), \ldots, (\mathbf{x}_n', Y_n').$$

Notons que formellement on arrive à (8.2) à partir du modèle $Y_i = g(\mathbf{x}_i)\xi_i$ de régression où les erreurs ξ_i interviennent de façon multiplicative et non pas additive (on a alors $\xi_i' = \ln \xi_i$).

Néanmoins, souvent la transformation logarithmique est utilisée sans mentionner cette nuance de manière explicite.

8.2. Méthode des moindres carrés

Une méthode usuelle et très répandue pour estimer le paramètre $\theta \in \mathbb{R}^p$ est celle des moindres carrés. Elle consiste à chercher une valeur $\theta = \hat{\theta}$ qui minimise la somme des carrés des déviations :

$$\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\hat{\theta})^2 = \min_{\theta\in\mathbb{R}^p}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\theta)^2.$$

Il est facile de voir qu'il existe toujours une solution $\hat{\theta}$ de ce problème de minimisation que l'on appelle **estimateur des moindres carrés** de θ. On écrit alors

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\mathbb{R}^p}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\theta)^2.$$

L'estimateur des moindres carrés n'est pas toujours unique. La condition de l'unicité est donnée dans la proposition suivante.

Proposition 8.1. *Supposons que la matrice*

$$B = \sum_{i=1}^{n}\mathbf{x}_i\,\mathbf{x}_i^T \in \mathbb{R}^{p\times p}$$

soit strictement positive. Alors, l'estimateur des moindres carrés est unique et il s'écrit sous la forme

$$\boxed{\hat{\theta} = B^{-1}\sum_{i=1}^{n}\mathbf{x}_i\,Y_i.}$$

Preuve. La condition nécessaire pour que $\hat{\theta}$ soit un point de minimum pour $h(\theta) = \sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\theta)^2$ est $(\partial h/\partial\theta_i)(\hat{\theta}) = 0$ pour tout $i = 1, \ldots, p$. Cette condition équivaut à

$$2\sum_{i=1}^{n}\mathbf{x}_i(Y_i - \mathbf{x}_i^T\hat{\theta}) = 0$$

ou encore

$$B\hat{\theta} = \sum_{i=1}^{n}\mathbf{x}_i\,Y_i. \tag{8.3}$$

C'est un système de p équations linéaires qui admet une solution unique car la matrice B est inversible. Cette solution vaut

$$\hat{\theta} = B^{-1}\sum_{i=1}^{n}\mathbf{x}_i\,Y_i.$$

Comme la fonction $h(\theta)$ est convexe et positive, ce vecteur $\hat{\theta}$ fournit le minimum global de h. ∎

Il est convenable d'écrire le modèle de régression linéaire sous la forme matricielle :

$$y = \mathbf{X}\,\theta + \xi,$$

où $y = (Y_1, \ldots, Y_n)^T$, $\theta = (\theta_1, \ldots, \theta_p)^T$, $\xi = (\xi_1, \ldots, \xi_p)^T$ et $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. Avec ces notations, on a $B = \mathbf{X}^T\mathbf{X}$, et on peut écrire l'estimateur des moindres carrés sous la forme

$$\boxed{\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y.}$$

Le système des équations linéaires (8.3) s'appelle **système des équations normales** pour la méthode des moindres carrés. On peut l'écrire sous la forme

$$B\theta = \mathbf{X}^T y.$$

Proposition 8.2. *La matrice*

$$B = \sum_{i=1}^{n} \mathbf{x}_i\,\mathbf{x}_i^T = \mathbf{X}^T\mathbf{X}$$

est toujours positive. Afin qu'elle soit strictement positive, il est nécessaire et suffisant que le rang de la matrice $\mathbf{X}$ *soit égal à* p.

Preuve. Notons d'abord que B est positive, car tout $v \in \mathbb{R}^p \setminus \{0\}$ vérifie l'inégalité

$$v^T B v = v^T \mathbf{X}^T \mathbf{X} v = w^T w = \sum_{i=1}^{p} w_i^2 \geq 0,$$

où $w = \mathbf{X}v = (w_1, \ldots, w_p)$. Il est évident que l'inégalité précédente devient égalité si et seulement si $w = \mathbf{X}v = 0$. Or, $\mathbf{X}v = 0$ pour un vecteur v différent de 0 implique que le rang de $\mathbf{X}$ est strictement inférieur à p. On a donc montré que si B n'est pas strictement positive, alors $\mathrm{Rang}(\mathbf{X}) < p$.

La preuve de la réciproque est similaire. Si $\mathrm{Rang}(\mathbf{X}) < p$, alors il existe un vecteur $v \in \mathbb{R}^p \setminus \{0\}$ tel que $\mathbf{X}v = 0$. Il en résulte que $v^T B v = v^T \mathbf{X}^T \mathbf{X} v = 0$. Par conséquent, la matrice B n'est pas strictement positive. ∎

