Master 1 - Mathématiques et Applications

Statistique appliquée

Tabea Rebafka

Partie I et II

Université Paris 6

2015

Table des matières

Partie 1

Introduction et Rappels

Chapitre 1

Introduction

L'objectif de la *statistique* est d'extraire des informations utiles des données. Les *données* sont issues de domaines très variés comme la médecine, l'économie, la sociologie, l'ingénierie, l'astrophysique, l'internet etc. Un statisticien cherche à les analyser et interpréter pour des objectifs concrets comme le contrôle de qualité, l'aide à la décision etc.

L'approche prise en statistique consiste à se donner un cadre mathématique, dans lequel la variabilité dans les données est expliquée par l'aléa. On adopte donc une *modélisation probabiliste* des données. On souligne qu'il n'est pas indispensable que le phénomène observé soit vraiment aléatoire, c'est-à-dire les données soient issue d'une expérience où intervient le hasard. La modélisation probabiliste n'est que le moyen pour prendre en compte la variabilité dans les données, et on doit toujours justifier et critiquer le choix d'un modèle. Par ailleurs, il est clair que tout modèle est faux, car il ne peut être qu'une approximation de la réalité. Néanmoins, on espère que le modèle choisi est approprié pour apporter des réponses en vue des objectifs concrets de l'application.

Prenons comme exemple les ventes dans une boutique de vêtements. Chaque jour on observe le montant dépensé par tous les clients passés à la caisse. Tous les jours, on observe alors un vecteur $\mathbf{x} = (x_1, \dots, x_n)$ représentant les dépenses des n clients de ce jour. Une modélisation probabiliste simple serait de considérer les données $\mathbf{x}$ comme la réalisation d'un vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ de loi $\mathbb{P}$ inconnue.

Afin d'"expliquer" les données, le statisticien cherche à déterminer cette loi de probabilité $\mathbb{P}$. En général, il est impossible de reconstituer $\mathbb{P}$ exactement, mais on essaye de l'approcher en utilisant au mieux les données observées $\mathbf{x}$. Ensuite, le statisticien essaye de donner des réponses aux questions issues de l'application. Dans notre exemple, on pourrait étudier la question si le client moyen dépense plus ou moins pendant les soldes comparé à la période normale. En s'appuyant sur des données d'un jour de soldes $\mathbf{x} = (x_1, \dots, x_n)$ de loi $\mathbb{P}_{\text{solde}}$ et des données d'un jour "normal" $\mathbf{y} = (y_1, \dots, y_m)$ de loi $\mathbb{P}_{\text{normal}}$, on pourrait comparer les lois $\mathbb{P}_{\text{solde}}$ et $\mathbb{P}_{\text{normal}}$ pour répondre à la question.

Ce cours comporte deux grandes parties, que l'on reconnaît dans ce petit exemple : la première partie porte sur l'identification de la loi $\mathbb{P}$ des données, plus précisément, sur le choix du modèle probabiliste et l'estimation. La deuxième partie présente des tests statistiques pour apporter des réponses à des questions d'intérêt pratique.

Puisque l'approche statistique repose toujours sur une modélisation probabiliste, nous commençons ce cours par un rappel sur la théorie des probabilités et la présentation de quelques outils probabilistes particulièrement utiles en statistique (Chapitre 2).

Le choix d'un modèle pour les données repose d'une part sur une connaissance partielle préalable du phénomène étudié, de la façon dont une expérience a été menée et d'autre part sur des représentations graphiques des données recueillies. Ces outils graphiques sont

connus sous le nom de *statistique descriptive*, et ils sont présentés dans le Chapitre 3. Cette démarche mène à définir des hypothèses sur la loi $\mathbb{P}$ des données et à déterminer une famille de lois à laquelle la loi $\mathbb{P}$ est susceptible d'appartenir.

Ensuite on cherche à identifier la loi $\mathbb{P}$ des données dans cette famille de lois, en construisant un *estimateur*. Chapitre 4 introduit des propriétés souhaitables d'estimateurs permettant d'évaluer et de comparer différents estimateurs. Chapitre 5 présente des approches classiques pour la construction d'estimateurs, notamment la méthode de substitution, la méthode des moments et la méthode du maximum de vraisemblance.

Chapitre 6 porte sur des modèles pertinents pour des nombreuses applications, groupés sous le nom de *modèles à variables latentes*. D'une part, ces modèles sont très importants pour la pratique, d'autre part, ils sont relativement difficiles d'un point de vue mathématique, en sorte que l'estimation nécessite des méthodes adaptées. Dans cette optique, l'algorithme EM est présenté.

D'autres modèles de grande utilité en statistique sont les modèles de régression, traités dans le Chapitre 7. Ces modèles permettent d'étudier la relation entre plusieurs variables, notamment l'impact de certaines variables sur une autre variable.

Cette partie du cours se termine par un chapitre sur l'*estimation par intervalle*, où notamment le bootstrap sera présenté (Chapitre 8).

De façon générale, nous ne fournissons que quelques éléments d'analyse théorique des propriétés d'estimateurs, car un intérêt particulier sera porté sur des solutions pratiques (notamment des algorithmes) pour le calcul des estimateurs.

Dans la partie sur les *tests statistiques*, nous présenterons, d'une part, des tests pour valider ou choisir un modèle approprié aux données. P. ex. on veut répondre à la question si l'hypothèse que les observations $\mathbf{x}$ suivent une loi normale est juste au vu des données. D'autre part, nous développerons des tests pour répondre aux questions issues de l'application, comme p. ex. la question si les achats par personne sont plus ou moins élevés pendant les soldes comparé à la période normale.

Chapitre 2

Rappel et compléments : Théorie des probabilités

Ce chapitre comprend un rappel sur la théorie des probabilités et la présentation de quelques outils probabilistes pour la statistique. Les preuves de nombreux résultats sont omis, car elles sont supposées connues de votre cours de probabilité. Par ailleurs, vous les trouverez dans la majorité des ouvrages classiques de la théorie des probabilités.

2.1 Variables aléatoires et principales lois de probabilité

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

La **fonction de répartition** d'une variable aléatoire X est la fonction $F : \mathbb{R} \to [0, 1]$ définie par $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\omega : X(\omega) \leq x)$. La fonction F sera aussi appelée la **loi** ou la **distribution** de X.

La fonction de répartition est une fonction monotone croissante, continue à droite et telle que $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to \infty} F(x) = 1$.

On a, pour tout $a < b$,

$$\begin{aligned}
&\mathbb{P}(X = a) = F(a) - \lim_{t \to a-} F(t) \,, \qquad (2.1)\\
&\mathbb{P}\left(X \in]a, b]\right) = F(b) - F(a) \,,\\
&\mathbb{P}(X \in [a, b]) = \mathbb{P}(X \in]a, b]) + \mathbb{P}(X = a) = F(b) - \lim_{t \to a-} F(t) \,,\\
&\mathbb{P}(X \in]a, b[) = \mathbb{P}(X \in]a, b]) - \mathbb{P}(X = b) = \lim_{t \to b-} F(t) - F(a) \,,\\
&\mathbb{P}(X \in [a, b[) = \mathbb{P}(X \in]a, b]) - \mathbb{P}(X = b) + \mathbb{P}(X = a) = \lim_{t \to b-} F(t) - \lim_{t \to a-} F(t) \,.
\end{aligned}$$

Notons p la **densité** de la loi F par rapport à une mesure de référence μ. Plus précisément, p est une fonction μ-mesurable positive telle que

$$F(x) = \int_{]-\infty, x]} p \, \mathrm{d}\mu \,, \qquad \text{pour tout } x \in \mathbb{R} \,.$$

Plus généralement, on a pour tout ensemble $B \in \mathcal{B}$,

$$\mathbb{P}(X \in B) = \int_B p \, \mathrm{d}\mu \, .$$

D'après le théorème de Radon-Nikodym, p est unique (à égalité μ-presque partout près). On note que

$$\int_{\mathbb{R}} p \, \mathrm{d}\mu = 1 \, .$$

Souvent on note F_X et p_X pour la fonction de répartition et la densité d'une variable aléatoire X.

Il existe deux principaux types de variables aléatoires : les variables discrètes qui admettent une densité par rapport à une mesure de comptage, et les variables continues qui admettent une densité par rapport à la mesure de Lebesgue. Les lois des variables discrètes sont entièrement définies par les probabilités $\mathbb{P}(X = \cdot)$ et les lois des variables continues par leur densité $f(\cdot)$.

2.1.1 Lois discrètes

On dit que X est une **variable aléatoire discrète** quand les valeurs de X appartiennent à un ensemble $\mathcal{V} = \{v_1, v_2, \ldots\}$ fini ou dénombrable. Plus précisément, on a

$$\mathbb{P}(X \in \mathcal{V}) = \sum_k \mathbb{P}(X = v_k) = 1 \, .$$

Autrement dit, la loi de X admet une densité p par rapport à la mesure $\mu_{\mathcal{V}}$ de comptage sur $\mathcal{V}$. Elle vérifie $p(v) = \mathbb{P}(X = v)$ pour tout $v \in \mathcal{V}$, et pour tout $x \in \mathbb{R}$

$$F(x) = \int_{]-\infty, x]} p \, \mathrm{d}\mu_{\mathcal{V}} = \sum_k \mathbb{P}(X = v_k) \mathbb{1}\{v_k \leq x\} = \sum_{k : v_k \leq x} \mathbb{P}(X = v_k) \, .$$

La fonction de répartition d'une variable aléatoire discrète est une fonction en escalier.

Principales lois discrètes

1. **Loi de Dirac** δ_x. On dit que la variable aléatoire X suit une loi de Dirac avec masse en $x \in \mathbb{R}$, si X vaut x avec probabilité 1, i.e.

   $$\mathbb{P}(X = x) = 1 \, .$$

   Cette loi modélise un phénomène déterministe (non aléatoire) puisque le résultat de l'expérience est presque sûrement égal à la valeur x. La variable X est donc une constante.

2. **Loi uniforme discrète**. Soit $\mathcal{V} = \{v_1, \ldots, v_m\}$ un ensemble de m nombres réels v_k. On dit que X suit une loi uniforme sur $\mathcal{V}$ si

   $$\mathbb{P}(X = v_k) = \frac{1}{m} \, , \quad k = 1, \ldots, m \, .$$

3. **Loi Bernoulli** $\mathcal{B}(p)$. La loi de Bernoulli de paramètre $p \in [0, 1]$ est donnée par une variable aléatoire X qui prend ses valeurs dans $\mathcal{V} = \{0, 1\}$ avec

   $$\mathbb{P}(X = 1) = p \quad \text{et} \quad \mathbb{P}(X = 0) = 1 - p \, .$$

Pour $p = 0$ ou $p = 1$, X est une constante (loi de Dirac).

La loi de Bernoulli est utilisée pour modéliser des expériences ayant seulement deux résultats possibles, en particulier des situations classifiées en échec (0) ou succès (1).

4. **Loi binomiale** $\mathcal{B}(n,p)$. Soient $X_1,\dots,X_n$ des variables aléatoires indépendantes de loi Bernoulli $\mathcal{B}(p)$. Alors, on dit que la somme $S_n = \sum_{k=1}^n X_k$ suit une loi binomiale $\mathcal{B}(n,p)$. Autrement dit, S_n est une variable aléatoire à valeurs dans $\mathcal{V} = \{0,\dots,n\}$ vérifiant
$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} , \quad k = 0,1,\dots,n .$$
La loi binomiale $\mathcal{B}(1,p)$ avec $n = 1$ coïncide avec la loi Bernoulli $\mathcal{B}(p)$.
La loi binomiale $\mathcal{B}(n,p)$ modélise le nombre de succès parmi n expériences Bernoulli (i.i.d.).

5. **Loi de Poisson** Poi(λ). La variable aléatoire X suit une loi de Poisson de paramètre $\lambda > 0$ si
$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} , \quad k = 0,1,2,\dots .$$
La loi de Poisson est liée à la loi binomiale. Plus précisément, soient $(Y_n)_{n\geq 1}$ une suite de variables aléatoires de loi binomiale $\mathcal{B}(n,p_n)$, où les paramètres p_n vérifient $np_n \to \lambda$ lorsque $n \to \infty$. Alors, pour tout $k \in \{0,1,\dots\}$,
$$\lim_{n\to\infty} \mathbb{P}(Y_n = k) = \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} = \mathbb{P}(X = k) .$$

6. **Loi géométrique** Geo(p). Soit $X_1, X_2, \dots$ une suite de variables aléatoires indépendantes de loi Bernoulli $\mathcal{B}(p)$. Considérons 0 comme un échec et 1 comme succès. Alors la variable aléatoire W définie comme le nombre d'échecs jusqu'au premier succès suit une loi géométrique Geo(p). Autrement dit, W est la longueur maximale de la suite $X_1,\dots,X_\ell$ telle que $X_1 = \cdots = X_\ell = 0$, i.e.
$$W = \max\{\ell : X_1 = \cdots = X_\ell = 0\} .$$
Donc, W est une variable aléatoire à valeurs dans $\mathcal{V} = \{0,1,\dots\}$ qui vérifie
$$\mathbb{P}(W = \ell) = p(1-p)^\ell , \quad \ell = 0,1,\dots .$$

2.1.2 Lois absolument continues

On dit que X est une **variable aléatoire (absolument) continue** lorsque sa loi admet une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$, i.e.
$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t ,$$
pour tout $x \in \mathbb{R}$. Dans ce cas, la fonction de répartition F de X est continue et différentiable presque partout sur $\mathbb{R}$ et la densité de probabilité de X est égale à la dérivée
$$f(x) = F'(x) \quad \text{presque partout.}$$
La densité f est positive ($f \geq 0$) et vérifie $\int_{\mathbb{R}} f(t)\ \mathrm{d}t = 1$.

Notons que par (2.1) et par la continuité de F, on a
$$\mathbb{P}(X = x) = F(x) - \lim_{t\to x-} F(t) = F(x) - F(x) = 0 , \text{ pour tout } x \in \mathbb{R} .$$

Par conséquent, pour tout $a < b$

$$\mathbb{P}(X \in]a,b]) = \mathbb{P}(X \in]a,b[) = \mathbb{P}(X \in [a,b[) = \mathbb{P}(X \in [a,b]) = \int_a^b f(x)\mathrm{d}x \ .$$

Des nombreuses familles de lois courantes peuvent être définies par un paramètre de translation et/ou un paramètre d'échelle. Plus précisément, soit $\mathcal{F}$ une famille de lois donnée et supposons que la loi de X est comprise dans $\mathcal{F}$. Si la loi de $X + \tau$ appartient à $\mathcal{F}$ pour tout τ dans un ensemble $\mathcal{T} \subset \mathbb{R}$, on dit que τ est un **paramètre de translation** ou de position. De même, si σX appartient à $\mathcal{F}$ pour tout σ d'un ensemble $\mathcal{S} \subset \mathbb{R}$, on dit que σ est un **paramètre d'échelle**.

Principales lois continues

1. **Loi uniforme** $\mathbf{U}(a,b)$. La loi uniforme sur l'intervalle $[a,b]$, $-\infty < a < b < \infty$, est la loi notée $U[a,b]$, de densité

$$f(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x) \ ,$$

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

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

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

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

avec $\mu \in \mathbb{R}$ et $\sigma > 0$. Si $\mu = 0$ et $\sigma = 1$, la loi $\mathcal{N}(0,1)$ est dite *loi normale standard*. Si X suit la loi normale standard $\mathcal{N}(0,1)$, alors la variable aléatoire $Y = \sigma X + \mu$ avec $\mu, \sigma \in \mathbb{R}$ suit la loi normale $\mathcal{N}(\mu,\sigma^2)$. Le paramètre μ est alors un paramètre de translation, et σ un paramètre d'échelle.

3. **Loi exponentielle** $\mathcal{E}(\lambda)$. La loi exponentielle $\mathcal{E}(\lambda)$ est la loi de densité

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

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

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

Si X suit la loi exponentielle $\mathcal{E}(1)$, alors aX avec $a > 0$ suit la loi exponentielle $\mathcal{E}(1/a)$.

4. **Loi Gamma** $\Gamma(a,b)$. La loi Gamma $\Gamma(a,b)$ est la loi de densité

$$f(x) = \frac{b^a}{\Gamma(a)}x^{a-1}\mathrm{e}^{-bx} \ , \quad x > 0 \ .$$

où $a > 0$ est un paramètre de forme et $b > 0$ un paramètre d'intensité, et $\Gamma(\cdot)$ dénote la fonction gamma définie par

$$\Gamma(t) = \int_0^\infty z^{t-1}\mathrm{e}^{-z}\mathrm{d}z \ , \quad t > 0 \ .$$

Pour $a = 1$, la loi $\Gamma(1, b)$ est la loi exponentielle $\mathcal{E}(b)$.
Une paramétrisation alternative (qui ne sera pas utilisée dans ce cours) repose sur un paramètre de forme $k > 0$ et un paramètre d'échelle $\lambda > 0$. Dans ce cas, la densité s'écrit

$$f(x) = \frac{x^{k-1}}{\lambda^k \Gamma(k)} \mathrm{e}^{-x/\lambda} , \quad x > 0 .$$

5. **Loi de Cauchy**. La densité de la loi de Cauchy $\mathrm{C}(\mu, \sigma)$ de paramètre de translation μ et de paramètre d'échelle σ est donnée par

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

2.2 Quelques caractéristiques des lois de probabilité

Il existe des nombreux indicateurs pour caractériser et comparer des lois de probabilité. Ces indicateurs sont des fonctionnelles de la fonction de répartition. Des exemples de telles fonctionnelles sont les moments et les quantiles.

2.2.1 Moments

Soit X une variable aléatoire de loi F et de densité p par rapport à une mesure de référence μ.

La **moyenne** (ou l'**espérance mathématique**) de X est définie par

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x \mathrm{d}F(x) = \int_{\mathbb{R}} x p(x) \mu(\mathrm{d}x) \\ &= \begin{cases} \sum_k v_k \mathbb{P}(X = v_k) & \text{si } X \text{ est une v.a. discrète} \\ \int_{-\infty}^{\infty} x f(x) \mathrm{d}x & \text{si } X \text{ est une v.a. continue} , \end{cases} \end{aligned}$$

pourvu que $\mathbb{E}[|X|] = \int_{-\infty}^{\infty} |x| \mathrm{d}F(x) < \infty$.

Plus généralement, on définit pour toute fonction h mesurable et intégrable

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) \mathrm{d}F(x) = \begin{cases} \sum_k h(v_k) \mathbb{P}(X = v_k) & \text{si } X \text{ est une v.a. discrète} \\ \int_{-\infty}^{\infty} h(x) f(x) \mathrm{d}x & \text{si } X \text{ est une v.a. continue} . \end{cases}$$

Avec $h(x) = x^k$, on obtient le **moment d'ordre** k ($k = 1, 2, \dots$) de X : $\mathbb{E}[X^k]$, ainsi que pour $h(x) = (x - \mathbb{E}[X])^k$ le **moment centré d'ordre** k : $\mathbb{E}[(X - \mathbb{E}[X])^k]$.

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

Une propriété utile en pratique est la suivante : $\mathbb{E}[|X|] = 0$ implique que $\mathbb{P}(X = 0) = 1$, i.e. $X = 0$ *p.s.*.

La donnée de certains ensembles de moments d'une variable aléatoire X déterminent sa loi, comme par exemple les moments $\mathbb{E}[\mathrm{e}^{itX}]$ pour tout t ou encore les moments $\mathbb{E}[h(X)]$ pour toute fonction h mesurable bornée. Plus précisément, on a les deux propositions suivantes.

Définissions d'abord la **fonction caractéristique** $\varphi_X : \mathbb{R} \to \mathbb{C}$ d'une variable aléatoire X par l'espérance

$$\varphi_X(t) = \mathbb{E}[\mathrm{e}^{itX}] , \quad \text{pour } t \in \mathbb{R} .$$

Notons que la fonction caractéristique est toujours bien définie.

Proposition 1. *Soient X et Y deux variables aléatoires. Si X et Y ont la même fonction caractéristique, c'est-à-dire si*

$$\varphi_X(t) = \varphi_Y(t) \ , \qquad \text{pour tout } t \in \mathbb{R} \ ,$$

alors X et Y ont la même loi.

Proposition 2. *Soit X une variable aléatoire continue. S'il existe une fonction q mesurable telle que pour tout h mesurable bornée on a*

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)q(x)\mathrm{d}x \ ,$$

alors q est la densité par rapport à la mesure de Lebesgue de la loi de X.

Caractéristiques des lois basées sur des moments

Tendance centrale. La moyenne $\mathbb{E}[X]$ est un *indicateur de la tendance centrale* ou de position de la loi de X.

Dispersion. La **variance** $\mathbf{Var}(X)$ est définie comme le moment centré d'ordre 2 :

$$\mathbf{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \ .$$

La racine carrée de la variance s'appelle **écart-type** de X : $\sigma = \sqrt{\mathbf{Var}(X)}$. On notera que l'écart-type s'exprime dans la même unité que X.

La variance et l'écart-type sont des *indicateurs de la dispersion* de la loi de X.

Symétrie. La loi F est dite **symétrique par rapport à** $\mu \in \mathbb{R}$ si $F(\mu+x) = 1-F(\mu-x)$ pour tout $x \in \mathbb{R}$.
Si F est symétrique par rapport à zéro, on dit tout simplement que F est **symétrique**. La définition est équivalente à $p(\mu + x) = p(\mu - x)$ pour tout x, où p dénote la densité de F par rapport à une mesure de référence.

Proposition 3. *Si F est symétrique et tous les moments absolus $\mathbb{E}[|X^k|]$ existent, alors les moments $\mathbb{E}[X^k]$ sont nuls pour tout k impair. Si F est symétrique par rapport à μ et tous les moments absolus $\mathbb{E}[|X^k|]$ existent, alors $\mathbb{E}[(X-\mu)^k] = 0$ pour tout k impair.*

En particulier, si X est d'espérance finie et symétrique par rapport à μ, alors $\mathbb{E}[X] = \mu$.

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

$$\alpha = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^{3/2}} \ .$$

On remarque que $\alpha = 0$ pour une fonction de répartition symétrique avec $\mathbb{E}[|X|^3] < \infty$. Notons que le réciproque n'est pas exacte : la condition $\alpha = 0$ n'implique pas la symétrie de la loi.

Le coefficient d'asymétrie α est invariant par rapport aux transformations affines (d'échelle et de position) de la variable aléatoire X. Autrement dit, les variables X et $Y = aX + b$ avec $a > 0$ et $b \in \mathbb{R}$ ont le même coefficient d'asymétrie, i.e.

$$\alpha_X = \alpha_{aX+b} \ .$$

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.

Aplatissement. Le **coefficient d'aplatissement** β (ou β_X) (en anglais **kurtosis**) est défini de la façon suivante, si le quatrième moment de X existe ($\mathbb{E}[X^4] < \infty$), alors

$$\beta = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^2} - 3 \, .$$

Le coefficient d'aplatissement β est invariant par rapport aux transformations affines (d'échelle et de position) de la variable aléatoire X. Autrement dit, les variables X et $Y = aX + b$ avec $a > 0$ et $b \in \mathbb{R}$ ont le même coefficient d'aplatissement, i.e.

$$\beta_X = \beta_{aX+b} \, .$$

Notons que, pour toute loi de probabilité, on a $\beta > -2$.

Le coefficient d'aplatissement de la loi normale $\mathcal{N}(\mu, \sigma^2)$ est nul.

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

$$Q(b) = \mathbb{P}(|X| > b)$$

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 : $Q(b) = O(\mathrm{e}^{-b^2/2})$, 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 $\beta < 0$, elles sont plus légères que celles de la loi normale.

2.2.2 Quantiles

Lorsque la fonction de répartition $F : \mathbb{R} \to]0, 1[$ est une application bijective, on peut définir son inverse $F^{-1} :]0, 1[\to \mathbb{R}$ de façon classique, c'est-à-dire F^{-1} est la fonction qui vérifie $F^{-1} \circ F = Id$ et $F \circ F^{-1} = Id$. La fonction de quantile est la généralisation de l'inverse à toute les fonctions de répartition F.

On appelle **fonction de quantile** la fonction $F^{-1} :]0, 1[\to \mathbb{R}$ définie par

$$F^{-1}(\alpha) = \inf\{t \in \mathbb{R} : F(t) \geq \alpha\} \, , \quad \text{pour } \alpha \in]0, 1[\, .$$

On appelle **quantile q_α d'ordre** α, $0 < \alpha < 1$, de la loi F la valeur

$$q_\alpha = F^{-1}(\alpha) \, .$$

Pour souligner qu'il s'agit des quantiles de la loi F, on peut aussi noter q_α^F au lieu de q_α.

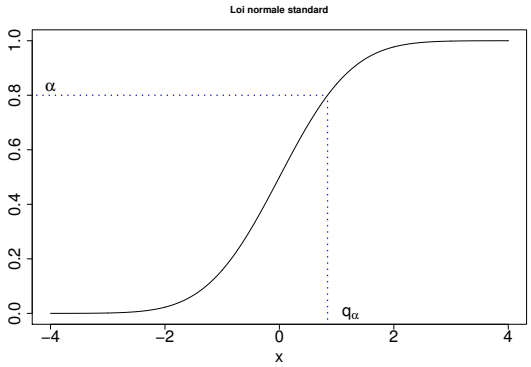
Théorème 1. *Pour toute fonction de répartition F on a*

(i) F^{-1} *est croissante.*

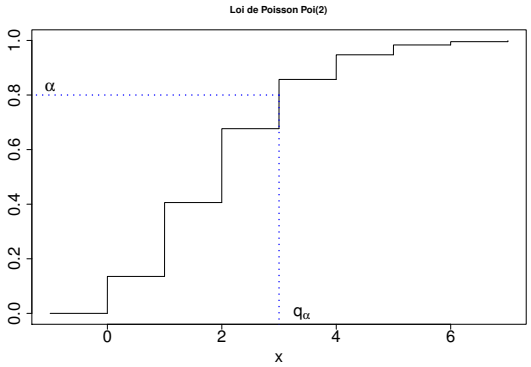
(ii) F^{-1} *est continue à gauche.*

(iii) $F(F^{-1}(\alpha)) \geq \alpha$ *pour* $\alpha \in]0, 1[$.

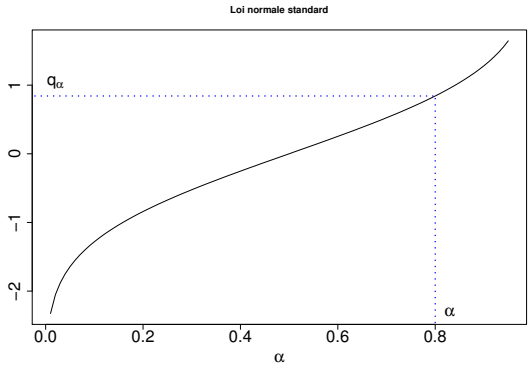
(iv) $F^{-1}(F(t)) \leq t$ *pour* $t \in \mathbb{R}$.



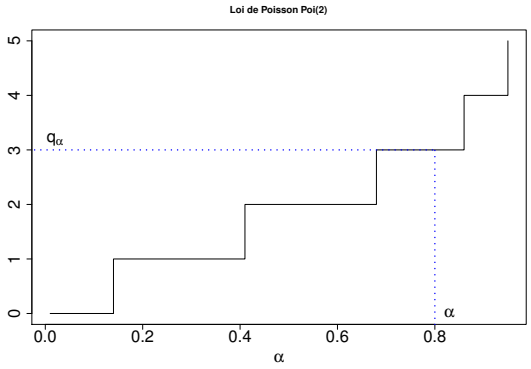
(a) Fonction de répartition $F(x)$ de la loi $\mathcal{N}(0,1)$



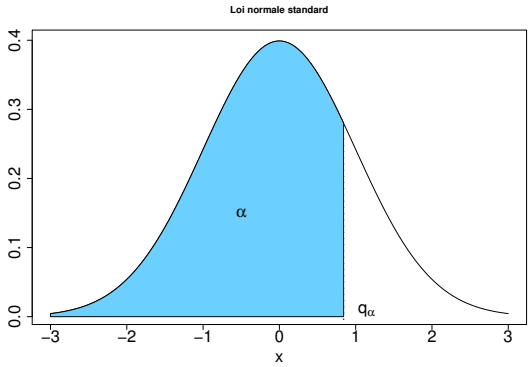
(b) Fonction de répartition $F(x)$ de la loi Poi(2)



(c) Fonction de quantile $F^{-1}(\alpha)$ de la loi $\mathcal{N}(0,1)$



(d) Fonction de quantile $F^{-1}(\alpha)$ de la loi Poi(2)



(e) Densité de la loi $\mathcal{N}(0,1)$

FIGURE 2.1 – Illustration du quantile q_α d'ordre $\alpha = 0.8$ de la loi de Poisson Poi(2) et de la loi normale standard.

Démonstration. Soient $\beta > \alpha$. Si pour $t \in \mathbb{R}$ on a $F(t) \geq \beta$, alors $F(t) \geq \alpha$. Donc, $\{t \in \mathbb{R} : F(t) \geq \beta\} \subset \{t \in \mathbb{R} : F(t) \geq \alpha\}$. Donc, $F^{-1}(\beta) = \inf\{t \in \mathbb{R} : F(t) \geq \beta\} \geq \inf\{t \in \mathbb{R} : F(t) \geq \alpha\} = F^{-1}(\alpha)$. D'où (i).

Comme F est une fonction croissante, l'ensemble $\{t \in \mathbb{R} : F(t) \geq \alpha\}$ est un intervalle soit de la forme $]a, +\infty[$, soit de la forme $[a, +\infty[$ pour un $a \in \mathbb{R}$. En tous cas, on a $F(a+\varepsilon) \geq \alpha$ pour tout $\varepsilon > 0$. Or, étant continue à droite, F vérifie $\lim_{\varepsilon \downarrow 0} F(a + \varepsilon) = F(a)$. D'où $F(a) \geq \alpha$ et $\{t \in \mathbb{R} : F(t) \geq \alpha\} = [a, +\infty[$. Autrement dit, $F^{-1}(\alpha) \in \{t \in \mathbb{R} : F(t) \geq \alpha\}$ (et ici $a = q_\alpha$). D'où (iii).

Soit $(\alpha_n)_{n\geq 1}$ une suite croissante qui converge vers α lorsque $n \to \infty$. Par (i), $(F^{-1}(\alpha_n))_{n\geq 1}$ est une suite croissante bornée par $F^{-1}(\alpha)$. Donc, $(F^{-1}(\alpha_n))_{n\geq 1}$ est une suite convergente, dont on note la limite a_0. Si $a_0 < F^{-1}(\alpha)$, alors il existe $\tilde{a} \in]a_0, F^{-1}(\alpha)[$ et on a, par monotonie de F, $F(\tilde{a}) \geq F(F^{-1}(\alpha_n)) \geq \alpha_n$ pour tout $n \geq 1$, ce qui implique que $F(\tilde{a}) \geq \alpha$. Donc, $\tilde{a} \in \{t \in \mathbb{R} : F(t) \geq \alpha\}$ et $\tilde{a} \geq F^{-1}(\alpha)$, ce qui est en contradiction avec $a_0 < F^{-1}(\alpha)$. Donc, $a_0 = F^{-1}(\alpha)$ et (ii).

On obtient (iv), parce que clairement $t \in \{s \in \mathbb{R} : F(s) \geq F(t)\}$ ce qui implique $F^{-1}(F(t)) \leq t$. □

Si $F(F^{-1}(\alpha)) > \alpha$, F a un saut en q_α et F^{-1} est constante sur un intervalle $]\alpha_1, \alpha_2]$ avec $\alpha_1 < \alpha < \alpha_2$. Si $F^{-1}(F(t)) < t$, il existe un intervalle $[a, b[$ avec $a < t < b$ sur lequel F est constante et F^{-1} a un saut en $F(t)$.

Si F admet une densité f par rapport à la mesure de Lebesgue, alors les quantiles sont caractérisés par l'aire sous la densité f. En effet, $\int_{-\infty}^{q_\alpha} f(t)\mathrm{d}t = \alpha$.

La Figure 2.1 illustre la relation de la fonction de répartition et la fonction de quantile dans le cas d'une loi discrète et d'une loi continue ainsi que le lien entre quantile et densité pour des lois continues.

Caractéristiques des lois basées sur des quantiles

Tendance centrale. La **médiane** Med de F est le quantile d'ordre $1/2$, autrement dit,

$$\text{Med} = q_{1/2} = F^{-1}\left(\frac{1}{2}\right) .$$

La médiane est un *indicateur de la tendance centrale* ou de position de la loi F.

Théorème 2. *Si la loi F est symétrique par rapport à μ, intégrable $\mathbb{E}|X| < \infty$ et si elle admet une densité f par rapport à la mesure de Lebesgue qui est nonnulle dans un voisinage de μ, alors sa médiane et sa moyenne coïncident. Plus précisément, on a*

$$\mu = \text{Med} = \mathbb{E}[X] .$$

Démonstration. Montrons d'abord que $\mu = \text{Med}$. Par symétrie de F, on a

$$F(\mu) = 1 - F(\mu) \Longleftrightarrow F(\mu) = \frac{1}{2} .$$

Donc $\mu \in \{t, F(t) \geq \frac{1}{2}\}$. De plus, pour tout $\varepsilon > 0$, on a

$$F(\mu - \varepsilon) = F(\mu) - \mathbb{P}(X \in]\mu - \varepsilon, \mu]) = \frac{1}{2} - \int_{\mu-\varepsilon}^{\mu} f(x)\mathrm{d}x .$$

Ce dernier intégrale est strictement positif, car f est nonnul dans un voisinage de μ. Donc, $F(\mu-\varepsilon)<\frac{1}{2}$. D'où $\mu=\inf\{t, F(t)\geq\frac{1}{2}\}=\text{Med}$.

Enfin, encore par symétrie de F, on a

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} xf(x)\mathrm{d}x = \int_{-\infty}^{\mu} xf(x)\mathrm{d}x + \int_{\mu}^{\infty} xf(x)\mathrm{d}x \\ &= \int_{0}^{\infty} (\mu-y)f(\mu-y)\mathrm{d}y + \int_{0}^{\infty} (\mu+y)f(\mu+y)\mathrm{d}y \\ &= \int_{0}^{\infty} (\mu-y)f(\mu+y)\mathrm{d}y + \int_{0}^{\infty} (\mu+y)f(\mu+y)\mathrm{d}y \\ &= 2\mu\int_{0}^{\infty} f(\mu+y)\mathrm{d}y = 2\mu\int_{\mu}^{\infty} f(y)\mathrm{d}y = 2\mu F(\mu) \\ &= \mu\ .\end{aligned}$$

□

Dispersion. On appelle **premier, deuxième** et **troisième quartile** le quantile $q_{1/4}$, la médiane Med et le quantile $q_{3/4}$.

L'**écart interquartile** est défini par la différence entre le troisième et le premier quartile :

$$EIQ = q_{3/4} - q_{1/4}\ .$$

L'écart interquartile est un *indicateur de la dispersion* de la loi F.

2.3 Relations entre variables aléatoires

Quand on considère plusieurs variables aléatoires à la fois, il est important de s'interroger sur la relation qui peut exister (ou pas) entre eux. Les valeurs $x=X(\omega)$ que prend une variable X ont-elles un impact sur les valeurs $y=Y(\omega)$ d'une autre variable Y définie sur le même espace de probabilité ?

2.3.1 Vecteurs aléatoires

Un vecteur aléatoire dans $\mathbb{R}^p$ est un vecteur $\mathbf{X}=(X_1,\ldots,X_p)^T$ dont toutes les composantes $X_1,\ldots,X_p$ sont des variables aléatoires réelles.

La fonction de répartition du vecteur aléatoire $\mathbf{X}$ est définie par

$$F_{\mathbf{X}}(t)=\mathbb{P}(X_1\leq t_1,\ldots,X_p\leq t_p)\ ,\quad t=(t_1,\ldots,t_p)^T\in\mathbb{R}^p\ .$$

La moyenne du vecteur aléatoire $\mathbf{X}$ est le vecteur des moyennes des variables aléatoires $X_1,\ldots,X_p$:

$$\mathbb{E}[\mathbf{X}]=(\mathbb{E}[X_1],\ldots,\mathbb{E}[X_p])^T\ ,$$

pourvu que $\mathbb{E}[|X_j|]<\infty$ pour tout j.

La **fonction caractéristique** $\varphi_{\mathbf{X}}:\mathbb{R}^p\to\mathbb{C}$ du vecteur aléatoire $\mathbf{X}$ est donnée par

$$\varphi_{\mathbf{X}}(t)=\mathbb{E}\left[\mathrm{e}^{itX}\right]\ ,\quad \text{pour } t\in\mathbb{R}^p\ .$$

Propriétés des vecteurs aléatoires dans le cas continu

Nous considérerons principalement le cas continu, c'est-à-dire nous supposerons que la loi de $\mathbf{X}$ admet une densité de probabilité $f_{\mathbf{X}}(\cdot)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$. Cela signifie que

$$F_{\mathbf{X}}(t) = \int_{-\infty}^{t_1} \dots \int_{-\infty}^{t_p} f_{\mathbf{X}}(u_1, \dots, u_p) \mathrm{d}u_p \dots \mathrm{d}u_1 \ ,$$

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

$$f_{\mathbf{X}}(t) = f_{\mathbf{X}}(t_1, \dots, t_p) = \frac{\partial^p F_{\mathbf{X}}(t)}{\partial t_1 \dots \partial t_p} \ ,$$

pour presque tout t. Toute densité de probabilité vérifie

$$f_{\mathbf{X}}(t) \geq 0 \ , \quad \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{X}}(t_1, \dots, t_p) \mathrm{d}t_1 \dots \mathrm{d}t_p = 1 \ .$$

Soit $\tilde{\mathbf{X}} = (X_1, \dots, X_k)^T$ (où $k < p$) un vecteur aléatoire, partie de $\mathbf{X}$. La **densité marginale** de $\tilde{\mathbf{X}}$ est donnée par

$$f_{\tilde{\mathbf{X}}}(t_1, \dots, t_k) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{X}}(t_1, \dots, t_k, u_{k+1} \dots, u_p) \mathrm{d}u_{k+1} \dots \mathrm{d}u_p \ .$$

Notons que la connaissance de toutes les densités marginales n'est pas suffisante pour déterminer la loi du vecteur aléatoire $\mathbf{X}$. Deux vecteurs aléatoires différents peuvent avoir les mêmes lois marginales.

Transformations des vecteurs aléatoires

Pour démontrer la formule de la densité d'une transformation d'un vecteur aléatoire, nous rappelons d'abord le résultat suivant de l'analyse.

Proposition 4. *Soit $h = (h_1, \dots, h_p)^T : \mathbb{R}^p \to \mathbb{R}^p$ une transformation avec $h(t_1, \dots, t_p) = (h_1(t_1, \dots, t_p), \dots, h_p(t_1, \dots, t_p))^T$, pour $t = (t_1, \dots, t_p)^T \in \mathbb{R}^p$. Supposons que*

(i) h est une bijection,

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

(iii) le Jacobien J_h de h défini par

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

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

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

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

Proposition 5. *Soit $\mathbf{X}$ un vecteur aléatoire dans $\mathbb{R}^p$ de densité $f_{\mathbf{X}}$. Soit $g : \mathbb{R}^p \to \mathbb{R}^p$ une transformation vérifiant les hypothèses de la Proposition 4. Alors, la densité $f_{\mathbf{Y}}$ du vecteur aléatoire $\mathbf{Y} = g(\mathbf{X})$ est donnée par*

$$f_{\mathbf{Y}}(u) = f_{\mathbf{X}}(g^{-1}(u)) |J_{g^{-1}}(u)| \ , \quad u \in \mathbb{R}^p \ .$$

Démonstration. Par (i), l'inverse $g^{-1}(\cdot)$ existe partout dans $\mathbb{R}^p$ et vérifie les conditions de la Proposition 4. En effet, d'après le Théorème de fonction inverse,

$$J_{g^{-1}}(g(u)) = \frac{1}{J_g(u)} , \quad J_{g^{-1}}(t) = \frac{1}{J_g(g^{-1}(t))} .$$

Soient $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$, $v = (v_1, \ldots, v_p)^T$ et $A_v = \{t \in \mathbb{R}^p : g_i(t) \leq v_i, i = 1, \ldots, p\}$. D'après la Proposition 1, la f.d.r. de $\mathbf{Y}$ s'écrit sous la forme

$$\begin{aligned} F_{\mathbf{Y}}(v) &= \mathbb{P}(Y_i \leq v_i, i = 1, \ldots, p) = \mathbb{P}(g_i(\mathbf{X}) \leq v_i, i = 1, \ldots, p) \\ &= \int_{A_v} f_{\mathbf{X}}(t)\mathrm{d}t = \int_{g(A_v)} f_{\mathbf{X}}(g^{-1}(u))|J_{g^{-1}}(u)|\mathrm{d}u . \end{aligned}$$

Or,

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

d'où on obtient

$$F_{\mathbf{Y}}(v) = \int_{-\infty}^{v_1} \ldots \int_{-\infty}^{v_p} f_{\mathbf{X}}(g^{-1}(u))|J_{g^{-1}}(u)|\mathrm{d}u ,$$

pour tout $v = (v_1, \ldots, v_p)^T \in \mathbb{R}^p$. Ceci signifie que la densité de $\mathbf{Y}$ est $f_{\mathbf{X}}(g^{-1}(u))|J_{g^{-1}}(u)|$. □

Corollaire 1. *Si $\mathbf{Y} = A\mathbf{X} + b$, où $\mathbf{X}$ est un vecteur aléatoire dans $\mathbb{R}^p$ de densité $f_{\mathbf{X}}$, $b \in \mathbb{R}^p$ est un vecteur déterministe et A est une matrice $p \times p$ telle que $\mathrm{Det}(A) \neq 0$, alors*

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

2.3.2 Indépendance

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

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) , \quad \text{pour tout } A \in \mathcal{B}, B \in \mathcal{B} .$$

Si $\mathbb{E}|X| < \infty, \mathbb{E}|Y| < \infty$, l'indépendance implique

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] . \tag{2.2}$$

Le réciproque n'est pas vraie : (2.2) n'est pas équivalent à l'indépendance de X et Y.