Une conséquence immédiate de cette proposition est la suivante : *si la taille d'échantillon n est strictement inférieure à la dimension p des observations, la matrice B est dégénérée.* En effet, $n < p$ implique que $\mathrm{Rang}(\mathbf{X}) < p$, car le rang d'une matrice M est le nombre maximal des lignes de M qui forment une famille de vecteurs libre. Une autre formulation de cette propriété est :

$$B > 0 \quad \Longrightarrow \quad n \geq p.$$

8.2.1. Interprétation géométrique de la méthode des moindres carrés. Le problème de minimisation de la somme des carrés des déviations peut s'écrire sous la forme suivante :

$$\min_{\theta \in \mathbb{R}^p} \|y - \mathbf{X}\theta\|^2 = \min_{v \in D} \|y - v\|^2 \tag{8.4}$$

où D désigne le sous-espace linéaire de $\mathbb{R}^n$ défini par

$$D = \big\{ v \in \mathbb{R}^n \,:\, v = \mathbf{X}\theta,\ \theta \in \mathbb{R}^p \big\}.$$

En mots, D est le sous-espace linéaire de $\mathbb{R}^n$ engendré par les p colonnes de la matrice $\mathbf{X}$. Si $\mathbf{X}$ est une matrice de rang p, ce qui est vrai lorsque $B > 0$, alors D est un sous-espace linéaire de dimension p :

$$\text{Rang}(\mathbf{X}) = p \quad \Longleftrightarrow \quad B > 0 \quad \Longleftrightarrow \quad \dim(D) = p.$$

Si $B > 0$, la solution du problème (8.4) est $\hat{v} = \mathbf{X}\hat{\theta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \stackrel{\text{déf}}{=} Ay$.

Définition 8.1. *Soit $B > 0$. La matrice*

$$\boxed{A = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \in \mathbb{R}^{n\times n}}$$

est dite **matrice chapeau** *("hat" matrice).*

Proposition 8.3. *Supposons que $B > 0$. Alors la matrice A est symétrique, idempotente, $\text{Rang}(A) = p$ et A est le projecteur dans $\mathbb{R}^n$ sur le sous-espace D.*

Preuve. Il vient

$$A^T = \mathbf{X}[(\mathbf{X}^T\mathbf{X})^{-1}]^T\mathbf{X}^T = \mathbf{X}[(\mathbf{X}^T\mathbf{X})^T]^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = A$$

et

$$A^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = A.$$

Donc A est symétrique et idempotente, ce qui signifie que A est un projecteur. En outre, pour tout $y \in \mathbb{R}^n$, on a $Ay = \mathbf{X}\hat{\theta} = \hat{v} \in D$. Donc A projette sur un sous-ensemble de D. Mais ce sous-ensemble coïncide avec D, car pour tout vecteur $v \in D$ il existe $\theta \in \mathbb{R}^p$ tel que $v = \mathbf{X}\theta$ et, par conséquent,

$$Av = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T v = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\theta = \mathbf{X}\theta = v.$$

Cela signifie que A est le projecteur sur D. Comme D est un sous-espace de $\mathbb{R}^n$ de dimension p, le rang de A est égal à p. ∎

8.3. Propriétés statistiques de la méthode des moindres carrés

Supposons que l'hypothèse suivante soit vérifiée.

Hypothèse (R).

(R1) Les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ appartenant à $\mathbb{R}^p$ sont déterministes.

(R2) La matrice B est strictement positive.

(R3) Le vecteur aléatoire ξ est de moyenne $E(\xi) = 0$ et de matrice de covariance $V(\xi) = \sigma^2 I_n$, où $\sigma^2 > 0$ et I_n est la matrice unité de dimension $n \times n$.

Théorème 8.1. *Sous l'Hypothèse (R), l'estimateur des moindres carrés est sans biais :*

$$E(\hat{\theta}) = \theta \tag{8.5}$$

et sa matrice de covariance $V(\hat{\theta}) = E\big[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\big]$ vaut

$$V(\hat{\theta}) = \sigma^2 B^{-1}.$$

Preuve. Il vient
$$\hat{\theta} = B^{-1}\mathbf{X}^T y = B^{-1}\mathbf{X}^T(\mathbf{X}\theta + \xi) = \theta + B^{-1}\mathbf{X}^T\xi, \tag{8.6}$$
d'où découle (8.5). En utilisant (8.6) on obtient aussi que
$$V(\hat{\theta}) = E\big[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\big] = E[(B^{-1}\mathbf{X}^T\xi)(\xi^T\mathbf{X}B^{-1})] = B^{-1}\mathbf{X}^T E[\xi\xi^T]\mathbf{X}B^{-1}.$$
Comme $V(\xi) = E[\xi\xi^T] = \sigma^2 I_n$, on obtient
$$B^{-1}\mathbf{X}^T E[\xi\xi^T]\mathbf{X}B^{-1} = \sigma^2 B^{-1}\mathbf{X}^T\mathbf{X}B^{-1} = \sigma^2 B^{-1}.$$
■

Théorème 8.2. *Sous l'Hypothèse (R), la statistique*
$$\hat{\sigma}^2 \overset{\text{déf}}{=} \frac{\|y - \mathbf{X}\hat{\theta}\|^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^n (Y_i - \mathbf{x}_i^T\hat{\theta})^2$$
est un estimateur sans biais de la variance σ^2 :
$$E(\hat{\sigma}^2) = \sigma^2.$$