On dit que les vecteurs aléatoires $\mathbf{X} \in \mathbb{R}^p$ et $\mathbf{Y} \in \mathbb{R}^q$ sont **indépendants** si

$$\mathbb{P}(\mathbf{X} \in A, \mathbf{Y} \in B) = \mathbb{P}(\mathbf{X} \in A)\mathbb{P}(\mathbf{Y} \in B) , \quad \text{pour tout } A \in \mathcal{B}(\mathbb{R}^p), B \in \mathcal{B}(\mathbb{R}^q) .$$

On peut caractériser l'indépendance par la fonction caractéristique. Plus précisément, $\mathbf{X}$ et $\mathbf{Y}$ sont indépendants si et seulement si

$$\varphi_{(\mathbf{X},\mathbf{Y})}(t) = \varphi_{\mathbf{X}}(t_1)\varphi_{\mathbf{Y}}(t_2) , \quad \text{pour tout } \ t = (t_1, t_2)^T \in \mathbb{R}^{p+q} .$$

Soient X et Y des variables aléatoires absolument continues de densité f_X et f_Y. Notons $f_{(X,Y)}$ la densité du vecteur (X, Y). Alors, $X \perp\!\!\!\perp Y$ si et seulement si $f_{(X,Y)}(x, y) = f_X(x) f_Y(y)$ pour tout $x, y \in \mathbb{R}$.

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

On dit que les variables aléatoires $X_1, \ldots, X_n$ sur $(\Omega, A, \mathbb{P})$ sont **(mutuellement) indépendantes** si

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n) \,, \qquad \text{pour tout } A_1, \ldots, A_n \in \mathcal{B} \,. \tag{2.3}$$

On dit que $(X_n)_{n\geq 1}$ est une suite infinie de variables aléatoires indépendantes si (2.3) est vérifié pour tout $n \geq 1$ entier. 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 $X_1, \ldots, X_n$ soient mutuellement indépendantes. En revanche, l'indépendance mutuelle implique l'indépendance deux à deux.

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

2.3.3 Covariance et corrélation

La covariance ainsi que la corrélation sont des mesures pour quantifier la relation linéaire entre deux variables aléatoires.

Soient X et Y deux variables aléatoires de carré intégrable, i.e. $\mathbb{E}[X^2] < \infty$ et $\mathbb{E}[Y^2] < \infty$. La **covariance** entre X et Y est la valeur

$$\mathbf{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \,.$$

Si $\mathbf{Cov}(X, Y) = 0$, on dit que X et Y sont *non-corrélées* ou *orthogonales* et on écrit $X \perp Y$.

Propriétés de la covariance

1. $\mathbf{Cov}(X, X) = \mathbf{Var}(X)$.
2. $\mathbf{Cov}(X, Y) = \mathbf{Cov}(Y, X)$.
3. $\mathbf{Cov}(aX, bY) = ab\mathbf{Cov}(X, Y)$ pour $a, b \in \mathbb{R}$.
4. $\mathbf{Cov}(X + a, Y + b) = \mathbf{Cov}(X, Y)$ pour $a, b \in \mathbb{R}$.
5. $\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y) + 2\mathbf{Cov}(X, Y)$. En effet,
$$\begin{aligned}\mathbf{Var}(X + Y) &= \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] - (\mathbb{E}[X])^2 - (\mathbb{E}[Y])^2 - 2\mathbb{E}[X]\mathbb{E}[Y] \,.\end{aligned}$$
6. Si X et Y sont indépendantes, $\mathbf{Cov}(X, Y) = 0$. Le réciproque n'est pas vrai.

Soit $\mathbf{Var}(X) > 0$ et $\mathbf{Var}(Y) > 0$. La **corrélation** ou le **coefficient de corrélation** entre X et Y est la quantité

$$\rho_{X,Y} = \frac{\mathbf{Cov}(X, Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}} \,.$$

On dit que la corrélation entre X et Y est **positive** si $\rho_{X,Y} > 0$ et qu'elle est **négative** si $\rho_{X,Y} < 0$.

Propriétés de la corrélation

1. $-1 \leq \rho_{X,Y} \leq 1$.
2. Si les variables aléatoires X et Y sont indépendantes, $\rho_{X,Y} = 0$.
3. $|\rho_{X,Y}| = 1$ si et seulement si il existe un lien linéaire déterministe entre X et Y : il existe $a \neq 0, b \in \mathbb{R}$ tels que $Y = aX + b$ *p.s.*.
4. La corrélation est invariante par rapport aux transformations affines : pour tout $a \neq 0$, $b \neq 0$ $c, d \in \mathbb{R}$,
$$|\rho_{aX+c,bY+d}| = \text{signe}(ab)|\rho_{X,Y}| \ ,$$
où $\text{signe}(u) = \mathbb{1}\{u > 0\} + \mathbb{1}\{u < 0\}$.

Interprétation géométrique de la corrélation

Soit $\langle \cdot, \cdot \rangle$ le produit scalaire et $\| \cdot \|$ la norme de $L_2(\mathbb{P})$. Alors,

$$\mathbf{Cov}(X, Y) = \langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle \ ,$$

et

$$\rho_{X,Y} = \frac{\langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle}{\|X - \mathbb{E}[X]\| \|Y - \mathbb{E}[Y]\|} \ .$$

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

Caractéristique de la relation linéaire entre deux variables aléatoires

La covariance ainsi que la corrélation sont toutes les deux des mesures pour quantifier la relation linéaire entre deux variables aléatoires. Étant donné que la corrélation est une caractéristique normalisée, dont les valeurs sont comprises dans $[-1, 1]$, elle se prête mieux à juger si le lien entre les deux variables est plutôt linéaire ($|\rho_{X,Y}|$ est près de 1) ou pas linéaire du tout ($\rho_{X,Y}$ est près de 0). La covariance ne permet pas les mêmes conclusions.

Il est bien sûr possible que X et Y sont liées par une relation plus compliquée que linéaire. Dans le modèle de régression, par exemple, on suppose qu'il existe une fonction f telle que

$$Y = f(X) + \varepsilon \ ,$$

où ε est une variable aléatoire, appelée *bruit*.

Matrices de covariance

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

$$\Sigma \stackrel{\text{déf}}{=} \mathbf{Var}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] = (\sigma_{ij})_{i,j} \ ,$$

où

$$\sigma_{ij} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbf{Cov}(X_i, X_j) \ .$$

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 $\mathbf{Y} \in \mathbb{R}^q$:

$$\mathbf{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T] \, .$$

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 $\mathbf{Cov}(\mathbf{X}, \mathbf{Y}) = 0$ (matrice $p \times q$ nulle).

Propriétés des matrices de covariance. Soient $\mathbf{X} \in \mathbb{R}^p$ et $\mathbf{Y} \in \mathbb{R}^q$ deux vecteurs aléatoires.

(i) $\mathbf{Var}(\mathbf{X}) = \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^T$.

(ii) $\mathbf{Var}(\mathbf{X})$ est une matrice symétrique et semi-définie positive.
En effet, la symétrie découle de $\mathbf{Cov}(X_i, X_j) = \mathbf{Cov}(X_j, X_i)$ pour tout i, j. De plus, pour tout $a \in \mathbb{R}^p$

$$a^T \mathbf{Var}(\mathbf{X}) a = \mathbb{E}[(a^T\mathbf{X} - \mathbb{E}[a^T\mathbf{X}])(\mathbf{X}^T a - \mathbb{E}[\mathbf{X}^T a])] = \mathbb{E}[(a^T\mathbf{X} - \mathbb{E}[a^T\mathbf{X}])^2] = \mathbf{Var}(a^T\mathbf{X}) \geq 0 \, ,$$

car $a^T\mathbf{X} = \mathbf{X}^T a \in \mathbb{R}$.

(iii) Soit A une matrice $q \times p$ et $b \in \mathbb{R}^q$. Alors $\mathbf{Var}(A\mathbf{X} + b) = A\mathbf{Var}(\mathbf{X})A^T$.
En effet, soit $\mathbf{Y} = A\mathbf{X} + b$, alors par linéarité de l'espérance,

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[A\mathbf{X} + b] = A\mathbb{E}[\mathbf{X}] + b \quad \text{et} \quad \mathbf{Y} - \mathbb{E}[\mathbf{Y}] = A(\mathbf{X} - \mathbb{E}[\mathbf{X}]) \, .$$

Par linéarité de l'espérance pour les matrices aléatoires,

$$\mathbf{Var}(\mathbf{Y}) = \mathbb{E}[A(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T A^T) = A\mathbf{Var}(\mathbf{X})A^T \, .$$

(iv) $\mathbf{Cov}(\mathbf{X}, \mathbf{X}) = \mathbf{Var}(\mathbf{X})$.

(v) $\mathbf{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{Cov}(\mathbf{Y}, \mathbf{X})^T$.

(vi) Pour deux vecteurs aléatoires $\mathbf{X}_1 \in \mathbb{R}^p$ et $\mathbf{X}_2 \in \mathbb{R}^p$, on a $\mathbf{Cov}(\mathbf{X}_1 + \mathbf{X}_2, \mathbf{Y}) = \mathbf{Cov}(\mathbf{X}_1, \mathbf{Y}) + \mathbf{Cov}(\mathbf{X}_2, \mathbf{Y})$.

(vii) Si A est une matrice $m \times p$ et B est une matrice $k \times q$, alors $\mathbf{Cov}(A\mathbf{X}, B\mathbf{Y}) = A\mathbf{Cov}(\mathbf{X}, \mathbf{Y})B^T$.

(viii) Si $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$, alors $\mathbf{Cov}(\mathbf{X}, \mathbf{Y}) = 0$ (matrice $p \times q$ nulle). L'implication inverse n'est pas vraie.

2.4 Convergence de suites de vecteurs aléatoires

2.4.1 Définitions et propriétés fondamentales

Soient $\mathbf{X}_1, \mathbf{X}_2, \ldots$ et $\mathbf{X}$ des vecteurs aléatoires de dimension p définis sur $(\Omega, A, \mathbb{P})$. On note $\|\cdot\|$ la norme euclidienne.

On dit que la suite $(\mathbf{X}_n)_{n\geq 1}$ **converge en probabilité** vers $\mathbf{X}$ quand $n \to \infty$ (et on écrit $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$) si

$$\lim_{n\to\infty} \mathbb{P}(\|\mathbf{X}_n - \mathbf{X}\| > \varepsilon) = 0 \, , \quad \text{pour tout } \varepsilon > 0 \, .$$

On dit que la suite $(\mathbf{X}_n)_{n\geq 1}$ **converge presque sûrement** (en abrégé *p.s.*) vers $\mathbf{X}$ quand $n \to \infty$ (et on écrit $\mathbf{X}_n \to \mathbf{X}$ *p.s.*) si

$$\mathbb{P}\left(\omega : \lim_{n\to\infty} \mathbf{X}_n(\omega) \neq \mathbf{X}(\omega)\right) = 0 .$$

La convergence presque sûre est équivalente à

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geq n} \|\mathbf{X}_k - \mathbf{X}\| \geq \varepsilon\right) = 0 , \quad \text{pour tout } \varepsilon > 0 .$$

On dit que la suite $(\mathbf{X}_n)_{n\geq 1}$ **converge en loi** (ou en distribution) vers $\mathbf{X}$ quand $n \to \infty$ (et on écrit $\mathbf{X}_n \xrightarrow{\mathcal{L}} \mathbf{X}$) si

$$\lim_{n\to\infty} F_{\mathbf{X}_n}(t) = F_{\mathbf{X}}(t) ,$$

pour chaque point $t \in \mathbb{R}^p$ de continuité de la fonction de répartition $F_{\mathbf{X}}(\cdot)$ de $\mathbf{X}$.

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

$$\lim_{n\to\infty} \mathbb{E}[h(\mathbf{X}_n)] = \mathbb{E}[h(\mathbf{X})] .$$

Relations entre les différents modes de convergence

convergence presque sûre $\Rightarrow$ convergence en probabilité $\Rightarrow$ convergence en loi

Remarquons que si $a \in \mathbb{R}^p$ est un vecteur déterministe, alors

$$\mathbf{X}_n \xrightarrow{\mathcal{L}} a \quad \Longleftrightarrow \quad \mathbf{X}_n \xrightarrow{P} a . \tag{2.4}$$

Sommes et produits de suites

Proposition 6. *Soient $X, X_1, X_2, \ldots, Y, Y_1, Y_2, \ldots$ des variables aléatoires.*
(i) Si $X_n \to X$ p.s. et $Y_n \to Y$ p.s., alors

$$X_n + Y_n \to X + Y \text{ p.s.} \quad \text{et} \quad X_n Y_n \to XY \text{ p.s.}$$

(ii) Si $X_n \xrightarrow{P} X$ et $Y_n \xrightarrow{P} Y$, alors

$$X_n + Y_n \xrightarrow{P} X + Y \quad \text{et} \quad X_n Y_n \xrightarrow{P} XY .$$

(iii) **(Théorème de Slutsky)** *Si $X_n \xrightarrow{\mathcal{L}} a$ et $Y_n \xrightarrow{\mathcal{L}} Y$ et $a \in \mathbb{R}$ est une constante, alors*

$$X_n + Y_n \xrightarrow{\mathcal{L}} a + Y \quad \text{et} \quad X_n Y_n \xrightarrow{\mathcal{L}} aY .$$

Sommes de variables indépendantes

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires. Si $\mathbb{E}|X_i| < \infty$ pour $i = 1, \ldots, n$,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] .$$

Si les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et si $\mathbb{E}[X^2] < \infty$ pour $i = 1, \ldots, n$, alors

$$\mathbf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbf{Var}(X_i) \; .$$

Proposition 7. *Soient $X_1, \ldots, X_n$ des v.a. i.i.d. telles que $\mathbb{E}[X_1] = \mu$ et $\mathbf{Var}(X_1) = \sigma^2 < \infty$. Alors la moyenne arithmétique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ vérifie $\mathbb{E}[\bar{X}_n] = \mu$ et $\mathbf{Var}(\bar{X}_n) = \frac{1}{n}\mathbf{Var}(X_1) = \sigma^2/n$.*

Proposition 8. (Loi forte des grands nombres de Kolmogorov) *Soient $X_1, X_2, \ldots$ des v.a. i.i.d. telles que $\mathbb{E}|X_1| < \infty$ et $\mathbb{E}[X_1] = \mu$. Alors,*

$$\bar{X}_n \to \mu \; p.s. \quad \textit{quand } n \to \infty \; .$$

Proposition 9. (Théorème central limite) *Soient $X_1, X_2, \ldots$ des v.a. i.i.d. telles que $\mathbb{E}[X^2] < \infty$ et $\sigma^2 = \mathbf{Var}(X_1) > 0$. Alors,*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{\mathcal{L}} Z \; , \quad \textit{quand } n \to \infty \; ,$$

où $\mu = \mathbb{E}[X]$ et $Z \sim \mathcal{N}(0,1)$.

2.4.2 Théorèmes de continuité

Théorème 3. (Premier théorème de continuité) *Soit $g : D \subset \mathbb{R}^p \to \mathbb{R}^q$ une fonction continue définie sur un sous-ensemble D de $\mathbb{R}^p$ et soient $\mathbf{X}_1, \mathbf{X}_2, \ldots$ et $\mathbf{X}$ des vecteurs aléatoires définis sur $(\Omega, A, \mathbb{P})$ à valeurs dans D. Alors,*

$$\begin{aligned}
&(i) && \mathbf{X}_n \longrightarrow \mathbf{X} \; p.s. && \Longrightarrow g(\mathbf{X}_n) \longrightarrow g(\mathbf{X}) \; p.s. \; , \\
&(ii) && \mathbf{X}_n \xrightarrow{P} \mathbf{X} && \Longrightarrow g(\mathbf{X}_n) \xrightarrow{P} g(\mathbf{X}) \; , \\
&(iii) && \mathbf{X}_n \xrightarrow{\mathcal{L}} \mathbf{X} && \Longrightarrow g(\mathbf{X}_n) \xrightarrow{\mathcal{L}} g(\mathbf{X}) \; ,
\end{aligned}$$

quand $n \to \infty$.

Démonstration. La partie (i) est évidente. Montrons (ii) sous l'hypothèse supplémentaire que $\mathbf{X} = a$, où a est un vecteur 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 $\varepsilon > 0$ il existe $\delta > 0$ tel que

$$\|\mathbf{X}_n - a\| < \delta \Rightarrow \|g(\mathbf{X}_n) - g(a)\| < \varepsilon \; .$$

En particulier, $\mathbb{P}(\|\mathbf{X}_n - a\| < \delta) \leq \mathbb{P}(\|g(\mathbf{X}_n) - g(a)\| < \varepsilon)$. Comme $\mathbf{X}_n \xrightarrow{P} a$, on a

$$\lim_{n\to\infty} \mathbb{P}(\|\mathbf{X}_n - a\| < \delta) = 1 \quad \text{pour tout } \delta > 0 \; ,$$

ce qui implique

$$\lim_{n\to\infty} \mathbb{P}(\|g(\mathbf{X}_n) - g(a)\| < \varepsilon) = 1 \quad \text{pour tout } \varepsilon > 0 \; .$$

(iii) Il suffit de démontrer que, pour toute fonction continue et bornée $h(\cdot)$, $\mathbb{E}[h(g(\mathbf{X}_n))] \to \mathbb{E}[h(g(\mathbf{X}))]$ quand $n \to \infty$. Comme g est continue, $f = h \circ g$ est aussi continue et bornée. Ceci démontre (iii), car $\mathbf{X}_n \xrightarrow{\mathcal{L}} \mathbf{X}$ signifie que $\mathbb{E}[f(\mathbf{X}_n)] \to \mathbb{E}[f(\mathbf{X})]$ quand $n \to \infty$ pour toute fonction f continue et bornée. □

Théorème 4. (Deuxième théorème de continuité *ou* Méthode delta) *Soit* $g = (g_1, \ldots, g_q)^T : D \subset \mathbb{R}^p \to \mathbb{R}^q$ *une fonction continûment différentiable. Soient* $\mathbf{X}_1, \mathbf{X}_2, \ldots$ *et* $\mathbf{X}$ *des vecteurs aléatoires à valeurs dans* D *tels que* $r_n(\mathbf{X}_n - m) \xrightarrow{\mathcal{L}} \mathbf{X}$*, où* $r_n \to \infty$ *est une suite réelle et* $m \in \mathbb{R}^p$ *un vecteur déterministe. Alors*

$$r_n \left(g(\mathbf{X}_n) - g(m)\right) \xrightarrow{\mathcal{L}} \nabla g(m)\mathbf{X} \ , \quad n \to \infty \ ,$$

où $\nabla g(t) = (\frac{\partial g_i(t)}{\partial t_j})_{i,j}$.

Démonstration. Nous montrons le résultat pour des variables aléatoires $X_1, X_2, \ldots$ réelles et $g : D \subset \mathbb{R} \to \mathbb{R}$ une fonction réelle.

Sous les hypothèses de la proposition, la fonction

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

est continue.

Notons d'abord que l'on a $1/r_n \to 0$, et donc d'après le Théorème de Slutsky,

$$X_n - m = \frac{1}{r_n} r_n (X_n - m) \xrightarrow{\mathcal{L}} 0 \ .$$

Par (2.4) on obtient $X_n \xrightarrow{P} m$. Le Premier théorème de continuité implique que

$$h(X_n) \xrightarrow{P} h(m) = g'(m) \ , \quad \text{quand } n \to \infty \ .$$

Or,

$$\sqrt{n} \left(g(X_n) - g(m)\right) = h(X_n)\sqrt{n}(X_n - m) \ .$$

On conclut en utilisant le Théorème de Slutsky. □

Le cas d'application le plus courant de la méthode delta concerne la moyenne empirique quand celle-ci converge en loi d'après le théorème central limite.

Corollaire 2. *Soit* $g(\cdot)$ *une fonction continûment différentiable et soient* $X_1, X_2, \ldots$ *des variables aléatoires i.i.d. telles que* $\mathbb{E}[X_1^2] < \infty$ *et* $\sigma^2 = \mathbf{Var}(X_1) > 0$*. Alors*

$$\sqrt{n} \left(\frac{g(\bar{X}_n) - g(\mu)}{\sigma} \right) \xrightarrow{\mathcal{L}} g'(\mu) Z \ , \quad \textit{quand } n \to \infty \ ,$$

où $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$*,* $\mu = \mathbb{E}[X_1]$ *et* $Z \sim \mathcal{N}(0, 1)$.

2.5 Quelques inégalités

Théorème 5. (Inégalité de Markov) *Soit* $h(\cdot)$ *une fonction positive croissante et soit* X *une variable aléatoire telle que* $\mathbb{E}[h(X)] < \infty$*. Alors pour tout* $a \in \mathbb{R}$ *tel que* $h(a) > 0$*,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[h(X)]}{h(a)} \ . \tag{2.5}$$

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

$$\begin{aligned}\mathbb{P}(X \geq a) \leq \mathbb{P}(h(X) \geq h(a)) &= \int \mathbb{1}\{h(x) \geq h(a)\} \mathrm{d}F(x) \\ &= \mathbb{E}[\mathbb{1}\{h(X) \geq h(a)\}] \leq \mathbb{E}\left[\frac{h(X)}{h(a)} \mathbb{1}\{h(X) \geq h(a)\}\right] \leq \mathbb{E}\left[\frac{h(X)}{h(a)}\right] .\end{aligned}$$

□

Théorème 6. (Inégalité de Tchebychev) *Soit X une variable aléatoire telle que $\mathbb{E}[X^2] < \infty$. Alors, pour tout $a > 0$,*

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[X^2]}{a^2} \quad \text{et} \quad \mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathbf{Var}(X)}{a^2} .$$

Démonstration. Il suffit de poser $h(t) = t^2$ et d'appliquer (2.5) aux variables aléatoires $|X|$ et $|X - \mathbb{E}[X]|$ respectivement. □

Théorème 7. (Inégalité de Hölder) *Soit $1 < r < \infty$, $\mathbb{1}/r + 1/s = 1$. Soient X et Y deux variables aléatoires telles que $\mathbb{E}[|X|^r] < \infty$ et $\mathbb{E}[|Y|^s] < \infty$. Alors $\mathbb{E}[|XY|] < \infty$ et*

$$\mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^r])^{1/r} (\mathbb{E}[|Y|^s])^{1/s} .$$

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

$$\frac{1}{r} \log a + \frac{1}{s} \log b \leq \log\left(\frac{a}{r} + \frac{b}{s}\right) ,$$

ce qui est équivalent à :

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

Posons ici $a = |X|^r / \mathbb{E}[|X|^r], b = |Y|^s / \mathbb{E}[|Y|^s]$ (on suppose pour l'instant que $\mathbb{E}[|X|^r] \neq 0$, $\mathbb{E}[|Y|^s] \neq 0$), ce qui donne

$$|XY| \leq \mathbb{E}[|X|^r]^{1/r} \mathbb{E}[|Y|^s]^{1/s} \left(\frac{|X|^r}{r\mathbb{E}[|X|^r]} + \frac{|Y|^s}{s\mathbb{E}[|Y|^s]}\right) .$$

On conclut en prenant l'espérance et en utilisant le fait que $1/r + 1/s = 1$. Si $\mathbb{E}[|X|^r] = 0$ ou $\mathbb{E}[|Y|^r] = 0$, alors $X = 0$ p.s. ou $Y = 0$ p.s., et l'inégalité est triviale. □

Théorème 8. (Inégalité de Jensen) *Soit $g(\cdot)$ une fonction convexe et soit X une variable aléatoire telle que $\mathbb{E}|X| < \infty$. Alors*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)] .$$

Démonstration. Par convexité de g, il existe une fonction $\dot{g}(\cdot)$ telle que

$$g(x) \geq g(x_0) + (x - x_0)\dot{g}(x_0) ,$$

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

$$g(X) \geq g(\mathbb{E}[X]) + (X - \mathbb{E}[X])\dot{g}(\mathbb{E}[X]) .$$

En prenant les espérances on obtient $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. □

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

$$|\mathbb{E}[X]| \leq \mathbb{E}|X| . \tag{2.6}$$

Théorème 9. (Inégalité de Cauchy-Schwarz) *Soient X et Y deux variables aléatoires telles que $\mathbb{E}[X^2] < \infty$ et $\mathbb{E}[Y^2] < \infty$. Alors $\mathbb{E}[|XY|] < \infty$ et*

$$(\mathbb{E}[XY])^2 \leq (\mathbb{E}[|XY|])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2] \,, \tag{2.7}$$

et les égalités dans (2.7) 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,

$$a_1 X + a_2 Y = 0 \,. \tag{2.8}$$

Démonstration. La seconde inégalité dans (2.7) est un cas particulier de l'inégalité de Hölder pour $r = s = 2$. La première inégalité dans (2.7) est une conséquence de (2.6). Si (2.8) est vrai, il est évident que

$$(\mathbb{E}[XY])^2 - \mathbb{E}[X^2]\mathbb{E}[Y^2] = 0 \,. \tag{2.9}$$

Réciproquement, si l'on a (2.9) et $\mathbb{E}[Y^2] \neq 0$, alors $\mathbb{E}[(X - aY)^2] = 0$ avec $a = \mathbb{E}[XY]/\mathbb{E}[Y^2]$, ce qui implique $X = aY$ presque sûrement. Le cas où $\mathbb{E}[Y^2] = 0$ est trivial. □

2.6 Exercices

Exercice 1. Lois continues
Pour toutes les familles de lois continues du paragraphe 2.1.2 tracer l'allure de la densité et étudier l'impact des paramètres sur la forme de la densité.

Exercice 2. Loi normale
Soient X et Y deux variables aléatoires de loi normale : $X \sim \mathcal{N}(\mu, \sigma^2)$ et $Y \sim \mathcal{N}(\nu, \rho^2)$.

1. Montrer que la fonction caractéristique de X est donnée par

$$\varphi_X(t) = \exp\left\{ it\mu - \frac{1}{2}\sigma^2 t^2 \right\} \,, \quad t \in \mathbb{R} \,.$$

2. Montrer que $aX + b$ suit une loi normale pour tout $a, b \in \mathbb{R}$.
3. Montrer que X et Y sont indépendants si et seulement si $\mathbf{Cov}(X, Y) = 0$.
4. Soient X et Y indépendants. Montrer que $X + Y$ suit une loi normale.

Exercice 3. Fonction gamma
La fonction gamma $\Gamma(\cdot)$ est définie par

$$\Gamma(p) = \int_0^\infty z^{p-1} \mathrm{e}^{-z} \mathrm{d}z \,, \quad p > 0 \,.$$

Montrer que

1. $\Gamma(1) = 1$ et $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.
2. $\Gamma(p+1) = p\Gamma(p)$ pour tout $p > 0$, et en particulier $\Gamma(n+1) = n!$ pour tout $n \in \mathbb{N}$.

Exercice 4. Montrer que $\mathbb{E}[|X|] = 0$ implique que $X = 0$ *p.s.*.

Exercice 5. 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 $\mathbb{E}[|X|^a] = \infty$ pour tout $a > 0$.

Exercice 6. Moments d'une loi symétrique
Démontrer la Proposition 3.

Exercice 7. Les coefficients d'asymétrie et d'aplatissement

1. Montrer que, pour la loi normale $\mathcal{N}(\mu, \sigma^2)$, les coefficients d'asymétrie et d'aplatissement valent $\alpha = 0$ et $\beta = 0$.
2. Donner un exemple de densité non-symétrique avec $\alpha = 0$.
3. Montrer que les coefficients d'asymétrie et d'aplatissement sont invariants par rapport aux transformations affines (d'échelle et de position) de la variable aléatoire X. Autrement dit, les variables X et $Y = aX + b$ avec $a > 0$ et $b \in \mathbb{R}$ ont le même coefficient d'asymétrie α et le même coefficient d'aplatissement β.

Exercice 8. Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes. Posons

$$X_{(1)} = \min(X_1, \ldots, X_n) \ , \quad X_{(n)} = \max(X_1, \ldots, X_n) \ .$$

1. Montrer que

$$\mathbb{P}(X_{(1)} \geq x) = \prod_{i=1}^{n} \mathbb{P}(X_i \geq x) \ , \quad \mathbb{P}(X_{(n)} < x) = \prod_{i=1}^{n} \mathbb{P}(X_i < x) \ .$$

2. Supposons, de plus, que $X_1, \ldots, X_n$ sont identiquement distribuées avec la loi uniforme $U[0, \theta]$. Calculer $\mathbb{E}[X_{(1)}], \mathbb{E}[X_{(n)}], \mathbf{Var}(X_{(1)})$ et $\mathbf{Var}(X_{(n)})$.

Exercice 9. Soit X une variable aléatoire positive avec la f.d.r. F et d'espérance finie. Démontrer que

$$\mathbb{E}[X] = \int_0^\infty (1 - F(x))\mathrm{d}x = \int_0^\infty \mathbb{P}(X > x)\mathrm{d}x \ .$$

Exercice 10. 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 11. Méthode de Box-Muller
Soient X, Y deux variables indépendantes et de même loi $\mathcal{N}(0, 1)$ et soient (R, θ) les coordonnées polaires de (X, Y).

1. Donner la loi du couple (R, θ).
2. Proposer une méthode permettant de simuler la loi de (X, Y), à partir de deux variables uniformément distribuées sur $[0, 1]$ et indépendantes.

Partie 2

Modélisation et Estimation

Chapitre 3

Statistique descriptive

L'objectif de la statistique consiste à extraire des informations utiles des données observées. On dénote les données par un vecteur $\mathbf{x} = (x_1, \ldots, x_n)$. Les x_i peuvent être des valeurs réelles, des vecteurs ou encore des matrices. Afin que l'analyse statistique d'un jeu de données ait un sens, il faut que les différents éléments de cette série d'observations représentent la même quantité mesurée sur des entités différentes. On appelle la suite $\mathbf{x} = (x_1, \ldots, x_n)$ **échantillon**, **données**, **jeu de données** ou encore **observations**. Le nombre n est dit **taille d'échantillon**. On écrit aussi $\mathbf{x}_n$ au lieu de $\mathbf{x}$ quand on veut mettre en avant la taille d'échantillon.

Des exemples de source de données sont les sondages, les expérience scientifiques (physiques, chimiques, médicales,...), les enregistrements historiques (météorologiques, socioéconomiques,...) etc. Dans certains cas, ces données sont volumineux et difficile à interpréter. On a alors besoin de les résumer et de trouver des outils pertinents pour les visualiser.

L'approche statistique repose sur l'introduction d'un modèle probabiliste pour les données. La première question qu'un statisticien doit alors traiter lors d'une analyse statistique d'un jeu de données $\mathbf{x}$ est celle du choix de modèle. Autrement dit, il s'agit d'identifier les hypothèses permettant de définir une famille de lois $\mathcal{P}$, appelée le **modèle**, à laquelle la loi $\mathbb{P}$ des données $\mathbf{x}$ est susceptible d'appartenir.

La démarche pour sélectionner un tel modèle $\mathcal{P}$ est complexe, et est un mélange d'expérience, de connaissance *a priori*, de considération sur les lois physiques ayant engendrés les données et bien sûr d'hypothèses de travail. Par ailleurs, on utilise des outils de la statistique descriptive pour représenter les données graphiquement et de les décrire ou résumer par des caractéristiques numériques dans le but d'identifier les hypothèses sur la loi $\mathbb{P}$ ayant générée l'observation $\mathbf{x}$.

Lorsque l'observation $\mathbf{x}$ est considérée comme la réalisation d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$, on se pose généralement trois questions sur les X_i :

- Sont-ils indépendants ?
- Ont-ils la même loi ?
- Quelle est la forme de la loi ?

Le plus souvent le statisticien s'appuie sur ses connaissances *a priori* du phénomène observé pour répondre aux deux premières questions. En effet, la plupart du temps nous supposerons dans ce cours que les X_i sont bien indépendants et de même loi. Pour répondre à la troisième question, les outils de la statistique descriptive sont employés. En effet, des représentations graphiques des données permettent de se faire une première idée de la distribution des données, de vérifier la pertinence de certaines hypothèses sur les données et leur distribution et de choisir un modèle probabiliste approprié.

Nous présenterons dans ce chapitre les outils les plus répandus de la statistique descriptive.

Tableau de données

-3.26	-1.04	-0.46	0.93	-0.99	0.25	-0.35	0.59	0.54	-1.04
0.06	-0.99	1.04	-0.34	0.24	-0.30	-1.72	1.57	0.58	1.09
0.45	-0.08	-1.06	0.95	0.19	-1.40	-0.13	0.61	0.88	-0.02
1.00	0.02	-0.69	0.27	-0.26	0.85	-0.58	0.89	1.15	-1.18
0.14	-0.67	-0.62	-1.39	1.31	-0.17	0.19	-1.01	-1.10	1.28
-1.12	0.26	1.24	-0.70	-0.61	-2.09	1.30	0.18	0.17	-0.66
2.42	1.53	-1.21	-1.75	0.85	-0.10	-0.69	-0.46	0.56	-1.54
1.61	-1.10	0.23	0.37	1.07	1.94	-1.23	0.86	-0.66	-0.85
0.89	0.21	-1.13	-0.21	-0.31	0.47	-0.98	1.75	0.29	0.41
-2.02	-8.91	4.76	3.34	1.16	-0.21	-5.36	-5.40	-0.78	8.04

TABLE 3.1 – 100 observations d'une variable continue.

3.1 Observations à valeurs réelles

Nous commençons par le cas d'observations à valeurs réelles, c'est-à-dire tout au long de ce paragraphe, nous considérons un jeu de données $\mathbf{x} = (x_1, \ldots, x_n)$ qui représente n valeurs réelles constituant les résultats d'une certaine expérience répétée n fois. On dit alors que $\mathbf{x}$ sont les valeurs d'une variable observée sur n individus. Par exemple, $\mathbf{x}$ peuvent être les durées de n personnes passées sur internet durant un mois, ou les températures journalières moyennes à Paris enregistrées au cours de l'année 2009. Mathématiquement, on considérera les données $\mathbf{x}$ comme la réalisation d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$. De plus, *nous supposerons que les X_i sont i.i.d.*, c'est-à-dire des variables indépendantes et de même loi.

On peut différencier deux types de données : le *cas discrète*, où la loi des X_i est une loi discrète, et le *cas continu*, où la loi des X_i est une loi absolument continue. Le statisticien doit décider au vu des données et grâce à sa connaissance sur l'expérience de quel type il s'agit. En fait, si le nombre de valeurs différentes parmi $\mathbf{x} = (x_1, \ldots, x_n)$ est petit devant n, il s'agit d'une loi discrète. Dans le cas contraire, on est dans le cas continu. Par exemple, la durée journalière passée sur internet de n personnes sont des observations d'une variable continue. En revanche, le nombre de e-mails envoyés par ces n personnes sont des observations discrètes.

Les données représentées dans le Tableau 3.1 forme un échantillon de taille 100, issue d'une expérience où on a mesuré la même entité sur 100 sujets différents. Le nombre de valeurs différentes étant presque n, il est clair que la loi des données est continue. Pour des données d'une loi discrète le Tableau 3.2 en est un exemple. C'est un échantillon de taille 100, composé de nombres entiers.

Notre objectif est de se faire une idée de la distribution de ces données. Cependant, sous forme d'un tableau, on a du mal à caractériser les données. Il convient alors, d'une part, de représenter les données par des différents outils graphiques, et d'autre part, de calculer des caractéristiques numériques qui donnent des informations sur p. ex. la tendance centrale ou la variabilité des observations.

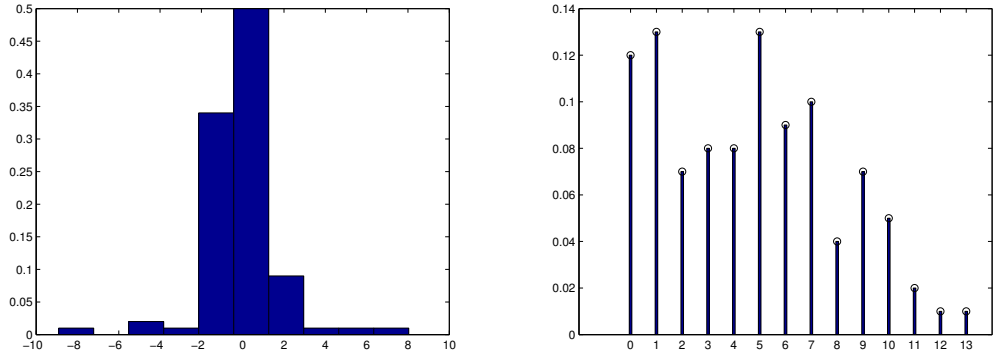
3.1.1 Histogramme

L'histogramme est une façon de représenter graphiquement les données lorsque la loi des X_i est *continue*. On le construit de façon suivante. Soit A un intervalle qui contient toutes les observations $\mathbf{x} = (x_1, \ldots, x_n)$ et soit $A_1, \ldots, A_m$ une partition de A en m sous-intervalles

Tableau de données

2	5	5	0	7	5	11	1	0	1	9	5	9	5	3	3	0	5	4	10
2	1	3	4	10	8	1	5	5	9	9	4	4	0	1	0	9	1	11	3
3	6	10	5	0	1	0	4	2	1	1	1	3	10	7	3	5	6	7	2
6	2	4	7	12	7	7	4	0	0	1	2	5	6	6	13	4	8	6	9
0	1	7	8	7	8	3	6	0	0	2	9	7	6	5	6	1	7	5	10

TABLE 3.2 – 100 observations d'une variable discrète.



(a) (b)

FIGURE 3.1 – (a) Histogramme des données du Tableau 3.1. (b) Diagramme en bâtons des données du Tableau 3.2.

de longueur h chacun. On définit $N_j = \#\{i : x_i \in A_j\} = \sum_{i=1}^n \mathbb{1}\{x_i \in A_j\}$ le nombre d'observations x_i dans l'intervalle A_j. L'**histogramme** est une fonction constante par morceaux définie par

$$\hat{f}^H(x) = \frac{1}{nh} \sum_{j=1}^m N_j \mathbb{1}\{x \in A_j\} , \quad x \in \mathbb{R} .$$

Figure 3.1 (a) montre l'histogramme pour les données du Tableau 3.1.

La forme de l'histogramme dépend du choix de la largeur h des sous-intervalles A_j. Une bonne valeur de h dépend de la taille n de l'échantillon et des observations $\mathbf{x}_n$.

Notons que l'histogramme est une fonction positive ($\hat{f}^H \geq 0$) et elle intègre à 1 :

$$\int_{\mathbb{R}} \hat{f}^H(x)\mathrm{d}x = \frac{1}{nh} \sum_{j=1}^m N_j \int_{\mathbb{R}} \mathbb{1}\{x \in A_j\}\mathrm{d}x = \frac{1}{nh} \sum_{j=1}^m N_j h = 1 .$$

Par conséquent, l'histogramme $\hat{f}^H$ est la densité de probabilité d'une loi continue.

En effet, on peut montrer que si les X_i sont i.i.d. de loi continue et de densité f l'histogramme $\hat{f}^H$ peut être considéré comme une approximation de f. Plus précisément, pour x_0 fixé, $\hat{f}^H(x_0)$ converge en probabilité vers $f(x_0)$ lorsque la taille d'échantillon n tend vers l'infini et quand on considère des histogrammes avec des sous-intervalles A_j de plus en plus courts ($h_n \to 0$) (cf. Feuille TD n°2 pour plus de détails).

La grande importance de l'histogramme pour la modélisation provient alors du fait qu'il donne une idée de la forme de la densité f que l'on pourra choisir comme modèle (pourvu qu'on est dans le cas i.i.d. et la loi est absolument continue).

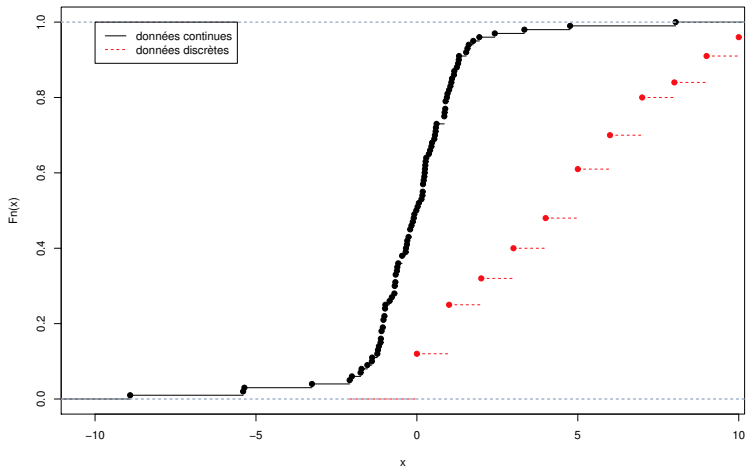


FIGURE 3.2 – Fonction de répartition empirique $\hat{F}$ pour les données continues du Tableau 3.1 (ligne continue) et pour les données discrètes du Tableau 3.2 (ligne pointillée).

Sous R, l'instruction pour tracer l'histogramme d'un jeu de données nommé `vec` est

```
hist(vec, freq=FALSE)
```

L'option `freq=FALSE` est indisponsable afin d'ajuster l'échelle de l'histogramme en sorte que ce soit une densité de probabilité. D'autres option de la fonction `hist()` permettent de modifier la partition en sous-intervalles.

3.1.2 Diagramme en bâtons

La représentation des données par histogramme est appropriée pour des données générée par une loi continue. Pour des données *discrètes*, il est préférable d'utiliser le **diagramme en bâtons**. Notons $\mathcal{V} = \{v_k, k = 1, \ldots, m\}$ l'ensemble de valeurs prises par les observations $\mathbf{x} = (x_1, \ldots, x_n)$. On a évidemment $m \leq n$. Pour tout k, on calcule la proportion d'observations de valeur v_k dans l'échantillon

$$p_k = \frac{\#\{i : x_i = v_k\}}{n} , \quad k = 1, \ldots, m .$$

Les p_k vérifie, bien évidemment, $\sum_{k=1}^{m} p_k = 1$.

Pour le diagramme en bâtons on trace des bâtons verticaux au niveau des v_k de hauteur p_k. Figure 3.1 (b) montre le diagramme en bâtons pour les données du Tableau 3.2.

Remarquons que quand les X_i sont i.i.d., p_k converge en probabilité vers la probabilité $\mathbb{P}(X_1 = v_k)$ lorsque n tend vers l'infini. Par conséquent, le diagramme en bâtons donne une idée de la loi de probabilité des X_i et peut être utile dans le choix d'un modèle.

Sous R, l'instruction pour tracer le diagramme en bâtons d'un jeu de données est

```
plot(table(vec)/length(vec))
```

3.1.3 Fonction de répartition empirique

Alternativement à l'histogramme et au diagramme en bâtons, on peut représenter les données par leur fonction de répartition empirique.