Preuve. Notons d'abord que les observations y proviennent du modèle $y = \mathbf{X}\theta + \xi$, ce qui implique que $y - \mathbf{X}\hat{\theta} = \mathbf{X}(\theta - \hat{\theta}) + \xi$. Vu (8.6), il en résulte que
$$y - \mathbf{X}\hat{\theta} = -\mathbf{X}B^{-1}\mathbf{X}^T\xi + \xi = (I_n - \mathbf{X}B^{-1}\mathbf{X}^T)\xi = (I_n - A)\xi. \tag{8.7}$$
Par conséquent,
$$E[\|y - \mathbf{X}\hat{\theta}\|^2] = E[\xi^T(I_n - A)^T(I_n - A)\xi] = E[\xi^T(I_n - A)^2\xi] = E[\xi^T(I_n - A)\xi],$$
où on a utilisé le fait que A est une matrice idempotente. Désignons par a_{ij} les éléments de A. On a alors
$$E[\xi^T(I_n - A)\xi] = \sum_{i,j=1}^n (\delta_{ij} - a_{ij})\, E[\xi_i\xi_j] = \sigma^2\sum_{i,j=1}^n (\delta_{ij} - a_{ij})\,\delta_{ij} = \sigma^2\sum_{i=1}^n (1 - a_{ii}) = \sigma^2(n - \mathrm{Tr}(A)),$$
où δ_{ij} est le symbole de Kronecker. Comme A est un projecteur, ses valeurs propres valent 0 ou 1. D'après la Proposition 8.3, $\mathrm{Rang}(A) = p$, donc il y a exactement p valeurs propres égales à 1. On en déduit que $\mathrm{Tr}(A) = p$, d'où le résultat. ■

8.4. Régression linéaire normale

Supposons maintenant que les variables aléatoires ξ_i suivent la loi normale $\mathcal{N}(0, \sigma^2)$. Dans ce cas la condition (R3) entraîne l'indépendance des variables aléatoires ξ_i.

Hypothèse (NR). *L'Hypothèse (R) est vérifiée et ξ est un vecteur gaussien.*

Sous l'Hypothèse (NR), $\hat{\theta}$ est l'estimateur du maximum de vraisemblance du paramètre θ (cf. Exercice 8.2).

Le théorème suivant permet de déduire la loi jointe de $(\hat{\theta}, \hat{\sigma}^2)$ sous l'Hypothèse (NR). Ce théorème est une généralisation multidimensionnelle de la Proposition 4.4.

Théorème 8.3. *Si l'Hypothèse (NR) est vérifiée, alors*

(i) $\hat{\theta} \sim \mathcal{N}_p(\theta, \sigma^2 B^{-1})$,

(ii) $\hat{\theta} \perp\!\!\!\perp y - \mathbf{X}\hat{\theta}$ *et* $y - \mathbf{X}\hat{\theta} \perp\!\!\!\perp \mathbf{X}(\hat{\theta} - \theta)$,

(iii) $\sigma^{-2}\|y - \mathbf{X}\hat{\theta}\|^2 \sim \chi^2_{n-p}$ *et* $\sigma^{-2}\|\mathbf{X}(\hat{\theta} - \theta)\|^2 \sim \chi^2_p$.

Preuve. D'après (8.6) et (8.7),

$$\hat{\theta} - \theta = B^{-1}\mathbf{X}^T\xi, \qquad y - \mathbf{X}\hat{\theta} = (I_n - A)\,\xi. \tag{8.8}$$

La première égalité, compte tenu du fait que B et $\mathbf{X}$ sont déterministes, implique que $\hat{\theta}$ est un vecteur gaussien. D'après le Théorème 8.1, la moyenne de ce vecteur est θ et sa matrice de covariance vaut $\sigma^2 B^{-1}$, d'où le résultat (i).

Vu (8.8), le vecteur aléatoire $\big(y - \mathbf{X}\hat{\theta}, \hat{\theta}\big) \in \mathbb{R}^{n+p}$ est gaussien comme transformation affine du vecteur gaussien ξ. De plus, la matrice de covariance entre $\hat{\theta}$ et $y - \mathbf{X}\hat{\theta}$ est

$$C(\hat{\theta}, y - \mathbf{X}\hat{\theta}) = E[(\hat{\theta} - \theta)(y - \mathbf{X}\hat{\theta})^T] = E[B^{-1}\mathbf{X}^T\xi\xi^T(I_n - A)] = \sigma^2(B^{-1}\mathbf{X}^T - B^{-1}\mathbf{X}^T A) = \mathbf{0}.$$

En utilisant la propriété (N6) de la loi normale multidimensionnelle démontrée au Chapitre 3, on obtient la première partie du résultat (ii). Sa deuxième partie en découle vu la préservation de l'indépendance par transformations mesurables.

Pour prouver le résultat (iii) du théorème, introduisons le vecteur aléatoire $\xi' = \xi/\sigma$ et appliquons le Théorème de Cochran (cf. Chapitre 3). D'après (8.8), $y - \mathbf{X}\hat{\theta} = \sigma(I_n - A)\xi'$ et $\mathbf{X}(\hat{\theta} - \theta) = \sigma\mathbf{X}B^{-1}\mathbf{X}^T\xi' = \sigma A\xi'$. Par ailleurs, la Proposition 8.3 implique que les matrices A et $I_n - A$ sont symétriques et idempotentes, $(I_n - A)A = 0$, $\mathrm{Rang}(A) = p$ et $\mathrm{Rang}(I_n - A) = n - p$. D'après le Théorème de Cochran, ceci entraîne le résultat (iii). ∎

8.5. Application au problème de prévision

Considérons d'abord un exemple de problème de prévision qui motive ce qui va suivre.