La **fonction de répartition empirique** $\hat{F}$ (ou $\hat{F}_n$) associée à l'échantillon $\mathbf{x} = (x_1, \dots, x_n)$ est définie par

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq t\} = \frac{\#\{i : x_i \leq t\}}{n} , \quad t \in \mathbb{R} . \tag{3.1}$$

Il est simple de voir que la fonction de répartition empirique $\hat{F}$ a les propriétés suivantes : $\hat{F}$ est une fonction définie sur tout $\mathbb{R}$, elle est croissante et continue à droite. Elle prend ses valeurs dans $[0, 1]$ et vérifie

$$\hat{F}(t) = 0 \ , \forall t < \min\{x_1, \dots, x_n\} \qquad \text{et} \qquad \hat{F}(t) = 1 \ , \forall t > \max\{x_1, \dots, x_n\} .$$

Plus précisément, $\hat{F}$ est une fonction en escalier avec des sauts en x_i. En fait, $\hat{F}$ est la fonction de répartition d'une loi discrète. Plus précisément, soit X une variable aléatoire de loi $\hat{F}$, alors X vérifie pour $i = 1, \dots, n$

$$\begin{aligned} \mathbb{P}(X = x_i) &= \int_{\{x_i\}} \mathrm{d}\hat{F}(u) = \hat{F}(x_i) - \lim_{u \to x_i -} \hat{F}(u) \\ &= \frac{\#\{k : x_k \leq x_i\}}{n} - \frac{\#\{k : x_k < x_i\}}{n} = \frac{\#\{k : x_k = x_i\}}{n} . \end{aligned}$$

Si les valeurs des observations sont deux à deux distinctes ($x_i \neq x_j$ pour tout $i \neq j$), alors

$$\mathbb{P}(X = x_i) = \frac{1}{n} .$$

Lorsque les x_i sont des réalisations i.i.d. de loi F, la fonction de répartition empirique $\hat{F}_n$ donne une approximation de F.

Théorème 10. *Soient $X_1, X_2, \dots$ une suite de variables aléatoires i.i.d. de loi F et $\hat{F}_n$ la fonction de répartition empirique de répartition associée à $(X_1, \dots, X_n)$.*

(i) $n\hat{F}_n(t) \sim \mathrm{Bin}(n, F(t))$.

(ii) $\hat{F}_n(t) \longrightarrow F(t)$ *p.s. quand* $n \to \infty$ *pour tout* $t \in \mathbb{R}$.

(iii) $\sqrt{n}(\hat{F}_n(t) - F(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(t)(1 - F(t)))$ *lorsque* $n \to \infty$.

(iv) **(Théroème de Glivenko-Cantelli)** $\hat{F}_n$ *converge uniformément presque sûrement vers F, c'est-à-dire*

$$\|\hat{F}_n - F\|_\infty := \sup \left\{ |\hat{F}_n(t) - F(t)|, t \in \mathbb{R} \right\} \longrightarrow 0 \ p.s. , \quad n \to \infty .$$

Démonstration. *(i)* Notons $Y_i = \mathbb{1}\{X_i \leq t\}$. Les Y_i sont i.i.d., car les X_i le sont. Comme Y_i prend ses valeurs dans $\{0, 1\}$, Y_i suit la loi de Bernoulli de paramètre $\mathbb{P}(Y_i = 1) = \mathbb{P}(X_i \leq t) = F(t)$. D'où $n\hat{F}_n(t) = \sum_{i=1}^n Y_i \sim \mathrm{Bin}(n, F(t))$.

(ii) Par la loi forte des grands nombres, $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n Y_i \longrightarrow \mathbb{E}[Y_1] = F(t)$ *p.s.* quand $n \to \infty$.

(iii) D'après le théorème central limite, $\quad n \to \infty$,

$$\sqrt{n}(\hat{F}_n(t) - F(t)) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} Y_i - \mathbb{E}[Y_1] \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{Var}(Y_1)) = \mathcal{N}(0, F(t)(1 - F(t))) .$$

(*iv*) Nous montrons le théorème de Glivenko-Cantelli dans le cas particulier où F est continue. Soit $0 < \varepsilon < 1$. Il existe une partition $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ telle que $F(t_j) - F(t_{j-1}) < \varepsilon$ pour tout $j = 1, \ldots, k$. Par (*ii*) on obtient la convergence uniforme sur un nombre fini de points, c'est-à-dire on a

$$\sup\left\{|\hat{F}_n(t) - F(t)|, t \in \{t_1, \ldots, t_{k-1}\}\right\} \longrightarrow 0 \ p.s. \ , \quad n \to \infty \ .$$

Or, pour tout $t \in [t_{j-1}, t_j]$, on a

$$\begin{aligned}\hat{F}_n(t) - F(t) &\leq \hat{F}_n(t_j) - F(t_j) + \varepsilon \ , \\ \hat{F}_n(t) - F(t) &\geq \hat{F}_n(t_{j-1}) - F(t_{j-1}) - \varepsilon \ .\end{aligned}$$

Donc,

$$\|\hat{F}_n - F\|_\infty = \sup\left\{|\hat{F}_n(t) - F(t)|, t \in \mathbb{R}\right\} \leq \sup\left\{|\hat{F}_n(t) - F(t)|, t \in \{t_1, \ldots, t_{k-1}\}\right\} + \varepsilon$$

On en déduit que $\limsup_{n\to\infty} \|\hat{F}_n - F\|_\infty \leq \varepsilon \ p.s.$ Ceci est vrai pour tout $\varepsilon > 0$, ce qui implique le résultat.

□

L'avantage de la fonction de répartition empirique comparé à l'histogramme ou au diagramme en bâtons, est le fait qu'elle est définie pour tout échantillon quelque soit son type : continue, discrète ou autre. Néanmoins, il est plus facile de repérer certaines caractéristiques de loi (comme p.ex. la symétrie) dans un histogramme ou un diagramme en bâtons. De même, il est plus facile de comparer deux histogrammes ou diagrammes en bâtons que des fonctions de répartition empiriques.

Figure 3.2 montre les fonctions de répartition empiriques pour les données des Tableaux 3.1 et 3.2. La fonction de répartition empirique des données discrètes est décalée vers la droite par rapport à la fonction de répartition empirique des données continues. Ceci indique que les valeurs de la variable discrète ont tendance à être plus élevées que celles de la variable continue. De plus, on observe que la fonction de répartition empirique des données continues à beaucoup plus de sauts que l'autre fonction de répartition empirique alors que les deux échantillons sont de même taille. Cette différence s'explique par le fait qu'une des variables est continue et l'autre discrète.

Sous `R`, on calcule la fonction de répartition empirique d'un vecteur nommé `vec` par l'instruction suivante :

```
ecdf(vec)
```

et pour tracer la fonction de répartition empirique on écrit

```
plot(ecdf(vec))
```

En fait, la fonction de répartition empirique est peu utilisée pour choisir un modèle. En revanche, elle intervient souvent dans la construction d'estimateurs ou de tests statistiques.

En effet, quand on ne connait pas la loi F, mais seulement un échantillon i.i.d. $\mathbf{x}$ de F, on peut utiliser les caractéristiques de la loi empirique $\hat{F}$ associée à $\mathbf{x}$ pour approcher les caractéristiques de la loi F. Par exemple, si on aimerait connaître la moyenne de la loi F, on peut calculer la moyenne de la loi $\hat{F}$ qui devrait être proche, puisque $\hat{F}$ est une approximation de F. Pour calculer la moyenne $\mathbb{E}[X]$ lorsque X suit la loi empirique $\hat{F}$. On rappelle que $\hat{F}$ est la loi discrète à valeurs dans $\{x_1, \ldots, x_n\}$ qui associe le poids $1/n$

à chaque observation. Plus précisément, $\mathbb{P}_{\hat{F}}(X = t) = \#\{i : x_i = t\}/n$ pour tout $t \in \mathbb{R}$. Notons $\{v_1, \ldots, v_k\}$ les valeurs différentes dans l'échantillon $(x_1, \ldots, x_n)$. On a alors

$$\mathbb{E}_{\hat{F}}[X] = \int u \mathrm{d}\hat{F}(u) = \sum_{j=1}^{k} v_j \mathbb{P}_{\hat{F}}(X = v_j) = \frac{1}{n} \sum_{j=1}^{k} v_j \#\{i : x_i = v_j\} = \frac{1}{n} \sum_{i=1}^{n} x_i =: \bar{x}_n \ .$$

On appelle $\bar{x}_n$ la **moyenne empirique**.

De même, on définit la **variance empirique** $s^2_{\mathbf{x}}$ par

$$\begin{aligned} s^2_{\mathbf{x}} &= \mathbf{Var}_{\hat{F}}(X) = \mathbb{E}_{\hat{F}}[X^2] - (\mathbb{E}_{\hat{F}}[X])^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \ , \end{aligned}$$

et l'**écart-type empirique** $s_{\mathbf{x}} = \sqrt{s^2_{\mathbf{x}}}$.

Ou encore, on définit le **coefficient d'asymétrie empirique** $\alpha_{\mathbf{x}}$ par

$$\alpha_{\mathbf{x}} = \frac{\mathbb{E}_{\hat{F}}[(X - \mathbb{E}_{\hat{F}}[X])^3]}{(\mathbb{E}_{\hat{F}}[(X - \mathbb{E}_{\hat{F}}[X])^2]))^{3/2}} = \frac{1}{n s^3_{\mathbf{x}}} \sum_{i=1}^{n} (x_i - \bar{x}_n)^3 \ .$$

ainsi que le **coefficient d'aplatissement empirique** $\beta_{\mathbf{x}}$ par

$$\beta_{\mathbf{x}} = \frac{\mathbb{E}_{\hat{F}}[(X - \mathbb{E}_{\hat{F}}[X])^4]}{(\mathbb{E}_{\hat{F}}[(X - \mathbb{E}_{\hat{F}}[X])^2]))^2} - 3 = \frac{1}{n s^4_{\mathbf{x}}} \sum_{i=1}^{n} (x_i - \bar{x}_n)^4 - 3 \ .$$

Pour éviter toute confusion, la fonction de répartition F de X sera appelée fonction de répartition *théorique* et ses caractéristiques (fonctionnelles) seront appelées *caractéristiques théoriques.* Les fonctionnelles respectives de $\hat{F}$ seront appelées *caractéristiques empiriques.*

On peut également utiliser la fonction de répartition empirique $\hat{F}$ pour définir des équivalents empiriques aux quantiles de la loi F. Plus précisément, on définit le **quantile empirique** $\hat{q}_\alpha$ (ou $\hat{q}^n_\alpha$) d'ordre α, $0 < \alpha < 1$, associé à $\mathbf{x}$ par

$$\hat{q}_\alpha = \hat{F}^{-1}(\alpha) \ ,$$

où $\hat{F}^{-1}$ désigne la fonction quantile associée à la loi empirique $\hat{F}$.

Afin de donner une expression simple des quantiles empiriques, nous introduisons la notion des statistique d'ordre. Les statistiques d'ordre $x_{(1)}, \ldots, x_{(n)}$ sont les valeurs x_i classées par ordre croissant :

$$x_{(j)} \in \{x_1, \ldots, x_n\} \ , \quad x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \ .$$

Le j-ième plus petit élément $x_{(j)}$ de l'échantillon $\mathbf{x}$ s'appelle la **j-ième statistique d'ordre** ou **statistique d'ordre de rang j**.

Clairement, $x_{(1)} = \min\{x_1, \ldots, x_n\}$ et $x_{(n)} = \max\{x_1, \ldots, x_n\}$.

Théorème 11. *Soit $\mathbf{x} = (x_1, \ldots, x_n)$ et $0 < \alpha < 1$. Le quantile empirique $\hat{q}_\alpha$ d'ordre α est donné par*

$$\hat{q}_\alpha = x_{(\lceil \alpha n \rceil)} \ ,$$

où $\lceil a \rceil$ désigne le plus petit entier supérieur ou égal à a.

Démonstration. cf. TD Feuille n°2. □

Le quantile empirique $\hat{q}_\alpha$ d'ordre α est tel que $\alpha 100\%$ des observations x_i sont inférieures à $\hat{q}_\alpha$ et $(1-\alpha)100\%$ des x_i sont supérieures à $\hat{q}_\alpha$. Plus précisément, on a

$$\frac{\#\{x_i : x_i \leq \hat{q}_\alpha\}}{n} \geq \alpha \quad \text{et} \quad \frac{\#\{x_i : x_i \geq \hat{q}_\alpha\}}{n} \geq 1-\alpha \, .$$

On peut également montrer que les quantiles empiriques $\hat{q}_\alpha^n$ converge en probabilité vers les quantiles théoriques q_α^F de la loi F lorsque n tend vers l'infini.

Théorème 12. *Soient $X_1, X_2, \ldots$ des v.a. i.i.d. de loi F. Notons $\hat{q}_\alpha^n$ le quantile empirique d'ordre α associé à $(X_1, \ldots, X_n)$. Alors,*

$$\hat{q}_\alpha^n \xrightarrow{P} q_\alpha^F \, , \qquad \textit{lorsque} \quad n \to \infty \, .$$

Démonstration. cf. TD Feuille n°2. □

En utilisant les quantiles empiriques, on définit les équivalents aux caractéristiques du Chapitre 2. Plus précisément, on appelle **médiane empirique** $\text{MED}^{\mathbf{x}}$ de l'échantillon $\mathbf{x}$ le quantile empirique d'ordre $1/2$:

$$\text{MED}^{\mathbf{x}} = \hat{q}_{1/2} \, .$$

Notons que la médiane empirique $\text{MED}^{\mathbf{x}}$ est la valeur telle que la moitié des observations est plus grandes que $\text{MED}^{\mathbf{x}}$ et la moitié des observations est plus petites que $\text{MED}^{\mathbf{x}}$.

On définit l'**écart interquartile** $\mathit{EIQ}^{\mathbf{x}}$ par

$$\mathit{EIQ}^{\mathbf{x}} = \hat{q}_{3/4} - \hat{q}_{1/4} \, ,$$

où $\hat{q}_{1/4}$ et $\hat{q}_{3/4}$ désignent les quantiles empiriques d'ordre $1/4$ et $3/4$, aussi appelé **le premier et le troisième quartile empirique**.

Pour résumer, on a les propriétés suivantes :

Théorème 13. *Soit $X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d. de loi F. On a, quand n tend vers l'infini,*

(i) $\text{MED}^{\mathbf{X}_n} \xrightarrow{P} \text{MED}^F$,

(ii) $\mathit{EIQ}^{\mathbf{X}_n} \xrightarrow{P} \mathit{EIQ}^F$,

(iii) $\bar{X}_n \xrightarrow{P} \mathbb{E}_F[X]$, *si* $\mathbb{E}_F[|X|] < \infty$,

(iv) $s^2_{\mathbf{X}_n} \xrightarrow{P} \mathbf{Var}_F(X)$, *si* $\mathbb{E}_F[X^2] < \infty$,

(v) $\alpha_{\mathbf{X}_n} \xrightarrow{P} \alpha_F$, *si* $\mathbb{E}_F[|X|^3] < \infty$,

(vi) $\beta_{\mathbf{X}_n} \xrightarrow{P} \beta_F$, *si* $\mathbb{E}_F[X^4] < \infty$.

Démonstration. cf. TD Feuille n° 2. □

Sous `R` on calcule la moyenne, la variance et l'écart type empirique du vecteur `vec` par les commandes

```
mean(vec)
var(vec)
sd(vec)
```

Notons que les deux dernières fonctions divisent la somme par $n-1$ au lieu de n. La médiane et les quantiles empiriques sont obtenus par les instructions suivantes

```
median(vec)
quantile(vec,0.25)
```

pour le quantile empirique d'ordre 0.25.

3.1.4 Indicateurs de la tendance centrale, dispersion et forme

Résumons les différentes caractéristiques pour décrire la loi d'un échantillon $\mathbf{x} = (x_1, \ldots, x_n)$.

Tendance centrale

La **tendance centrale** ou la **position** de la loi de $\mathbf{x} = (x_1, \ldots, x_n)$ est un point autour duquel se concentre une grande partie des observations. On peut la mesurer par la *moyenne empirique* $\bar{x}_n$ ou par la **médiane empirique** $\mathrm{MED}^{\mathbf{x}}$.

En général, la moyenne empirique et la médiane empirique ne sont pas identiques. Par exemple, le salaire net moyen en France est de 2082 € (en 2010), alors que le salaire médian est de 1674 € seulement. La différence s'explique par le fait que quelques salaires très élevés ont un fort impact sur la moyenne empirique, mais pas sur la médiane empirique.

Un troisième estimateur de la tendance centrale est la *moyenne tronquée* que l'on peut considérer comme une version robuste de la simple moyenne empirique. En effet, il s'agit de calculer la moyenne empirique de seulement une partie des observations en ignorant les observations aberrantes. Pour cela on fixe une proportion $\gamma \in (0, 1)$ et on écarte $\gamma 100\%$ des observations les plus éloignées de l'échantillon $(x_1, \ldots, x_n)$ (les $\frac{\gamma}{2}100\%$ des plus petites valeurs ainsi que les $\frac{\gamma}{2}100\%$ des plus grandes valeurs). Ainsi pour $n = 100$ et $\gamma = 0.04$ on supprime les deux plus petites et les deux plus grandes observations :

$$\underbrace{x_{(1)} \leq x_{(2)}}_{\text{à supprimer}} \leq \underbrace{x_{(3)} \leq \cdots \leq x_{(98)}}_{\text{à conserver}} \leq \underbrace{x_{(99)} \leq x_{(100)}}_{\text{à supprimer}}$$

$$\Downarrow$$

$$\bar{x}_{\mathrm{tronq}(0.04)} = \frac{1}{96} \sum_{i=3}^{98} x_{(i)} \ .$$

Plus généralement, la *moyenne tronquée* avec une troncature à $\gamma 100\%$ est définie par

$$\bar{x}_{\mathrm{tronq}(\gamma)} = \frac{1}{n - 2\lceil \gamma n/2 \rceil} \sum_{\lceil \gamma n/2 \rceil + 1}^{n - \lceil \gamma n/2 \rceil} x_{(i)} \ ,$$

où $x_{(i)}$ désigne la i-ème statistique d'ordre de $(x_1, \ldots, x_n)$.

Pour une comparaison de la moyenne, la médiane et la moyenne tronquée voir Chapitre 8.

Dispersion

Bien que deux échantillons peuvent avoir la même tendance centrale (en terme de moyenne empirique ou de médiane empirique), la **dispersion** ou **variabilité** des données, c'est-à-dire la concentration des observations autour de cette valeur, n'est pas nécessairement identique. La dispersion est quantifiée par la *variance empirique* $s^2_{\mathbf{x}}$, l'*écart-type empirique* $s_{\mathbf{x}}$ ou l'*écart interquartile EIQ*$^{\mathbf{x}}$.

Symétrie

Un échantillon $\mathbf{x}_n = (x_1, \ldots, x_n)$ symétrique est caractérisé par le fait que les observations sont distribuées (plus ou moins) symétriquement autour d'un point μ. La symétrie d'un

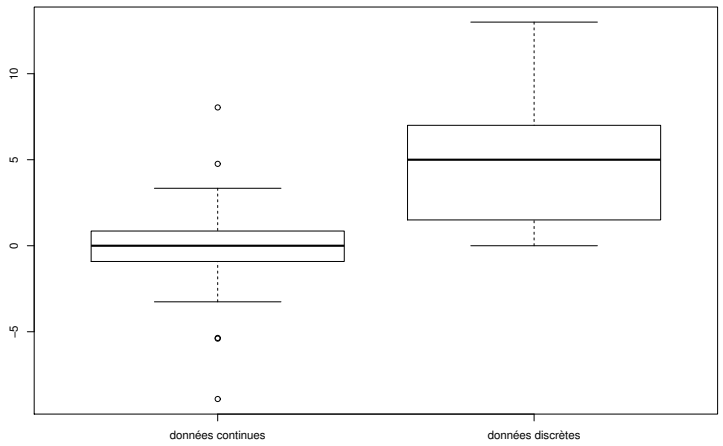


FIGURE 3.3 – Boxplot pour les données du Tableau 3.1 (à gauche) et les données du Tableau 3.2 (à droite).

échantillon se traduit par le *coefficient d'asymétrie empirique* $\alpha_{\mathbf{x}}$ près de 0. Notons que le réciproque n'est pas exacte : la condition $\alpha_{\mathbf{x}} \approx 0$ n'implique pas la symétrie des données.

On peut qualifier les lois asymétriques comme étant "proches" ou "éloignées" de distributions symétriques.

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

APLATISSEMENT

Le *coefficient d'aplatissement empirique* $\beta_{\mathbf{x}}$ est le plus souvent calculé pour avoir une idée intuitive sur les "queues" de la distribution des données $\mathbf{x}_n$. Plus précisément, il s'agit de dire quelque chose sur la présence d'observations "éloignées" des autres observations. On dit qu'il s'agit d'une loi avec des "queues lourdes" si l'échantillon contient des observations aberrantes, i.e. des observations loin de la tendance centrale. On dit que "les queues sont légères" si les observations sont bien concentrées autour de la tendance centrale. En effet, si $\beta_{\mathbf{x}} > 0$, l'échantillon contient plus d'"observations éloignées" qu'une loi normale. Si $\beta_{\mathbf{x}} < 0$, il en contient moins.

Notons aussi qu'en tout cas on a $\beta_{\mathbf{x}} > -2$.

3.1.5 BOXPLOT

Un **boxplot** ou une **boîte à moustaches** est une représentation graphique des observations $\mathbf{x} = (x_1, \ldots, x_n)$ basée sur des caractéristiques de la tendance centrale, de la dispersion et de la forme. Plus précisément, le boxplot permet de repérer le centre des données (représenté par la médiane empirique $\text{MED}^{\mathbf{x}}$), la dispersion (intervalle interquartile $EIQ^{\mathbf{x}}$), la symétrie ou dissymétrie des données (par la localisation de la médiane par rapport aux quartiles) et la présence des observations aberrantes. Figure 3.3 montre les

boxplots associés aux données du Tableau 3.1 et du Tableau 3.2.

Les limites du rectangle d'un boxplot sont données par les quartiles empiriques $\hat{q}_{1/4}$ et $\hat{q}_{3/4}$ de l'échantillon $\mathbf{x}$. La longueur du rectangle correspond donc à l'écart interquartile $EIQ^{\mathbf{x}}$. Par conséquent, la moitié des observations $\mathbf{x}$ est contenue dans le rectangle du boxplot.

Le trait horizontal dans le rectangle est la médiane empirique $\text{MED}^{\mathbf{x}}$.

En général, les moustaches se terminent aux observations minimale $x_{(1)}$ et maximale $x_{(n)}$ (voir Figure 3.3 à droite), sauf si l'échantillon contient des observations aberrantes. On dit que x_j est une **observation aberrante** (en anglais **outlier**) si

$$x_j > q^{\mathbf{x}}_{3/4} + \tau EIQ^{\mathbf{x}} \quad \text{ou} \quad x_j < q^{\mathbf{x}}_{1/4} - \tau EIQ^{\mathbf{x}} \ ,$$

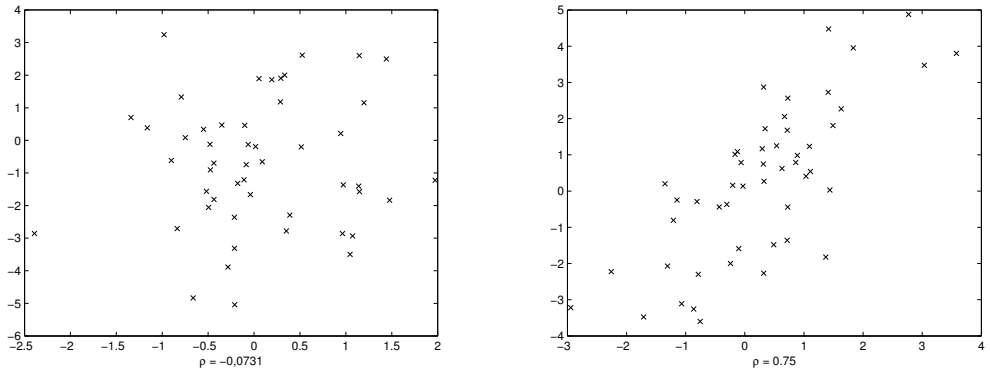
où τ est typiquement choisie égal à 1,5. Si les données contiennent des observations aberrantes, chacune est représentée dans le boxplot par un point isolé aux extrémités du graphique, et les moustaches ont une longueur maximale de $\tau EIQ^{\mathbf{x}}$. Plus précisément, la moustache inférieure se termine à la plus petite observation x_j qui vérifie $x_j > q^{\mathbf{x}}_{1/4} - \tau EIQ^{\mathbf{x}}$. De même, la moustache supérieure se termine à la plus grande observation x_j qui vérifie $x_j < q^{\mathbf{x}}_{3/4} + \tau EIQ^{\mathbf{x}}$.

La présence ou non d'observations aberrantes donne une indication sur la loi de l'échantillon, notamment s'il s'agit d'une loi avec queues légères (comme la loi normale) ou une loi avec des queues lourdes (comme la loi de Pareto). Le boxplot à gauche dans la Figure 3.3 représente un échantillon contenant quatre valeurs aberrantes.

Le boxplot permet alors de repérer très rapidement plusieurs caractéristiques de la loi qui a engendrée les données. Il est aussi très pratique pour comparer deux échantillons comme nous voyons dans la Figure 3.3. On observe tout de suite des nombreuses différences entre les deux distributions, notamment en termes de tendance centrale, de variance et de présence de valeurs aberrantes.

Sous `R` il y a plusieurs possibilités pour tracer des boxplots. On utilise par exemple la fonction `boxplot()` :

```
boxplot(vec)
```



(a) (b)

(c) (d)

FIGURE 3.4 – Différents exemples pour des nuages des points avec une corrélation quasi nulle (a), une corrélation positive $\hat{\rho}_{\mathbf{xy}} = 0{,}75$ (b), une relation presque linéaire $\hat{\rho}_{\mathbf{xy}} = -0{,}975$ (c), et encore une corrélation quasi nulle (d).

3.2 Observations d'un couple de variables

Considérons maintenant le cas où on observe un couple de variables, ce qui donne lieu à un échantillon avec des observations bidimensionnelles

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix} ,$$

correspondant aux valeurs de deux variables prélevées sur le même individu. Par exemple, x_i et y_i peuvent constituer la taille et le poids d'une personne, le budget d'état dédié au sport de n pays et leur nombre de médailles olympiques remportées, la température moyenne et le niveau de pollution à Paris un jour donné,...

3.2.1 Nuage des points

Il est naturel de représenter des données bidimensionnelles par un nuage de points, où chaque observation est représentée par un point de cordonnées (x_i, y_i). Figure 3.4 montre quatre exemples de nuages des points de différentes formes.

Il est courant d'étudier la nature de la relation entre les observations x_i avec les y_i. On veut savoir si, par exemple, il existe un lien entre la température et la pollution de l'air à Paris.

Sous R, on trace le nuage des points de (vecx, vecy) par la commande

```
plot(vecx,vecy)
```

3.2.2 Corrélation empirique

Si on considère les observations (x_i, y_i) comme des réalisiations i.i.d. d'un vecteur aléatoir (X, Y), on peut approcher la loi du couple (X, Y) par la fonction de répartition empirique bidimensionnelle $\hat{F}$ définie par

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \leq t_1, y_i \leq t_2\} = \frac{\#\{i : x_i \leq t_1, y_i \leq t_2\}}{n} , \quad t = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \in \mathbb{R}^2 .$$

Comme dans le cas unidimensionnelle, la fonction de répartition empirique $\hat{F}$ est la loi discrète à valeurs dans $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ qui associe à chaque points (x_i, y_i) le poids $1/n$. Plus précisément, si (X, Y) suit la loi empirique $\hat{F}$, on a

$$\mathbb{P}_{\hat{F}}((X, Y) = (x_i, y_i)) = \frac{\#\{k : x_k = x_i, y_k = y_i\}}{n} .$$

En utilisant la loi empirique $\hat{F}$, on peut définir des équivalents empiriques à la covariance et de la corrélation. On appelle **covariance empirique** des observations $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$ la valeur

$$s_{\mathbf{xy}} = \mathbf{Cov}_{\hat{F}}(X, Y) = \mathbb{E}_{\hat{F}}[XY] - \mathbb{E}_{\hat{F}}[X]\mathbb{E}_{\hat{F}}[Y] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n) ,$$

où $\bar{x}_n$ et $\bar{y}_n$ sont respectivement la moyenne empirique de $\mathbf{x} = (x_1, \ldots, x_n)$ et celle de $\mathbf{y} = (y_1, \ldots, y_n)$.

On appelle **coefficient de corrélation (linéaire) empirique** ou **corrélation empirique** des observations $(\mathbf{x}, \mathbf{y})$ la valeur

$$\hat{\rho}_{\mathbf{xy}} = \rho_{\hat{F}} = \frac{s_{\mathbf{xy}}}{s_{\mathbf{x}} s_{\mathbf{y}}} ,$$

où $s_{\mathbf{x}}$ et $s_{\mathbf{y}}$ sont respectivement l'écart-type empirique de $\mathbf{x}$ et celui de $\mathbf{y}$. Par convention, on pose $\hat{\rho}_{\mathbf{xy}} = 0$ si au moins l'un des deux écart-types $s_{\mathbf{x}}$, $s_{\mathbf{y}}$ est nul.

Pour une suite $(X_1, Y_1), (X_2, Y_2), \ldots$ de vecteurs aléatoires i.i.d., on peut montrer que le coefficient de corrélation empirique $\hat{\rho}_{\mathbf{X}_n \mathbf{Y}_n}$ converge en probabilité vers le coefficient de corrélation théorique ρ_{X_1, Y_1} lorsque n tend vers l'infini (pourvu que X_1 et Y_1 sont de carré intégrable).

Rappelons que la corrélation est une mesure quantitative pour évaluer la relation *linéaire* entre les deux variables. À la différence de la covariance, les valeurs du coefficient de corrélation $\hat{\rho}_{\mathbf{xy}}$ sont restreinte à l'intervalle fermé $[-1, 1]$. Ceci permet de parler d'une "forte (ou faible) corrélation linéaire" quand $|\hat{\rho}_{\mathbf{xy}}|$ est près de 1 (ou près de 0). Étant non normalisée, la covariance ne donne pas le même renseignement et elle est quasi inutile pour l'interprétation de la relation entre les données.

Théorème 14. *Le coefficient de corrélation empirique est toujours entre -1 est +1 :*

$$-1 \leq \hat{\rho}_{\mathbf{xy}} \leq 1 .$$

De plus, $|\hat{\rho}_{\mathbf{xy}}| = 1$ si et seulement si les observations $\mathbf{x} = (x_1, \ldots, x_n)$ et $\mathbf{y} = (y_1, \ldots, y_n)$ sont liées par une relation affine, i.e. il existe a et b tels que $x_i = a y_i + b$ pour tout $i = 1, \ldots, n$.

Démonstration. En utilisant l'inégalité de Cauchy-Schwarz, on vérifie que

$$|s_{\mathbf{xy}}| \leq \frac{1}{n}\sum_{i=1}^{n}|(x_i-\bar{x}_n)(y_i-\bar{y}_n)| \leq \frac{1}{n}\left(\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\sum_{i=1}^{n}(y_i-\bar{y}_n)^2\right)^{1/2} = s_{\mathbf{x}}s_{\mathbf{y}}\ .$$

Cela implique que le coefficient de corrélation $\hat{\rho}_{\mathbf{xy}} = s_{\mathbf{xy}}/(s_{\mathbf{x}}s_{\mathbf{y}})$ est toujours entre -1 et +1.

De plus, l'inégalité de Cauchy-Schwarz est une égalité si et seulement si $x_i - \bar{x}_n = a(y_i - \bar{y}_n) + b$ pour tout $i = 1, \ldots, n$, ce qui entraîne la seconde assertion de la proposition. □

Les quatre exemples de la Figure 3.4 sont des nuages des points avec des différents coefficients de corrélation empiriques. Dans le premier nuage des points, on ne peut pas apercevoir de tendance linéaire entre les différents points, et effectivement, la corrélation $\hat{\rho}_{\mathbf{xy}}$ est quasi nulle. Dans le deuxième graphique, la corrélation est positive, et on observe une nette tendance croissante dans le nuage de points. Dans le troisième exemple, les points sont presque alignés sur une droite décroissante, ce qui correspond à une corrélation $\hat{\rho}_{\mathbf{xy}}$ proche de -1. Dans le dernier exemple, on observe une tendance décroissante sur la première partie du nuage des points, et puis une tendance croissante sur la deuxième partie. Cependant, la corrélation est nulle. En fait, le coefficient de corrélation quantifie la relation *linéaire* entre les points, mais il est inapproprié pour quantifier des relations plus complexes, comme par exemple des relations polynômiales.

Sous `R`, on calcule la covariance empirique et la corrélation empirique de `vecx` et `vecy` par les commandes

```
cov(vecx,vecy)
cor(vecx,vecy)
```

3.3 Comparaison de distributions

Comparaison de deux échantillons

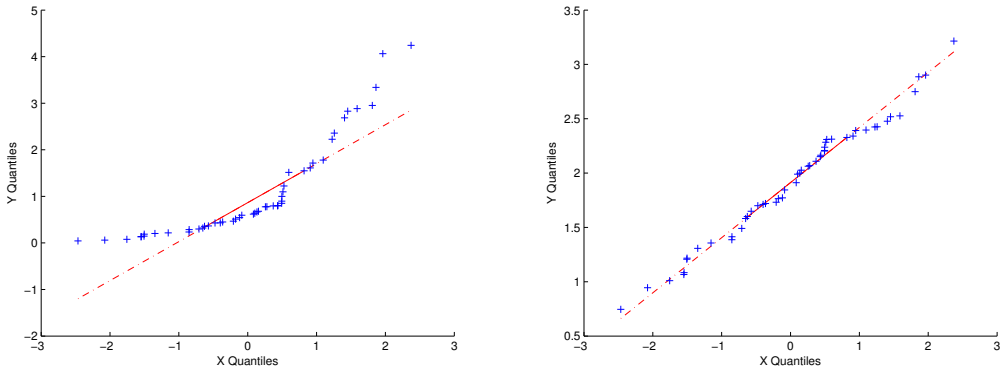
Considérons deux échantillons $\mathbf{x}_n = (x_1, \ldots, x_n)$ et $\mathbf{y}_m = (y_1, \ldots, y_m)$ (qui ne sont pas nécessairement de même taille). Le **Q-Q plot** ou **diagramme quantile-quantile** permet de voir rapidement si les distributions des échantillons $\mathbf{x}_n$ et $\mathbf{y}_m$ se ressemblent ou pas.

Le Q-Q plot compare les quantiles empiriques $\hat{q}^{\mathbf{x}}_{j/r}$ d'ordre j/r de l'échantillon $\mathbf{x}_n$ aux quantiles empiriques $\hat{q}^{\mathbf{y}}_{j/r}$ de l'échantillon $\mathbf{y}_m$, où $r = \min(n, m)$. Plus précisément, on trace le nuage des points

$$(\hat{q}^{\mathbf{x}}_{j/r}, \hat{q}^{\mathbf{y}}_{j/r}) = (x_{(\lceil jn/r \rceil)}, y_{(\lceil jm/r \rceil)}) \quad \text{pour } j = 1, \ldots, r = \min(n, m)\ .$$

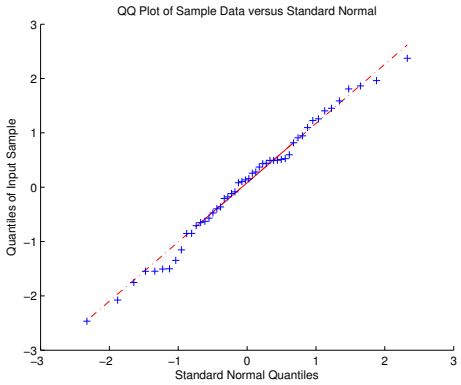
Lorsque les deux échantillons ont la même distribution, les points du Q-Q plot s'alignent sur la première bissectrice.

Si les points s'alignent sur une droite quelconque, alors les deux échantillons ont la même distribution à une transformation affine près.



(a)

(b)



(c)

FIGURE 3.5 – Exemples de différents Q-Q plot. (a) et (b) sont chacun basés sur deux échantillons. (c) compare un échantillon avec la loi normale standard $\mathcal{N}(0,1)$.

Comparaison d'un échantillon à une loi théorique

Le Q-Q plot permet également de comparer la distribution d'un échantillon $\mathbf{x}_n$ à une loi théorique donnée, disons F_0. Plus précisément, le Q-Q plot permet de vérifier une hypothèse de loi pour les données, par exemple on peut tester s'il est envisageable d'utiliser une loi normale pour modéliser la loi des données observées.

Dans ce cas, on trace les points $(\hat{q}^{\mathbf{x}}_{j/n}, q^{F_0}_{j/n})$ pour $j = 1, \ldots, n$, où $q^{F_0}_{j/n}$ désigne le quantile (théorique) d'ordre j/n de la loi F_0.

L'interprétation du Q-Q plot est la même que dans le cas de comparaison de deux échantillons.

Figure 3.5 montre plusieurs exemples de Q-Q plots. Dans le graphique (a), les points ne s'alignent pas du tout sur une droite. On conclut que les lois des deux échantillons ne se ressemblent pas. Le graphique (b) montre un nuage où les points s'alignent sur une droite. En revanche, cette droite n'est pas la première bissectrice. Par conséquent, on conclut que les lois des deux échantillons sont les mêmes à une transformation affine près. Le dernier exemple montre une comparaison d'un échantillon avec les quantiles de la loi normale standard $\mathcal{N}(0, 1)$. L'adéquation semble parfaite.

Sous `R`, on obtient le QQ-plot `vecx` et `vecy` par la commande

```
qqplot(vecx,vecy)
```

Pour obtenir le QQ-plot qui compare un vecteur d'observation `vecx` à la loi normale standard, on exécute la commande

```
qqnorm(vecx)
```

3.4 Exercices

Exercice 1. Interprétation de graphiques
Les graphiques de la Figure 3.6 représentent quatre échantillons i.i.d. de taille 100. Pour chaque graphique

(i) déduire des caractéristiques de la loi de l'échantillon,

(ii) proposer une loi (ou famille de lois) qui est susceptible d'avoir généré les données.

Exercice 2. Diagramme en bâtons
Soient $X_1, X_2, \ldots$ une suite de variables aléatoires discrètes i.i.d. à valeurs dans $\mathcal{V} = \{v_k, k = 1, \ldots, m\}$. Notons

$$p_{k,n} = \frac{\#\{i = 1, \ldots, n : X_i = v_k\}}{n} , \quad k = 1, \ldots, m .$$

Montrer que

1. pour tout n on a $\sum_{k=1}^{m} p_{k,n} = 1$.
2. pour tout k on a $p_{k,n} \to \mathbb{P}(X_1 = v_k)$ *p.s.* lorsque n tend vers l'infini.

Exercice 3. Fonction de répartition empirique
Soient $X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d. de loi F et $\hat{F}_n$ la fonction de répartition empirique de répartition associée à $X_1, \ldots, X_n$. Montrer que pour tout $x_0 \in \mathbb{R}$ on a

$$\hat{F}_n(x_0) \longrightarrow F(x_0) \ p.s. , \quad \text{lorsque } n \to \infty .$$

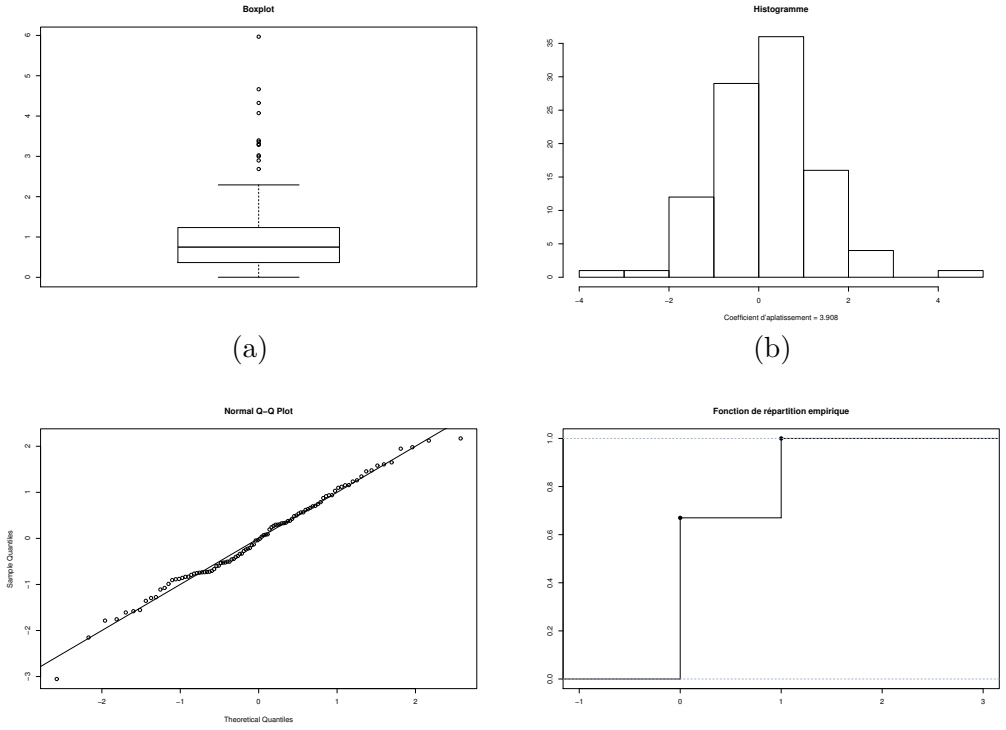


FIGURE 3.6 – Représentations graphiques de quatre échantillons différents i.i.d. de taille 100.

Exercice 4. Histogramme

Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de densité f et de fonction de répartition F. Soit h un réel positif et $x_0 \in \mathbb{R}$ un point fixé.

1. On suppose que F soit dérivable en x_0 avec $F'(x_0) = f(x_0)$. Quelle est la limite de

$$\frac{F(x_0 + h/2) - F(x_0 - h/2)}{h}$$

lorsque h tend vers 0 ?

Soit $A =]x_0 - h/2, x_0 + h/2]$ l'intervalle centré en x_0 de longueur h. L'histogramme des observations $X_1, \ldots, X_n$ au point $x_0 \in \mathbb{R}$ peut être défini par

$$\hat{f}^H(x_0) = \frac{\#\{i : X_i \in A\}}{nh} = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}\{X_i \in A\} .$$

2. Montrer que

$$\hat{f}^H(x_0) = \frac{\hat{F}(x_0 + h/2) - \hat{F}(x_0 - h/2)}{h} ,$$

où $\hat{F}$ désigne la fonction de répartition empirique associée à $(X_1, \ldots, X_n)$.

3. Donner la loi de la variable aléatoire $nh\hat{f}^H(x_0)$.
4. Calculer l'espérance et la variance de $\hat{f}^H(x_0)$ et en déduire la valeur du risque quadratique $\mathbb{E}\left[\left(\hat{f}^H(x_0) - f(x_0)\right)^2\right]$.
5. Considérons le comportement asymptotique de $\hat{f}^H(x_0) = \hat{f}^H_n(x_0)$ lorsque n tend vers l'infini. Plus précisément, on pose $h = h_n = n^{-1/3}$ et, pour tout n, on dénote $A_n =]x_0 - h_n/2, x_0 + h_n/2]$ un intervalle de longueur h_n qui contient x_0. On suppose que

la densité f est une fonction Lipschitzienne de constante Lipschitz $L > 0$. Montrer que le risque quadratique est de l'ordre de $n^{-2/3}$. Et en déduire que

$$\hat{f}_n^H(x_0) \xrightarrow{P} f(x_0) \ , \quad n \to \infty \ .$$

Exercice 5. Statistiques d'ordre
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 définit les statistiques d'ordre $X_{(1)}, \ldots, X_{(n)}$ où les $X_{(i)}$ sont les valeurs X_i classées dans l'ordre croissant :

$$X_{(i)} \in \{X_1, \ldots, X_n\} \ , \quad X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \ .$$

1. Calculer les fonctions de répartition des variables aléatoires $X_{(n)}$ et $X_{(1)}$. Déterminer la fonction de répartition de $X_{(k)}$, notée $G_k(x)$.
2. Déterminer la densité jointe $g(x_1, ..., x_n)$ de $(X_{(1)}, \ldots, X_{(n)})$ ainsi que la densité marginale $g_k(x)$ de $X_{(k)}$.
3. Donner la loi du couple $(X_{(1)}, X_{(n)})$ et la loi de la statistique $W = X_{(n)} - X_{(1)}$ (appelée *étendue*). Les variables $X_{(1)}$ et $X_{(n)}$ sont–elles indépendantes ?
4. Définissons les variables aléatoires :

   $$Y_k = F(X_{(k)}) \quad \text{et} \quad Z_k = G_k(X_{(k)}).$$

   Déterminer la loi de Y_k et de Z_k (nous supposerons que F et G_k sont des fonctions strictement croissantes).

Chapitre 4

Estimation ponctuelle

4.1 Problème d'estimation

L'approche statistique consiste à introduire un *modèle probabiliste* pour les données. En d'autres terme, on suppose que l'observation $\mathbf{x}$ est la réalisation d'une variable (ou vecteur) aléatoire $\mathbf{X}$ de loi inconnue $\mathbb{P}$. Un des objectifs principaux de la théorie statistique consiste à *déterminer cette loi $\mathbb{P}$ ou des caractéristiques de $\mathbb{P}$.*

Grâce à la connaissance *a priori* du phénomène observé et à l'aide de la statistique descriptive, on parvient à choisir ce qu'on appelle un **modèle statistique** : la donnée d'une famille de lois de probabilité $\mathcal{P}$. Ensuite, on cherche à trouver la meilleure loi $\mathbb{P}^*$ dans $\mathcal{P}$ appropriée pour engendrer l'observation $\mathbf{x}$. Pour cela, il est courant et utile de paramétrer l'ensemble $\mathcal{P}$.

Paramétrisation du modèle

Pour décrire un modèle statistique $\mathcal{P}$, il est pratique de définir une paramétrisation, c'est-à-dire une application $\theta \mapsto \mathbb{P}_\theta$ définie de l'*ensemble de paramètres* Θ dans l'ensemble $\mathcal{P}$; nous écrirons $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$. Par exemple $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0\}$ est l'ensemble de lois normales avec $\theta = (\mu, \sigma^2)$ et $\Theta = \mathbb{R} \times \mathbb{R}_+$.

Remarquons qu'il existe en général de multiples manières de définir une paramétrisation. N'importe quelle transformation bijective sur Θ permet en particulier de définir une nouvelle paramétrisation. Par exemple, nous pourrions choisir de paramétrer la loi normale par $(\mu, \mu^2 + \sigma^2)$ plutôt que (μ, σ^2). La paramétrisation que nous choisissons est en général naturellement dictée par le phénomène que nous modélisons, bien que la paramétrisation qui semble la plus naturel ne soit pas toujours nécessairement celle qui se prête le mieux à l'analyse mathématique.

Problème d'estimation

Lorsque la loi $\mathbb{P}$ de l'observation $\mathbf{x}$ appartient à un modèle paramétré $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ on peut écrire $\mathbb{P} = \mathbb{P}_{\theta_0}$ pour un paramètre $\theta_0 \in \Theta$. On dit que θ_0 est la *vraie valeur du paramètre θ*. Le **problème d'estimation** de la loi $\mathbb{P}_{\theta_0}$ consiste alors à déterminer le paramètre θ_0 à partir de l'observation $\mathbf{x}$. Nous insistons sur le fait que le paramètre θ_0 est inconnu mais que l'ensemble Θ des paramètres possibles est connu, car la famille de lois

$\{\mathbb{P}_\theta, \theta \in \Theta\}$ est connue. La démarche statistique consiste à extraire de l'information sur le paramètre θ_0 – on parle d'inférence et donc de *statistique inférentielle* – en s'appuyant sur l'observation $\mathbf{x}$.

Identifiabilité du modèle

Pour que le problème d'estimation du paramètre θ_0 soit bien défini, il faut que la paramétrisation soit identifiable, c'est-à-dire qu'elle vérifie, pour tout $\theta_1, \theta_2 \in \Theta$,

$$\theta_1 \neq \theta_2 \quad \Rightarrow \quad \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2} \ .$$

Si cette hypothèse n'est pas vérifiée, deux valeurs différentes de θ peuvent donner des lois de probabilité identiques, auquel cas l'unicité de la vraie valeur du paramètre θ_0 est compromise. Dans la suite, nous supposerons toujours que la paramétrisation est identifiable, et on dit que $\mathcal{P}$ est un **modèle identifiable**.

Domination

Nous nous placerons souvent dans le cadre suivant. Nous dirons qu'un modèle statistique $\mathcal{P}$ est **dominé** s'il existe une mesure σ-finie μ sur $\mathcal{B}(\mathcal{X})$ telle que, pour tout $\mathbb{P} \in \mathcal{P}$, $\mathbb{P} \ll \mu$ (i.e. $\mu(A) = 0$ implique $\mathbb{P}(A) = 0$ pour tout $A \in \mathcal{B}(\mathcal{X})$).

La notion de modèle dominé sous-tend la notion de densité qui lui est associée par le théorème de Radon-Nikodym. En effet, ce dernier indique que tout modèle dominé s'écrit $\mathcal{P} = \{p\mu : p \in D\}$, où D est une famille de densités pour la mesure σ-finie μ, c'est-à-dire une famille de fonctions positives p définie μ-presque partout sur $\mathcal{X}$ telles que $\int p(x)\mu(\mathrm{d}x) = 1$. L'intérêt des modèles dominés est de pouvoir travailler directement sur une famille de densités au lieu d'une famille de mesures de probabilité. Nous noterons $p(\cdot, \theta)$ ou $p_\theta(\cdot)$ la densité de la loi $\mathbb{P}_\theta$ par rapport à une mesure de référence μ.

Nous considérons en particulier les modèles suivants.

1. Le modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$ est dominé par la mesure de Lebesgue sur $\mathbb{R}^n$,

$$\mathbb{P}_\theta(A) = \int_A p_\theta(x)\mathrm{d}x \ ,$$

   pour $A \in \mathcal{B}(\mathbb{R}^n)$. On parlera de *cas continu*, et nous noterons aussi $f_\theta, f(\cdot, \theta)$ ou f les densités par rapport à la mesure de Lebesgue.

2. L'espace $\mathcal{X}$ est discret, et le modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$ est dominé par la mesure de comptage sur $\mathcal{X}$,

$$\mathbb{P}_\theta(A) = \sum_{x \in A} p(x, \theta) \ .$$

   Il s'agit du *cas discret*.

Estimateurs

Comment estimer le paramètre θ_0 ? Ou une caractéristique $q(\theta_0)$ de la loi $\mathbb{P}_{\theta_0}$? Nous ne disposons que de l'observation $\mathbf{x}$, et la seule liberté que nous pouvons nous permettre pour estimer le paramètre θ_0 (ou une quantité $q(\theta_0)$) est de composer des fonctions appropriées de l'observation $\mathbf{x}$, ce qui nous amène à la notion fondamentale suivante : On appelle **statistique** toute fonction borélienne de l'observation $S = S(\mathbf{x}) = S(x_1, \ldots, x_n)$.

On voit immédiatement que cette définition n'est pas très contraignante : par exemple, l'échantillon $\mathbf{x} = (x_1, \ldots, x_n)$ est une statistique, la fonction $S(\mathbf{x}) = 0$ l'est aussi, mais ces deux statistiques sont sans intérêt, car elles ne nous approchent pas de la connaissance de caractéristiques de la loi $\mathbb{P}_{\theta_0}$ sous-jacente.

Une statistique est aussi appelée **estimateur** si elle est utilisée pour approcher le paramètre θ_0 (ou d'autres caractéristiques $q(\theta_0)$) de la loi de probabilité $\mathbb{P}_{\theta_0}$. Nous apprendrons rapidement à mesurer les performances des différents estimateurs, ce qui exclura les estimateurs "fantaisistes" comme $S(\mathbf{x}) = 0$.

Remarquons qu'il existe des méthodes d'estimation différentes en fonction de la taille de la famille de loi $\mathcal{P}$. En effet, s'il existe une paramétrisation $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ tel que $\Theta \subset \mathbb{R}^d$ avec $d < \infty$, le modèle est dit **paramétrique**, et le problème d'estimation du paramètre θ_0 se résume à estimer un *vecteur de dimension finie*. Dans le cas contraire, s'il est impossible de paramétrer $\mathcal{P}$ par un paramètre de dimension finie, le modèle est dit **non paramétrique**. Citons comme exemple pour le cas non paramétrique l'ensemble de lois qui admet une densité continue par rapport à la mesure de Lebesgue λ,

$$\mathcal{P} = \left\{ \mathbb{P} : f = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\lambda} \text{ existe et } f \text{ est continue} \right\} .$$

En effet, une spécification très précise de la structure du modèle $\mathcal{P}$, si les hypothèses sur lesquelles elle repose sont justifiées, permet en général de simplifier les procédures d'estimation des quantités inconnues du modèles. Le danger est que, si le modèle est mal spécifié, nos analyses, bien que correctes sur le plan mathématique, soient erronées pour interpréter les mesures considérées. Il est donc quelquefois nécessaire d'évaluer la robustesse des estimateurs à des variations du modèle ou de mettre en œuvre des procédures de sélection de modèle.

Nous présenterons maintenant quelques propriétés d'estimateurs grâce auxquelles on pourra évaluer et comparer des estimateurs, et éventuellement en choisir un pour l'utiliser dans la pratique.

4.2 Propriétés d'un estimateur

Rappelons qu'un estimateur $\hat{\theta}$ est une fonction associant une valeur $\hat{\theta}(\mathbf{x})$ à une observation $\mathbf{x}$ que l'on espère proche de la vraie valeur θ_0. Pour évaluer si l'approximation est bonne, on ne veut pas se limiter à étudier un cas particulier d'un échantillon $\mathbf{x}$ fixé. En revanche, on s'intéresse à la règle générale à partir laquelle est définie la statistique $\hat{\theta} = \hat{\theta}(\mathbf{x})$ pour une réalisation $\mathbf{x}$ *quelconque* de la loi $\mathbb{P}_{\theta_0}$. Donc, au lieu de vérifier si $\hat{\theta}(\mathbf{x})$ est près de θ_0 pour un jeu de données $\mathbf{x}$ fixé, on considère la *variable aléatoire* $\hat{\theta}(\mathbf{X})$ où $\mathbf{X}$ suit la loi $\mathbb{P}_{\theta_0}$ et on étudie la distance entre $\hat{\theta}(\mathbf{X})$ et θ_0. Pour distinguer $\hat{\theta}(\mathbf{x})$ de $\hat{\theta}(\mathbf{X})$, parfois on appelle $\hat{\theta}(\mathbf{X})$ l'**estimateur** et $\hat{\theta}(\mathbf{x})$ une **estimation**. Par ailleurs, la vraie valeur θ_0 étant inconnue, on s'intéresse au comportement de $\hat{\theta}(\mathbf{X})$ par rapport à θ_0 quelque soit la loi $\mathbb{P}_{\theta_0}$ de $\mathbf{X}$ *pour tout* $\theta_0 \in \Theta$.

Nous considérons dans ce chapitre le problème d'estimation dans un contexte paramétrique. Nous supposons donc que l'observation $\mathbf{x} = (x_1, \ldots, x_n)$ est la réalisation d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathbb{P}_{\theta_0}$ appartenant à une famille paramétrique de lois $\{\mathbb{P}_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^d$ et $d < \infty$. On va essentiellement traiter le cas d'un échantillon i.i.d., c'est-à-dire quand les X_i sont des variables aléatoires indépendantes et de même loi.

Par souci de clarté, nous parlerons toujours de l'estimation du paramètre θ_0. En revanche,

les propriétés présentées, et plus tard les méthodes développées, s'appliquent également à l'estimation d'une caractéristique $q(\theta_0)$ de la loi $\mathbb{P}_{\theta_0}$ (si $q(\theta_0)$ est de dimension finie) ou d'une partie du vecteur θ_0.

Nous présenterons ici trois propriétés d'estimateur : une propriété minimale pour un bon estimateur (la consistance), un critère pour comparer des estimateurs à n fini (le risque quadratique) et une propriété asymptotique (la loi limite et la vitesse de convergence).

4.2.1 Consistance

En général, et notamment dans le cas i.i.d., un estimateur $\hat{\theta}$ est bien défini quelque soit la taille d'échantillon n (voir p. ex. la moyenne empirique $\bar{x}_n$). Pour mettre en avant la dépendance de $\hat{\theta}$ de n, on note $\hat{\theta}_n$.

Intuitivement, il devrait être plus facile d'estimer le paramètre θ_0 si on dispose d'un grand échantillon $\mathbf{x}_n = (x_1, \ldots, x_n)$ (n grand) que si n est petit, car chaque observation x_i apporte de l'information sur la loi $\mathbb{P}_{\theta_0}$ et donc sur le paramètre à estimer θ_0. Or, lorsque l'échantillon $\mathbf{x}_n$ croît indéfiniment, c'est-à-dire quand n tend vers l'infini, on attend d'un estimateur "raisonnable" à ce que $\hat{\theta}_n(\mathbf{x}_n)$ converge vers θ_0.

Cette réflexion mène à la notion de la consistance d'un estimateur, propriété minimale que l'on exigera de tout estimateur. Mais avant de définir cette propriété asymptotique en toute rigueur mathématique, faisons la remarque suivante pour comprendre comment le nombre d'observations n peut intervenir dans la description d'un modèle.

Pour l'instant nous avons considéré une réalisation $\mathbf{x}$ d'un vecteur aléatoire $\mathbf{X}$ de loi $\mathbb{P}_{\theta_0}$ appartenant à un modèle $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$. Afin d'étudier des propriétés asymptotiques d'un estimateur $\hat{\theta}_n$, introduisons le nombre d'observation n dans les notations. Désormais, notons $\mathbb{P}_{\theta_0}$ la loi de la variable aléatoire X_i (et non la loi de tout le vecteur aléatoire $\mathbf{X}$). Dans le cas où les X_i sont i.i.d., la loi du vecteur aléatoire $\mathbf{X}_n = (X_1, \ldots, X_n)$ est donnée par la mesure produit $\mathbb{P}_{\theta_0,n} = \mathbb{P}_{\theta_0}^{\otimes n}$. Rappelons que $\mathbb{P}_{\theta_0}^{\otimes n}$ vérifie

$$\mathbb{P}_{\theta_0}^{\otimes n}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}_{\theta_0}(X_i \in A_i) \ , \quad \text{pour tout } A_i \in \mathcal{B}(\mathcal{X}) \ ,$$

où $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ désigne l'espace de probabilités sur lequel la variable aléatoire X_i est définie. De la même façon, on peut introduire, pour tout n, un modèle $\mathcal{P}_n = \{\mathbb{P}_\theta^{\otimes n}, \theta \in \Theta\}$ sur l'espace d'observations $(\mathcal{X}^n, \mathcal{B}(\mathcal{X}^n))$. Notons que dans cette suite de modèles $(\mathcal{P}_n)_n$ l'espace de paramètres Θ est le même pour tout n.

Or, quand on étudiera les propriétés asymptotiques d'un estimateur $\hat{\theta} = \hat{\theta}_n$ lorsque n tend vers l'infini, le mot *estimateur* désignera aussi, pour abréger, une suite d'estimateurs $(\hat{\theta}_n(\mathbf{X}_n))_{n\geq 1}$ ou bien la règle à partir de laquelle est définie la statistique $\hat{\theta}_n(\mathbf{X}_n)$ pour tout n donné.

Un estimateur $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n)$ de θ_0 est dit **convergent** ou **consistant** si

$$\hat{\theta}_n \xrightarrow{P} \theta_0 \ , \quad \text{pour tout } \theta_0 \in \Theta \ .$$

Dans cette définition, la convergence doit avoir lieu pour tout $\theta_0 \in \Theta$, ce qui garantit qu'elle a lieu pour la vraie valeur inconnue θ_0 des observations $\mathbf{x}_n$. La consistance est une propriété liée au modèle statistique : un estimateur $\hat{\theta}_n$ peut être consistant pour un modèle et non-consistant pour un autre.

Si l'on a la convergence presque sûre : $\hat{\theta}_n \to \theta_0$ *p.s.* au lieu de la convergence en probabilité, on dit que l'estimateur $\hat{\theta}_n$ est **fortement consistant**.

La consistance est une propriété assez faible. Cette notion 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 d'estimateurs que l'on doit étudier. En effet, les estimateurs non consistants doivent être avec certitude exclus de toute considération.

4.2.2 Risque quadratique

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** ou **erreur quadratique moyenne** de l'estimateur $\hat{\theta}$ au point $\theta \in \Theta$ la quantité

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[\|\hat{\theta} - \theta\|^2 \right] .$$

Le risque quadratique est bien défini pour tout estimateur $\hat{\theta}$. Il peut, en particulier, prendre la valeur $R(\theta, \hat{\theta}) = +\infty$. Le risque permet de mesurer la distance entre l'estimateur $\hat{\theta}$ et la valeur θ_0.

Théorème 15. *Si $R(\theta_0, \hat{\theta}_n) \to 0$ quand $n \to \infty$ pour tout $\theta_0 \in \Theta$, alors $\hat{\theta}_n$ est un estimateur consistant de θ_0.*

Démonstration. Le théorème découle directement de l'inégalité de Tchebychev (Théorème 6). □

Un calcul élémentaire montre que le risque $R(\theta, \hat{\theta})$ admet la décomposition suivante

$$\begin{aligned} R(\theta, \hat{\theta}) &= \mathbb{E}_\theta \left[\|\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}] + \mathbb{E}_\theta[\hat{\theta}] - \theta\|^2 \right] \\ &= \mathbb{E}_\theta \left[\|\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]\|^2 \right] + \|\mathbb{E}_\theta[\hat{\theta}] - \theta\|^2 + 2 \left\langle \mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_\theta[\hat{\theta}], \mathbb{E}_\theta[\hat{\theta}] - \theta \right\rangle \\ &= \left(\|\mathbb{E}_\theta[\hat{\theta}] - \theta\| \right)^2 + \mathbb{E}_\theta \left[\|\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]\|^2 \right] \\ &=: b^2(\theta, \hat{\theta}) + \sigma^2(\theta, \hat{\theta}) . \end{aligned}$$

Le terme $b^2(\theta, \hat{\theta})$ représente la partie déterministe de l'erreur d'estimation, alors que $\sigma^2(\theta, \hat{\theta})$ mesure la contribution de sa partie stochastique.

Si $\Theta \subset \mathbb{R}$, on appelle $b(\theta, \hat{\theta})$ **biais** de l'estimateur $\hat{\theta}$ et $\sigma^2(\theta, \hat{\theta})$ **variance** de $\hat{\theta}$, et on note aussi $\sigma^2(\theta, \hat{\theta}) = \mathbf{Var}_\theta(\hat{\theta})$.

On dit qu'un estimateur $\hat{\theta}$ est **sans biais** si $\mathbb{E}_{\theta_0}[\hat{\theta}] = \theta_0$ (i.e. $b(\theta_0, \hat{\theta}) = 0$) pour tout $\theta_0 \in \Theta$. Dans le cas contraire, on dit que $\hat{\theta}$ est **biaisé**.

Si $\mathbb{E}_{\theta_0}[\hat{\theta}_n] \to \theta_0$ lorsque $n \to \infty$, on dit que $\hat{\theta}_n$ est **asymptotiquement sans biais**.

Plus la valeur du risque est petite, plus l'estimateur $\hat{\theta}$ est performant. Afin de comparer deux estimateurs, on peut comparer leurs risques quadratiques. Soient $\hat{\theta}^{(1)}$ et $\hat{\theta}^{(2)}$ deux estimateurs de θ_0 dans le modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Si

$$R(\theta, \hat{\theta}^{(1)}) \leq R(\theta, \hat{\theta}^{(2)}) \quad \text{pour tout } \theta \in \Theta ,$$

et si, de plus, il existe $\theta' \in \Theta$ tel que l'inégalité est stricte, alors on dit que $\hat{\theta}^{(1)}$ est **plus efficace** que $\hat{\theta}^{(2)}$ (ou meilleur que $\hat{\theta}^{(2)}$) et que $\hat{\theta}^{(2)}$ est **inadmissible**.

4.2.3 Loi limite et vitesse de convergence

Pour tout estimateur consistant la suite $\hat{\theta}_n - \theta_0$ converge en probabilité vers 0. En revanche, la vitesse à laquelle cette convergence a lieu peut être différente d'un estimateur à l'autre. Il est donc intéressant de comparer des différents estimateurs par leur vitesse de convergence. Il est courant de considérer à cet égard plutôt la converge en loi, ce qui nous amène à introduire la notion de la loi limite d'un estimateur.

Soit $\hat{\theta}_n$ un estimateur consistant de θ_0. On appelle G_{θ_0} **loi limite** de l'estimateur $\hat{\theta}_n$, s'il existe une suite $(r_n)_{n\geq 1}$ déterministe positive telle que $r_n \to \infty$ et

$$r_n(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \eta \sim G_{\theta_0} \ , \quad \text{pour tout } \theta_0 \in \Theta \ , \tag{4.1}$$

lorsque $n \to \infty$. On appelle $1/r_n$ la **vitesse de convergence** de l'estimateur $\hat{\theta}_n$.

La vitesse de convergence $1/r_n$ n'est pas unique, car si une suite $(r_n)_n$ vérifie (4.1), alors pour tout $\lambda > 0$ la suite $(\lambda r_n)_n$ tend vers l'infini et elle vérifie

$$\lambda r_n(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \lambda\eta \ .$$

Typiquement, on donne la vitesse de convergence sous la forme $1/r_n = n^{-\gamma}$ avec $\gamma > 0$. Dans le contexte de l'estimation paramétrique, la vitesse de convergence la plus fréquente est $1/r_n = n^{-1/2}$.

La variance $\mathbf{Var}(\eta)$ de la loi limite G_{θ_0} (si elle existe) est dite **variance limite**. Elle peut dépendre du paramètre θ_0 ou pas.

Un cas particulier et récurrent est le cas où la loi limite est une loi normale.

Soient $\Theta \subset \mathbb{R}$ et $\hat{\theta}_n$ un estimateur consistant de θ_0. Si

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \eta \sim \mathcal{N}(0, \sigma^2_{\theta_0}) \ , \quad \text{pour tout } \theta_0 \in \Theta \ ,$$

où $0 < \sigma^2_{\theta_0} < \infty$, l'estimateur $\hat{\theta}_n$ est dit **asymptotiquement normal**.

Entre deux estimateurs $\hat{\theta}^{(1)}$ et $\hat{\theta}^{(2)}$ de θ_0, on préfère celui dont la vitesse de convergence est la plus petite. Si elles sont identiques, on compare leurs variances limites, et on préfère l'estimateur avec la plus petite variance limite.

4.3 Exercices

Exercice 1. Risque quadratique et consistance
Montrer la Proposition 15.

Exercice 2. Estimation de la variance
Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de variance σ^2. On pose

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

1. Montrer que s^2 est l'estimateur de σ^2 que l'on obtient par la méthode de substitution.
2. Le but est de calculer le risque quadratique de l'estimateur s^2 de σ^2. On suppose que les $(X_i)_{1\leq i\leq n}$ ont un moment fini d'ordre 4.
   2.1 Montrer que l'on peut supposer sans perte de généralité que les X_i sont centrées. On fera cette hypothèse dans la suite.

2.2 Démontrer que

$$s^2 = \frac{n-1}{n^2}\sum_{i=1}^{n} X_i^2 - \frac{2}{n^2}\sum_{k<l} X_k X_l \, .$$

2.3 En déduire que le biais vaut $b(s^2) = \mathbb{E}(s^2) - \sigma^2 = -\sigma^2/n$ et la variance

$$\mathbf{Var}(s^2) = \frac{n-1}{n^3}\left((n-1)\mathbb{E}[X_1^4] - (n-3)\sigma^4\right) \, .$$

On montrera d'abord que

$$\mathbf{Cov}\left(\sum_{i=1}^{n} X_i^2, \sum_{k<l} X_k X_l\right) = 0 \, , \qquad \mathbf{Var}\left(\sum_{k<l} X_k X_l\right) = n(n-1)\sigma^4/2 \, .$$

2.4 Calculer la valeur du risque quadratique de s^2. Et montrer que s^2 est un estimateur consistant de σ^2.

3. Introduisons un autre estimateur de σ^2 donné par

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

Supposons que les X_i suivent une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$. Dans ce cas, quel estimateur de σ^2 est préférable ?

Exercice 3. Ampoules défaillantes
Un statisticien observe n fois le nombre x_i d'ampoules défaillantes à la sortie d'une chaîne de fabrication. Il veut estimer la probabilité de n'avoir aucune ampoule défaillante.

Il considère les observations $x_1, \ldots, x_n$ comme des réalisations des variables aléatoires $X_1, \ldots, X_n$ i.i.d. à valeurs dans $\{0, 1, \ldots\}$. Le but consiste à estimer la probabilité $p = \mathbb{P}(X_1 = 0)$ de n'avoir aucune ampoule défaillante.

1. Dans un premier temps, le statisticien compte le nombre N_n de x_i, $i = 1, \ldots, n$, égaux à 0. Il propose d'estimer p par

$$\hat{p}_1 = \frac{N_n}{n} \, .$$

Montrer que l'estimateur $\hat{p}_1$ est obtenue par la méthode de substitution. Est-il un estimateur consistant de p et sans biais ? Calculer son risque quadratique, et donner sa loi limite.

Maintenant on introduit comme modèle statistique la famille de lois de Poisson. Plus précisément, on suppose en plus que $X_i \sim \mathcal{P}(\lambda)$ avec $\lambda > 0$ inconnu.

2. Le statisticien propose comme estimateur de p :

$$\hat{p}_2 = \mathrm{e}^{-\bar{X}_n},$$

où $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Expliquer sa démarche. Montrer que $\hat{p}_2$ est consistant. Calculer sa variance et son biais. Déterminer des équivalents asymptotiques des quantités précédentes. On pourra d'abord déterminer la loi de $\sum_{i=1}^{n} X_i$.

3. Montrer que l'on peut choisir t_n tel que $\hat{p}_3 = \mathrm{e}^{-t_n \bar{X}_n}$ soit sans biais.

4. Lequel de $\hat{p}_1$, $\hat{p}_2$ et $\hat{p}_3$ choisiriez-vous pour estimer p ?

Exercice 4. Modèle d'autorégression

On considère une suite de variables aléatoires $X_0, \ldots, X_n$, issues du *modèle d'autorégression* :

$$X_0 = 0,$$

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

où les $\{\varepsilon_i\}_{1 \leq i \leq n}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et $a \in \mathbb{R}$.

1. Écrire X_i en fonction de la série des perturbations $(\varepsilon_1, \ldots, \varepsilon_n)$. En déduire, selon les valeurs du paramètre a, la loi, l'espérance μ et la variance σ_i^2 de X_i.
2. Calculer le coefficient de corrélation entre X_i et X_{i+1} et montrer que celui ne dépend pas de σ^2. Les variables X_i et X_{i+1} sont-elles indépendantes ?
3. Donner la loi de X_{i+1} sachant $X_i, X_{i-1}, \ldots, X_1$. En déduire la densité de $(X_n, \ldots, X_1)$.
4. Calculer l'estimateur du maximum de vraisemblance de a et σ^2 à partir d'une réalisation $\mathbf{x} = (x_n, \ldots, x_1)$ du vecteur aléatoire $\mathbf{X} = (X_n, \ldots, X_1)$.

Chapitre 5

Méthodes d'estimation classiques

Dans ce chapitre nous présenterons des approches classiques d'estimation de paramètre, notamment la méthode de substitution, la méthode des moments et la méthode du maximum de vraisemblance.

5.1 Méthode de substitution

Dans ce paragraphe, nous supposons que l'on observe un échantillon i.i.d. $\mathbf{x} = (x_1, \dots, x_n)$, c'est-à-dire les x_i sont des réalisations indépendantes d'une variable aléatoire X de loi $\mathbb{P}_{\theta_0}$ appartenant à un modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$.

La méthode de substitution est une approche assez générale pour estimer le paramètre θ_0 ou plus généralement une caractéristique $q(\theta_0)$ de la loi $\mathbb{P}_{\theta_0}$. En fait, nous l'avons déjà utilisée au Chapitre 3 à plusieurs reprises pour estimer des caractéristiques de la loi des observations. En effet, la quantité d'intérêt $q(\theta_0)$ apparaît souvent comme une fonctionnelle de sa fonction de répartition F_{θ_0} :

$$q(\theta_0) = T(F_{\theta_0}) \ .$$

Puisque la fonction de répartition F_{θ_0} peut être approchée par la fonction de répartition empirique $\hat{F}$ définie en (3.1), il est naturel de considérer comme estimateur de $q(\theta_0)$ la statistique

$$S(\mathbf{x}) = T(\hat{F}) \ .$$

L'idée de construction de cet estimateur est appelée **méthode de substitution** ou **principe du plug-in** : on substitue $\hat{F}$ à F.

Prenons comme exemple le moment d'ordre r d'une variable aléatoire X de loi $\mathbb{P}_{\theta_0}$ donné par

$$\mathbb{E}[X^r] = \int x^r \mathrm{d}F_{\theta_0}(x) \ .$$

Selon le principe du *plug-in*, on obtient un estimateur de $\mathbb{E}[X^r]$ par la quantité

$$m_r = \int x^r \mathrm{d}\hat{F}(x) \ .$$

Rappelons que la loi empirique $\hat{F}$ est la loi discrète à valeurs dans $\{x_1, \dots, x_n\}$ qui associe

à chaque observations le poids $1/n$. On obtient alors

$$m_r = \int x^r \mathrm{d}\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^r \ . \tag{5.1}$$

Pour un échantillon $\mathbf{X}_n = (X_1, \ldots, X_n)$ grandissant, on a par la loi des grands nombres,

$$m_r = m_r(\mathbf{X}_n) \longrightarrow \mu_r^* \ p.s \ , \quad \text{lorsque } n \to \infty \ .$$

En effet, sous des hypothèses assez générales sur la fonctionnelle T,

$$T(\hat{F}_n) \longrightarrow T(F_{\theta_0}) \ p.s. \ , \quad \text{quand } n \to \infty \ ,$$

ce qui justifie l'application de la méthode de substitution.

Exemple : Groupes sanguins Rhésus

Il existe deux groupes sanguins Rhésus, le groupe sanguin + et -. Chaque personne est porteur de deux gènes qui sont soit de type +, soit de type -. Une personne qui est génétiquement +/+ est de groupe sanguin Rhésus +. Ainsi, une personne qui est génétiquement -/- est de groupe sanguin Rhésus -. Dans le cas mixte, c'est-à-dire une personne qui est génétiquement +/- est de groupe sanguin Rhésus +.

D'un point de vue technique il est plus facile de déterminer le groupe sanguin Rhésus d'une personne que le type des deux gènes correspondants. Or, nous nous intéressons à estimer la proportion p de gènes de type - dans une population en utilisant les groupes sanguins Rhésus de n personnes.

Pour cela, nous faisons certaines hypothèses : Les groupes sanguins des n personnes sont indépendants, ainsi que le type des deux gènes chez une personne. Introduisons le vecteur aléatoire $Z = (Z_1, Z_2)^T \in \{0,1\}^2$ pour le type des deux gènes d'une personne (où 0 correspond au type - et 1 au type 1). Notons $R = \max\{Z_1, Z_2\} \in \{0,1\}$ la variable aléatoire pour le groupe sanguin résultant.

Il est clair que les Z_k sont i.i.d. de loi Bernoulli de paramètre $1-p$. Ainsi, R suit également une loi Bernoulli de paramètre

$$\mathbb{P}(R=1) = 1 - \mathbb{P}(R=0) = 1 - \mathbb{P}(Z_1 = 0, Z_2 = 0) = 1 - p^2 \ .$$

Or, nous disposons de n observations indépendantes $r_1, \ldots, r_n$ de la variable aléatoire R pour estimer le paramètre p.

Afin d'appliquer la méthode de substitution, il faut exprimer le paramètre à estimer en fonction de la fonction de répartition des observations. En fait, on a

$$p = \sqrt{1 - \mathbb{P}(R=1)} = \sqrt{1 - \int_{\{1\}} \mathrm{d}F_R}$$

On obtient un estimateur de p en remplaçant F_R par la fonction de répartition empirique $\hat{F}$ associée aux observations $r_1, \ldots, r_n$. On trouve

$$\hat{p} = \sqrt{1 - \int_{\{1\}} \mathrm{d}\hat{F}} = \sqrt{1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{r_i = 1\}} = \sqrt{1 - \bar{r}_n} \ ,$$

où $\bar{r}_n$ dénote la moyenne empirique de $r_1, \ldots, r_n$.

On montre facilement que l'estimateur $\hat{p}$ de p est consistant et asymptotiquement normal.

5.2 Méthode des moments

La méthode des moments a été proposée par Karl Pearson en 1894. Elle repose sur la méthode de substitution.

Dans ce paragraphe, nous supposons que l'on observe un échantillon i.i.d. $\mathbf{x} = (x_1, \ldots, x_n)$, c'est-à-dire les x_i sont des réalisations indépendantes d'une variable aléatoire X de loi $\mathbb{P}_{\theta_0}$ appartenant à un modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^d$. De plus, nous supposons que le moment d'ordre d de X existe pour tout $\theta_0 \in \Theta$. Notons μ_r les fonctions des moments d'ordre r pour $r = 1, \ldots, d$ données par

$$\mu_r(\theta) = \mathbb{E}_\theta[X^r] = \int x^r \mathrm{d}F_\theta(x) \ , \quad r = 1, \ldots, d \ ,$$

où la notation $\mathbb{E}_\theta$ indique que la variable aléatoire X dans l'espérance suit la loi $\mathbb{P}_\theta$ de paramètre θ. Supposons que les fonctions μ_r sont connues explicitement.

Si les vraies valeurs $\mu_r^* = \mu_r(\theta_0)$, $r = 1, \ldots, d$, étaient disponibles, on pourrait résoudre le système de d équations

$$\mu_r(\theta) = \mu_r^* \ , \quad r = 1, \ldots, d \ ,$$

pour trouver la valeur du vecteur θ_0. Or, ces valeurs sont inconnues et nous disposons seulement d'un échantillon i.i.d. $\mathbf{x}$ de la loi $\mathbb{P}_{\theta_0}$. L'idée consiste à remplacer dans ce système d'équations les valeurs inconnues μ_r^* par les valeurs correspondantes m_r obtenues par la méthode de substitution en (5.1). Comme les m_r sont censés d'approcher les valeurs μ_r^*, on peut espérer qu'une solution par rapport à θ du système d'équations

$$\mu_r(\theta) = m_r \ , \quad r = 1, \ldots, d \ , \tag{5.2}$$

soit proche de θ_0. En fait, on appelle **estimateur par la méthode des moments (EMM)** du paramètre θ_0 dans le modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$ toute statistique $\hat{\theta}^{MM}$ à valeurs dans Θ solution du système de d équations (5.2). Autrement dit, $\hat{\theta}^{MM}$ vérifie

$$\mu_r(\hat{\theta}^{MM}) = m_r \ , \quad r = 1, \ldots, d \ .$$

Il est clair que l'EMM peut ne pas exister, et, s'il existe, il est possible qu'il ne soit pas unique.

Au lieu d'utiliser les d premiers moments pour construire un estimateur de $\theta_0 \in \mathbb{R}^d$, on peut utiliser d moments quelconques $\mu_{r_1}, \ldots, \mu_{r_d}$ (pourvu qu'ils soient finis). Ou encore, on peut utiliser des moments de la forme $\mathbb{E}_\theta[\varphi_r(X)]$ avec des fonction φ_r quelconques et intégrables. Plus précisément, on calcule les fonctions

$$\tilde{\mu}_r(\theta) = \mathbb{E}_\theta[\varphi_r(X)] \ , \quad r = 1, \ldots, d$$

ainsi que les équivalents empiriques

$$\tilde{m}_r = \frac{1}{n} \sum_{i=1}^{n} \varphi_r(x_i) \ , \quad r = 1, \ldots, d \ .$$

Ensuite, on cherche la valeur $\hat{\theta}^{MG} \in \Theta$ (si elle existe) qui vérifie

$$\tilde{\mu}_r(\hat{\theta}^{MG}) = \tilde{m}_r \ , \quad r = 1, \ldots, d \ .$$

On appelle $\hat{\theta}^{MG}$ l'estimateur par la **méthode des moments généralisée**.

Sous des conditions assez générales, cet estimateur (s'il existe) est consistant et asymptotiquement normal.

Pour montrer la consistance de l'estimateur de la méthode des moments (généralisée) dans un cas particulier, on utilise typiquement la loi forte des grands nombres et le premier théorème de continuité, car dans l'estimateur intervient des sommes de variables aléatoires i.i.d.. La loi limite est généralement obtenue par le théorème central limite en combinaison avec le deuxième théorème de continuité.

Exemple : Loi uniforme

1. Soit $\mathbf{x} = (x_1, \ldots, x_n)$ un échantillon i.i.d. de loi uniforme $U[0, \theta]$ de densité de probabilité $f_\theta = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}$ avec $\theta > 0$. Dans ce modèle statistique, l'ensemble de paramètres Θ est $]0, \infty] \subset \mathbb{R}$ et donc $d = 1$. On a pour $X \sim U[0, \theta]$

$$\mu_1(\theta) = \mathbb{E}_\theta[X] = \int x f_\theta(x) \mathrm{d}x = \frac{1}{\theta}\int_0^\theta x \mathrm{d}x = \frac{\theta}{2} \, .$$

   La méthode des moments suggère de chercher la solution de

$$\mu_1(\theta) = \bar{x}_n \quad \Longleftrightarrow \quad \theta = 2\bar{x}_n \, ,$$

   où $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ désigne la moyenne empirique. L'estimateur par la méthode des moments de θ est donc $\hat{\theta}^{MM} = 2\bar{x}_n$. En effet, on vérifie que $\hat{\theta}^{MM} \in \Theta$ avec probabilité 1, car dans ce modèle toutes les observations x_i sont positives presque sûrement.

   La consistance de l'estimateur est triviale. En plus, l'EMM est asymptotiquement normal par le théorème central limite.

2. Considérons maintenant le modèle statistique des lois uniformes sur $U[-\theta, \theta]$ de densité $f_\theta(x) = \frac{1}{2\theta}\mathbb{1}_{[-\theta,\theta]}$ avec encore $\Theta =]0, \infty]$. Cette fois, on a pour $X \sim U[-\theta, \theta]$,

$$\mu_1(\theta) = \mathbb{E}_\theta[X] = \int x f_\theta(x) \mathrm{d}x = 0 \, , \quad \text{pour tout } \theta > 0 \, .$$

   On remarque que l'équation $\mu_1(\theta) = \bar{x}_n \Leftrightarrow 0 = \bar{x}_n$ n'a pas de solution. Donc, dans ce cas, l'EMM n'existe pas.

   En revanche, on peut essayer avec le moment d'ordre 2. On calcule

$$\mu_2(\theta) = \mathbb{E}_\theta[X^2] = \int x^2 f_\theta(x) \mathrm{d}x = \frac{1}{2\theta}\int_{-\theta}^\theta x^2 \mathrm{d}x = \frac{\theta^2}{3} \, .$$

   L'EMM basé sur le moment d'ordre 2 est donc $\hat{\theta}^{MM} = \sqrt{\frac{3}{n}\sum_{i=1}^n x_i^2}$.

   Alternativement, on peut estimer θ par la méthode des moment généralisée. Par exemple, on peut choisir la fonction $\varphi(x) = |x|$. On obtient alors

$$\tilde{\mu}(\theta) = \mathbb{E}_\theta[|X|] = \int |x| f_\theta(x) \mathrm{d}x = 2\int_0^\theta \frac{x}{2\theta} \mathrm{d}x = \frac{\theta}{2} \, ,$$

   ce qui implique l'estimateur $\hat{\theta}^{MG} = \frac{2}{n}\sum_{i=1}^n |x_i|$.

5.3 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).

Intuition

Pour comprendre l'intuition de la méthode du maximum de vraisemblance (MV) considérons le problème d'une pièce de monnaie et la question si cette pièce est équilibrée ou pas. Autrement dit, on veut savoir, en jouant à pile ou face, quelle est la probabilité d'obtenir 'pile'. On modélise les sorties d'une suite de n lancées $\mathbf{x} = (x_1, \dots, x_n)$ par une expérience Bernoulli, i.e. $\mathbf{x}$ est la réalisation d'un vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ de v.a. X_i i.i.d. de loi Bernoulli de paramètre $p \in (0, 1)$ (et on identifie l'événement 'pile' avec 1, et 'face' avec 0). Si la pièce de monnaie est équilibrée p vaut $1/2$.

L'approche de MV consiste à étudier la fonction de vraisemblance $\mathcal{L}(\mathbf{x}; p)$ définie par

$$p \mapsto \mathcal{L}(\mathbf{x}; p) = \mathbb{P}_p(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \ ,$$

où la notation $\mathbb{P}_p$ signifie que les v.a. X_i suivent la loi Bernoulli de paramètre p.

Notons que $\mathcal{L}(\mathbf{x}; p)$ est la probabilité d'obtenir la suite $\mathbf{x}$ lorsque la vraie valeur du paramètre de la loi de Bernoulli est p. La méthode de MV cherche la valeur de p qui maximise cette probabilité. Autrement dit, on cherche le paramètre qui rend cette suite d'observations le plus vraisemblable.

Soyons encore plus concret. Supposons que l'on observe $\mathbf{x} = (1, 1, 1, 1, 0, 1, 1, 1)$. Clairement, avec une pièce équilibrée, il est très peu probable d'obtenir la suite observée (même s'il n'est pas impossible). En revanche, avec une pièce fortement déséquilibrée, il serait bien plus vraisemblable d'observer ces valeurs. En fait, pour cet exemple, on a $\mathcal{L}(\mathbf{x}; p) = p^7(1 - p)$ et on montre facilement que cette fonction est maximale en $p = 7/8$. Donc, c'est avec une pièce de monnaie dont la probabilité d'obtenir pile est de $7/8$, qu'on a la plus de chance d'obtenir la suite $\mathbf{x} = (1, 1, 1, 1, 0, 1, 1, 1)$. La méthode de MV propose de considérer $\hat{p}^{MV} = 7/8$ comme estimateur de p.