EXEMPLE 8.4. *Prévision dans le modèle de régression sur le temps.* Supposons que l'on dispose des données statistiques (Y_i, x_i), $i = 1, \dots, n$, où $x_i = i\Delta$ et $\Delta > 0$ est un nombre fixé, telles que $Y_i = \theta x_i + \xi_i$, $i = 1, \dots, n$, avec $\theta \in \mathbb{R}$. On peut penser à Y_i comme à la valeur à l'instant $i\Delta$ d'une variable Y évoluant dans le temps de manière aléatoire (exemples : la température, le niveau de l'eau dans un fleuve, le cours d'une option financière, etc). Le problème de prévision consiste à donner un estimateur $\hat{Y}_0$ qui approche bien la valeur de la fonction de régression $g(x_0) = \theta x_0$ à l'instant donné x_0 tel que $x_0 > x_n = n\Delta$. Une méthode très répandue est de chercher une prévision *linéaire* de la forme $\hat{Y}_0 = \bar{\theta}x_0$, où $\bar{\theta}$ est un estimateur convenable de θ. Le plus souvent on utilise $\bar{\theta} = \hat{\theta}$, l'estimateur des moindres carrés de θ.

Considérons maintenant le cas général quand les $\mathbf{x}_i$ sont multidimensionnels. Soit $\mathbf{x}_0 \in \mathbb{R}^p$ un vecteur donné. Le problème est formulé de manière similaire : trouver une prévision $\hat{Y}_0$ de $g(\mathbf{x}_0) = \theta^T\mathbf{x}_0$, étant donné un échantillon $(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$ provenant du modèle de régression linéaire

$$Y_i = \theta^T\mathbf{x}_i + \xi_i, \qquad i = 1, \dots, n.$$

La recherche d'une prévision linéaire de la forme $\hat{Y}_0 = \bar{\theta}^T \mathbf{x}_0$ revient à la recherche d'un estimateur $\bar{\theta}$ du paramètre θ. Un choix possible est $\bar{\theta} = \hat{\theta}$, l'estimateur des moindres carrés de θ. La valeur $\hat{Y}_0 = \hat{\theta}^T \mathbf{x}_0$ est donc une prévision de $g(\mathbf{x}_0)$. Les propriétés de cette prévision sont données dans le théorème suivant.

Théorème 8.4.

(i) Si l'Hypothèse (R) est vérifiée,

$$E(\hat{Y}_0) = \theta^T \mathbf{x}_0 \quad et \quad \mathrm{Var}(\hat{Y}_0) = \sigma^2 \mathbf{x}_0^T B^{-1} \mathbf{x}_0.$$

(ii) Si l'Hypothèse (NR) est vérifiée,

$$\hat{Y}_0 \sim \mathcal{N}(\theta^T \mathbf{x}_0, \sigma^2 \mathbf{x}_0^T B^{-1} \mathbf{x}_0) \quad et \quad \hat{Y}_0 - \theta^T \mathbf{x}_0 \perp\!\!\!\perp y - X\hat{\theta}.$$

Preuve. Elle est immédiate d'après les Théorèmes 8.1 et 8.3. ∎

La seconde partie de ce théorème nous permet de construire un intervalle de confiance pour $g(\mathbf{x}_0) = \theta^T \mathbf{x}_0$. En effet, d'après la partie (ii) du Théorème 8.4, si l'Hypothèse (NR) est satisfaite,

$$\eta \stackrel{\text{déf}}{=} \frac{\hat{Y}_0 - \theta^T \mathbf{x}_0}{\sqrt{\sigma^2 \mathbf{x}_0^T B^{-1} \mathbf{x}_0}} \;\sim\; \mathcal{N}(0,1).$$

Cette relation implique, en particulier, que

$$P(g(\mathbf{x}_0) \in [\underline{g}, \overline{g}]) = 1 - \alpha,$$

où

$$\underline{g} = \hat{Y}_0 - \sqrt{\sigma^2 \mathbf{x}_0^T B^{-1} \mathbf{x}_0}\;\; q^N_{1-\alpha/2},$$
$$\overline{g} = \hat{Y}_0 + \sqrt{\sigma^2 \mathbf{x}_0^T B^{-1} \mathbf{x}_0}\;\; q^N_{1-\alpha/2}.$$

Donc, dans le cas où la variance σ est connue, l'intervalle $[\underline{g}, \overline{g}]$ est un intervalle de confiance de taille exacte $1 - \alpha$ pour $g(\mathbf{x}_0)$.

Lorsque la variance σ^2 est inconnue, il est naturel de la remplacer par son estimateur sans biais $\hat{\sigma}^2$ défini dans le Théorème 8.2. Pour pouvoir construire un intervalle de confiance exacte, il nous faut connaître la loi de la v. a.

$$t \stackrel{\text{déf}}{=} \frac{\hat{Y}_0 - \theta^T \mathbf{x}_0}{\sqrt{\hat{\sigma}^2 \mathbf{x}_0^T B^{-1} \mathbf{x}_0}}\;.$$

D'après le Théorème 8.4, les variables aléatoires η et $\chi \stackrel{\text{déf}}{=} (n-p)\hat{\sigma}^2/\sigma^2 = \|y - \mathbf{X}\hat{\theta}\|^2/\sigma^2$ sont indépendantes. Par conséquent, la variable aléatoire t peut être représentée sous la forme

$$t = \frac{\eta}{\sqrt{\chi/(n-p)}}\;,$$

où $\eta \sim \mathcal{N}(0,1)$, $\chi \sim \chi^2_{n-p}$ et $\eta \perp\!\!\!\perp \chi$. Il en résulte que t suit la loi de Student t_{n-p} avec $n - p$ degrés de liberté. On en déduit que $[\underline{g}', \overline{g}']$ est un intervalle de confiance de taille exacte $1 - \alpha$

pour $g(\mathbf{x}_0)$ si

$$\underline{g}' = \hat{Y}_0 - \sqrt{\hat{\sigma}^2 \mathbf{x}_0^T B^{-1} \mathbf{x}_0}\ q_{1-\alpha/2}(t_{n-p}),$$
$$\overline{g}' = \hat{Y}_0 + \sqrt{\hat{\sigma}^2 \mathbf{x}_0^T B^{-1} \mathbf{x}_0}\ q_{1-\alpha/2}(t_{n-p}).$$

Soulignons que l'hypothèse de normalité des erreurs ξ_i est cruciale pour que $[\underline{g}', \overline{g}']$ soit un intervalle de confiance de taille exacte $1 - \alpha$.

8.6. Application aux tests sur le paramètre θ

Dans ce paragraphe, on supposera que les erreurs ξ_i du modèle de régression sont normales et que l'Hypothèse (NR) est vérifiée. Notre premier objectif est de tester l'hypothèse

$$H_0 \ : \ \theta_j = a$$

contre l'hypothèse alternative

$$H_1 \ : \ \theta_j \neq a,$$

où $a \in \mathbb{R}$ est une valeur donnée et θ_j est la $j^{\text{ème}}$ coordonnée du vecteur θ. Désignons par $\hat{\theta}_j$ la $j^{\text{ème}}$ coordonnée de l'estimateur des moindres carrés $\hat{\theta}$ et par b_j le $j^{\text{ème}}$ élément diagonal de la matrice B^{-1}. L'Hypothèse (R2) implique que $b_j > 0$ pour $j = 1, \dots, p$.

Corollaire 8.1. *Si l'Hypothèse (NR) est vérifiée,*

$$\frac{\hat{\theta}_j - \theta_j}{\sigma\sqrt{b_j}} \ \sim \ \mathcal{N}(0, 1).$$

Preuve. D'après le Théorème 8.3, $\hat{\theta} - \theta \sim \mathcal{N}(0, \sigma^2 B^{-1})$. Soit v_j le vecteur de $\mathbb{R}^p$ dont toutes les coordonnées sont nulles sauf la $j^{\text{ème}}$ qui vaut 1. La v. a. $(\hat{\theta}_j - \theta_j)$ est donc égale à $(\hat{\theta} - \theta)^T v_j$, ce qui entraîne qu'elle est suit une loi gaussienne. Afin d'identifier cette loi, il suffit de calculer sa moyenne et sa variance :

$$E(\hat{\theta}_j - \theta_j) = E[(\hat{\theta} - \theta)^T v_j] = 0,$$
$$\mathrm{Var}(\hat{\theta}_j - \theta_j) = E\big[\big((\hat{\theta} - \theta)^T v_j\big)^2\big] = v_j^T E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] v_j = \sigma^2 v_j^T B^{-1} v_j = \sigma^2 b_j.$$

On a alors $\hat{\theta}_j - \theta \sim \mathcal{N}(0, \sigma^2 b_j)$ ou encore $(\sigma^2 b_j)^{-1/2}(\hat{\theta}_j - \theta) \sim \mathcal{N}(0, 1)$. ∎

Si le paramètre σ est inconnu, nous ne pouvons pas utiliser la statistique $(\sigma^2 b_j)^{-1/2}(\hat{\theta}_j - \theta)$. Dans ce cas, il faut la modifier en remplaçant σ par son estimateur $\hat{\sigma}$ défini au Paragraphe 8.3.

Corollaire 8.2. *Si l'Hypothèse (NR) est vérifiée,*

$$\frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}\sqrt{b_j}} \ \sim \ t_{n-p}.$$

Preuve. Soit $\eta \stackrel{\text{déf}}{=} (\sigma^2 b_j)^{-1/2}(\hat{\theta}_j - \theta)$ et $\chi \stackrel{\text{déf}}{=} (n-p)\hat{\sigma}^2/\sigma^2 = \|y - \mathbf{X}\hat{\theta}\|^2/\sigma^2$. D'après le Théorème 8.3 et le Corollaire 8.1, $\eta \sim \mathcal{N}(0,1)$, $\chi \sim \chi^2_{n-p}$ et $\eta \perp\!\!\!\perp \chi$. Par ailleurs,

$$\frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}\sqrt{b_j}} = \frac{\eta}{\sqrt{\chi/(n-p)}},$$

d'où le résultat. ∎

Ce corollaire implique que sous l'hypothèse $H_0 \,:\, \theta_j = a$, la loi de la v. a.