La méthode en général

Considérons un modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$, où $\Theta \subset \mathbb{R}^d$ et un échantillon $\mathbf{x} = (x_1, \dots, x_n)$ de loi $\mathbb{P}_{\theta_0}$. Supposons que le modèle soit dominée par une mesure μ, et notons $p_\theta = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}$ la densité de $\mathbb{P}_\theta$ par rapport à μ. Définissons la **fonction de vraisemblance** de $\mathbf{x}$ par

$$\theta \mapsto \mathcal{L}(\mathbf{x}; \theta) = p_\theta(\mathbf{x}) \ .$$

Si $\mathbf{x}$ est un échantillon i.i.d. de densité p_{θ_0}, alors $\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n p_\theta(x_i)$. Dans le cas i.i.d. continu, on a $\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n f_\theta(x_i)$, et dans le cas i.i.d. discret, on a $\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i)$.

On appelle **estimateur du maximum de vraisemblance (EMV)** du paramètre θ_0 dans le modèle statistique $\{P_\theta, \theta \in \Theta\}$ toute statistique $\hat{\theta}^{MV} \in \Theta$ telle que

$$\mathcal{L}(\mathbf{x}; \hat{\theta}^{MV}) = \max_{\theta \in \Theta} \mathcal{L}(\mathbf{x}; \theta) \ . \tag{5.3}$$

Autrement dit, $\hat{\theta}^{MV}$ est tel que

$$\hat{\theta}^{MV} = \arg\max_{\theta \in \Theta} \mathcal{L}(\mathbf{x}; \theta) \ .$$

L'EMV peut ne pas exister, car le problème de maximisation (5.3) n'admet pas de solution (dans Θ), et si un EMV existe, il est possible qu'il ne soit pas unique.

Si le support des densités $x \mapsto f_\theta(x)$ ne dépend pas de θ (c'est-à-dire l'ensemble $\{x : f_\theta(x) > 0\}$ est le même pour tout $\theta \in \Theta$), on définit la **fonction de log-vraisemblance** $\ell(\theta)$ par

$$\ell(\theta) = \log(\mathcal{L}(\mathbf{x}; \theta)) \ .$$

Remarquons que

$$\hat{\theta}^{MV} = \arg\max_{\theta \in \Theta} \ell(\theta) \ .$$

Par conséquent, au lieu de maximiser la fonction de vraisemblance $\mathcal{L}(\mathbf{x}; \theta)$, on peut maximiser la fonction de log-vraisemblance $\ell(\theta)$ pour trouver l'EMV, ce qui s'avère en général beaucoup plus facile.

Si le maximum de $\mathcal{L}(\mathbf{x}; \theta)$ (ou de $\ell(\theta)$) n'est pas atteint sur la frontière de Θ et si l'application $\theta \mapsto \mathcal{L}(\mathbf{x}; \theta)$ est différentiable, une condition nécessaire de maximum est l'annulation du gradient :

$$\nabla_\theta \mathcal{L}(\mathbf{x}; \theta)|_{\theta = \hat{\theta}^{MV}} = 0 \ , \tag{5.4}$$

ce qui représente un système de d équations, car $\theta \in \mathbb{R}^d$. De façon similaire, une condition nécessaire de maximum de la fonction de log-vraisemblance est

$$\nabla \ell(\theta) = 0 \ . \tag{5.5}$$

On appelle (5.5) l'**équation de vraisemblance** si $\theta \in \mathbb{R}$ et **système des équations de vraisemblance** si $\theta \in \mathbb{R}^d, d > 1$.

On appelle **racine de l'équation de vraisemblance** (REV) dans le modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$, avec $\Theta \in \mathbb{R}^d$, toute statistique $\hat{\theta}^{RV}$ à valeurs dans Θ solution du système de d équations (5.5). Autrement dit,

$$\nabla \ell(\hat{\theta}^{RV}) = 0 \ .$$

Notons qu'en résolvant le système (5.5) on obtient tous les maxima et tous les minima locaux de $\ell(\cdot)$, 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 tous les EMV soient des REV et vice versa, il faut essentiellement que la fonction $\ell(\cdot)$ atteint son minimum global pour tous les θ tels que $\nabla \ell(\theta) = 0$. Cette condition est très restrictive : on ne peut effectivement la vérifier que si la fonction ℓ est convexe et son minimum global n'est pas atteint sur la frontière de Θ. L'équivalence des EMV et des REV n'a donc lieu que dans une situation très particulière. Il s'agit essentiellement de deux estimateurs différents, sauf cas exceptionnel.

Dans des nombreux cas, on vérifie que l'EMV (s'il existe) est consistant, en revanche, sa loi limite n'est pas nécessairement une loi normale.

La théorie statistique fournit de nombreux résultats concernant l'EMV. En particulier, sous des conditions de régularité du modèle statistique on peut montrer que l'EMV est efficace, c'est-à-dire qu'il est optimal dans un sens précis. La présentation de ces résultats fait l'objet du cours *Statistique mathématique*. Cependant, ce cours de *Statistique appliquée* apporte un regard approfondi sur des aspects pratiques du calcul de l'EMV dans des modèles pertinents dans des applications.

Exemple : Loi uniforme

1. Soit $\mathbf{x} = (x_1, \ldots, x_n)$ un échantillon i.i.d. de loi uniforme $U[0, \theta]$ de densité de probabilité $f_\theta(x) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$ avec $\theta > 0$. La fonction de vraisemblance s'écrit

$$\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]}(x_{(1)})\mathbb{1}_{[0,\theta]}(x_{(n)}) = \frac{1}{\theta^n}\mathbb{1}\{\theta \geq x_{(n)}\} .$$

   En fait, il est clair que $X_i > 0$ *p.s.* pour $X_i \sim U[0, \theta]$ pour tout $\theta \in \Theta =]0, \infty[$, et donc $X_{(1)} > 0$ *p.s.* et $X_{(n)} > 0$ *p.s.*.
   On constate que, d'une part, la fonction $\theta \mapsto \mathbb{1}\{\theta \geq x_{(n)}\}$ est maximale sur $[x_{(n)}, \infty]$, et nulle ailleurs. D'autre part, la fonction $\theta \mapsto \frac{1}{\theta^n}$ est strictement décroissante. Par conséquence, la fonction de vraisemblance $\theta \mapsto \mathcal{L}(\mathbf{x}; \theta)$ est maximale en $\theta = x_{(n)}$. L'EMV de θ est donc $\hat{\theta}^{MV} = x_{(n)}$.
   On voit que l'EMV est unique et presque sûrement bien défini, c'est-à-dire $\hat{\theta}^{MV} = X_{(n)} \in \Theta$ *p.s.*.

2. Supposons maintenant que les observations $\mathbf{x} = (x_1, \ldots, x_n)$ sont n réalisations indépendantes d'une variable aléatoire X de loi uniforme $U[\theta, \theta + 1]$ de paramètre $\theta \in \mathbb{R}$ inconnu.
   La densité de $X \sim U[\theta, \theta + 1]$ est donnée par $f_\theta(x) = \mathbb{1}_{[\theta,\theta+1]}(x)$. D'où

$$\begin{aligned}\mathcal{L}(\mathbf{x}; \theta) &= \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \mathbb{1}_{[\theta,\theta+1]}(x_i) \\ &= \mathbb{1}\left\{x_{(1)} \geq \theta, x_{(n)} \leq \theta + 1\right\} = \mathbb{1}\left\{x_{(n)} - 1 \leq \theta \leq x_{(1)}\right\} .\end{aligned}$$

   On voit que la fonction de vraisemblance $\theta \mapsto \mathcal{L}(\mathbf{x}; \theta)$ ne prend que deux valeurs : 0 et 1. On en déduit que tout point de l'intervalle $[x_{(n)} - 1, x_{(1)}]$ maximise la fonction de vraisemblance $\mathcal{L}(\mathbf{x}; \theta)$. Par conséquence, tout point de l'intervalle $[x_{(n)} - 1, x_{(1)}]$ est EMV de θ. Dans ce cas, on n'a pas d'unicité de l'EMV.

5.4 Optimisation d'une fonction

Afin de déterminer l'EMV il faut savoir résoudre des problèmes d'optimisation. Plus précisément, il est question de déterminer le point $\hat{\theta}$ où la fonction de (log-)vraisemblance atteint son maximum global (s'il existe).

Dans ce paragraphe, nous rappelons d'abord les techniques d'optimisation classiques d'analyse. En revanche, en pratique il est rare que l'on peut exhiber la formule explicite de l'estimateur du maximum de vraisemblance. En effet, dans très peu de cas seulement, il est possible de calculer l'EMV explicitement. Le plus souvent, et notamment dans des modèles pertinents pour la pratique, le problème de maximisation (5.3) n'admet pas de solution explicite. Par conséquent, il est nécessaire de recourir à des méthodes numériques. Une des méthodes numériques les plus répandues en statistique est la méthode de Newton-Raphson, qui sera présentée à la fin de cette section.

5.4.1 Rappel : Techniques d'optimisation classiques

Rappelons qu'il existe des fonctions non bornées, qui n'ont donc pas de maximum global. Il existe alors des problèmes de maximisation qui n'admettent pas de solution.

Fonction concave

Soit $f : I \subset \mathbb{R} \to \mathbb{R}$ une fonction deux fois dérivable définie sur un intervalle I. Si

$$f''(x) < 0 \,, \quad \text{pour tout } x \in I \,,$$

alors f est strictement concave. Par conséquence, f admet un maximum global et celui est unique. Pour le trouver, il suffit de chercher la solution de $f'(x) = 0$. Si $f'(x) = 0$ n'a pas de solutions dans I, le maximum se trouve aux bords de l'intervalle I.

Fonction deux fois dérivable

Soit $f : I \subset \mathbb{R} \to \mathbb{R}$ une fonction deux fois dérivable, mais pas nécessairement concave. Alors on calcule tous les points critiques de la fonction f, i.e. toutes les solutions de $f'(x) = 0$, et on étudie le comportement de f aux bords de l'intervalle I.

Si un point critique x^* est tel que $f''(x^*) < 0$, alors x^* est un maximum local.

Si un point critique x^* vérifie $f''(x^*) > 0$, il s'agit d'un minimum local.

Si un point critique x^* est tel que $f''(x^*) = 0$, il peut s'agir d'un point d'inflexion ou d'un maximum ou d'un minimum. Dans ce cas, on peut étudier le comportement de la dérivée f' dans un voisinage V de x^*. Si

$$f'(x) > 0 \,, \quad \forall x < x_0 \text{ et } x \in V \qquad \text{et} \qquad f'(x) < 0 \,, \quad \forall x > x_0 \text{ et } x \in V \,, \tag{5.6}$$

alors x^* est bien un maximum (local ou global).

Une fois qu'on a déterminé tous les maxima locaux, on détermine le maximum global en comparant les valeurs de f en ses maxima locaux et en prenant en compte le comportement de f aux bords de l'intervalle I. Rappelons qu'il est possible que f tend vers l'infini aux bords de I. Dans ce cas, f n'admet pas de maximum global.

Fonction dérivable

Si f n'est dérivable qu'une fois, on détermine tous les points critiques de f, i.e. toutes les solutions de $f'(x) = 0$. Puis on détermine les maxima locaux en vérifiant (5.6). Enfin, afin de trouver le maximum global (et afin de voir s'il existe), on compare les valeurs de f en ces maxima et on prend en compte le comportement de f aux bords de l'intervalle I.

Fonction non dérivable

Si f n'est pas dérivable, on peut établir le tableau de variation de la fonction et/ou tracer l'allure de la fonction pour trouver le maximum global de f (s'il existe).

Fonction de plusieurs variables

Soit $f : D \subset \mathbb{R}^d \to \mathbb{R}$ une fonction deux fois dérivable définie sur D. Si la matrice hessienne

$$H(x) = \nabla^2 f(x) < 0 \,, \quad \text{pour tout } x \in D \,,$$

alors f est concave et la solution de $\nabla f(x) = 0$ (si elle existe) est le maximum global de f.

Si f est deux fois dérivable mais pas concave, alors un point x_0 tel que

$$\nabla f(x_0) = 0 \quad \text{et} \quad H(x_0) < 0 \,,$$

est un maximum local de f.

Rappelons une propriété de l'algèbre sur les matrice symétrique $A = (a_{i,j})_{1 \leq i,j \leq r}$ de taille $r \times r$. Notons $A_s = (a_{i,j})_{1 \leq i,j \leq s}$. On appelle *s-ième mineur principal dominant* de A le déterminant de A_s, $\det(A_s)$. La matrice A est définie négative, notée $A < 0$, si et seulement si tous les mineurs principaux dominants avec s pair sont strictement positifs, et tous les mineurs principaux dominants avec s impair sont négatifs.

Maximisation sous contrainte

La *méthode des multiplicateurs de Lagrange* permet de trouver les points stationnaires (maximum, minimum...) d'une fonction dérivable d'une ou plusieurs variables, sous contraintes. Soient $f : D \subset \mathbb{R}^d \to \mathbb{R}$ une fonction dérivable définie sur D et $\psi = (\psi_1, \ldots, \psi_p)^T : D \to \mathbb{R}^p$ une fonction dérivable qui exprime les contraintes. Le problème à résoudre est de trouver le maximum suivant :

$$\max_{x \in G} f(x) \quad \text{avec} \quad G = \{x \in D : \psi(x) = 0\} \,.$$

On introduit les multiplicateurs de Lagrange $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ et la fonction de Lagrange

$$L(x, \lambda) = f(x) + \sum_{k=1}^{p} \lambda_k \psi_k(x) \,.$$

Les points critiques de la fonction de Lagrange $(x, \lambda) \mapsto L(x, \lambda)$ (pour $x \in D$) donnent les points critiques de f. Plus précisément, on cherche les points (x, λ) tels que

$$\frac{\partial L(x,\lambda)}{\partial x_i} = 0 \,, \quad i = 1, \ldots, d \quad \text{et} \quad \frac{\partial L(x,\lambda)}{\partial \lambda_k} = 0 \,, \quad k = 1, \ldots, p \,.$$

Il reste à vérifier si un tel point critique est vraiment un maximum de f.

5.4.2 Méthode de Newton-Raphson

La méthode de Newton-Raphson est une procédure numérique pour déterminer un point critique d'une fonction f. En appliquant cette méthode à la fonction de log-vraisemblance $\ell(\theta)$, on peut espérer de trouver son maximum et donc le l'estimateur de maximum de vraisemblance. Bien évidemment, trouver un point critique n'est pas équivalent à déterminer le point du maximum global d'une fonction, car ce point peut être un maximum local seulement, un minimum ou un point d'inflexion. Néanmoins, dans des nombreux cas, cette méthode donne des résultats satisfaisants pour détecter le point maximum d'une fonction.

La méthode de Newton-Raphson

La méthode de Newton-Raphson est une procédure itérative pour trouver des points critiques d'une fonction réelle $f : \mathcal{X} \to \mathbb{R}$ où $\mathcal{X} \subset \mathbb{R}^d$. On suppose que f soit deux fois dérivable. Elle repose sur le développement de Taylor, plus précisément sur l'approximation linéaire du gradient $\nabla f(x)$ par

$$\nabla f(x) = \nabla f(\xi) + H(\xi)(x - \xi) + r(x, \xi) \,, \tag{5.7}$$

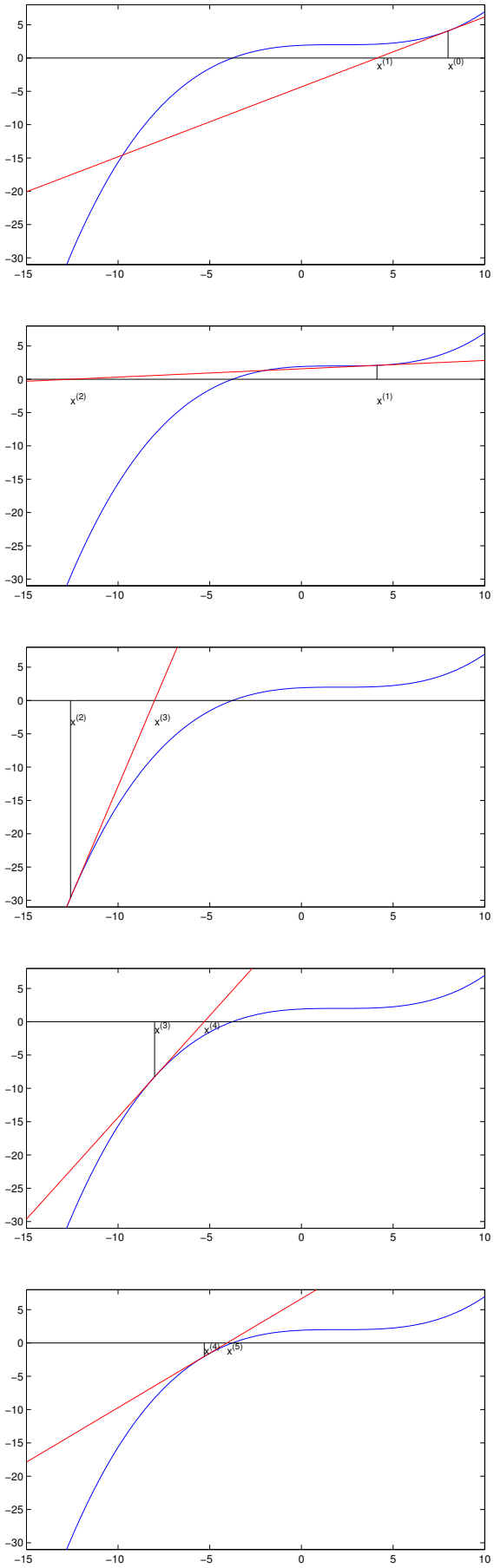


FIGURE 5.1 – Illustration de 5 itérations de la méthode de Newton-Raphson.

où $\xi \in \mathcal{X}$, $H(\xi) = \nabla^2 f(\xi)$ est la matrice hessienne de f en ξ et $r(x, \xi)$ est un terme de reste.

Si ξ est près de x, le reste r est négligeable comparé au terme linéaire. Au lieu de résoudre $\nabla f(x) = 0$ directement, la méthode de Newton-Raphson consiste à négliger le terme r en (5.7) et résoudre

$$\nabla f(\xi) + H(\xi)(x - \xi) = 0 \tag{5.8}$$

par rapport à x. En fait, on procède itérativement : on se donne un point initial $\xi = x^{(0)}$, puis on résout l'équation (5.8) par rapport à x et on appelle la solution $x^{(1)}$. Ensuite, on pose $\xi = x^{(1)}$ et on recommence. Plus généralement, l'itération t consiste à calculer

$$x^{(t)} = x^{(t-1)} - \left[H(x^{(t-1)})\right]^{-1} \nabla f(x^{(t-1)}) \ , \tag{5.9}$$

où $x^{(t-1)}$ est le résultat de l'itération précédente.

Pour que l'algorithme soit bien défini, il est nécessaire que l'inverse $\left[H(x^{(t)})\right]^{-1}$ existe pour tout t.

Interprétation géométrique

Pour une interprétation géométrique de la méthode de Newton-Raphson remarquons que le terme à gauche de l'équation (5.8) est la tangente à $\nabla f(x)$ au point $x = \xi$. Au lieu de chercher le point où le gradient $\nabla f(x)$ s'annule, on cherche donc le zéro de la tangente. Étant donné que la tangente est une fonction linéaire, il est beaucoup plus facile de déterminer ce point que de trouver un zéro de $\nabla f(x)$.

Figure 5.1 illustre les cinq premières étapes de la méthode de Newton-Raphson dans un exemple. La suite des $x^{(t)}$ avec point initial $x^{(0)} = 8$ est la suivante :

$x^{(0)} = 8$ $\quad \nabla f(x^{(0)}) = 4{,}07$
$x^{(1)} = 4{,}12$ $\quad \nabla f(x^{(1)}) = 2{,}08$
$x^{(2)} = -12{,}49$ $\quad \nabla f(x^{(2)}) = -29{,}61$
$x^{(3)} = -8{,}01$ $\quad \nabla f(x^{(3)}) = -8{,}28$
$x^{(4)} = -5{,}31$ $\quad \nabla f(x^{(4)}) = -2{,}02$
$x^{(5)} = -4{,}06$ $\quad \nabla f(x^{(5)}) = -0{,}32$
$x^{(6)} = -3{,}7800$ $\quad \nabla f(x^{(6)}) = -0{,}0145$
$x^{(7)} = -3{,}7659$ $\quad \nabla f(x^{(7)}) = -3{,}48\ 10^{-5}$
$x^{(8)} = -3{,}7659$ $\quad \nabla f(x^{(8)}) = -2{,}02\ 10^{-10}$
$x^{(9)} = -3{,}7659$ $\quad \nabla f(x^{(9)}) = 4{,}44\ 10^{-16}$

On observe que l'algorithme converge après quelques itérations seulement. En effet, à toute itération (sauf la deuxième), on s'approche du zéro de la fonction $\nabla f(x)$.

Critères d'arrêt

Plusieurs critères d'arrêt sont envisageable pour cet algorithme. Les deux plus courants sont les suivants. Soit $\varepsilon > 0$ un seuil fixé.

— On arrête dès que $|x^{(t)} - x^{(t-1)}| < \varepsilon$.
— On arrête dès que $|\nabla f(x^{(t)})| < \varepsilon$.

Quelque soit le critère d'arrêt, il est possible que la condition est vérifiée dans des points qui ne correspondent pas à des zéros de $\nabla f(x)$.

En général, il est difficile de garantir que la suite $(x^{(t)})_t$ converge. En revanche, et c'est ce qui rend cette méthode attractive, *si* elle converge, elle converge assez vite comme dans l'exemple ci-dessus.

Vitesse de convergence

Supposons que x^* est une solution de $\nabla f(x) = 0$ et que $\|x^{(t)} - x^*\|$ est petit. Alors par (5.9)

$$\begin{aligned} x^{(t+1)} - x^* &= x^{(t)} - x^* - \left[H(x^{(t)})\right]^{-1} \nabla f(x^{(t)}) \\ &= x^{(t)} - x^* - \left[H(x^{(t)})\right]^{-1} (\nabla f(x^{(t)}) - \nabla f(x^*)) \ , \end{aligned}$$

car $\nabla f(x^*) = 0$. Or, par le développement limité (5.7) on a

$$\nabla f(x^*) = \nabla f(x^{(t)}) + H(x^{(t)})(x^* - x^{(t)}) + r(x^*, x^{(t)}) \ .$$

D'où

$$\begin{aligned} x^{(t+1)} - x^* &= x^{(t)} - x^* + \left[H(x^{(t)})\right]^{-1} \left\{H(x^{(t)})(x^* - x^{(t)}) + r(x^*, x^{(t)})\right\} \\ &= \left[H(x^{(t)})\right]^{-1} r(x^*, x^{(t)}) \ . \end{aligned}$$

D'après le théorème de Taylor, $r(x^*, x^{(t)}) \leq c\|x^{(t)} - x^*\|^2$, où c dénote le maximum de $\|\nabla^3 f(x)\|$ sur le rectangle dont les extrémités sont déterminées par $x^{(t)}$ et x^*. Plus précisément, dans le cas unidimensionnel où $x \in \mathbb{R}$, ce rectangle est l'intervalle $[x^{(t)}, x^*]$ si $x^{(t)} < x^*$, ou bien $[x^*, x^{(t)}]$ si $x^* < x^{(t)}$.

On obtient alors

$$\|x^{(t+1)} - x^*\| \leq c \left\| \left[H(x^{(t)})\right]^{-1} \right\| \|x^* - x^{(t)}\|^2 \ .$$

Ceci montre qu'en une itération l'erreur diminue de façon quadratique : on passe de $\|x^* - x^{(t)}\|$ à un terme d'ordre $\|x^* - x^{(t)}\|^2$. On dit que la vitesse de convergence de la méthode de Newton-Raphson est quadratique.

Exemple : Loi de Cauchy

Considérons la loi de Cauchy centrée ($\mu = 0$) de paramètre $\sigma > 0$ inconnu dont la densité est donnée par

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

Soient $\mathbf{x} = (x_1, \ldots, x_n)$ des réalisations i.i.d. d'une variable aléatoire X de loi de Cauchy$(0, \sigma)$. Essayons de calculer l'EMV de σ.

La fonction de vraisemblance s'écrit

$$\mathcal{L}(\mathbf{x}; \sigma) = \prod_{i=1}^n f_\sigma(x_i) = \frac{\sigma^n}{\pi^n \prod_{i=1}^n (\sigma^2 + x_i^2)} \ .$$

On peut considérer la log-vraisemblance

$$\ell(\sigma) = \log \mathcal{L}(\mathbf{x}; \sigma) = n \log \sigma - n \log \pi - \sum_{i=1}^n \log(\sigma^2 + x_i^2) \ .$$

On dérive

$$\ell'(\sigma) = \frac{n}{\sigma} - 2\sum_{i=1}^{n} \frac{\sigma}{\sigma^2 + x_i^2} \, .$$

Résoudre l'équation $\ell'(\sigma) = 0$ est équivalent à trouver les zéros d'un polynôme d'ordre $2n$. Ce n'est pas faisable par un calcul explicite, et donc l'EMV n'est pas explicite dans ce modèle.

En revanche, on peut appliquer la méthode de Newton-Raphson pour trouver des points critiques de la fonction de log-vraisemblance $\ell(\sigma)$ afin d'approcher l'EMV numériquement. Pour cela, on calcule la dérivée seconde de $\ell(\sigma)$

$$\ell''(\sigma) = -\frac{n}{\sigma^2} - 2\sum_{i=1}^{n} \frac{\sigma^2 + x_i^2 - 2\sigma^2}{(\sigma^2 + x_i^2)^2} = -\frac{n}{\sigma^2} - 2\sum_{i=1}^{n} \frac{x_i^2 - \sigma^2}{(\sigma^2 + x_i^2)^2}$$

Or, la méthode de Newton-Raphson consiste à calculer iterativement (pour un point initial $\sigma^{(0)}$ choisi par l'utilisateur) pour $t = 0, 1, \ldots$

$$\begin{aligned}
\sigma^{(t+1)} &= \sigma^{(t)} - \frac{\ell'(\sigma^{(t)})}{\ell''(\sigma^{(t)})} \\
&= \sigma^{(t)} + \frac{\frac{n}{\sigma^{(t)}} - 2\sum_{i=1}^{n} \frac{\sigma^{(t)}}{(\sigma^{(t)})^2 + x_i^2}}{\frac{n}{(\sigma^{(t)})^2} + 2\sum_{i=1}^{n} \frac{x_i^2 - (\sigma^{(t)})^2}{[(\sigma^{(t)})^2 + x_i^2]^2}} \\
&= \frac{n\sigma^{(t)} - 2(\sigma^{(t)})^5 \sum_{i=1}^{n} \frac{1}{[(\sigma^{(t)})^2 + x_i^2]^2}}{\frac{n}{2} + (\sigma^{(t)})^2 \sum_{i=1}^{n} \frac{x_i^2 - (\sigma^{(t)})^2}{[(\sigma^{(t)})^2 + x_i^2]^2}} \, .
\end{aligned}$$

Cet exemple sera mise en œuvre et étudié en pratique en TP.

5.5 Exercices

Exercice 1. Groupes sanguins Rhésus

1. Montrer que l'estimateur $\hat{p}$ de p par la méthode de substitution est consistant et asymptotiquement normal.
2. Calculer les estimateurs de p par la méthode des moments et par la méthode du maximum de vraisemblance pour l'exemple des groupes sanguins Rhésus.

Exercice 2. EMM
Calculer l'estimateur $\hat{\theta}^{MM}$ par la méthode des moments pour des observations $\mathbf{x} = (x_1, \ldots, x_n)$ que l'on considère comme n réalisations indépendantes d'une variable aléatoire X de loi uniforme $U[\theta, \theta + 1]$ de paramètre $\theta \in \mathbb{R}$ inconnu. Cet estimateur, donne-t-il presque sûrement des valeurs raisonnables au vu des données $\mathbf{x}$? Comparer à l'EMV.

Exercice 3. Chaîne de montage
Une chaîne de montage produit des objets dont on veut estimer la durée moyenne de fabrication. On suppose que les durées de fabrication T_i sont indépendantes et de loi exponentielle de paramètre λ (c'est à dire de loi $\Gamma(1, \lambda)$. Le n-ieme objet est donc fabriqué à la date $T_1 + \cdots + T_n$, et on observe le nombre d'objets N_t fabriqués à une unique date $t > 0$.

1. Montrer que $\mathbb{P}(N_t \leq n) = \mathbb{P}(T_1 + \cdots + T_{n+1} > t)$.

2. Quelle est la loi de $T_1 + \cdots + T_n$? 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 λ 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.

Chapitre 6

Modèle de mélange et algorithme EM

Dans ce chapitre, nous présenterons un des modèles le plus utilisé en pratique : le modèle de mélange. Nous monterons son utilité et son importance pour la pratique, avant de présenter un algorithme qui permet d'approcher l'estimateur de maximum de vraisemblance dans des modèles de mélange et même dans un groupe de modèles plus large (les modèles à variables latentes). Mais tout d'abord, nous introduisons la notion de la loi conditionnelle, qui est un outil nécessaire pour manipuler des modèles de mélange.

6.1 Loi conditionnelle

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

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

Si A et B sont des événements indépendants, on a

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

Le théorème de Bayes dit que

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \ ,$$

si $\mathbb{P}(A) > 0$ et $\mathbb{P}(B) > 0$.

Soit $B_1, B_2, \ldots$ une partition de $\mathcal{A}$ telle que $\mathbb{P}(B_i) > 0$ pour tout i. Alors, d'après la **formule des probabilités totales** on a

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i) \ .$$

Soient $\mathbf{X} \in \mathbb{R}^p$ et $\mathbf{Y} \in \mathbb{R}^q$ des vecteurs aléatoires. Notons par $p_{(\mathbf{X},\mathbf{Y})}$ la densité jointe du vecteur aléatoire $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{p+q}$ par rapport à une mesure $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$ sur $\mathbb{R}^{p+q}$. Notons $p_{\mathbf{X}}$ et $p_{\mathbf{Y}}$ les densités marginales de $\mathbf{X}$ et de $\mathbf{Y}$ données par

$$p_{\mathbf{X}}(x) = \int_{\mathbb{R}^q} p_{(\mathbf{X},\mathbf{Y})}(x, y)\mu_Y(\mathrm{d}y) \quad \text{et} \quad p_{\mathbf{Y}}(y) = \int_{\mathbb{R}^p} p_{(\mathbf{X},\mathbf{Y})}(x, y)\mu_X(\mathrm{d}x) \ .$$

Pour $x \in \mathbb{R}^p$ fixée, on définit la fonction $y \mapsto p_{\mathbf{Y}|\mathbf{X}}(y|x)$ comme

$$p_{\mathbf{Y}|\mathbf{X}}(y|x) = \begin{cases} \frac{p_{(\mathbf{X},\mathbf{Y})}(x,y)}{p_{\mathbf{X}}(x)} , & \text{si } p_{\mathbf{X}}(x) > 0 , \\ p_{\mathbf{Y}}(y) , & \text{sinon.} \end{cases}$$

On remarque que $p_{\mathbf{Y}|\mathbf{X}}(\cdot|x)$ est une densité de probabilité par rapport à la mesure μ_Y pour tout $x \in \mathbb{R}^p$, car

$$p_{\mathbf{Y}|\mathbf{X}}(y|x) \geq 0 \, , \forall y \in \mathbb{R}^q \quad \text{et} \quad \int_{\mathbb{R}^q} p_{\mathbf{Y}|\mathbf{X}}(y|x)\mu_Y(\mathrm{d}y) = 1 \, .$$

On appelle $p_{\mathbf{Y}|\mathbf{X}}(\cdot|x)$ la **densité conditionnelle de Y sachant que X** $= x$.

La **loi conditionnelle de** Y **sachant que** $X = x$ est donnée par la formule

$$\mathbb{P}(\mathbf{Y} \in A|\mathbf{X} = x) = \int_A p_{\mathbf{Y}|\mathbf{X}}(y|x)\mu_Y(\mathrm{d}y) \, , \quad A \in \mathcal{B}, x \in \mathbb{R} \, .$$

On définit l'**espérance conditionnelle de Y sachant que X** $= x$ par

$$\mathbb{E}[\mathbf{Y}|\mathbf{X} = x] = \int_{\mathbb{R}^q}^{\infty} y p_{\mathbf{Y}|\mathbf{X}}(y|x)\mu_Y(\mathrm{d}y) \, .$$

La condition $\mathbb{E}[\|\mathbf{Y}\|] < \infty$ est suffisante pour assurer l'existence de l'espérance conditionnelle $\mathbb{E}[\mathbf{Y}|\mathbf{X} = x]$ pour tout x.

On peut également définir $F_{\mathbf{Y}|\mathbf{X}}(\cdot|x)$, la **fonction de répartition conditionnelle de Y sachant que X** $= x$: c'est la f.d.r. qui correspond à la mesure de probabilité $\mathbb{P}(\mathbf{Y} \in \cdot|\mathbf{X} = x)$. Elle est donnée par

$$F_{\mathbf{Y}|\mathbf{X}}(y|x) = \int_{-\infty}^{y} p_{\mathbf{Y}|\mathbf{X}}(t|x)\mu_Y(\mathrm{d}t) \, .$$

Dans le cas discret, quand $\mathbf{X}$ et $\mathbf{Y}$ sont deux vecteurs aléatoires discrets à valeurs dans $\mathcal{V} = \{v_1, v_2, \ldots\}$ resp. $\mathcal{W} = \{w_1, w_2, \ldots\}$, la loi conditionnelle de $\mathbf{Y}$ sachant que $\mathbf{X} = v$, pour $v \in \mathcal{V}$ fixé, est donnée par les probabilités

$$\mathbb{P}(\mathbf{Y} = w_k|\mathbf{X} = v) = \frac{\mathbb{P}(\mathbf{Y} = w_k, \mathbf{X} = v)}{\mathbb{P}(\mathbf{X} = v)} \, , \quad \forall k \geq 1 \, .$$

Dans le cas continu, où $\mathbf{X}$ et $\mathbf{Y}$ sont deux vecteurs aléatoires de densité jointe $f_{(\mathbf{X},\mathbf{Y})}(x, y)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^{p+q}$, la densité conditionnelle $f_{\mathbf{Y}|\mathbf{X}}$ de $\mathbf{Y}$ sachant $\mathbf{X}$ est donnée par

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

où $f_{\mathbf{X}}$ et $f_{\mathbf{Y}}$ dénotent les densités marginales de $\mathbf{X}$ et $\mathbf{Y}$.

6.2 Modèle de mélange

Le modèle de mélange est très fréquemment utilisé dans des applications. Il permet de modéliser le comportement de plusieurs groupes ou populations à la fois.

TABLE 6.1 – Longueurs des ailes en mm de 381 passereaux.

Longueur	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	98
Fréquence	5	3	12	36	55	45	21	13	15	34	59	48	16	12	6	1

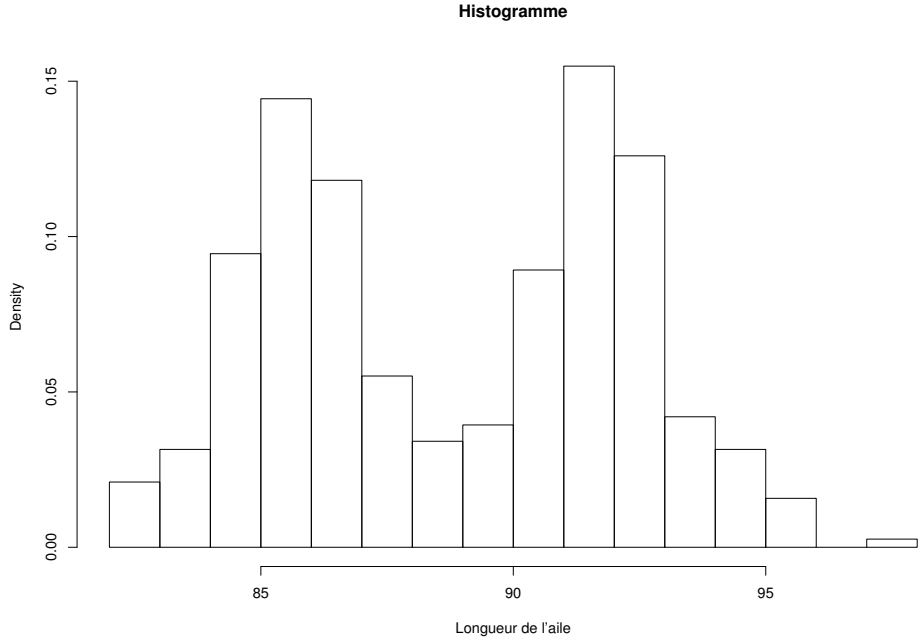


FIGURE 6.1 – Histogramme des longueurs des ailes.

6.2.1 Exemple : Longueurs des ailes de passereaux

Les données suivantes proviennent d'une étude sur la migration des passereaux. Pour ne pas perturber les oiseaux, seules quelques mesures rapides sont effectuées. La Table 6.1 reporte les mesures de la longueur d'une aile (en mm) de 381 passereaux et la Figure 6.1 représente les données graphiquement. La forme bimodale de l'histogramme laisse penser à la présence de deux populations différentes dans l'échantillon. En effet, lors de cette étude, il est clair que mâles et femelles ont été mélangés, cette variable n'ayant pu être mesurée.

Pour modéliser une telle situation où on observe deux populations de comportement différent, il convient d'associer une loi à chaque population. Autrement dit, on introduit une loi $\mathbb{P}_F$ pour les longueurs d'aile des femelles et une loi $\mathbb{P}_M$ pour les ailes des mâles. Vu la forme bimodale de l'histogramme, on pourrait choisir des lois normales pour $\mathbb{P}_F$ et $\mathbb{P}_M$ (de paramètres différents).

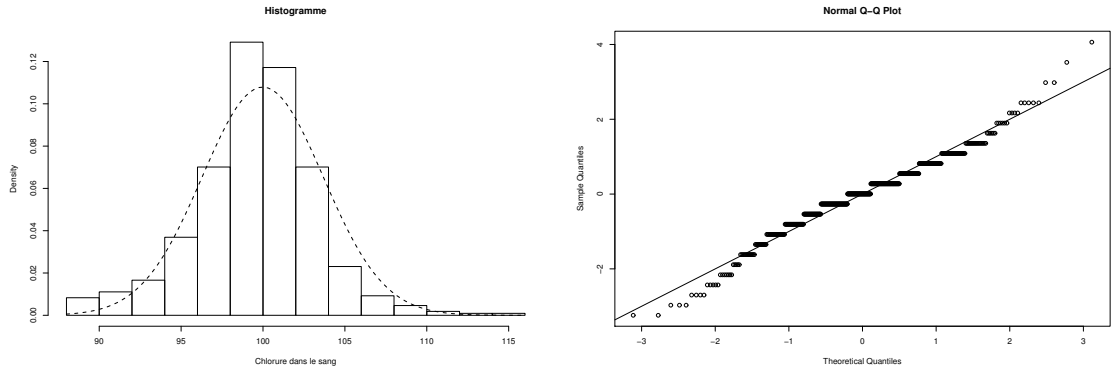
Si on avait noté le sexe de chaque oiseau, on pourrait faire deux sous-échantillons pour estimer les paramètres des lois $\mathbb{P}_F$ et $\mathbb{P}_M$ séparément. Malheureusement, ce n'est pas le cas. Par conséquent, il faut intégrer dans le modèle le fait que l'on ignore si la i-ème observation x_i est la longueur d'aile d'une femelle ou d'un mâle.

6.2.2 Exemple : Niveau de chlorure dans le sang

Une autre étude porte sur le niveau de chlorure dans le sang chez des adultes. Nous disposons d'un échantillon de taille 542 (cf. Table 6.2). La Figure 6.2 (a) montre l'histogramme qui est de forme unimodale et relativement symétrique. La densité d'une loi normale su-

TABLE 6.2 – Niveau de chlorure dans le sang (mmol/L) pour 542 individus.

Niveau	88	89	90	91	92	93	94	95	96	97	98	99	100
Fréquence	2	3	4	5	7	5	13	13	27	36	40	72	68
Niveau	101	102	103	104	105	106	107	108	109	111	113	115	
Fréquence	80	47	43	33	19	6	6	4	5	2	1	1	



(a) (b)

FIGURE 6.2 – (a) Histogramme des données de chlorure dans le sang et la densité de la loi normale $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ où $\hat{\mu}$ et $\hat{\sigma}^2$ sont l'EMV. (b) QQ-plot des données standardisées en comparaison à la loi normale standard.

perposée à l'histogramme est en bonne adéquation, bien que pas parfait. En revanche, le QQ-plot (Figure 6.2 (b)) qui compare les données (standardisées) à la loi normale standard est nettement moins favorable à l'hypothèse d'une simple loi normale.

En fait, ici, il est plus réaliste de supposer que la population observée contient une grande proportion π d'individus en bonne santé et une petite proportion $(1-\pi)$ d'individus malades. On peut faire l'hypothèse que les individus sains suivent une loi normale $\mathcal{N}(\mu_1, \sigma_1^2)$ alors que les individus malades suivent une loi normale $\mathcal{N}(\mu_2, \sigma_2^2)$ avec à peu près même moyenne ($\mu_1 \approx \mu_2$) mais avec une plus grande variance que les personnes en bonne santé ($\sigma_1^2 < \sigma_2^2$). Nous soulignons que nous ignorons qui sont les personnes malades dans cette étude. Il est alors nécessaire d'utiliser un modèle qui modélise à la fois le comportement des personnes malades comme des personnes en bonne santé.

6.2.3 DÉFINITION DU MODÈLE DE MÉLANGE

Les deux exemples précédents appartiennent à la famille des modèles de mélange, qui est la modélisation du comportement de plusieurs populations différentes. La définition d'une population, classe ou groupe dépend de l'application : Parfois il est justifié de distinguer le comportement des femmes et des hommes, parfois non. On pourrait également former des groupes par tranche d'âge, milieu social, nationalité, antécédents médicaux etc.