$$t = \frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}\sqrt{b_j}}$$

est t_{n-p} (loi de Student avec $n-p$ degrés de liberté). Par conséquent, si l'on définit la région critique du test par

$$R = \left\{ \left| \frac{\hat{\theta}_j - a}{\hat{\sigma}\sqrt{b_j}} \right| > c_\alpha \right\}$$

avec une constante $c_\alpha > 0$ convenablement choisie, alors le risque de première espèce est

$$\sup_{\theta \in \Theta_0} P_\theta(R) = \sup_{\theta \in \Theta_0} P_\theta\left(\left| \frac{\hat{\theta}_j - a}{\hat{\sigma}\sqrt{b_j}} \right| > c_\alpha \right),$$

où $\Theta_0 = \big\{ \theta \in \mathbb{R}^p \,:\, \theta_j = a \big\}$ (soulignons que H_0 est une hypothèse composite, car on peut la réécrire comme $H_0 \,:\, \theta \in \Theta_0$). Sur l'ensemble Θ_0 le paramètre θ_j vaut a, donc la variable t suit la loi de Student t_{n-p}. On a alors

$$\sup_{\theta \in \Theta_0} P_\theta\left(\left| \frac{\hat{\theta}_j - a}{\hat{\sigma}\sqrt{b_j}} \right| > c_\alpha \right) = \sup_{\theta \in \Theta_0} P\big(|t_{n-p}| > c_\alpha\big) = P\big(|t_{n-p}| > c_\alpha\big).$$

Pour avoir le risque de première espèce égal à α, il faut choisir la valeur critique $c_\alpha = q_{1-\alpha/2}(t_{n-p})$. Ainsi, on obtient la région critique du test de niveau (et de taille) α :

$$\boxed{R = \left\{ \left| \frac{\hat{\theta}_j - a}{\hat{\sigma}\sqrt{b_j}} \right| > q_{1-\alpha/2}(t_{n-p}) \right\}.} \tag{8.9}$$

On rejette donc l'hypothèse H_0 si

$$\left| \frac{\hat{\theta}_j - a}{\hat{\sigma}\sqrt{b_j}} \right| > q_{1-\alpha/2}(t_{n-p})$$

et on ne la rejette pas dans le cas contraire.

Dans les applications, on est souvent confronté aux tests des hypothèses plus générales, en particulier, de l'hypothèse

$$H_0 \;:\; \theta_{j_1} = a_1, \ldots, \theta_{j_m} = a_m$$

contre l'alternative

$$H_1 \;:\; \exists\, k \in \{1, \ldots, m\} \quad \text{tel que } \theta_{j_k} \neq a_k,$$

où $\{j_1, \ldots, j_m\}$ est un sous-ensemble de $\{1, \ldots, p\}$. Notons que H_1 est le complémentaire de H_0.

EXEMPLE 8.5. Test de “sélection des variables” dans la régression polynomiale :

$$Y_i = g(\mathbf{x}_i) + \xi_i = \theta_1 + \theta_2 Z_i + \cdots + \theta_p Z_i^{p-1} + \xi_i, \quad i = 1, \ldots, n.$$

On veut tester l'hypothèse

$$H_0 \ : \ \theta_{j+l} = 0, \quad l = 1, \ldots, p - j.$$

contre l'alternative H_1 : il existe $l \geq 1$ tel que $\theta_{j+l} \neq 0$. Pour ce faire, on peut utiliser le test de Bonferroni.

8.6.1. Test de Bonferroni. Ce test doit son nom à l'inégalité suivante que l'on appelle inégalité de Bonferroni : soient $A_1, \ldots, A_m$ des événements aléatoires, alors

$$P\Big(\bigcup_{i=1}^{m} A_i\Big) \leq \sum_{i=1}^{m} P(A_i).$$

Supposons maintenant que l'on souhaite tester l'hypothèse

$$H_0 \ : \ \theta_{j_1} = a_1, \ldots, \theta_{j_m} = a_m$$

contre l'alternative

$$H_1 \ : \ \exists \ k \in \{1, \ldots, m\} \quad \text{tel que } \theta_{j_k} \neq a_k,$$

où $J = \{j_1, \ldots, j_m\}$ est un sous-ensemble de $\{1, \ldots, p\}$ (notons que l'hypothèse H_0 ainsi que l'alternative H_1 sont composites). Considérons la région critique

$$R = \bigcup_{i=1}^{m} R_i \quad \text{avec} \quad R_i = \left\{ \left| \frac{\hat{\theta}_{j_i} - a_i}{\hat{\sigma}\sqrt{b_{j_i}}} \right| > q_{1-\alpha/(2m)}(t_{n-p}) \right\}. \tag{8.10}$$

La région R définit un test de niveau α. En effet, d'après l'inégalité de Bonferroni,

$$\sup_{\theta \in \Theta_0} P_\theta(R) \leq \sum_{i=1}^{m} \sup_{\theta \in \Theta_0} P_\theta(R_i) = m \cdot \alpha/m = \alpha,$$

où $\Theta_0 = \{\theta \in \mathbb{R}^p : \ \theta_{j_i} = a_i, \, i = 1, \ldots, m\}$. On appelle le test basé sur la région critique (8.10) **test de Bonferroni**.

REMARQUE. A la différence du test (8.9) pour une seule coordonnée, le test de Bonferroni est de niveau α mais il n'est pas de taille α.

Une autre approche pour traiter des situations similaires et même plus générales est la suivante.

8.6.2. Hypothèse linéaire générale. F-test. Supposons que l'on souhaite tester l'hypothèse

$$H_0 \ : \ G\theta = b$$

contre l'alternative

$$H_1 \ : \ G\theta \neq b,$$

où G est une matrice $m \times p$ et b est un vecteur de $\mathbb{R}^m$. En particulier, si l'on pose

$$G = \left.\begin{pmatrix} \underbrace{\begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix}}_{p-m} & \underbrace{\begin{matrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{matrix}}_{m} \end{pmatrix}\right\} m \quad , \qquad b = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \in \mathbb{R}^m,$$

on obtient l'hypothèse et l'alternative décrites dans l'Exemple 8.5.