Soit $m \geq 2$ le nombre de sous-populations différentes dont nous cherchons à modéliser le comportement commun. Notons $\mathbb{P}_j$ la loi associée à la j-ème classe. Pour simplifier les notations, nous supposons que toutes les lois $\mathbb{P}_j$ appartiennent à une même famille de lois $\mathcal{H} = \{h_\phi, \phi \in \Phi\}$ où $\Phi \subset \mathbb{R}^d$ et h_ϕ désignent des densités par rapport à une mesure de références μ. On notera $\phi_j \in \Phi$ le paramètre de la loi $\mathbb{P}_j$, autrement dit, $\mathbb{P}_j$ est la loi de densité h_{ϕ_j}. En plus, notons π_j la proportion d'individus de la j-ème classe dans la population totale. Supposons que $\pi_j \in [0,1]$ pour tout $j = 1, \ldots, m$ et $\sum_{j=1}^m \pi_j = 1$.

Pour définir une variable aléatoire X qui représente m populations différentes, il convient d'introduire d'abord une variable aléatoire U pour modéliser l'appartenance d'un individu à une des m populations. Plus précisément, la loi de U est discrète à valeurs dans $\{1, \ldots, m\}$ avec

$$\mathbb{P}(U = j) = \pi_j \ , \quad j = 1, \ldots, m \ .$$

Par ailleurs, on introduit des variables aléatoires V_j de densité h_{ϕ_j} avec $\phi_j \in \Phi$ pour $j = 1, \ldots, m$. On suppose que les variables aléatoires $U, V_1, \ldots, V_m$ sont mutuellement indépendantes. Enfin, on définit la variable aléatoire X par

$$X = \sum_{j=1}^{m} \mathbb{1}\{U = j\} V_j \ .$$

Dans les deux exemples précédents, on peut considérer les données $\mathbf{x} = (x_1, \ldots, x_n)$ comme des réalisations i.i.d. d'une telle variable aléatoire X.

Calculons la loi de X. Par la formule des probabilités totales et l'indépendance des variables V_j et U, on a

$$\begin{aligned} F_X(x) = \mathbb{P}(X \leq x) &= \sum_{k=1}^{m} \mathbb{P}(X \leq x | U = k) \mathbb{P}(U = k) \\ &= \sum_{k=1}^{m} \pi_k \mathbb{P}\left(\left(\sum_{j=1}^{m} \mathbb{1}\{U = j\} V_j \right) \leq x \middle| U = k \right) \\ &= \sum_{k=1}^{m} \pi_k \mathbb{P}(V_k \leq x | U = k) = \sum_{k=1}^{m} \pi_k \mathbb{P}(V_k \leq x) = \sum_{k=1}^{m} \pi_k F_{V_k}(x) \ . \end{aligned}$$

Comme les lois des V_j admettent des densités par rapport à une mesure μ, on en déduit que la loi de X admet également une densité p_X par rapport à μ. En dérivant F_X on obtient pour la densité

$$p_X(x) = \sum_{j=1}^{m} \pi_j h_{\phi_j}(x) \ .$$

La densité p_X est dite **densité de mélange**. On appelle h_{ϕ_j} le **j-ième composant du mélange** et π_j son **poids**. Notons que, si $\mathcal{H} = \{h_\phi, \phi \in \Phi\}$ est une famille de lois continues (ou discrètes), p_θ est également continue (ou discrète).

Les paramètres du modèle de mélange sont, d'une part, les paramètres $\phi_1, \ldots, \phi_m \in \Phi$ des différentes composantes du mélange, et d'autre part, les probabilités discrètes $\pi_1, \ldots, \pi_m$ de la loi de U. Puisque $\sum_{j=1}^{m} \pi_j = 1$, la valeur de π_m est déterminée par les valeurs de $\pi_1, \ldots, \pi_{m-1}$. Il en résulte que le vecteur de paramètres θ d'un modèle de mélange comme décrit ci-dessus est donné par

$$\theta = (\phi_1, \ldots, \phi_m, \pi_1, \ldots, \pi_{m-1}) \ .$$

Si les ϕ_j sont des nombres réels, le modèle de mélange contient alors $2m - 1$ paramètres inconnus.

Le nombre m de populations ou classes est appelé **ordre du mélange**. Dans ce cours, on suppose que l'ordre m du modèle de mélange soit connu. Cependant, ce n'est pas toujours le cas en pratique. En effet, il existe toute une théorie et des méthodes diverses pour sélectionner l'ordre m d'un mélange, mais cela dépasse le cadre de ce cours.

Pour simuler des réalisations d'une variable X dans un modèle de mélange, on peut utiliser la construction de X par des variables aléatoires $U, V_1, \ldots, V_m$.

TABLE 6.3 – Composition des 12 mélanges gaussiens.

(a)	Densité normale standard	$\mathcal{N}(0,1)$
(b)	Densité unimodale dissymétrique	$\frac{1}{5}\mathcal{N}(0,1) + \frac{1}{5}\mathcal{N}(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}\mathcal{N}(\frac{13}{15},(\frac{5}{9})^2)$
(c)	Densité fortement dissymétrique	$\sum_{k=0}^{7}\frac{1}{8}\mathcal{N}(3((\frac{2}{3})^k-1),(\frac{2}{3})^{2k})$
(d)	Densité unimodale leptocurtique	$\frac{2}{3}\mathcal{N}(0,1) + \frac{1}{3}\mathcal{N}(0,(\frac{1}{10})^2)$
(e)	Densité avec outlier	$\frac{1}{10}\mathcal{N}(0,1) + \frac{9}{10}\mathcal{N}(0,(\frac{1}{10})^2)$
(f)	Densité bimodale	$\frac{1}{2}\mathcal{N}(-1,(\frac{2}{3})^2) + \frac{1}{2}\mathcal{N}(1,(\frac{2}{3})^2)$
(g)	Densité bimodale séparée	$\frac{3}{4}\mathcal{N}(-\frac{3}{2},(\frac{1}{2})^2) + \frac{1}{4}\mathcal{N}(\frac{3}{2},(\frac{1}{2})^2)$
(h)	Densité bimodale asymétrique	$\frac{3}{4}\mathcal{N}(0,1) + \frac{1}{4}\mathcal{N}(\frac{3}{2},(\frac{1}{3})^2)$
(i)	Densité trimodale	$\frac{9}{20}\mathcal{N}(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}\mathcal{N}(\frac{6}{5},(\frac{3}{5})^2) + \frac{1}{10}\mathcal{N}(0,(\frac{1}{4})^2)$
(j)	Griffe	$\frac{49}{100}\mathcal{N}(-1,(\frac{2}{3})^2) + \frac{49}{100}\mathcal{N}(1,(\frac{2}{3})^2) + \sum_{k=0}^{6}\frac{1}{350}\mathcal{N}((k-3)/2,(\frac{1}{100})^2)$
(k)	Griffe asymétrique	$\frac{1}{2}\mathcal{N}(0,1) + \sum_{k=-2}^{2}\frac{2^{1-k}}{31}\mathcal{N}(k+\frac{1}{2},(2^{-k}/10)^2)$
(l)	Peigne	$\sum_{k=0}^{5}\frac{2^{5-k}}{63}\mathcal{N}((65-96(\frac{1}{2})^k)/21,(\frac{32}{63})^2/2^{2k})$

Notons que le modèle de mélange est adéquat quand on ne dispose pas de l'information sur l'appartenance de groupe de chaque individu observé, c'est-à-dire quand la variable aléatoire U, dite **étiquette**, n'est pas observée. Cette manque d'information est parfois dû à un oubli pendant l'acquisition des données. Parfois il est impossible (ou très cher) d'obtenir cette information. Ou encore, avant l'étude, on n'est pas conscient que la variable observée s'explique le mieux en utilisant plusieurs populations.

Au-delà de l'utilisation des modèles de mélange pour modéliser le comportement de plusieurs sous-populations, les modèles de mélange définissent des nouvelles classes de lois de probabilité qui sont très grandes et parfois très intéressantes en elles-mêmes. En d'autres termes, on peut utiliser un modèle de mélange même si le phénomène observé ne permet pas de parler de groupes ou sous-populations. Dans ce cas, un modèle de mélange est juste utilisé pour approcher la loi du phénomène observé, quand les familles de lois classiques ne sont pas appropriées.

Pour illustrer la diversité des mélanges de lois normales, nous en donnons quelques exemples : Les mélanges présentés dans la Table 6.3 sont représentés graphiquement dans les Figures 6.3 et 6.4. Le nombre de classes varie entre 2 et 9. Ces figures nous montrent la variété de densités qu'il est possible d'obtenir à partir de mélanges gaussiens.

En fait, il est possible d'approcher au sens de la norme L_1 toute densité continue à l'aide d'un mélange gaussien.

Théorème 16. *Soit g une densité continue. Pour tout $\varepsilon > 0$ il existe un mélange gaussien de densité $\bar{g}$ donnée par*

$$\bar{g}(x) = \sum_{j=1}^{m} \pi_j f_{\mathcal{N}(\mu_j,\sigma_j^2)}(x) \ ,$$

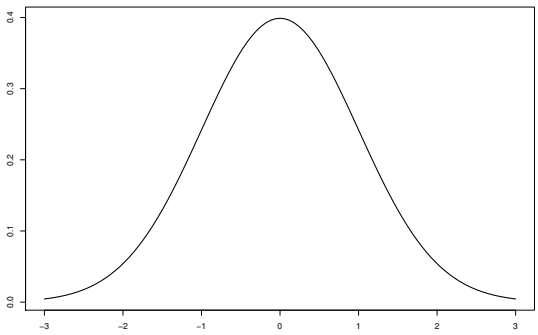
avec $\mu_j \in \mathbb{R}$, $\sigma_j > 0$, $\pi_j > 0$ tel que $\sum_{j=1}^{m}\pi_j = 1$, tel que

$$\|g - \bar{g}\|_1 := \int_{\mathbb{R}} |g(x) - \bar{g}(x)| \mathrm{d}x < \varepsilon \ .$$

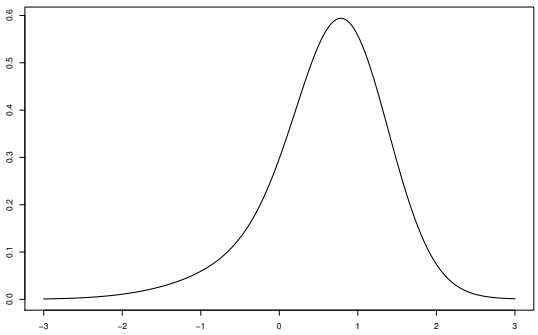
La démonstration de ce théorème repose sur le résultat suivant (que nous n'avons pas le temps de démontrer ici).

Lemme 1. *Soit g et f des densités continues. Notons*

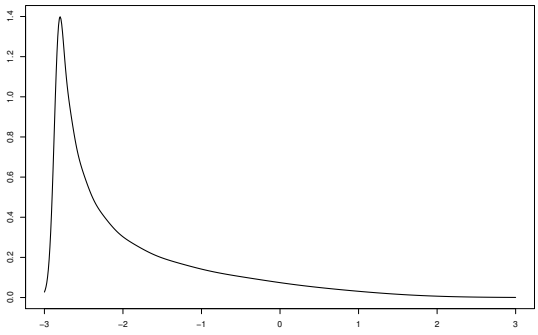
$$\mathcal{H} = \left\{ h_{(\mu,\sigma)}(x) := \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right), \mu \in \mathbb{R}, \sigma > 0 \right\}$$



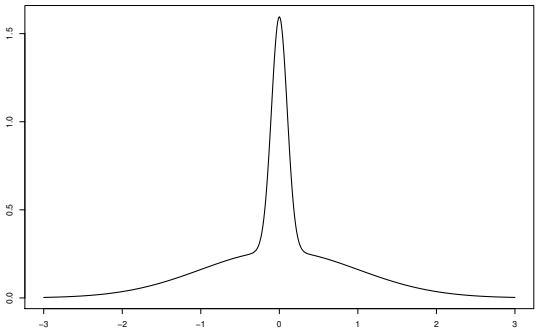
(a) Densité normale standard



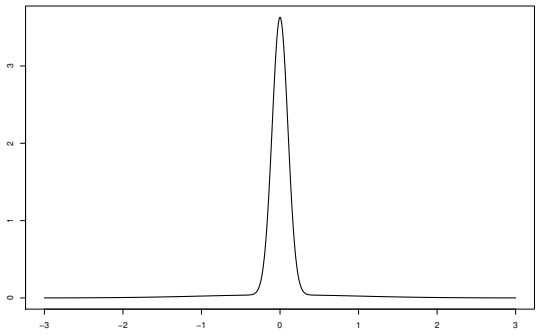
(b) Densité unimodale dissymétrique



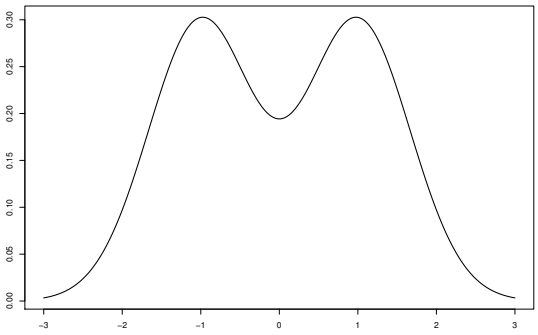
(c) Densité fortement dissymétrique



(d) Densité unimodale leptocurtique

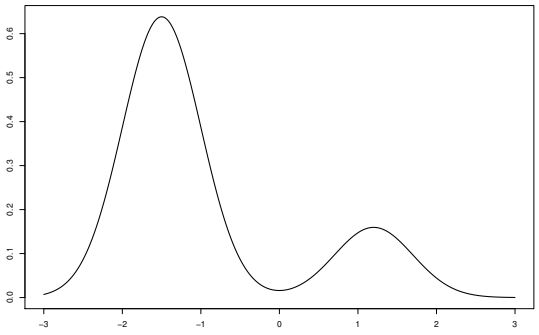


(e) Densité avec outlier

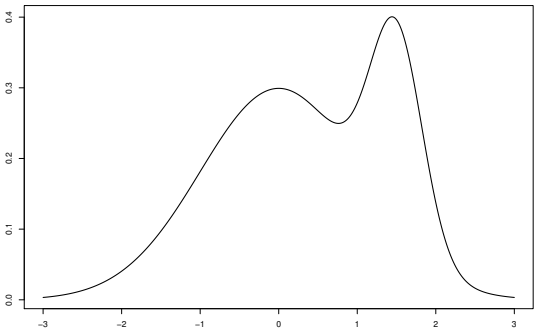


(f) Densité bimodale

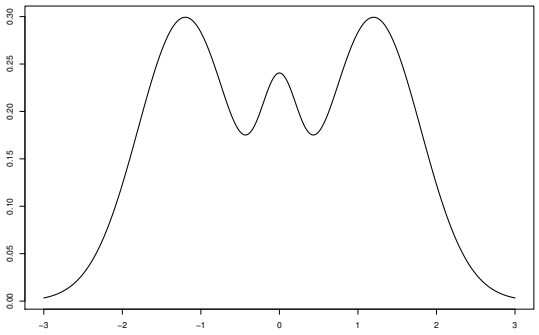
FIGURE 6.3 – Graphes de densités de mélanges gaussiens.



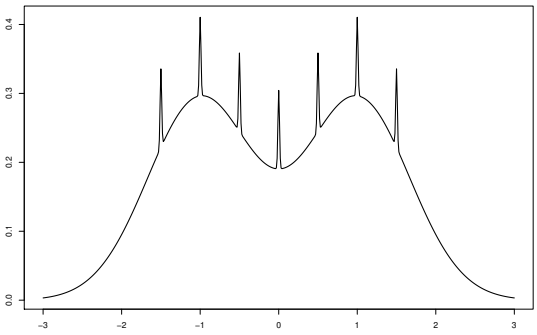
(g) Densité bimodale séparée



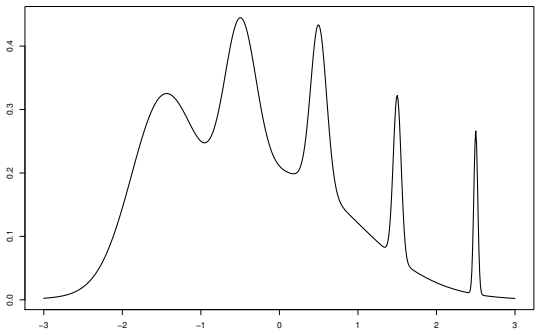
(h) Densité bimodale asymétrique



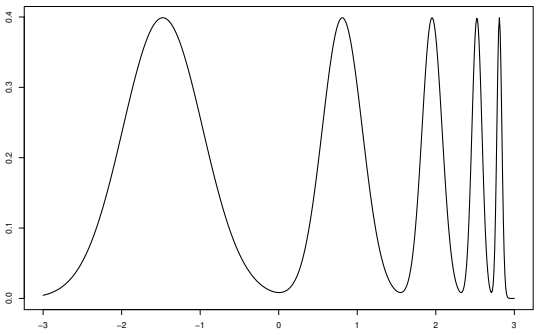
(i) Densité trimodale



(j) Griffe



(k) Griffe asymétrique



(l) Peigne

FIGURE 6.4 – Graphes de densités de mélanges gaussiens (suite).

l'ensemble des densités obtenues par transformation affine de la loi de f. Alors pour tout $\varepsilon > 0$ et tout compact $C \subset \mathbb{R}$, il existe un mélange $\bar{g}$ de densités de $\mathcal{H}$ tel que

$$\sup_{x \in C} |g(x) - \bar{g}(x)| < \varepsilon \ ,$$

où

$$\bar{g}(x) = \sum_{j=1}^{m} \pi_j h_{(\mu_j, \sigma_j)}(x) \ ,$$

avec $m \in \mathbb{N}$, $\mu_j \in \mathbb{R}, \sigma_j > 0$, $h_{(\mu_j,\sigma_j)} \in \mathcal{H}$ pour $j = 1, \ldots, m$, et $\pi_j > 0$ avec $\sum_{j=1}^{m} \pi_j = 1$.

Démonstration du Théorème 16. Soit $\varepsilon > 0$ fixé. Puisque g est intégrable, il existe un compact C_ε tel que

$$\int_{\mathbb{R} \setminus C_\varepsilon} g(x) \mathrm{d}x < \frac{\varepsilon}{4} \ .$$

D'après Lemme 1, il existe un mélange gaussien $\bar{g}$ tel que

$$\sup_{x \in C_\varepsilon} |g(x) - \bar{g}(x)| < \frac{\varepsilon}{4|C_\varepsilon|} \ ,$$

où $|C_\varepsilon|$ désigne la mesure de Lebesgue de C_ε. Donc, $\int_{C_\varepsilon} |g(x) - \bar{g}(x)| \mathrm{d}x < \frac{\varepsilon}{4}$. Or,

$$\|g - \bar{g}\|_1 = \int_{\mathbb{R}} |g(x) - \bar{g}(x)| \mathrm{d}x = \int_{\mathbb{R} \setminus C_\varepsilon} |g(x) - \bar{g}(x)| \mathrm{d}x + \int_{C_\varepsilon} |g(x) - \bar{g}(x)| \mathrm{d}x \ .$$

Pour terminer la preuve, on constate que

$$\begin{aligned}
\int_{\mathbb{R} \setminus C_\varepsilon} |g(x) - \bar{g}(x)| \mathrm{d}x &\leq \int_{\mathbb{R} \setminus C_\varepsilon} g(x) \mathrm{d}x + \int_{\mathbb{R} \setminus C_\varepsilon} \bar{g}(x) \mathrm{d}x \\
&= \int_{\mathbb{R} \setminus C_\varepsilon} g(x) \mathrm{d}x + 1 - \left[\int_{C_\varepsilon} (\bar{g}(x) - g(x)) \mathrm{d}x + \int_{C_\varepsilon} g(x) \mathrm{d}x \right] \\
&= 2 \int_{\mathbb{R} \setminus C_\varepsilon} g(x) \mathrm{d}x - \int_{C_\varepsilon} (\bar{g}(x) - g(x)) \mathrm{d}x \\
&\leq 2 \int_{\mathbb{R} \setminus C_\varepsilon} g(x) \mathrm{d}x + \int_{C_\varepsilon} |\bar{g}(x) - g(x)| \mathrm{d}x \\
&< 2 \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{3}{4} \varepsilon \ .
\end{aligned}$$

□

Identifiabilité

Le modèle de mélange ainsi défini n'est pas identifiable. Il est clair que tout mélange de m classes peut être représenté par un mélange de $m + 1$ classes, soit en ajoutant une population avec poids $\pi_j = 0$, soit en "coupant" une population en deux avec le même comportement, c'est-à-dire avec le même paramètre ϕ_j. Pour obtenir un modèle de mélange avec exactement m populations différentes, il faut ajouter les contraintes suivantes sur les paramètres : $\pi_j > 0$ pour tout j, et $\phi_j \neq \phi_{j'}$ pour tout $j \neq j'$.

Cependant, même sous ces contraintes, le modèle n'est toujours pas identifiable. En effet, il est possible de permuter les indices. Plus précisément, on peut aussi bien associer le couple de paramètres (ϕ_1, π_1) au comportement des femmes que les paramètres (ϕ_2, π_2), car il n'y

a pas de règle pour définir quelle population est le premier, deuxième... composant du mélange. Si $\Phi \subset \mathbb{R}$, on peut obtenir l'identifiabilité du modèle de mélange par les deux contraintes suivantes

$$\phi_1 < \phi_2 < \dots \phi_m \ , \quad \text{et} \quad \pi_j > 0 \quad \text{pour } j = 1, \dots, m \ .$$

Sous ces contraintes la forme de l'ensemble de paramètres Θ devient compliquée, ce qui peut entraîner des problèmes sérieux au niveau du calcul d'estimateurs. En effet, certains algorithmes d'optimisation ne peuvent pas prendre en compte de telles contraintes sur les paramètres.

Estimation de paramètres

Revenons aux exemples des longueurs d'aile de passereaux et du chlorure dans le sang. Notons $\mathbf{x} = (x_1, \dots, x_n)$ un échantillon i.i.d. d'un modèle de mélange avec deux populations ($m = 2$), où chaque composant suit une loi normale. La densité de mélange f_θ s'écrit donc comme

$$\begin{aligned} f_\theta(x) &= p f_{\mathcal{N}(\mu_1, \sigma_1^2)} + (1-p) f_{\mathcal{N}(\mu_2, \sigma_2^2)} \\ &= \frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\} + \frac{1-p}{\sqrt{2\pi\sigma_2^2}} \exp\left\{ -\frac{(x-\mu_2)^2}{2\sigma_2^2} \right\} \ , \end{aligned} \tag{6.1}$$

avec $p \in]0,1[$. On cherche à estimer les paramètres inconnus $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p)$.

Pour la méthode du maximum de vraisemblance, on calcule la fonction de vraisemblance par

$$\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n f_\theta(x_i) = \frac{1}{(2\pi)^{n/2}} \prod_{i=1}^n \left(\frac{p}{\sigma_1} \exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\} + \frac{1-p}{\sigma_2} \exp\left\{ -\frac{(x-\mu_2)^2}{2\sigma_2^2} \right\} \right) \ .$$

et la fonction de log-vraisemblance

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^n \log f_\theta(x_i) \\ &= -\frac{n}{2} \log(2\pi) + \sum_{i=1}^n \log \left(\frac{p}{\sigma_1} \exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\} + \frac{1-p}{\sigma_2} \exp\left\{ -\frac{(x-\mu_2)^2}{2\sigma_2^2} \right\} \right) \ . \end{aligned}$$

Or,

$$\begin{aligned} \frac{\partial}{\partial p} \ell(\theta) &= \sum_{i=1}^n \frac{\frac{1}{\sigma_1} \exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\} - \frac{1}{\sigma_2} \exp\left\{ -\frac{(x-\mu_2)^2}{2\sigma_2^2} \right\}}{\frac{p}{\sigma_1} \exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\} + \frac{1-p}{\sigma_2} \exp\left\{ -\frac{(x-\mu_2)^2}{2\sigma_2^2} \right\}} \\ \frac{\partial}{\partial \mu_1} \ell(\theta) &= \sum_{i=1}^n \frac{\frac{p}{\sigma_1} (x_i - \mu_1) \exp\left\{ -\frac{(x_i-\mu_1)^2}{2\sigma_1^2} \right\}}{\frac{p}{\sigma_1} \exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\} + \frac{1-p}{\sigma_2} \exp\left\{ -\frac{(x-\mu_2)^2}{2\sigma_2^2} \right\}} \\ \frac{\partial}{\partial \mu_2} \ell(\theta) &= \dots \end{aligned}$$

Il est clair que l'équation $\nabla \ell(\theta) = 0$ n'admet pas de solution explicite. En effet, le fait que la fonction de vraisemblance s'écrit comme un produit de sommes rend sa maximisation assez compliquée. Le calcul de l'estimateur du maximum de vraisemblance dans de modèles de mélange nécessite généralement des méthodes numériques.

6.2.4 Modèles à variables latentes

Le modèle de mélange fait partie d'une famille de modèles plus large. En effet, il y a d'autres modèles qui font intervenir des variables cachées comme l'étiquette U dans le modèle de mélange (la variable qui désigne l'appartenance de groupe). On parle aussi de **variables latentes** ou de **variables manquantes**, quand il y a des variables du modèle qui ne sont pas observées, et on appelle ces modèles des **modèles à variables latentes**.

Dans la suite nous utiliserons les notations suivantes. Soit $\mathbf{x}$ un échantillon i.i.d. de densité p_{θ_0} dans le modèle statistique $\{p_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^d$. On dit que $\mathbf{x}$ sont les **données incomplètes** du modèle. On dénote $\mathbf{u}$ les variables latentes du modèle. On dit que $(\mathbf{x}, \mathbf{u})$ sont les **données complètes** du modèle. On considère $(\mathbf{x}, \mathbf{u})$ comme une réalisation de densité q_{θ_0} dans un modèle statistique $\{q_\theta, \theta \in \Theta\}$.

Typiquement, le modèle des données incomplètes $\{p_\theta, \theta \in \Theta\}$ est très compliqué tel que les estimateurs classiques (EMM ou EMV) ne sont pas calculables. L'objectif de l'introduction de variables latentes dans le modèle est de passer à un modèle dans lequel les calculs se passent mieux. En fait, on se rend compte facilement dans l'exemple d'un mélange gaussien que l'EMV est explicite si on disposait des données complètes $(\mathbf{x}, \mathbf{u})$.

Dans le paragraphe suivant nous présenterons une méthode numérique pour approcher l'EMV dans des modèles à variables latentes, qui exploite justement le fait que l'EMV dans le modèle des données complètes est abordable.

6.3 Algorithme EM

L'algorithme EM de Dempster, Laird et Rubin (1977) est une procédure itérative pour calculer l'EMV lorsque celui n'admet pas d'expression explicite. Il est notamment approprié pour des nombreux modèles à variables latentes.

6.3.1 Contexte d'application

Soit $\mathbf{x}$ un échantillon de la densité p_{θ_0} dans le modèle statistique $\{p_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^d$. Notons $\mathbf{u}$ l'observation des variables latentes $\mathbf{U}$. On dénote q_{θ_0} la densité des données complètes $(\mathbf{x}, \mathbf{u})$.

Nous considérons le cadre où la maximisation de la log-vraisemblance $\theta \mapsto \log p_\theta(\mathbf{x})$ n'a pas de solution explicite, et l'introduction des variables latentes $\mathbf{u}$ (choisies judicieusement) donne lieu à une fonction de log-vraisemblance $\theta \mapsto \log q_\theta(\mathbf{x}, \mathbf{u})$ facile à maximiser.

6.3.2 L'algorithme EM

Si on connaissait les variables latentes $\mathbf{u}$, i.e. si on observait $(\mathbf{x}, \mathbf{u})$, on calculerait l'EMV de θ_0 en maximisant la log-vraisemblance $\theta \mapsto \log q_\theta(\mathbf{x}, \mathbf{u})$ sur Θ. Maintenant, les variables latentes $\mathbf{u}$ sont inconnues. L'idée est d'utiliser une procédure itérative, pour "estimer" les variables latentes $\mathbf{u}$ en considérant l'espérance conditionnelle de la log-vraisemblance $\log q_\theta(\mathbf{x}, \mathbf{u})$ sachant les données observées $\mathbf{x}$ et l'actuel paramètre $\theta^{(t)}$ obtenu dans l'itération précédente. Autrement dit, l'algorithme EM consiste à itérer les étapes suivantes :

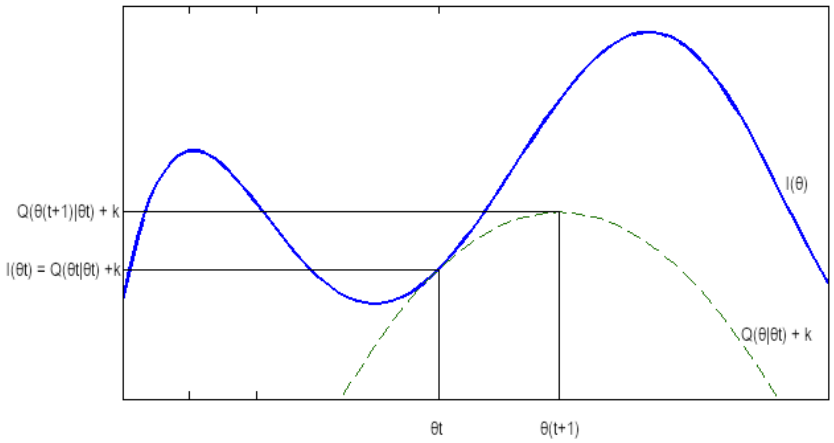


FIGURE 6.5 – Illustration d'une itération de l'algorithme EM. La courbe solide représente la log-vraisemblance $\theta \mapsto \ell(\theta) = \log p_\theta(\mathbf{x})$, qui est minorée $\theta \mapsto Q(\theta|\theta^{(t)}) + k$ (courbe hachurée). Les deux courbes se croisent en la valeur actuelle $\theta^{(t)}$ de θ. En maximisant $\theta \mapsto Q(\theta|\theta^{(t)})$, on augmente $\theta \mapsto \ell(\theta)$.

Étape E. Calculer la fonction

$$\theta \mapsto Q(\theta|\theta') = \mathbb{E}_{\theta'}[\log q_\theta(\mathbf{X}, \mathbf{U})|\mathbf{X} = \mathbf{x}] \ ,$$

où $\theta^{(t)}$ est le résultat de l'itération précédente.

Étape M. Maximiser la fonction $\theta \mapsto Q(\theta|\theta^{(t)})$. Plus précisément, calculer la nouvelle valeur de θ par

$$\theta^{(t+1)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(t)}) \ .$$

En effet, cette maximisation est généralement faisable car nous avons supposé qu'il est facile de maximiser la log-vraisemblance $\log q_\theta(\mathbf{x}, \mathbf{u})$.

La première étape est dite étape d'*espérance*, la deuxième étape de *maximisation*, d'où le nom de l'algorithme EM (en anglais *expectation-maximisation*).

6.3.3 PROPRIÉTÉS DE L'ALGORITHME EM

L'objectif de l'algorithme EM est de calculer l'EMV. Bien qu'on ne peut pas garantir que ce but soit toujours atteint, on peut montrer des propriétés importantes de cet algorithme. Notamment, à chaque itération la log-vraisemblance $\ell(\theta) = \log p_\theta(\mathbf{x})$ est augmentée.

Théorème 17. *Soit $(\theta^{(t)})_{t\geq 1}$ une suite obtenue par l'algorithme EM. La log-vraisemblance $\ell(\theta)$ vérifie pour tout t,*

$$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)}) \ .$$

Démonstration. Il suffit de montrer que pour tout θ il existe une constante $k_{\theta'}$ telle que

(i) la log-vraisemblance $\ell(\theta)$ est minorée par $Q(\theta|\theta') + k_{\theta'}$ pour tout θ, θ' et

(ii) $Q(\theta'|\theta') + k_{\theta'} = \ell(\theta')$.

En effet, avec ces deux propriétés de $Q(\theta|\theta')$ on a pour tout t

$$\begin{aligned}\ell(\theta^{(t)}) &\overset{(ii)}{=} Q(\theta^{(t)}|\theta^{(t)}) + k_{\theta^{(t)}} \\ &\leq \max_{\theta\in\Theta} Q(\theta|\theta^{(t)}) + k_{\theta^{(t)}} \\ &= Q(\theta^{(t+1)}|\theta^{(t)}) + k_{\theta^{(t)}} \\ &\overset{(i)}{\leq} \ell(\theta^{(t+1)}) \ .\end{aligned}$$

Autrement dit, en maximisant l'espérance conditionnelle $Q(\theta|\theta^{(t)})$ dans l'étape M, on obtient forcément une valeur $\theta^{(t+1)}$ où la log-vraisemblance est plus élevée qu'au point $\theta^{(t)}$, c'est-à-dire $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$. Ce phénomène est illustré graphiquement dans la Figure 6.5.

Montrons donc (i) et (ii). Ce résultat repose essentiellement sur l'inégalité de Jensen, d'après laquelle pour toute fonction h concave et toute v.a. Z on a $\mathbb{E}[h(Z)] \leq h(\mathbb{E}[Z])$. En effet, pour p et q deux densités par rapport à une même mesure μ, on obtient par la concavité du logarithme

$$\begin{aligned}\mathbb{E}_p[\log q(Z)] - \mathbb{E}_p[\log p(Z)] &= \mathbb{E}_p\left[\log \frac{q(Z)}{p(Z)}\right] \\ &\leq \log\left(\mathbb{E}_p\left[\frac{q(Z)}{p(Z)}\right]\right) = \log\left(\int \frac{q(z)}{p(z)} p(z)\mu(\mathrm{d}z)\right) = 0 \ ,\end{aligned}$$

d'où $\mathbb{E}_p[\log q(Z)] \leq \mathbb{E}_p[\log p(Z)]$. Nous appliquons cette inégalité aux densités conditionnelles des données complètes sachant les données observées, i.e. on pose

$$p(\mathbf{u}) = \frac{q_{\theta'}(\mathbf{x},\mathbf{u})}{p_{\theta'}(\mathbf{x})} \quad \text{et} \quad q(\mathbf{u}) = \frac{q_{\theta}(\mathbf{x},\mathbf{u})}{p_{\theta}(\mathbf{x})} \ .$$

On obtient alors

$$\begin{aligned}Q(\theta|\theta') - \ell(\theta) &= \mathbb{E}_{\theta'}[\log q_\theta(\mathbf{x},\mathbf{U})|\mathbf{X}=\mathbf{x}] - \log p_\theta(\mathbf{x}) \\ &= \mathbb{E}_{\theta'}\left[\log \frac{q_\theta(\mathbf{x},\mathbf{U})}{p_\theta(\mathbf{x})}\middle|\mathbf{X}=\mathbf{x}\right] \\ &\leq \mathbb{E}_{\theta'}\left[\log \frac{q_{\theta'}(\mathbf{x},\mathbf{U})}{p_{\theta'}(\mathbf{x})}\middle|\mathbf{X}=\mathbf{x}\right] \\ &= Q(\theta'|\theta') - \ell(\theta') \ ,\end{aligned}$$

avec égalité pour $\theta = \theta'$. Par conséquent, $\theta \mapsto Q(\theta|\theta') - Q(\theta'|\theta') + \ell(\theta')$ est un minorant de $\theta \mapsto \ell(\theta)$, et en posant $k_{\theta'} = -Q(\theta'|\theta') + \ell(\theta')$ on obtient (i) et (ii). □

Bien qu'on a montré que l'algorithme EM augmente la log-vraisemblance à chaque itération, cela n'est pas une garantie que l'algorithme atteint le maximum global à la fin. En effet, on peut montrer sous des conditions assez générales, que l'algorithme EM converge toujours vers un point stationnaire de la log-vraisemblance $\ell(\theta)$, mais il est possible que ceci est seulement un maximum local ou un point d'inflexion.

6.3.4 Aspects pratiques

Initialisation

Si le résultat de l'algorithme EM est l'EMV recherché ou non, dépend essentiellement de l'initialisation de l'algorithme, c'est-à-dire du choix de $\theta^{(0)}$. Il n'y a pas de règle générale

comment choisir convenablement $\theta^{(0)}$, il faut regarder au cas par cas. Parfois on peut initialiser avec un estimateur simple (et pas très précis) de θ_0. Si la convergence de l'algorithme n'est pas trop lente, il est envisageable de lancer l'algorithme plusieurs fois avec des différents points initiaux $\theta^{(0)}$ choisis au hasard. Puis on choisit comme estimateur de θ_0, le résultat avec la plus grande vraisemblance.

Critère d'arrêt

Après combien d'itérations covient-t-il d'arrêter l'algorithme EM ? Comment savoir si l'algorithme a convergé ? Il n'y a pas de nombre d'itérations qui soit convenable pour tous les modèles. Il y a des cas où l'algorithme converge rapidement, et d'autres où la convergence est très lente. De manière générale, plus le nombre de variables latentes est important, plus la convergence est lente.

On peut observer la suite $(\theta^{(t)})_{t\geq 1}$ et arrêter l'algorithme quand la différence entre deux $\theta^{(t)}$ successifs est petits, c'est-à-dire lorsque $\|\theta^{(t+1)} - \theta^{(t)}\| < \varepsilon$, où $\|\cdot\|$ désigne ou la norme euclidienne ou la norme sup et $\varepsilon > 0$ est un seuil fixé par avance.

Alternativement, on peut fonder le critère d'arrêt sur la fonction de vraisemblance, plus précisément sur la convergence de la suite $(\mathcal{L}(\mathbf{x}, \theta^{(t)}))_{t\geq 1}$. On arrête dès que l'augmentation entre deux itérations est trop faible, c'est-à-dire quand $\mathcal{L}(\mathbf{x}, \theta^{(t+1)}) - \mathcal{L}(\mathbf{x}, \theta^{(t)}) < \varepsilon$ pour un seuil $\varepsilon > 0$ donné.

Dans les deux cas, il est possible que l'on arrête alors que l'algorithme n'a pas encore convergé. Cela arrive si la convergence est très lente ou si la fonction de vraisemblance est relativement plat.

Grâce à son caractère général, l'algorithme EM s'applique à des problèmes très variés et son utilisation est très répandue en pratique. Au fil du temps, de nombreuses variantes de cet algorithme sont nées. Nous nous contentons d'étudier des cadres d'application classiques, comme le modèle de mélange et le modèle de censure.

6.3.5 Exemple : Mélange gaussien

Reprenons l'exemple d'un mélange de deux lois normales, dont la densité de mélange f_θ est donnée par (6.1). Autrement dit, $\mathbf{x} = (x_1, \ldots, x_n)$ est un échantillon i.i.d. de la variable aléatoire X définie par

$$X = \mathbb{1}\{U = 1\}V_1 + \mathbb{1}\{U = 2\}V_2 \ ,$$

où U, V_1, V_2 sont des variables aléatoires indépendantes telles que V_j suit la loi $\mathcal{N}(\mu_j, \sigma_j^2)$ pour $j = 1, 2$ et U vérifie $\mathbb{P}(U = 1) = 1 - \mathbb{P}(U = 2) = p$. L'étiquette U n'est pas observée, il s'agit de la variable latente du modèle. Dénotons u_i la réalisation de U associée à l'observation x_i, et $\mathbf{u} = (u_1, \ldots, u_n)$.

Il est clair que la loi conditionnelle de X sachant que $U = u$ est la loi normale $\mathcal{N}(\mu_u, \sigma_u^2)$, i.e.

$$f_{X|U}(x|u) = f_{\mathcal{N}(\mu_u, \sigma_u^2)}(x) \ , \quad \text{pour tout } x \in \mathbb{R}, u \in \{1, 2\} \ .$$

Or, la densité jointe $p_{(X,U)}$ de (X, U) par rapport à la mesure $\nu = \lambda \otimes \delta_{\{1,2\}}$, où λ désigne la mesure de Lebesgue sur $\mathbb{R}$ et $\delta_{\{1,2\}}$ la mesure de comptage sur $\{1, 2\}$, est donnée par

$$p_{(X,U)}(x, u) = f_{X|U}(x|u)p_U(u) = p^{\mathbb{1}\{u=1\}}(1-p)^{\mathbb{1}\{u=2\}} f_{\mathcal{N}(\mu_u, \sigma_u^2)}(x) \ , \quad u \in \{1, 2\}, x \in \mathbb{R} \ .$$

La densité jointe q_θ de l'échantillon $(\mathbf{x}, \mathbf{u}) = (x_1, \dots, x_n, u_1, \dots, u_n)$ est donc

$$q_\theta(\mathbf{x}, \mathbf{u}) = \prod_{i=1}^{n} p_{(X,U)}(x_i, u_i) = p^{\sum_{i=1}^n \mathbb{1}\{u_i=1\}}(1-p)^{n-\sum_{i=1}^n \mathbb{1}\{u_i=1\}} \prod_{i=1}^{n} f_{\mathcal{N}(\mu_{u_i}, \sigma^2_{u_i})}(x_i) \ .$$

On en déduit la fonction Q de l'algorithme EM

$$\begin{aligned}
Q(\theta|\theta') &= \mathbb{E}_{\theta'}\left[\log q_\theta(\mathbf{x}, \mathbf{U})|\mathbf{X} = \mathbf{x}\right] \\
&= \mathbb{E}_{\theta'}\left[\sum_{i=1}^{n} \mathbb{1}\{U_i = 1\} \log p + \left(n - \sum_{i=1}^{n} \mathbb{1}\{U_i = 1\}\right) \log(1-p) + \sum_{i=1}^{n} \log\left(f_{\mathcal{N}(\mu_{U_i}, \sigma^2_{U_i})}(x_i)\right) \middle| \mathbf{X} = \mathbf{x}\right] \\
&= \log p \sum_{i=1}^{n} \mathbb{E}_{\theta'}\left[\mathbb{1}\{U = 1\}|X = x_i\right] + \log(1-p)\left(n - \sum_{i=1}^{n} \mathbb{E}_{\theta'}\left[\mathbb{1}\{U = 1\}|X = x_i\right]\right) + \\
&\quad + \sum_{i=1}^{n} \mathbb{E}_{\theta'}\left[\log\left(f_{\mathcal{N}(\mu_U, \sigma^2_U)}(x_i)\right) \middle| X = x_i\right] \ .
\end{aligned}$$

D'une part, $\mathbb{E}_{\theta'}\left[\mathbb{1}\{U = 1\}|X = x_i\right] = \mathbb{P}_{\theta'}(U = 1|X = x_i) =: \pi_{\theta'}(x_i)$. D'autre part,

$$\begin{aligned}
\mathbb{E}_{\theta'}\left[\log\left(f_{\mathcal{N}(\mu_U, \sigma^2_U)}(x_i)\right) \middle| X = x_i\right] &= \mathbb{E}_{\theta'}\left[\log\left(\frac{1}{\sqrt{2\pi}\sigma_U} \exp\left\{-\frac{(x_i - \mu_U)^2}{2\sigma^2_U}\right\}\right) \middle| X = x_i\right] \\
&= -\log\sqrt{2\pi} - \mathbb{E}_{\theta'}[\log(\sigma_U)|X = x_i] - \mathbb{E}_{\theta'}\left[\frac{(x_i - \mu_U)^2}{2\sigma^2_U} \middle| X = x_i\right] \\
&= -\log\sqrt{2\pi} - [\log(\sigma_1)\pi_{\theta'}(x_i) + \log(\sigma_2)(1 - \pi_{\theta'}(x_i))] \\
&\quad - \frac{1}{2}\left\{\frac{(x_i - \mu_1)^2}{\sigma_1^2}\pi_{\theta'}(x_i) + \frac{(x_i - \mu_2)^2}{\sigma_2^2}(1 - \pi_{\theta'}(x_i))\right\} \ .
\end{aligned}$$

Il suffit de déterminer les probabilités conditionnelles $\pi_{\theta'}(x_i)$ pour tout $i = 1, \dots, n$, pour connaitre entièrement la fonction $Q(\theta|\theta')$. Or,

$$\pi_{\theta'}(x) = \mathbb{P}_{\theta'}(U = 1|X = x) = \frac{p_{(X,U)}(x, 1)}{f_X(x)} = \frac{p' f_{\mathcal{N}(\mu'_1, \sigma'^2_1)}(x)}{p' f_{\mathcal{N}(\mu'_1, \sigma'^2_1)}(x) + (1-p') f_{\mathcal{N}(\mu'_2, \sigma'^2_2)}(x)} \ . \tag{6.2}$$

Notons que $0 < \pi_{\theta'}(x) < 1$ pour tout x et θ'.

Enfin, on obtient

$$\begin{aligned}
Q(\theta|\theta') &= \log p \sum_{i=1}^{n} \pi_{\theta'}(x_i) + \log(1-p)\left(n - \sum_{i=1}^{n} \pi_{\theta'}(x_i)\right) + \\
&\quad - n\log\sqrt{2\pi} - \frac{1}{2}\log(\sigma_1^2)\sum_{i=1}^{n} \pi_{\theta'}(x_i) - \frac{1}{2}\log(\sigma_2^2)\sum_{i=1}^{n}(1 - \pi_{\theta'}(x_i)) \\
&\quad - \frac{1}{2\sigma_1^2}\sum_{i=1}^{n} \pi_{\theta'}(x_i)(x_i - \mu_1)^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{n}(1 - \pi_{\theta'}(x_i))(x_i - \mu_2)^2
\end{aligned}$$

Pour la maximisation de la fonction $\theta \mapsto Q(\theta|\theta')$ dans l'étape M, il est utile de constater que l'on peut décomposer la maximisation en trois problèmes indépendants, car $Q(\theta|\theta')$ s'écrit comme la somme de trois fonctions : $Q(\theta|\theta') = Q_1(p|\theta') + Q_2(\mu_1, \sigma_1|\theta') + Q_3(\mu_2, \sigma_2|\theta')$. On

obtient pour les dérivées partielles,

$$\begin{aligned}
\frac{\partial}{\partial p}Q(\theta|\theta') &= \frac{1}{p}\sum_{i=1}^{n}\pi_{\theta'}(x_i) - \frac{1}{1-p}\left(n - \sum_{i=1}^{n}\pi_{\theta'}(x_i)\right)\\
\frac{\partial^2}{\partial^2 p}Q(\theta|\theta') &= -\frac{1}{p^2}\sum_{i=1}^{n}\pi_{\theta'}(x_i) - \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^{n}\pi_{\theta'}(x_i)\right) < 0\\
\frac{\partial}{\partial \mu_1}Q(\theta|\theta') &= \frac{1}{\sigma_1^2}\sum_{i=1}^{n}(x_i-\mu_1)\pi_{\theta'}(x_i) = \frac{1}{\sigma_1^2}\left[\sum_{i=1}^{n}x_i\pi_{\theta'}(x_i) - \mu_1\sum_{i=1}^{n}\pi_{\theta'}(x_i)\right]\\
\frac{\partial^2}{\partial^2 \mu_1}Q(\theta|\theta') &= -\frac{1}{\sigma_1^2}\sum_{i=1}^{n}\pi_{\theta'}(x_i) < 0\\
\frac{\partial}{\partial \sigma_1^2}Q(\theta|\theta') &= -\frac{1}{2\sigma_1^2}\sum_{i=1}^{n}\pi_{\theta'}(x_i) + \frac{1}{2\sigma_1^4}\sum_{i=1}^{n}\pi_{\theta'}(x_i)(x_i-\mu_1)^2\\
\frac{\partial^2}{\partial^2 \sigma_1^2}Q(\theta|\theta') &= \frac{1}{2\sigma_1^4}\sum_{i=1}^{n}\pi_{\theta'}(x_i) - \frac{1}{\sigma_1^6}\sum_{i=1}^{n}\pi_{\theta'}(x_i)(x_i-\mu_1)^2
\end{aligned}$$

et similaire pour les dérivées partielles par rapport à μ_2 et σ_2^2.

Les fonctions $p \mapsto Q(\theta|\theta')$ et $\mu_1 \mapsto Q(\theta|\theta')$ étant strictement concave, on trouve leur maxima par les points critiques qui sont uniques. Ainsi

$$\begin{aligned}
\frac{\partial}{\partial p}Q(\theta|\theta') = 0 &\iff p = \frac{1}{n}\sum_{i=1}^{n}\pi_{\theta'}(x_i)\\
\frac{\partial}{\partial \mu_1}Q(\theta|\theta') = 0 &\iff \mu_1 = \frac{\sum_{i=1}^{n}x_i\pi_{\theta'}(x_i)}{\sum_{i=1}^{n}\pi_{\theta'}(x_i)} =: \hat{\mu}_1\ .
\end{aligned}$$

Maintenant, maximisant $\sigma_1^2 \mapsto Q(\theta|\theta')\big|_{\mu_1=\hat{\mu}_1}$. D'abord on trouve pour le point critique

$$\frac{\partial}{\partial \sigma_1^2}Q(\theta|\theta')\bigg|_{\mu_1=\hat{\mu}_1} = 0 \iff \sigma_1^2 = \frac{\sum_{i=1}^{n}\pi_{\theta'}(x_i)(x_i-\hat{\mu}_1)^2}{\sum_{i=1}^{n}\pi_{\theta'}(x_i)} =: \hat{\sigma}_1^2\ ,$$

ensuite on vérifie qu'on a bien

$$\frac{\partial^2}{\partial^2 \sigma_1^2}Q(\theta|\theta')\bigg|_{\mu_1=\hat{\mu}_1,\sigma_1=\hat{\sigma}_1^2} = -\frac{\left(\sum_{i=1}^{n}\pi_{\theta'}(x_i)\right)^2}{2\left(\sum_{i=1}^{n}\pi_{\theta'}(x_i)(x_i-\hat{\mu}_1)^2\right)^3} < 0\ .$$

Donc, il s'agit bien d'un maximum local. De plus, étant l'unique point critique, il s'agit du maximum local de la fonction.

Les calculs pour μ_2 et σ_2^2 sont analogs.

Enfin, l'algorithme EM consiste à calculer successivement pour tout $t = 1, 2, \ldots$

$$p^{(t+1)} = \frac{\sum_{i=1}^{n}\pi_{\theta^{(t)}}(x_i)}{n}\ ,\quad \mu_1^{(t+1)} = \frac{\sum_{i=1}^{n}x_i\pi_{\theta^{(t)}}(x_i)}{\sum_{i=1}^{n}\pi_{\theta^{(t)}}(x_i)}\ ,\quad \mu_2^{(t+1)} = \frac{\sum_{i=1}^{n}x_i(1-\pi_{\theta^{(t)}}(x_i))}{n-\sum_{i=1}^{n}\pi_{\theta^{(t)}}(x_i)}$$

$$\sigma_1^{2(t+1)} = \frac{\sum_{i=1}^{n}\pi_{\theta^{(t)}}(x_i)(x_i-\mu_1^{(t+1)})^2}{\sum_{i=1}^{n}\pi_{\theta^{(t)}}(x_i)}\ ,\quad \sigma_2^{2(t+1)} = \frac{\sum_{i=1}^{n}(1-\pi_{\theta^{(t)}}(x_i))(x_i-\mu_2^{(t+1)})^2}{\sum_{i=1}^{n}(1-\pi_{\theta^{(t)}}(x_i))}\ ,$$

où les $\pi_{\theta^{(t)}}(x_i)$ sont donnés par (6.2).

Revenons aux exemples des passereaux et du chlorure pour illustrer cet algorithme. Les Figures 6.6 et 6.7 (a) montrent les densités d'un mélange de deux lois normales où les paramètres ont été calculé par l'algorithme EM décrit ci-dessus. On observe la bonne adéquation des densités estimées avec les données. Dans le cas du chlorure, nous voyons que les points du QQ-plot de la Figure 6.7 (b) qui compare les données au mélange gaussien estimé donne un meilleur résultat que dans le cas d'une simple loi normale (Figure 6.2 (b)).

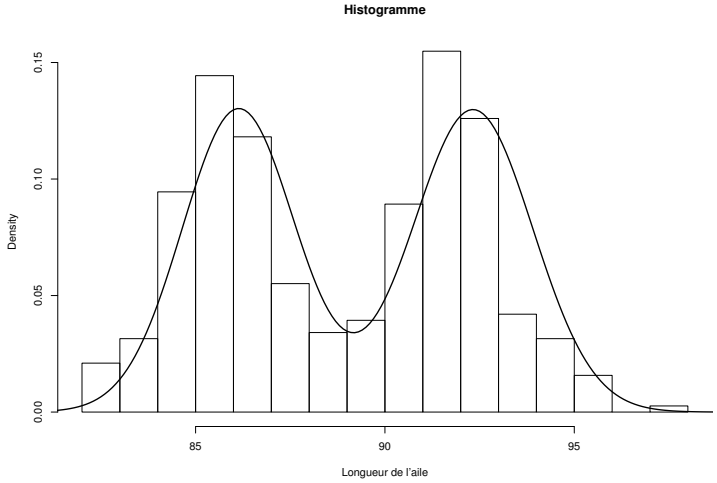
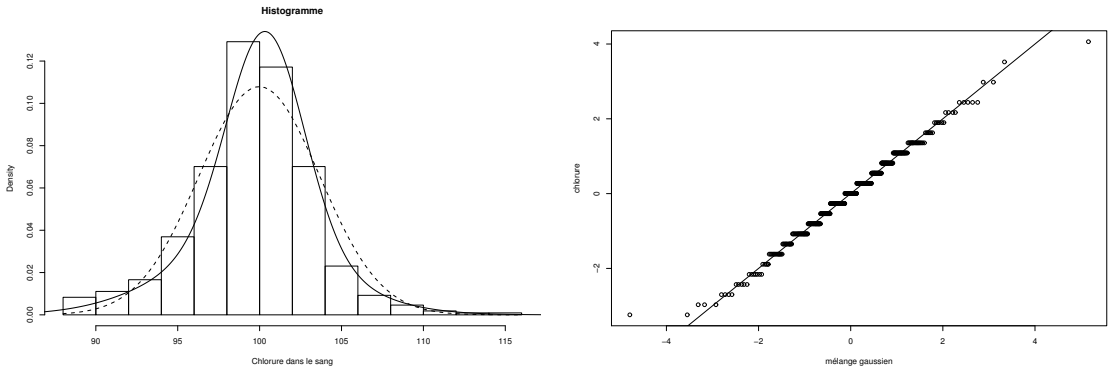


FIGURE 6.6 – Histogramme des longueurs des ailes et densité d'un mélange de deux lois normales dont les paramètres sont estimés par l'algorithme EM.



(a) (b)

FIGURE 6.7 – (a) Histogramme des données de chlorure dans le sang, la densité de la loi normale $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ où $\hat{\mu}$ et $\hat{\sigma}^2$ sont l'EMV (en trait pointillé) et la densité d'un mélange de deux lois normales dont les paramètres sont estimés par l'algorithme EM. (b) QQ-plot des données standardisées en comparaison au mélange gaussien estimé.

Exercice 4. Simulation Proposer une méthode pour générer des observations d'un modèle de censure à partir des réalisations de la loi uniforme $U[0,1]$.

Exercice 5. EMM dans le modèle de mélange
Une île est peuplée par 2 espèces d'oiseaux très proches, indiscernables à l'oeil nu. L'envergure (distance entre les extrémités des ailes) en centimètres des oiseaux de l'espèce A peut être modélisée par une loi gaussienne $\mathcal{N}(15, \sigma_1^2)$ avec $\sigma_1 > 0$, et celle des oiseaux de l'espèce B par une loi gaussienne $\mathcal{N}(15, \sigma_2^2)$ avec $\sigma_2 > 0$. Notons $p \in]0,1[$ la proportion d'oiseaux de l'espèce A vivant sur l'île. On capture n oiseaux au hasard parmi tous ceux vivant sur l'île et on mesure leur envergure. On considère la taille du i^{e} oiseau comme une réalisation d'une v.a. X_i, les $(X_i)_{1 \leq i \leq n}$ étant supposées i.i.d.

1. Quelle est la loi de X_i ? Donner sa densité.
2. Quelle difficulté rencontre-t-on pour calculer l'estimateur par la méthode des moments pour estimer les paramètres inconnus σ_1^2, σ_2^2 et p ?
3. Supposons désormais que $\sigma_1 = 1$ et $\sigma_2 = 2$. Expliciter $\hat{p}_n$, l'estimateur de p obtenu à l'aide de la méthode des moments.
4. Montrer que l'estimateur $\hat{p}_n$ est consistant et déterminer la loi limite de $\sqrt{n}(\hat{p}_n - p)$.

Exercice 6. Mélange gaussien
Soient $\mathbf{x} = (x_1, \ldots, x_n)$ des observations i.i.d. d'une variable aléatoire X d'un mélangé d'ordre m de lois normales $\mathcal{N}(\mu, \sigma^2)$.

1. Quelle est la densité de X que l'on notera p_θ ? On précisera le vecteur de paramètres θ et l'ensemble de paramètres Θ. Exprimer X en fonction d'une variable latente U à valeurs dans $\{1, \ldots, m\}$ avec

$$\mathbb{P}(U = j) = \pi_j \ , \quad j \in \{1, \ldots, m\} \ .$$

2. Donner la loi jointe des données complètes (X, U). On notera q_θ sa densité par rapport à la mesure $\lambda \times \delta_{\{1,\ldots,m\}}$ où λ désigne la mesure de Lebesgue et $\delta_{\{1,\ldots,m\}}$ la mesure de comptage sur $\{1, \ldots, m\}$.
3. La loi conditionnelle de U sachant que $X = x$ pour $x \in \mathbb{R}$ fixé est donnée par

$$\mathbb{P}_\theta(U = j | X = x) = \frac{q_\theta(x, j)}{p_\theta(x)} = \frac{\pi_j f_{\mathcal{N}(\mu_j, \sigma_j^2)}(x)}{p_\theta(x)} \ , \quad j \in \{1, \ldots, m\} \ .$$

   On notera par la suite $\alpha_{j|i} = \mathbb{P}_{\theta'}(U = j | X = x_i)$. Calculer la fonction

$$Q(\theta | \theta') = \mathbb{E}_{\theta'}[\log q_\theta(\mathbf{x}, \mathbf{U}) | \mathbf{X} = \mathbf{x}] \ .$$

4. Calculer le point maximum de la fonction $\theta \mapsto Q(\theta | \theta')$ sur Θ.
5. Préciser les étapes d'un programme qui met en œuvre l'algorithme EM pour le mélange gaussien.

Chapitre 7

Modèles de régression

7.1 Motivation et Définition

Dans des nombreuses applications, l'objectif est l'étude de l'effet de certains facteurs de variabilité d'un phénomène. Autrement dit, on cherche à comprendre comment un facteur ou une variable (la *variable à expliquer*) dépend d'autres facteurs ou variables (les *variables explicatives*).

Exemple : Pluie à Paris

Prenons comme exemple les données du Tableau 7.1, qui donne le nombre de jours de pluie et la hauteur de pluie en mm, observés pendant toute l'année à Paris de 1956 à 1995. Une représentation des données sur un graphique (voir Figure 7.1) avec en abscisse le nombre de jours de pluie et en ordonnée la hauteur de pluie permet de constater que l'ensemble des points forme un nuage allongé et que la quantité de pluie augmente lorsque le nombre de jours de pluie augmente. Ici, on considère le facteur *hauteur de pluie* comme un facteur à expliquer par le facteur explicatif *nombre de jours de pluie*.

Notons les observations (x_i, y_i) où x_i désigne le nombre de jours de pluie et y_i la hauteur de pluie dans la même année. La forme du nuage de points suggère une relation affine entre les deux variables, i.e. il existe des constantes $a, b \in \mathbb{R}$ telles que

$$y_i \approx a + bx_i \ , \quad \text{pour tout } i = 1, \dots, n \ . \tag{7.1}$$

Il est clair qu'en (7.1) on n'a pas d'égalité stricte, mais si les coefficients a et b sont bien choisis, les écarts des points (x_i, y_i) de la droite sont pas très grands. Pour obtenir l'égalité, on introduit des erreurs e_i que l'on considère comme des réalisations des variables aléatoires η_i :

$$y_i = \underbrace{a + bx_i}_{\text{partie déterministe}} + \underbrace{e_i}_{\text{partie aléatoire}} \ , \quad \text{pour tout } i = 1, \dots, n \ . \tag{7.2}$$

L'idée consiste à supposer que la partie linéaire, i.e. le terme $a + bx_i$ explique une partie substantielle de l'observation y_i, et par conséquent, les erreurs e_i sont petites. Autrement dit, les y_i dépendent essentiellement des x_i par une relation linéaire. Le fait qu'il y a des petits écarts est dû soit à l'aléa du phénomène soit à l'influence d'autres facteurs, qui ne sont pas pris en compte dans le modèle linéaire.

Pluie à Paris

Année	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
Jours	154	161	193	131	198	152	159	159	146	196
Hauteur	545	536	783	453	739	541	528	559	521	880
Année	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
Jours	192	161	176	173	199	141	170	156	198	164
Hauteur	834	592	634	618	631	508	740	576	668	658
Année	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Jours	135	179	171	172	170	197	173	177	177	163
Hauteur	417	717	743	729	690	746	700	623	745	501
Année	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Jours	176	180	167	140	149	140	154	155	192	162
Hauteur	611	707	734	573	501	472	645	663	699	670

TABLE 7.1 – Nombre de jours de pluie par an à Paris et hauteur totale de pluie pour les années 1956 - 1995.

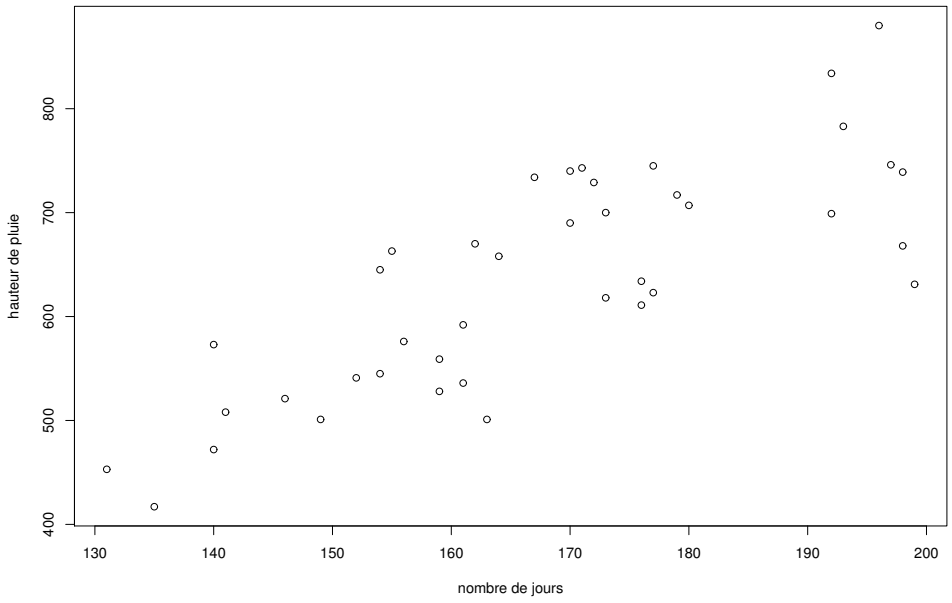


FIGURE 7.1 – Nuage des points pour les données de la pluie à Paris, Tableau 7.1.

Les questions qui nous préoccupent dans ce gendre de modèle sont entre autre : Comment *vérifier* si le choix du modèle (ici une droite) est approprié ? Comment *estimer* les paramètres a et b de la droite ? Peut-on *prédire* des nouvelles valeurs, p.ex. quelle est la hauteur de pluie pour un nombre de jours de pluie donné ?

Modèle de régression

De manière générale, le **modèle de régression** est de la forme

$$y_i = g(\mathbf{x}_i) + e_i \ , \quad \text{pour tout } i = 1, \dots, n \ ,$$

où

- $y_i \in \mathbb{R}$ désigne la i-ème réalisation de la **variable à expliquer**, aussi appelée **réponse** ou **variable dépendante**,
- $\mathbf{x}_i \in \mathbb{R}^p$ désigne un vecteur de p **variables explicatives**, aussi appelées **régresseurs** ou **variables indépendantes**,
- $g : \mathbb{R}^p \to \mathbb{R}$ est dite **fonction de régression**, **fonction de lien** ou encore **fonction de lissage**,
- e_i est l'**erreur**.

On suppose que les e_i sont des réalisations des variables aléatoires η_i centrée, i.e. $\mathbb{E}[\eta_i] = 0$. Par conséquent, l'observation y_i peut être considérée comme la réalisation d'une variable aléatoire Y_i donnée par

$$Y_i = \underbrace{g(\mathbf{x}_i)}_{\text{partie déterministe}} + \underbrace{\eta_i}_{\text{partie aléatoire}} \ . \tag{7.3}$$

Le modèle de régression est un modèle statistique qui distingue parmi les facteurs de variabilité les facteurs contrôlés (c'est la partie déterministe en (7.3)) et les facteurs aléatoires (la partie aléatoire de (7.3)).

L'objectif consiste à estimer la fonction de régression g à partir des observations $(x_i, y_i), i = 1, \dots, n$.

Remarquons que la fonction de régression g vérifie

$$\mathbb{E}[Y_i] = g(\mathbf{x}_i) \ ,$$

car les η_i sont centrées.

Il est possible d'introduire une modélisation probabilistes des variables explicatives $\mathbf{x}_i$, i.e. de supposer que les $\mathbf{x}_i$ sont des réalisations d'une lois $\mathbb{P}_{\mathbf{x}}$. Dans ce cours, on se contente de considérer les $\mathbf{x}_i$ comme des vecteurs déterministes.

7.2 Modèle linéaire et méthode des moindres carrés

7.2.1 Définition du modèle linéaire

Le cas le plus simple est de supposer que la fonction de régression g est une fonction linéaire en $\mathbf{x} \in \mathbb{R}^p$, qui s'écrit

$$g(\mathbf{x}) = \mathbf{x}^T \theta \ ,$$

pour un vecteur de paramètres $\theta = (\theta_1, \ldots, \theta_p)^T \in \mathbb{R}^p$. Le modèle de régression ainsi défini s'appelle **modèle de régression linéaire multidimensionnelle** (ou multivariée).

Le problème d'estimation de g se réduit alors à l'estimation du vecteur θ de dimension p. Il s'agit donc d'un problème d'estimation paramétrique.

L'importance de ce modèle pour les applications statistiques s'explique d'une part par sa relative simplicité et d'autre part par le fait qu'il permet d'inclure comme cas particuliers un certain nombre de modèles qui semblent, à première vue, non-linéaires.

Forme matricielle

Il est convenable d'écrire le modèle de régression linéaire sous forme matricielle

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{e} \,,$$

où $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $\theta = (\theta_1, \ldots, \theta_p)^T \in \mathbb{R}^p$, $\mathbf{e} = (e_1, \ldots, e_p)^T \in \mathbb{R}^n$ et

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T = \begin{pmatrix} x_{1,1} & \ldots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times p} \,,$$

où $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^T \in \mathbb{R}^p$. L'observation $\mathbf{y}$ est donc une réalisation du vecteur aléatoire $\mathbf{Y} = (Y_1, \ldots, Y_n)$ défini par

$$\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\eta} \,,$$

où $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ est le vecteur aléatoire des erreurs.

7.2.2 Cas particuliers

Modèle de translation

Lorsque $p = 1$ et

$$\mathbf{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \mathbb{1} \,,$$

le paramètre $\theta \in \mathbb{R}$, et les erreurs η_i sont i.i.d. centrées de loi F_η, on a

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \theta\mathbb{1} + \boldsymbol{\eta} \quad \text{et} \quad y_i = \theta + e_i \,,$$

ce qui revient à supposer que les y_i sont des réalisations indépendantes de la loi $F_\eta(\cdot - \theta)$ avec $\theta \in \mathbb{R}$. Autrement dit, le paramètre θ est un **paramètre de translation** qui vérifie

$$\mathbb{E}[Y_i] = \mathbb{E}[\theta + \eta_i] = \theta \,, \quad i = 1, \ldots, n \,.$$

Comparaison des moyennes de plusieurs populations

Supposons que l'on dispose de m échantillons $(y_{j,1} \ldots, y_{j,n_j})$ en provenance de m populations différentes. Soit F_η la densité d'une loi centrée. On suppose que les observations

$y_{j,1} \ldots, y_{j,n_j}$ sont des réalisations indépendantes de la loi $F_\eta(\cdot - \theta_j)$ pour un paramètre $\theta_j \in \mathbb{R}$.

Pour écrire ce modèle sous la forme d'un modèle de régression linéaire, on concatène les m échantillons pour former le vecteur

$$\mathbf{y} = (y_{1,1} \ldots, y_{1,n_1}, \ldots, y_{m,1} \ldots, y_{m,n_m})^T \in \mathbb{R}^{n_1+\cdots+n_m} ,$$

et on considère la matrice $\mathbf{X}$ de taille $n_1 + \cdots + n_m \times m$, telle que

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & 0 & \ldots & \vdots \\ 0 & \vdots & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_1 \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_2 \\ \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_m \end{matrix}$$

Pour les erreurs η_i, on suppose qu'elles sont des variables aléatoires i.i.d. de loi F_η.

Ce modèle est similaire au modèle de mélange du Chapitre 6 à la différence qu'ici on connaît la population de chaque observations $y_{i,j}$. En plus, dans ce modèle les paramètres θ_j sont des paramètres de translation, alors que dans le modèle de mélange il n'y pas cette restriction.

Régression linéaire simple

Reprenons l'exemple de la pluie à Paris. Posons $\theta = (a, b)^T$ et

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Le modèle linéaire et les observations y_i s'écrivent alors

$$\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\eta} \quad \text{et} \quad y_i = a + x_i b + e_i ,$$

Ceci est le modèle de **régression linéaire simple**, et la fonction $x \mapsto a + xb$ est dite **droite de régression**.

Régression polynomiale

Soit $z \in \mathbb{R}$ une variable explicative. 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. C'est-à-dire on suppose

$$y_i = \theta_1 + \theta_2 z + \theta_3 z^2 + \cdots + \theta_p z^{p-1} + e_i ,$$

où $p \geq 1$ est un entier et $\theta_1, \ldots, \theta_p$ sont des coefficients inconnus. Pour écrire ce modèle sous la forme d'un modèle de régression linéaire multidimensionnelle, on définit les vecteurs $\mathbf{x}_i = (1, z_i, z_i^2, \ldots, z_i^{p-1})^T$ et $\theta = (\theta_1, \ldots, \theta_p)^T$, et on obtient

$$y_i = \mathbf{x}_i^T \theta + e_i .$$

On voit donc que la **régression polynomiale** est un cas particulier de la régression linéaire multidimensionnelle.

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 soit de la forme

$$g(\mathbf{x}) = a\mathrm{e}^{\nu^T \mathbf{x}} , \quad \text{avec } \mathbf{x}, \nu \in \mathbb{R}^k ,$$

où ν est un vecteur de paramètres inconnus et $a > 0$ est une constante inconnue. Des fonctions de régression de ce type sont utilisées, par exemple, dans les applications en économie, pour modéliser la productivité des entreprises. En prenant les logarithmes, on obtient

$$\log g(\mathbf{x}) = \log a + \nu^T \mathbf{x} .$$

Afin de se ramener à une régression linéaire, on pose $\theta = (\log a, \nu^T)^T$, $\mathbf{x}' = (1, \mathbf{x}^T)^T$ et on obtient

$$y'_i = \log y_i = \theta^T \mathbf{x}'_i + e'_i , \quad i = 1, \dots, n . \tag{7.4}$$

C'est un modèle de régression linéaire par rapport à l'échantillon transformé

$$(\mathbf{x}'_1, y'_1), \dots, (\mathbf{x}'_n, y'_n) .$$

Notons que formellement on arrive à (7.4) à partir du modèle $y_i = g(\mathbf{x}_i)e_i$ de régression où les erreurs e_i interviennent de façon multiplicative et non pas additive (on a alors $e'_i = \log e_i$). Néanmoins, la transformation logarithmique est souvent utilisée sans mentionner cette nuance de manière explicite.

7.2.3 Méthode des moindres carrés

La méthode des moindres carrés est la technique d'estimation de paramètres la plus ancienne. Initialement proposée par Gauss en 1795 pour l'étude du mouvement des planètes, elle fut formalisée par Legendre en 1810. Elle occupe, aujourd'hui encore, une place centrale dans l'arsenal des méthodes d'estimation : son importance pratique est considérable.

La méthode des moindres carrés s'applique au modèle de régression linéaire multidimensionnelle de la forme

$$\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\eta} ,$$

où $\boldsymbol{\eta}$ est un vecteur aléatoire centré de matrice de covariance $\mathbf{Var}(\boldsymbol{\eta}) = \sigma^2 I_n$ avec $\sigma^2 > 0$ inconnu. La matrice $\mathbf{X}$ est supposée connue, ainsi que les observations $\mathbf{y}$. En revanche, les erreurs ne sont pas observables. Pour estimer le vecteur de paramètres θ on ne peut s'appuyer que sur $\mathbf{y}$ et $\mathbf{X}$.

Revenons à l'exemple de la pluie à Paris, Tableau 7.1, où on cherche la droite $x \mapsto a + bx$ qui approche le nuage de points "au mieux". La méthode des moindres carrés consiste à chercher les coefficients $\hat{a}$ et $\hat{b}$ qui minimisent la somme des carrés des déviations

$$\sum_{i=1}^{n} \Big(y_i - (\hat{a} + \hat{b}x_i) \Big)^2 = \min_{a,b \in \mathbb{R}} \sum_{i=1}^{n} (y_i - (a + bx_i))^2 .$$

Figure 7.2 montre la droite de régression obtenue par cette méthode ainsi que les déviations des points de cette droite pour notre exemple de la pluie.

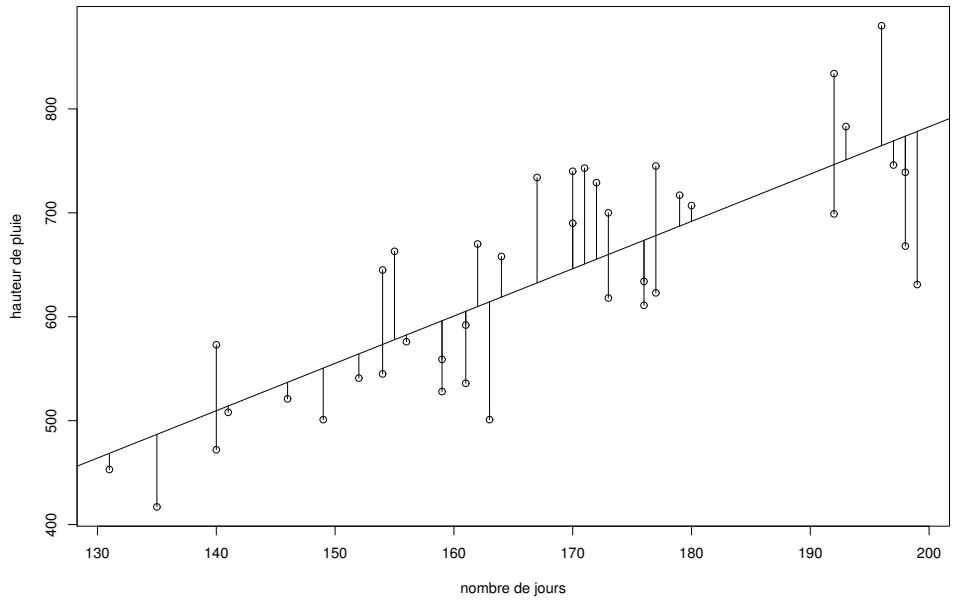


FIGURE 7.2 – Droite de régression et déviations des points de la droite de régression pour les données de la pluie à Paris, Tableau 7.1.

Plus généralement, **la méthode des moindres carrés** pour estimer le paramètre $\theta \in \mathbb{R}^p$ d'un modèle de régression linéaire multidimensionnelle consiste à chercher une valeur $\hat{\theta}$ qui minimise la somme des carrés des déviations :

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\theta})^2 = \min_{\theta\in\mathbb{R}^p}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\theta)^2 \ ,$$

ou en notation matricielle

$$\left\|\mathbf{y} - \mathbf{X}\hat{\theta}\right\|^2 = \min_{\theta\in\mathbb{R}^p}\|\mathbf{y} - \mathbf{X}\theta\|^2 \ .$$

Il est facile de voir qu'il existe toujours une solution $\hat{\theta}$ à ce problème de minimisation, que l'on appelle **estimateur des moindres carrés** de θ. On écrit alors

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^p}\|\mathbf{y} - \mathbf{X}\theta\|^2 \ .$$

L'estimateur des moindres carrés n'est pas toujours unique. La condition de l'unicité est donnée dans la proposition suivante.

Théorème 18. *Supposons que la matrice $\mathbf{X}^T\mathbf{X}$ soit strictement positive. Alors, l'estimateur des moindres carrés est unique et il s'écrit sous la forme*

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \ .$$

Démonstration. 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

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\hat{\theta} \; . \tag{7.5}$$

C'est un système de p équations linéaires qui admet une solution unique car la matrice $\mathbf{X}^T\mathbf{X}$ est inversible. Cette solution vaut

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \; .$$

Comme la fonction $h(\theta)$ est convexe et positive, ce vecteur $\hat{\theta}$ fournit le minimum global de h. □

Le système des équations linéaires (7.5) s'appelle **système des équations normales** pour la méthode des moindres carrés.

Proposition 10. *La matrice $\mathbf{X}^T\mathbf{X}$ 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 $\mathbf{X}^T\mathbf{X}$ est positive, car tout $v \in \mathbb{R}^p \backslash \{0\}$ vérifie l'inégalité

$$v^T\mathbf{X}^T\mathbf{X}v = w^Tw = \sum_{i=1}^{p} w_i^2 \geq 0 \; ,$$

où $w = \mathbf{X}v = (w_1, \ldots, w_p)$. Il est évident que l'inégalité précédente une 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 $\mathbf{X}^T\mathbf{X}$ n'est pas strictement positive, alors $\text{Rang}(\mathbf{X}) < p$. La preuve de la réciproque est similaire. Si $\text{Rang}(\mathbf{X}) < p$, alors il existe un vecteur $v \in \mathbb{R}^p \backslash \{0\}$ tel que $\mathbf{X}v = 0$. Il en résulte que $v^T\mathbf{X}^T\mathbf{X}v = 0$. Par conséquent, la matrice $\mathbf{X}^T\mathbf{X}$ 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 $\mathbf{X}^T\mathbf{X}$ est dégénérée. En effet, $n < p$ implique que $\text{Rang}(\mathbf{X}) < p$, 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 :

$$\mathbf{X}^T\mathbf{X} > 0 \Longrightarrow n \geq p \; .$$

Interprétation géométrique de la méthode des moindres carrés

La méthode des moindres carrés consiste à calculer la projection orthogonale de $\mathbf{y}$ sur l'espace engendré par les colonnes de $\mathbf{X}$.

Plus précisément, le problème de minimisation de la somme des carrés des déviations peut s'écrire sous la forme suivante :

$$\min_{\theta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\theta\|^2 = \min_{v \in D} \|\mathbf{y} - v\|^2 \; , \tag{7.6}$$

où D désigne le sous-espace linéaire de $\mathbb{R}^n$ défini par

$$D = \{v \in \mathbb{R}^n : \exists \theta \in \mathbb{R}^p, v = \mathbf{X}\theta\} \; .$$

Autrement dit, 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 $\mathbf{X}^T\mathbf{X} > 0$, alors D est un sous-espace linéaire de dimension p :

$$\text{Rang}(\mathbf{X}) = p \Longleftrightarrow \mathbf{X}^T\mathbf{X} > 0 \Longleftrightarrow \dim(D) = p \; .$$

Si $\mathbf{X}^T\mathbf{X} > 0$, la solution du problème (7.6) est $\hat{v} = \mathbf{X}\hat{\theta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \stackrel{\text{déf}}{=} H\mathbf{y}$.

La matrice $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \in \mathbb{R}^{n\times n}$ est dite **matrice chapeau** (en anglais *hat matrix*).

Proposition 11. *Supposons que $\mathbf{X}^T\mathbf{X} > 0$. Alors la matrice chapeau*

$$H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \in \mathbb{R}^{n\times n}$$

est symétrique, idempotente, Rang$(H) = p$ et H est le projecteur orthogonal dans $\mathbb{R}^n$ sur le sous-espace D.

Démonstration. Il vient

$$H^T = \mathbf{X}[(\mathbf{X}^T\mathbf{X})^{-1}]^T\mathbf{X}^T = \mathbf{X}[(\mathbf{X}^T\mathbf{X})^T]^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = H\ ,$$

et

$$H^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = H\ .$$

Donc H est symétrique et idempotente, ce qui signifie que H est un projecteur orthogonal. En outre, pour tout $\mathbf{y} \in \mathbb{R}^n$, on a $H\mathbf{y} = \mathbf{X}\hat{\theta} = \hat{v} \in D$. Donc H 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,

$$Hv = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Tv = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\theta = \mathbf{X}\theta = v\ .$$

Cela signifie que H est le projecteur orthogonal sur D. Comme D est un sous-espace de $\mathbb{R}^n$ de dimension p, le rang de H est égal à p. □

Propriétés statistiques de l'estimateur des moindres carrés

Théorème 19. *Supposons que*

- *la matrice $\mathbf{X}^T\mathbf{X}$ est déterministe et strictement positive et*
- *le vecteur aléatoire $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ est de moyenne $\mathbb{E}[\boldsymbol{\eta}] = 0$ et de matrice de covariance $\mathbf{Var}(\boldsymbol{\eta}) = \mathbb{E}[(\boldsymbol{\eta} - \mathbb{E}[\boldsymbol{\eta}])(\boldsymbol{\eta} - \mathbb{E}[\boldsymbol{\eta}])^T] = \sigma^2 I_n$, où $\sigma^2 > 0$ et I_n est la matrice unité de dimension $n \times n$.*

Alors, l'estimateur des moindres carrés est sans biais :

$$\mathbb{E}_\theta[\hat{\theta}] = \theta\ , \tag{7.7}$$

et sa matrice de covariance $\mathbf{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^T]$ vaut

$$\mathbf{Var}_\theta(\hat{\theta}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\ .$$

Démonstration. Il vient

$$\hat{\theta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\theta + \boldsymbol{\eta}) = \theta + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\eta}\ , \tag{7.8}$$

d'où découle (7.7). En utilisant (7.8) on obtient aussi que

$$\mathbf{Var}(\hat{\theta}) = \mathbb{E}\left[\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\eta}\right)\left(\boldsymbol{\eta}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)\right] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^T]\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\ .$$

Comme $\mathbf{Var}(\boldsymbol{\eta}) = \mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^T] = \sigma^2 I_n$, on obtient

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^T]\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\ .$$

□

L'estimateur par la méthode des moindres carrés $\hat{\theta}$ est à variance minimale parmi tous les estimateurs linéaires, sans biais de θ.

Théorème 20 (Théorème de Gauss-Markov)**.** *Sous les hypothèses du Théorème 19, soit $\tilde{\theta}$ un estimateur de θ de la forme $\tilde{\theta} = \mathbf{B}\mathbf{y}$ pour une matrice déterministe $\mathbf{B}$ de taille $p \times n$. Si $\tilde{\theta}$ est un estimateur sont biais, alors*

$$\mathbf{Var}_\theta(\tilde{\theta}) = \sigma^2 \mathbf{B}\mathbf{B}^T \geq \mathbf{Var}_\theta(\hat{\theta}) \ ,$$

où $\hat{\theta}$ désigne l'estimateur par la méthode des moindres carrés.

Démonstration. Comme $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\eta}$, on a $\mathbb{E}_\theta[\tilde{\theta}] = \mathbb{E}_\theta[\mathbf{B}\mathbf{Y}] = \mathbf{B}\mathbf{X}\theta + \mathbf{B}\mathbb{E}_\theta[\boldsymbol{\eta}] = \mathbf{B}\mathbf{X}\theta$. Puisque $\tilde{\theta}$ est un estimateur sans biais, on a

$$\mathbb{E}_\theta[\tilde{\theta}] = \theta \ , \quad \forall \theta \in \mathbb{R}^p \quad \Leftrightarrow \quad \mathbf{B}\mathbf{X}\theta = \theta \ , \quad \forall \theta \in \mathbb{R}^p \quad \Leftrightarrow \quad \mathbf{B}\mathbf{X} = I_p \ .$$

Il en découle que

$$\tilde{\theta} = \mathbf{B}\mathbf{Y} = \mathbf{B}\mathbf{X}\theta + \mathbf{B}\boldsymbol{\eta} = \theta + \mathbf{B}\boldsymbol{\eta} \ . \tag{7.9}$$

D'où $\mathbf{Var}_\theta(\tilde{\theta}) = \mathbf{Var}_\theta(\mathbf{B}\boldsymbol{\eta}) = \sigma^2 \mathbf{B}\mathbf{B}^T$. Or,

$$\begin{aligned}
\mathbf{Var}_\theta(\tilde{\theta}) &= \mathbb{E}_\theta\left[(\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T\right] \\
&= \mathbb{E}_\theta\left[(\tilde{\theta} - \hat{\theta} + \hat{\theta} - \theta)(\tilde{\theta} - \hat{\theta} + \hat{\theta} - \theta)^T\right] \\
&= \mathbb{E}_\theta\left[(\tilde{\theta} - \hat{\theta})(\tilde{\theta} - \hat{\theta})^T\right] + \mathbb{E}_\theta\left[(\tilde{\theta} - \hat{\theta})(\hat{\theta} - \theta)^T\right] + \mathbb{E}_\theta\left[(\hat{\theta} - \theta)(\tilde{\theta} - \hat{\theta})^T\right] + \mathbf{Var}_\theta(\hat{\theta}) \ .
\end{aligned}$$

En utilisant (7.9) et (7.8) on obtient

$$\begin{aligned}
\mathbb{E}_\theta\left[(\tilde{\theta} - \hat{\theta})(\hat{\theta} - \theta)^T\right] &= \mathbb{E}_\theta\left[(\mathbf{B}\boldsymbol{\eta} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\eta})(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\eta})^T\right] \\
&= (\mathbf{B} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T)\mathbb{E}_\theta\left[\boldsymbol{\eta}\boldsymbol{\eta}^T\right]\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{B}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right) \\
&= 0 \ ,
\end{aligned}$$

car $\mathbf{B}\mathbf{X} = I_p$. De même, on montre que $\mathbb{E}_\theta\left[(\hat{\theta} - \theta)(\tilde{\theta} - \hat{\theta})^T\right] = 0$. Enfin, il est clair que $\mathbb{E}_\theta\left[(\tilde{\theta} - \hat{\theta})(\tilde{\theta} - \hat{\theta})^T\right]$ est une matrice semi-définie positive, d'où le théorème. □

Théorème 21. *Sous les hypothèses du Théorème 19, la statistique*

$$\hat{\sigma}^2 \stackrel{\text{def}}{=} \frac{\|\mathbf{y} - \mathbf{X}\hat{\theta}\|^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\theta})^2$$

est un estimateur sans biais de la variance σ^2, i.e.