Proposition 8.4. *Si l'Hypothèse (NR) est vérifiée,*

$$G\hat{\theta} \sim \mathcal{N}_m(G\theta, \sigma^2 G B^{-1} G^T).$$

Preuve. Elle est immédiate d'après le Théorème 8.3. ∎

D'après cette proposition, sous l'hypothèse $H_0 : G\theta = b$ on a :

$$G\hat{\theta} \sim \mathcal{N}_m(b, D) \quad \text{avec} \quad D = \sigma^2 G B^{-1} G^T.$$

Soit $D > 0$. Définissons la variable aléatoire

$$\eta \stackrel{\text{déf}}{=} (G\hat{\theta} - b)^T D^{-1} (G\hat{\theta} - b).$$

D'après la Proposition 3.6,

$$\eta \sim \chi^2_m.$$

Si σ^2 est inconnu, on ne peut pas se servir de η pour définir la région critique du test. C'est pourquoi on replace σ^2 par son estimateur $\hat{\sigma}^2$. On obtient ainsi l'estimateur de la matrice de covariance D suivant :

$$\hat{D} = \hat{\sigma}^2 G B^{-1} G^T \quad \text{avec} \quad \hat{\sigma}^2 = \frac{\|y - \mathbf{X}\hat{\theta}\|^2}{n-p} \ .$$

Introduisons maintenant la variable aléatoire

$$F \stackrel{\text{déf}}{=} \frac{(G\hat{\theta} - b)^T \hat{D}^{-1} (G\hat{\theta} - b)}{m}$$

que l'on appelle **F-statistique** et définissons la région critique du test basé sur cette statistique :

$$R = \{F > c_\alpha\}.$$

Ici $c_\alpha > 0$ est à choisir de façon que le test soit de niveau α. On peut remarquer que F est une sorte de distance entre $G\hat{\theta}$ et b. On décidera donc de rejeter H_0 si cette distance F est assez grande ($> c_\alpha$).

En utilisant le Théorème 8.3, on peut facilement vérifier que sous H_0 la v. a. F suit la loi de Fisher-Snedecor à degrés de liberté m et $n - p$, ce qui nous conduit au choix suivant de la valeur critique : $c_\alpha = q_{1-\alpha}(m, n-p)$, où $q_{1-\alpha}(m, n-p)$ désigne le quantile d'ordre $1 - \alpha$ de la loi de Fisher-Snedecor $F_{m,n-p}$ à degrés de liberté m et $n - p$. On obtient finalement la région critique

$$\boxed{R = \Big\{F > q_{1-\alpha}(m, n-p)\Big\}.} \tag{8.11}$$

Le test basé sur la région critique (8.11) est appelé **F-test**.

8.7. Exercices

EXERCICE 8.1.
1°. Soit $\mathbf{x} \in \mathbb{R}^p$. Quel est le rang de la matrice $\mathbf{x}\mathbf{x}^T$? Quels sont ses vecteurs propres et valeurs propres ?
2°. On suppose maintenant que $\mathbf{x} \in \mathbb{R}^p$ est un vecteur aléatoire. Mêmes questions que 1° pour la matrice $E(\mathbf{x}\mathbf{x}^T)$.
3°. Considérons le cas particulier de 2° quand

$$\mathbf{x} = (1, Z, \dots, Z^{p-1})^T,$$

où Z est une variable aléatoire. Montrer que la matrice $E(\mathbf{x}\mathbf{x}^T)$ est strictement positive si la loi de Z admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$.

EXERCICE 8.2. Soient $\xi_1, \dots, \xi_n$ des variables aléatoires i.i.d. de densité $f(\cdot)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$, et soit $X_i \in \mathbb{R}$, $i = 1, \dots, n$. On observe les couples (X_i, Y_i), $i = 1, \dots, n$, issus du modèle de régression linéaire

$$Y_i = \theta X_i + \xi_i,$$

où $\theta \in \mathbb{R}$ est un paramètre inconnu.

1°. On suppose d'abord que les X_i sont déterministes (modèle de *régression à effets fixes*).
1.1°. Expliciter la densité jointe de $Y_1, \dots, Y_n$.
1.2°. Montrer que si la loi de ξ_i est $\mathcal{N}(0,1)$, la densité des $(Y_1, \dots, Y_n)$ est

$$\frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (Y_i - \theta X_i)^2 \right).$$

En déduire l'estimateur du maximum de vraisemblance $\hat{\theta}^{MV}$ de θ. Quelle est la loi de $\hat{\theta}^{MV}$? Son risque quadratique ?
1.3°. Dans le cadre énoncé en 1.2°, on étudie le cas particulier de *régression sur le temps* : $X_i = i$. Quelle est la vitesse de convergence du risque quadratique vers 0 dans ce cas ? Proposer la prévision linéaire de Y_{n+1} basée sur $(Y_1, \dots, Y_n)$. Donner l'intervalle de confiance de taille exacte $1 - \alpha$ pour Y_{n+1}.

2°. On suppose maintenant que les X_i sont des variables aléatoires i.i.d. (modèle de *régression à effets aléatoires*) et que X_i est indépendant de ξ_i, pour tout i. On note f_X la densité de X_1.
2.1°. Chercher la densité conditionnelle de $(Y_1, \dots, Y_n)$ sachant $(X_1, \dots, X_n)$, puis la densité jointe de $(X_1, \dots, X_n, Y_1, \dots, Y_n)$. Vérifier que l'estimateur du maximum de vraisemblance $\hat{\theta}^{MV}$ de θ ne dépend pas de la loi des X_i.
2.2°. Soit $\hat{\theta}_n$ l'estimateur des moindres carrés de θ. En supposant que les ξ_i sont de moyenne $E(\xi_1) = 0$ et de variance $E(\xi_1^2) = \sigma_\xi^2$ et que $E(X_1^2) = \sigma_X^2$, donner la loi asymptotique de $\sqrt{n}(\hat{\theta}_n - \theta)$ quand $n \to \infty$.
2.3°. En déduire un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour θ et un test de niveau asymptotique α de l'hypothèse $H_0 : \theta = 0$ contre l'alternative $H_1 : \theta > 0$.

3°. On suppose que les X_i sont déterministes et que les ξ_i sont de densité f. Montrer que l'estimateur du maximum de vraisemblance est le même que celui trouvé en 2.1°.

EXERCICE 8.3. Soient Z, ε, η des variables aléatoires gaussiennes mutuellement independantes de loi $\mathcal{N}(0,1)$ et soit $\theta \in \mathbb{R}$. On définit X et Y par

$$X = Z + \varepsilon, \qquad Y = \theta Z + \eta.$$

Supposons que l'on dispose de n observations i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$, où (X_i, Y_i) suit la même loi que (X, Y), et que l'on veut estimer le paramètre θ à partir de ces observations. Soit

$$\hat{\theta}_n = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

l'estimateur des moindres carrés de θ. Montrer que $\hat{\theta}_n$ n'est pas consistant. Modifier $\hat{\theta}_n$ pour obtenir un estimateur consistant que l'on notera $\hat{\theta}_n^*$. Chercher la loi limite de $\sqrt{n}(\hat{\theta}_n^* - \theta)$ quand $n \to \infty$.

EXERCICE 8.4. Soit le modèle de régression linéaire simple

$$Y_i = \theta_1 + \theta_2 X_i + \xi_i, \quad i = 1, \ldots, n,$$

où ξ_i sont des variables aléatoires gaussiennes indépendantes, de moyenne 0 et de variance $\sigma^2 > 0$ inconnue, $X_i \in \mathbb{R}$ sont des valeurs déterministes et θ_1 et θ_2 sont des paramètres réels inconnus. On note

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \qquad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i,$$

et on suppose dans la suite que

$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 > 0.$$

1°. Expliciter $\hat{\theta}_1$ et $\hat{\theta}_2$, les estimateurs des moindres carrés de θ_1 et θ_2 respectivement. Expliciter également l'estimateur $\hat{\sigma}^2$ de σ^2.
2°. Trouver les variances de $\hat{\theta}_1$, $\hat{\theta}_2$, ainsi que les covariances $\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2)$ et $\mathrm{Cov}(\bar{Y}, \hat{\theta}_2)$. Montrer que $\hat{\theta}_1 \perp\!\!\!\perp \hat{\theta}_2$ si et seulement si $\bar{X} = 0$.
3°. Donner la loi de la statistique

$$\frac{(\hat{\theta}_2 - \theta_2) S}{\hat{\sigma}}.$$

4°. Soit $0 < \alpha < 1$. Proposer un test de taille exacte α de l'hypothèse $H_0 : \ \theta_2 > 0$ contre l'alternative $H_1 : \ \theta_2 \leq 0$.

EXERCICE 8.5. Soit $\theta \in \mathbb{R}$ un paramètre inconnu. Supposons que l'on dispose de n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ telles que

$$Y_i = \theta X_i + \xi_i, \quad i = 1, \ldots, n,$$

où les ξ_i sont des variables aléatoires i.i.d. $\mathcal{N}(0,1)$ et les X_i sont des variables aléatoires i.i.d. de loi de Rademacher :

$$X_i = \begin{cases} 1 \ \text{avec la probabilité} & 1-p, \\ -1 \ \text{avec la probabilité} & p, \end{cases}$$

où $0 < p \leq 1/2$. Supposons de plus que $(X_1, \ldots, X_n)$ est indépendant de $(\xi_1, \ldots, \xi_n)$.
1°. Soit $\hat{\theta}_n^{MC}$ l'estimateur des moindres carrés de θ. Est-il biaisé ? Déterminer la loi de $\hat{\theta}_n^{MC}$.
2°. Considérons l'estimateur de θ défini par

$$\hat{\theta}_n^* = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} \, I\Big\{ \sum_{i=1}^n X_i \neq 0 \Big\}.$$

Calculer le biais de $\hat{\theta}_n^*$ lorsque $n = 2$.
3°. Trouver la loi limite de $\sqrt{n}(\hat{\theta}_n^* - \theta)$ quand $n \to \infty$ et $0 < p < 1/2$. Quelle difficulté rencontre-t-on au cas $p = 1/2$?
4°. Comparer les variances asymptotiques de $\hat{\theta}_n^{MC}$ et $\hat{\theta}_n^*$. Lequel de ces deux estimateurs est asymptotiquement le plus efficace ?