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \ .$$

Démonstration. Notons d'abord que les observations $\mathbf{y}$ proviennent du modèle $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\eta}$, ce qui implique que $\mathbf{Y} - \mathbf{X}\hat{\theta} = \mathbf{X}(\theta - \hat{\theta}) + \boldsymbol{\eta}$. Vu (7.8), il en résulte que

$$\mathbf{Y} - \mathbf{X}\hat{\theta} = -\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\eta} + \boldsymbol{\eta} = (I_n - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T)\boldsymbol{\eta} = (I_n - H)\boldsymbol{\eta} \ , \tag{7.10}$$

avec $H = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$. En utilisant que $I_n - H$ est un projecteur orthogonal, on obtient

$$\mathbb{E}[\|\mathbf{Y} - \mathbf{X}\hat{\theta}\|^2] = \mathbb{E}[\boldsymbol{\eta}^T(I_n - H)^T(I_n - H)\boldsymbol{\eta}] = \mathbb{E}[\boldsymbol{\eta}^T(I_n - H)^2\boldsymbol{\eta}] = \mathbb{E}[\boldsymbol{\eta}^T(I_n - H)\boldsymbol{\eta}] \ .$$

Désignons par h_{ij} les éléments de H. On a alors

$$\begin{aligned}\mathbb{E}[\boldsymbol{\eta}^T(I_n - H)\boldsymbol{\eta}] &= \sum_{i,j=1}^{n}(\delta_{ij} - h_{ij})\mathbb{E}[\eta_i\eta_j] = \sigma^2\sum_{i,j=1}^{n}(\delta_{ij} - h_{ij})\delta_{ij} \\ &= \sigma^2\sum_{i=1}^{n}(1 - h_{ii}) = \sigma^2(n - \mathrm{Tr}(H)) \ ,\end{aligned}$$

où δ_{ij} est le symbole de Kronecker. Comme H est un projecteur, ses valeurs propres valent 0 ou 1. D'après la Proposition 11, $\mathrm{Rang}(H) = p$, donc il y a exactement p valeurs propres égales à 1. On en déduit que $\mathrm{Tr}(H) = p$, d'où le résultat. □

7.3 Vecteurs gaussiens

7.3.1 Définition et propriétés des vecteurs gaussiens

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

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

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

$$\phi(t) = \exp\left\{i\mu t - \frac{\sigma^2 t^2}{2}\right\} \ , \quad t \in \mathbb{R} \ ,$$

en particulier $\phi(t) = \mathrm{e}^{-t^2/2}$ pour la loi normale standard $\mathcal{N}(0, 1)$. 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. ξ suit la loi de Dirac en μ si $\mathbb{P}(\xi = \mu) = 1$, $\phi(t) = \mathrm{e}^{i\mu t}$.

On dit que $\xi = (\xi_1, \ldots, \xi_p)^T$ suit une **loi normale (multivariée)** dans $\mathbb{R}^p$ si et seulement si toute combinaison linéaire $a^T\xi$ est suit une loi normale dans $\mathbb{R}$.

On dit aussi que ξ est un **vecteur normal** dans $\mathbb{R}^p$ ou bien un **vecteur gaussien** dans $\mathbb{R}^p$.

Théorème 22. *Soit $\xi = (\xi_1, \ldots, \xi_p)^T$ un vecteur aléatoire avec $\mu = \mathbb{E}[\xi] = (\mathbb{E}[\xi_1], \ldots, \mathbb{E}[\xi_p])^T$ et $\Sigma = \mathbf{Var}(\xi) = (\mathbf{Cov}(\xi_i, \xi_j))_{i,j\in\{1,\ldots,p\}}$, alors ξ est un vecteur gaussien si et seulement si sa fonction caractéristique s'écrit pour tout $t \in \mathbb{R}^p$,*

$$\phi_\xi(t) = \mathbb{E}[\mathrm{e}^{it^T\xi}] = \exp\left\{it^T\mu - \frac{t^T\Sigma t}{2}\right\} \ .$$

La démonstration est une conséquence immédiate de la définition et du calcul de la fonction caractéristique d'une gaussienne réelle.

Une conséquence de cette proposition est que toute loi normale dans $\mathbb{R}^p$ est entièrement déterminée par la donnée de sa moyenne et de sa matrice de covariance, comme dans le cas d'une variable gaussienne réelle. Ceci explique que par la suite on utilisera la notation

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

pour un vecteur aléatoire normal ξ de moyenne μ et de matrice de covariance Σ.

Théorème 23. *Si la matrice de covariance d'un vecteur gaussien se sépare par blocs, alors les sous-vecteurs correspondants sont mutuellement indépendants : Soit* $\xi = (\xi_1^T, \dots, \xi_k^T)^T$ *un vecteur gaussien dans* $\mathbb{R}^p$ *avec des sous-vecteurs* $\xi_j \in \mathbb{R}^{m_j}$ *et* $\sum_{j=1}^k m_j = p$. *Si les matrices de covariance*

$$\mathbf{Cov}(\xi_j, \xi_l) = \mathbb{E}[(\xi_j - \mathbb{E}[\xi_j])(\xi_l - \mathbb{E}[\xi_l])^T] = \mathbf{0} \ , \qquad \textit{pour tout } j \neq l \ ,$$

où $\mathbf{0}$ *désigne la matrice nulle, alors les vecteurs* $\xi_1, \dots, \xi_k$ *sont mutuellement indépendants.*

La démonstration est immédiate avec la fonction caractéristique.

Théorème 24. *Toute transformation affine d'un vecteur normal est un vecteur normal : Soient* $\xi \sim \mathcal{N}_p(\mu, \Sigma)$, B *est une matrice déterministe* $q \times p$ *et* $c \in \mathbb{R}^q$, *alors*

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

Démonstration. La loi de $\eta = B\xi + c$ est normale car toute combinaison linéaire de $a^T\eta$ est une v.a. normale réelle. En effet, pour tout $a \in \mathbb{R}^q$,

$$a^T\eta = a^T B\xi + a^T c = b^T\xi + d \ ,$$

où $b = B^T a \in \mathbb{R}^p$ et $d = a^T c$. D'après la définition on obtient que les combinaisons linéaires $b^T\xi$ sont des v.a. normales pour tout $b \in \mathbb{R}^p$. Il s'ensuit que les combinaisons linéaires $a^T\eta$ sont normales pour tout $a \in \mathbb{R}^q$ et alors η est un vecteur normal dans $\mathbb{R}^q$. Sa moyenne et sa matrice de covariance sont données par

$$\mathbb{E}[\eta] = B\mu + c \ , \qquad \mathbf{Var}(\eta) = \mathbf{Var}(B\xi + c) = \mathbf{Var}(B\xi) = B\Sigma B^T \ . \qquad \square$$

Corollaire 3. *Tout sous-ensemble de coordonnées d'un vecteur normal est un vecteur normal : soit* $\xi = (\xi_1^T, \xi_2^T)^T$ *un vecteur normal dans* $\mathbb{R}^p$, *où* $\xi_1 \in \mathbb{R}^k$ *et* $\xi_2 \in \mathbb{R}^{p-k}$, *alors* ξ_1 *et* ξ_2 *sont des vecteurs normaux.*

Preuve. Soit B la matrice $k \times p$ de la forme $B = (I_k, 0)$, où I_k est la matrice unité $k \times k$. Clairement, $\xi_1 = B\xi$, et donc par la Proposition 24 ξ_1 est un vecteur gaussien. De même, on a $\xi_2 = B\xi$ pour $B = (0, I_{p-k})$, ce qui entraîne la normalité de ξ_2. $\square$

Théorème 25. *Soit* Σ *une matrice de covariance et* μ *un vecteur dans* $\mathbb{R}^p$. *Alors il existe une matrice déterministe* A *de taille* $p \times p$ *telle que pour un vecteur* η *de loi normale standard,*

$$A\eta + \mu \sim \mathcal{N}_p(\mu, \Sigma) \ .$$

Démonstration. Toute matrice de covariance Σ est symétrique et semi-définie positive, donc elle s'écrit sous la forme

$$\Sigma = MDM^T \ ,$$

où M est une matrice orthogonale et D est une matrice diagonale, non négative. On peut donc prendre la racine de $D = D^{1/2}$, puis celle de Σ sous la forme $\Sigma^{1/2} = MD^{1/2}M^T = A$. On remarque que $A = A^T$ et $AA^T = \Sigma$. On en conclut que, $A\eta + \mu$ est un vecteur normal de moyenne μ et de matrice de covariance $\mathbf{Var}(A\eta + \mu) = A\mathbf{Var}(\eta)A^T = AA^T = \Sigma$. $\square$

Théorème 26. *Si la matrice de covariance* Σ *est définie positive,* $\Sigma > 0$ *(*$\Leftrightarrow \mathrm{Det}(\Sigma) > 0$*), la loi normale* $\mathcal{N}_p(\mu, \Sigma)$ *admet la densité* f *par rapport à la mesure de Lebesgue dans* $\mathbb{R}^p$ *de la forme*

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

Démonstration. Soit $\eta = (\eta_1, \ldots, \eta_p)^T$ un vecteur de loi normale standard $\mathcal{N}_p(0, I_p)$. La forme de la fonction caractéristique de η implique que $\eta_1, \ldots, \eta_p$ sont des v.a. réelles indépendantes de loi normale standard $\mathcal{N}(0,1)$. Par conséquent, la loi de η admet pour densité

$$\varphi(t) = \prod_{i=1}^{n} f_{\mathcal{N}(0,1)}(t_i) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{t^T t}{2}\right\} , \quad t = (t_1, \ldots, t_p)^T \in \mathbb{R}^p .$$

D'après la Proposition 25, le vecteur $A\eta + \mu$, où $A = MD^{1/2}$, suit la loi normale $\mathcal{N}_p(\mu, \Sigma)$. Enfin, il suffit d'appliquer la formule du changement de variable (Corollaire 1). □

7.3.2 Lois dérivées de la loi normale

Loi chi-deux

Si $\eta = (\eta_1, \ldots, \eta_p)^T$ est un vecteur de loi normale standard $\mathcal{N}_p(0, I_p)$, alors la loi de la variable aléatoire $Y = \|\eta\|^2 = \sum_{i=1}^p \eta_i^2$ est dite la **loi chi-deux** à p degrés de liberté, et on écrit $Y \sim \chi_p^2$. La densité de la loi χ_p^2 est

$$f(y) = \frac{1}{2^{p/2}\Gamma(p)} y^{p/2-1} \mathrm{e}^{-y/2} \mathbb{1}\{y > 0\} ,$$

où $\Gamma(\cdot)$ est la fonction gamma $\Gamma(x) = \int_0^\infty u^{x-1}\mathrm{e}^{-u}\mathrm{d}u$, $x > 0$. La loi χ_p^2 est donc la loi Gamma $\Gamma(p/2, 1/2)$.

Pour $Y \sim \chi_p^2$, on a $\mathbb{E}[Y] = p$, $\mathbf{Var}(Y) = 2p$.

Loi de Fisher-Snedecor

Soit $U \sim \chi_p^2$, $V \sim \chi_q^2$, deux v.a. indépendantes. La **loi de Fisher-Snedecor** à p et q degrés de liberté est la loi de la variable aléatoire

$$Y = \frac{U/p}{V/q} .$$

On écrit alors $Y \sim F_{p,q}$.

La densité de la loi $F_{p,q}$ est

$$f(y) = \frac{p^{p/2} q^{q/2}}{B(p/2, q/2)} \frac{y^{p/2-1}}{(q + py)^{\frac{p+q}{2}}} \mathbb{1}\{y > 0\} .$$

où $B(\cdot,\cdot)$ est la fonction Beta $B(r,s) = \Gamma(r)\Gamma(s)/\Gamma(r+s)$.

On peut montrer que cette densité converge vers une densité de type χ_p^2 quand $q \to \infty$.

Loi de Student

Soit $\eta \sim \mathcal{N}(0,1)$, $\xi \sim \chi_q^2$ 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 la loi t_q est

$$f(y) = \frac{1}{\sqrt{q}B(1/2, q/2)}(1 + y^2/q)^{-(q+1)/2} , \quad y \in \mathbb{R} .$$

Cette loi est symétrique.

Pour $q = 1$, la loi de Student t_1 est la loi de Cauchy.

Si $Y \sim t_q$, alors Y^2 suit la loi $F_{1,q}$.

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.

7.3.3 Théorème de Cochran

Théorème 27 (Théorème de Cochran)**.** *Soit $\xi \sim \mathcal{N}_p(0, I)$ et soient $A_1, \ldots, A_J$, $J \leq p$, des matrices $p \times p$ telles que $A_j^2 = A_j$, A_j est symétrique,* $\mathrm{Rang}(A_j) = N_j$*, $A_j A_k = 0$ pour $j \neq k$ et $\sum_{j=1}^J N_j \leq p$. Alors,*

(i) les vecteurs aléatoires $A_j\xi, j = 1, \ldots, J$, sont mutuellement indépendants de lois $\mathcal{N}_p(0, A_j), j = 1, \ldots, J$, respectivement,

(ii) les variables aléatoires $\|A_j\xi\|^2, j = 1, \ldots, J$, sont mutuellement indépendantes de lois $\chi^2_{N_j}, j = 1, \ldots, J$, respectivement.

Démonstration. (i) Notons d'abord que la loi jointe du vecteur aléatoire

$$\begin{pmatrix} A_1\xi \\ \vdots \\ A_J\xi \end{pmatrix} = \begin{pmatrix} A_1 \\ \vdots \\ A_J \end{pmatrix} \xi$$

est normale. De plus, $\mathbb{E}[A_j\xi] = 0$ et

$$\mathbf{Var}(A_j\xi) = A_j\mathbf{Var}(\xi)A_j^T = A_jA_j^T = A_j^2 = A_j .$$

Par ailleurs,

$$\mathbf{Cov}(A_k\xi, A_j\xi) = \mathbb{E}[A_k\xi\xi^T A_j^T] = A_k\mathbf{Var}(\xi)A_j^T = A_kA_j^T = A_kA_j = 0 ,$$

pour $j \neq k$. D'après la Proposition 23, on obtient alors que $A_1\xi, \ldots, A_J\xi$ sont mutuellement indépendants.

(ii) Comme A_j est symétrique, il existe une matrice Γ orthogonale telle que $A_j = \Gamma\Lambda\Gamma^T$, où $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_p)$ est la matrice diagonale des valeurs propres de A_j. Alors,

$$\|A_j\xi\|^2 = \xi^T A_j^T A_j\xi = \xi^T A_j\xi = (\xi^T\Gamma)\Lambda(\Gamma^T\xi) = \eta^T\Lambda\eta = \sum_{i=1}^p \lambda_i\eta_i^2 ,$$

où $\eta = \Gamma^T\xi = (\eta_1, \ldots, \eta_p)^T$. En utilisant l'orthogonalité de Γ, on vérifie que η est un vecteur normal de loi $\mathcal{N}_p(0, I)$. En effet,

$$\mathbb{E}[\eta] = \Gamma^T\mathbb{E}[\xi] = 0 \quad \text{et} \quad \mathbf{Var}(\eta) = \Gamma^T\mathbf{Var}(\xi)\Gamma = \Gamma^T\Gamma = I_p .$$

Or, A_j est un projecteur orthogonal, donc $\lambda_j \in \{0, 1\}$ et $\mathrm{Card}\{j : \lambda_j = 1) = \mathrm{Rang}(A_j) = N_j$, d'où il découle que $\|A_j\xi\|^2 \sim \chi^2_{N_j}$. Finalement, la partie (i) du théorème et le fait que les transformations mesurables préservent l'indépendance impliquent que les variables aléatoires $\|A_1\xi\|^2, \ldots, \|A_J\xi\|^2$ sont mutuellement indépendantes.

□

7.4 Modèle linéaire gaussien

Le modèle de régression linéaire multidimensionnelle est dit **modèle linéaire gaussien** si les erreurs η_i sont des variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma^2)$. Autrement dit, on suppose que $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ est un vecteur gaussien de loi

$$\boldsymbol{\eta} \sim \mathcal{N}_n(0, \sigma^2 I_n) \; .$$

Dans le modèle linéaire gaussien, l'estimateur par la méthode des moindres carrés coïncide avec l'estimateur du maximum de vraisemblance du paramètre θ.

Le théorème suivant permet de déduire la loi jointe de $(\hat{\theta}, \hat{\sigma}^2)$ dans le modèle linéaire gaussien.

Théorème 28. *Sous les hypothèses du Théorème 19, si en plus $\boldsymbol{\eta}$ est un vecteur gaussien, alors*

(i) $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$,

(ii) $\hat{\theta} \perp\!\!\!\perp \mathbf{Y} - \mathbf{X}\hat{\theta}$ *et* $\mathbf{Y} - \mathbf{X}\hat{\theta} \perp\!\!\!\perp \mathbf{X}(\hat{\theta} - \theta)$,

(iii) $\sigma^{-2}\|\mathbf{Y} - \mathbf{X}\hat{\theta}\|^2 \sim \chi^2_{n-p}$ *et* $\sigma^{-2}\|\mathbf{X}(\hat{\theta} - \theta)\|^2 \sim \chi^2_p$.

Démonstration. D'après (7.8) et (7.10),

$$\hat{\theta} - \theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta} \; , \qquad \mathbf{Y} - \mathbf{X}\hat{\theta} = (I_n - H)\boldsymbol{\eta} \; . \tag{7.11}$$

La première égalité, compte tenu du fait que $(\mathbf{X}^T\mathbf{X})$ et $\mathbf{X}$ sont déterministes, implique que $\hat{\theta}$ est un vecteur gaussien. D'après le Théorème 19, la moyenne de ce vecteur est θ et sa matrice de covariance vaut $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, d'où le résultat *(i)*.

Vu (7.11), le vecteur aléatoire $\left(\mathbf{Y} - \mathbf{X}\hat{\theta}, \hat{\theta}\right) \in \mathbb{R}^{n+p}$ est gaussien comme transformation affine du vecteur gaussien $\boldsymbol{\eta}$. De plus, la matrice de covariance entre $\hat{\theta}$ et $\mathbf{Y} - \mathbf{X}\hat{\theta}$ est

$$\begin{aligned}
\mathbf{Cov}(\hat{\theta}, \mathbf{Y} - \mathbf{X}\hat{\theta}) &= \mathbb{E}\left[(\hat{\theta} - \theta)(\mathbf{Y} - \mathbf{X}\hat{\theta})^T\right] \\
&= \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta}\boldsymbol{\eta}^T(I_n - H)\right] \\
&= \sigma^2((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T H) \\
&= 0 \; ,
\end{aligned}$$

ceci implique l'indépendance de $\hat{\theta}$ et $\mathbf{Y} - \mathbf{X}\hat{\theta}$, ce qui est 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 $\boldsymbol{\eta}' = \boldsymbol{\eta}/\sigma$ et appliquons le Théorème de Cochran (Théorème 27). D'après (7.11), $\mathbf{Y} - \mathbf{X}\hat{\theta} = \sigma(I_n - H)\boldsymbol{\eta}'$ et $\mathbf{X}(\hat{\theta} - \theta) = \sigma\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta}' = \sigma H\boldsymbol{\eta}'$. Par ailleurs, la Proposition 11 implique que les matrices H et $I_n - H$ sont symétriques et idempotentes, $(I_n - H)H = 0$, $\mathrm{Rang}(H) = p$ et $\mathrm{Rang}(I_n - H) = n - p$. D'après le Théorème de Cochran,

$$\frac{\|\mathbf{Y} - \mathbf{X}\hat{\theta}\|^2}{\sigma^2} = \|(I_n - H)\boldsymbol{\eta}'\|^2 \sim \chi^2_{n-p} \; , \qquad \frac{\|\mathbf{X}(\hat{\theta} - \theta)\|^2}{\sigma^2} = \|H\boldsymbol{\eta}'\|^2 \sim \chi^2_p \; .$$

□

Chapitre 8

Estimation par intervalle

Dans les chapitres précédents, nous avons mis en évidence le rôle d'un estimateur en tant que pourvoyeur d'une "approximation" de la valeur inconnue du paramètre θ. Cela étant, une estimation sans degré de précision est douteuse, dans la mesure où elle est variable (il suffit d'ajouter ou de retrancher une observation pour changer sa valeur). Ainsi, lorsqu'un statisticien propose, au vu des observations $x_1, \ldots, x_n$, une estimation $\hat{\theta}_n$ de θ, quelle confiance peut-il avoir en son résultat ? Lorsque l'estimateur est consistant, tout ce qu'on sait, c'est que plus n est grand, plus $\hat{\theta}_n$ a des chances d'être voisin de θ.

Prenons comme exemple la moyenne empirique comme estimateur de l'espérance μ d'une loi (intégrable) F. Supposons que la moyenne empirique vaut 5,2. Intuitivement, nous accordons à cette estimation beaucoup plus de confiance lorsque la taille d'échantillon n est grande (p. ex. $n = 100\,000$) que quand elle est petite (p. ex. $n = 3$). À part de la taille d'échantillon, d'autres facteurs peuvent influencer la précision d'une estimation $\hat{\theta}_n$, comme par exemple la variance ou la forme de la loi F. Nous voyons l'importance de savoir quantifier la précision ou volatilité d'un estimateur.

Dans ce chapitre nous présentons d'abord l'*erreur standard* d'un estimateur, qui est une mesure d'incertitude d'un estimateur, et son calcul en pratique par le *bootstrap*. Ensuite, nous introduisons la notion d'*intervalle de confiance*. L'idée consiste à calculer tout un intervalle (par opposition à un estimateur ponctuel) qui est susceptible de contenir la vraie valeur du paramètre θ avec une certaine probabilité prescrite. Plusieurs techniques de calcul d'intervalles de confiance par le bootstrap sont également présentés.

8.1 Erreur standard

Considérons une loi F et un paramètre ou une quantité $\theta = t(F)$ qui dépend de la loi F. Soit $\mathbf{x}$ une réalisation du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ où les variables aléatoires X_i sont i.i.d. de loi F. Notons $\hat{\theta}$ un estimateur *consistant* de θ. Rappelons que tout estimateur est une fonction (mesurable) définie sur les données $\mathbf{x}$, et de ce fait, on peut également écrire $\hat{\theta}(\mathbf{x})$ ou $\mathbf{x} \mapsto \hat{\theta}(\mathbf{x})$ pour désigner l'estimateur. Pour souligner que l'on considère la variable aléatoire $\hat{\theta}$ on écrit $\hat{\theta}(\mathbf{X})$. Une mesure de la variabilité ou incertitude de l'estimateur $\hat{\theta}$ est donnée par l'**erreur standard** ou l'**erreur type** (en anglais *standard error*), notée $\mathrm{se}(\hat{\theta})$, définie par

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathbf{Var}(\hat{\theta})} = \sqrt{\mathbf{Var}(\hat{\theta}(\mathbf{X}))}\,.$$

Parfois nous notons $\mathrm{se}_F(\hat{\theta}(\mathbf{X}))$ pour souligner qu'il s'agit de l'estimateur $\hat{\theta}$ appliquer à un échantillon $\mathbf{X}$ de la loi F.

L'interprétation de l'erreur standard d'un estimateur consistant est simple : Plus l'erreur standard $\mathrm{se}(\hat{\theta})$ est petite, plus $\hat{\theta}$ est précis comme estimateur de θ.

8.1.1 Cas de la moyenne empirique

Considérons d'abord le cas particulier de la moyenne empirique. Soit F une loi de carré intégrable, de moyenne $\mu = \mathbb{E}_F[X]$ et de variance $\sigma^2 = \mathbf{Var}_F(X) < \infty$. L'erreur standard de la moyenne empirique $\bar{X}_n$, qui est un estimateur consistant de μ, est donnée par

$$\mathrm{se}(\bar{X}_n) = \sqrt{\mathbf{Var}(\bar{X}_n)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{Var}(X_i)} = \frac{\sigma}{\sqrt{n}} \; . \tag{8.1}$$

Clairement, on observe que l'erreur standard $\mathrm{se}(\bar{X}_n)$ est décroissante en n. Par conséquent, plus le nombre n d'observations est grand, plus la moyenne empirique est un estimateur précis de μ. L'impact de la variance σ^2 sur la valeur de l'erreur standard $\mathrm{se}(\bar{X}_n)$ est également évident : Plus la variance de la loi F des observations est élevée, moins l'estimateur $\bar{X}_n$ est précis.

La formule (8.1) indique comment choisir la taille d'échantillon n pour atteindre une certaine précision. Par exemple, afin de diviser l'erreur standard de l'estimateur $\bar{X}_n$ par deux, il faut quatre fois plus de données.

L'erreur standard $\mathrm{se}(\bar{X}_n)$ peut aussi servir pour établir un intervalle qui contient la vraie valeur de μ avec une certaine probabilité. En fait, la probabilité que l'écart entre μ et $\bar{X}_n$ ne dépasse pas la valeur de l'erreur standard $\mathrm{se}(\bar{X}_n)$ est de 0.68. Ou encore, la chance que la moyenne empirique $\bar{X}_n$ se trouve à une distance d'au plus deux fois la valeur de l'erreur standard $\mathrm{se}(\bar{X}_n)$ est de 95%. En langage probabiliste,

$$\mathbb{P}\left(\mu \in [\bar{X}_n - \mathrm{se}(\bar{X}_n), \bar{X}_n + \mathrm{se}(\bar{X}_n)]\right) \approx 0.68 \tag{8.2}$$

$$\mathbb{P}\left(\mu \in [\bar{X}_n - 2\mathrm{se}(\bar{X}_n), \bar{X}_n + 2\mathrm{se}(\bar{X}_n)]\right) \approx 0.95 \; . \tag{8.3}$$

Montrons (8.2) et (8.3). En effet, dans le cas particulier où F est la loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ on sait que

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} = \frac{\bar{X}_n - \mu}{\mathrm{se}(\bar{X}_n)} \sim \mathcal{N}(0,1) \; .$$

Dans le cas plus général, lorsque la loi F n'est pas gaussienne, le théorème central limite implique que $\bar{X}_n$ suit approximativement une loi gaussienne lorsque n est grand. Plus précisément,

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} = \frac{\bar{X}_n - \mu}{\mathrm{se}(\bar{X}_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \; .$$

De ce fait on déduit que, lorsque n est grand,

$$\mathbb{P}\left(\mu \in [\bar{X}_n - \mathrm{se}(\bar{X}_n), \bar{X}_n + \mathrm{se}(\bar{X}_n)]\right) = \mathbb{P}\left(\frac{|\bar{X}_n - \mu|}{\mathrm{se}(\bar{X}_n)} < 1\right) \approx \mathbb{P}(|Z| < 1) = 0.68269 \; ,$$

où Z est une v.a. de loi $\mathcal{N}(0,1)$. De même,

$$\mathbb{P}(|\bar{X}_n - \mu| < 2\mathrm{se}(\bar{X}_n)) \approx \mathbb{P}(|Z| < 2) = 0.95450 \; .$$

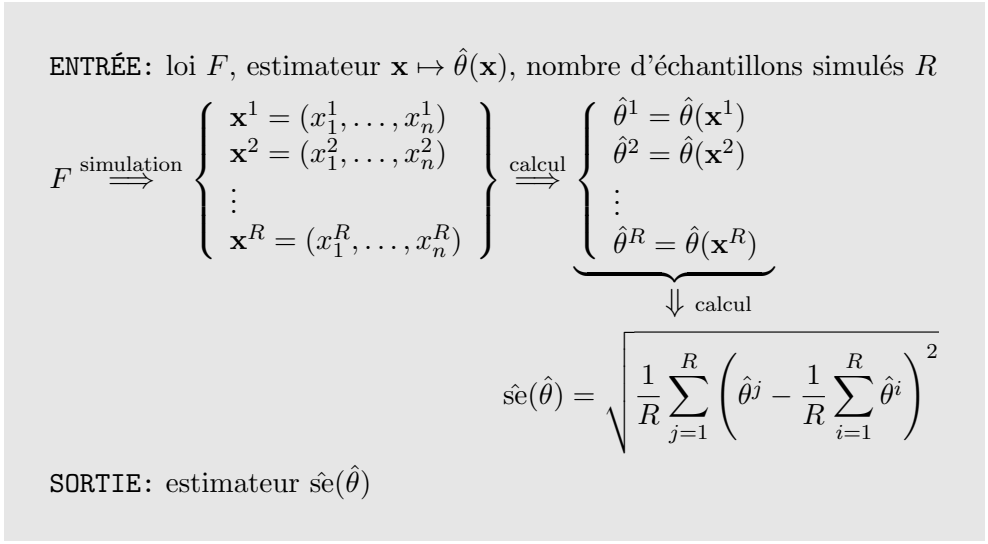


FIGURE 8.1 – Schéma des simulations de Monte Carlo pour l'estimation de l'erreur standard $\text{se}(\hat{\theta})$ d'un estimateur $\hat{\theta}$.

8.1.2 Erreur standard par simulation de Monte Carlo

Le cas de la moyenne empirique est particulier dans le sens que c'est quasi l'unique estimateur pour lequel l'erreur standard se calcule de façon explicite. En général, il s'avère très difficile voir impossible de trouver une expression explicite de l'erreur standard. Pensez par exemple à la médiane empirique, dont le calcul des moments n'est pas évident.

Une façon très pratique et simple à mettre en œuvre sur un ordinateur pour obtenir une *approximation* de l'erreur standard $\text{se}(\hat{\theta})$ d'un estimateur $\hat{\theta}$ repose sur des simulations de Monte Carlo. L'idée consiste à considérer la variable aléatoire $\hat{\theta}(\mathbf{X})$ et de générer un grand échantillon de cette variable aléatoire, disons $(\hat{\theta}^1, \ldots, \hat{\theta}^R)$. Ensuite, il suffit de calculer l'écart type empirique de cet échantillon pour obtenir une approximation de l'erreur standard $\text{se}(\hat{\theta})$. Plus précisément, on calcule

$$\hat{\text{se}}_R(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{j=1}^{R}\left(\hat{\theta}^j - \frac{1}{R}\sum_{i=1}^{R}\hat{\theta}^i\right)^2} \,.$$

Reste à savoir comment fabriquer des réalisations $\hat{\theta}^j$ de $\hat{\theta}(\mathbf{X})$ quand la loi de $\hat{\theta}(\mathbf{X})$ n'est pas forcément connue. La réponse est de générer des réalisations $\mathbf{x}$ de $\mathbf{X}$ et puis évaluer l'estimateur $\hat{\theta}$ sur ces réalisations $\mathbf{x}$. Figure 8.1 explique la démarche par un schéma.

D'après la loi des grands nombres, l'approximation $\hat{\text{se}}_R(\hat{\theta})$ obtenue par la méthode de Monte Carlo converge en probabilité vers la vraie valeur de l'erreur standard $\text{se}(\hat{\theta})$ lorsque le nombre R tend vers l'infini (pourvu que la loi de $\hat{\theta}(\mathbf{X})$ est de carré intégrable) :

$$\hat{\text{se}}_R(\hat{\theta}) \xrightarrow{P} \text{se}(\hat{\theta}) \,, \qquad R \to \infty \,.$$

En choissisant un nombre R très important, l'utilisateur peut rendre l'erreur d'approximation de $\text{se}(\hat{\theta})$ par $\hat{\text{se}}_R(\hat{\theta})$ négligeable. En revanche, plus R est grand, plus le temps de calcul est long. En général, $R = 1000$ est convenable.

Exemple

Supposons que l'on observe $\mathbf{x} = (x_1, \ldots, x_n)$ où les x_i sont des réalisations i.i.d. de la variable aléatoire $X = T + \theta$ où T suit la loi de Student t_q avec q degrés de liberté et le paramètre de position $\theta \in \mathbb{R}$ est inconnu. Notons que la loi de X est symétrique par rapport à θ, puisque la loi de Student est symétrique.

Au Chapitre 3 nous avons vu plusieurs estimateurs d'un paramètre de position, notamment la moyenne empirique $\bar{x}_n$, la médiane empirique $x_{(\lceil n/2 \rceil)}$ et la moyenne tronquée $\bar{x}_{\text{tronq}(\gamma)}$. Fixons $\gamma = 0.1$ et notons les trois estimateurs de θ par

$$\hat{\theta} = \bar{x}_n \ , \qquad \tilde{\theta} = x_{(\lceil n/2 \rceil)} \ , \qquad \check{\theta} = \bar{x}_{\text{tronq}(0.1)} \ .$$

Afin de comparer ces trois estimateurs de θ par leurs erreurs standard, nous procédons de la façon suivante :

1. Tout d'abord on fixe les valeurs des paramètres θ et q et la taille d'échantillon n. Posons $i = 1$.
2. Ensuite on génère un échantillon, noté $\mathbf{x}^i = (x_1^i, \ldots, x_n^i)$, de taille n de la loi de $X = T + \theta$ où T suit la loi de Student t_q.
3. On évalue les trois estimateurs $\hat{\theta}, \tilde{\theta}, \check{\theta}$ pour l'échantillon $\mathbf{x}^i$:

   $$\hat{\theta}^i = \hat{\theta}(\mathbf{x}^i) \ , \quad \tilde{\theta}^i = \tilde{\theta}(\mathbf{x}^i) \ , \quad \check{\theta}^i = \check{\theta}(\mathbf{x}^i) \ .$$

   Posons $i = i + 1$.
4. On répète 2. et 3. R fois pour une valeur de R assez grande, donnant lieu à trois échantillons de taille R des estimateurs $\hat{\theta}, \tilde{\theta}, \check{\theta}$:

   $$(\hat{\theta}^1, \ldots, \hat{\theta}^R) \quad (\tilde{\theta}^1, \ldots, \tilde{\theta}^R) \quad (\check{\theta}^1, \ldots, \check{\theta}^R) \ .$$

5. On calcule les écarts types empiriques associés à ces trois échantillons d'estimateurs :

   $$\hat{\text{se}}_R(\hat{\theta}) = \sqrt{\frac{1}{R} \sum_{j=1}^{R} \left(\hat{\theta}^j - \frac{1}{R} \sum_{i=1}^{R} \hat{\theta}^i \right)^2} \ .$$

   De même on calcule $\hat{\text{se}}_R(\tilde{\theta})$ et $\hat{\text{se}}_R(\check{\theta})$.

La mise en œuvre des simulations de Monte Carlo sous `R` est assez simple :

```
se.MC ← function(theta, df.t, nb.obs, REP){
# calcul des approximations des erreurs standard associees a la
   moyenne empirique, la mediane et la moyenne tronquee quand les
   observations sont des realisations d'une loi de Student decalee
# ENTREE: theta - parametre de position ; df.t - degre de liberte
   de la loi de Student ; nb.obs - taille d'echantillon ; REP -
   nombre d'echantillons simules
# SORTIE : erreurs standard associees a la moyenne empirique, la
   mediane et la moyenne tronquee

   # pour la moyenne tronquee :
   gamma = .1
   nb.tronq ← round(gamma/2* nb.obs)
   ind.tronq ← (nb.tronq+1):(nb.obs-nb.tronq)

   estim.moy ← rep(0,REP)
   estim.med ← rep(0,REP)
   estim.tronq ← rep(0,REP)
```

```
   for (i in 1:REP)
   {  # generer un echantillon de la loi de Student + theta :
      data <- rt(nb.obs,df.t)+theta
      # calcul des 3 estimateurs :
      estim.moy[i] <- mean(data)
      estim.med[i] <- median(data)
      data <- sort(data)
      estim.tronq[i] <- mean(data[ind.tronq])
   }
   # calcul des erreurs standard approchees :
   std.err.moy <- sd(estim.moy)
   std.err.med <- sd(estim.med)
   std.err.tronq <- sd(estim.tronq)
   return(c(std.err.moy,std.err.med,std.err.tronq))
}
```

Dans les résultats ci-dessous, le paramètre θ est fixé à 5, la taille d'échantillon n vaut 100 et on génère `REP`=1000 échantillons à chaque appel. Quant au degré de liberté q de la loi de Student, il varie entre 1 et 50.

```
> se.MC(5, 1, 100, 1000)
[1] 23.2060113  0.1598392  0.3338510
> se.MC(5, 1, 100, 1000)
[1] 46.6537950  0.1634105  0.3283516
> se.MC(5, 3, 100, 1000)
[1] 0.1840373 0.1318830 0.1327447
> se.MC(5, 3, 100, 1000)
[1] 0.1726544 0.1379129 0.1335570
> se.MC(5, 5, 100, 1000)
[1] 0.1264699 0.1312139 0.1155959
> se.MC(5, 10, 100, 1000)
[1] 0.1114037 0.1229353 0.1086758
> se.MC(5, 50, 100, 1000)
[1] 0.1016971 0.1246134 0.1022908
```

En fait, pour $q = 1$ la loi de Student conincïde avec la loi de Cauchy, qui est une loi à queues lourdes et qui n'est pas intégrable. Seulement à partir de $q > 2$ la loi de Student est de variance finie. De plus, quand q tend vers l'infini la loi de Student converge vers la loi normale standard qui est une loi à queues légères.

Les résultats ci-dessus montrent que les trois estimateurs ont des comportements différents en fonction des queues de la loi. En effet, la médiane est nettement le meilleur estimateur pour la loi de Cauchy en terme d'erreur standard, alors que la moyenne empirique a une petite avance sur les deux autres pour des degrés de liberté élevées, c'est-à-dire pour des lois à queues légères. La moyenne tronquée a une très bonne performance dans tous les cas. On peut la considérer comme la version robuste de la moyenne adaptée aux lois à queues lourdes.

Sachant que la variance de T avec $T \sim t_q$ est donnée par $\mathbf{Var}(T) = q/(q-2)$, l'erreur standard de la moyenne empirique dans cet exemple est donnée par

$$\mathrm{se}(\bar{X}_n) = \sqrt{\frac{q}{(q-2)n}}\,.$$

Comparons les valeurs obtenues par des simulations de Monte Carlo ci-dessus avec les valeurs théoriques des erreurs standard. On observe que les estimations $\hat{\mathrm{se}}_R(\hat{\theta})$ sont très proche des vraies valeurs :

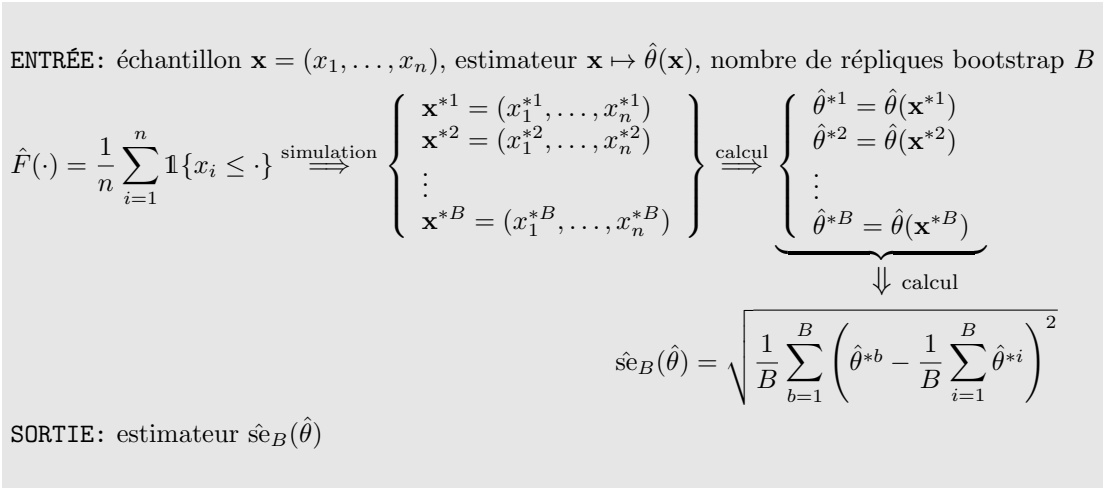


FIGURE 8.2 – Schéma bootstrap pour estimer l'erreur standard $\mathrm{se}(\hat{\theta})$ d'un estimateur $\hat{\theta}$.

```
> err.typ.moy ← function(df.t,nb.obs) return(sqrt(df.t/(df.t-2)/
    nb.obs))
> err.typ.moy(3,100)
[1] 0.1732051
> err.typ.moy(5,100)
[1] 0.1290994
> err.typ.moy(10,100)
[1] 0.1118034
> err.typ.moy(50,100)
[1] 0.1020621
```

8.1.3 Erreur standard par le Bootstrap

Nous venons de voir que lorsque la loi F des données est *connue*, on peut procéder par des simulations de Monte Carlo pour approcher l'erreur standard d'un estimateur. Mais comment faire quand F est inconnue, ce qui est généralement le cas dans des applications ? Pour revenir à l'exemple précédent : que faire quand le degré de liberté q de la loi de Student est inconnu ? Comment savoir le quel des trois estimateurs considérés est préférable sur un échantillon observé ? En fait, on peut adapter la méthode de Monte Carlo à ce scénario. Cette procédure est appelée le **bootstrap** et elle est décrite dans ce paragraphe.

En toute généralité, le bootstrap est une *méthode de rééchantillonnage* pour approcher des caractéristiques de la loi d'un estimateur quand un calcul explicite s'avère difficile. Ici nous verrons qu'elle permet d'approcher l'erreur standard d'un estimateur.

Comme précédemment, notons $\mathbf{x} = (x_1, \ldots, x_n)$ un échantillon i.i.d. de loi F et $\hat{\theta}$ un estimateur d'un paramètre θ. Maintenant la loi F est supposée inconnue. Le but est de déterminer l'erreur standard de $\hat{\theta}$ à partir d'un échantillon $\mathbf{x}$. Bien que F soit inconnue, on peut approcher la loi F par la fonction de répartition empirique $\hat{F}$ associée aux observations $\mathbf{x}$. L'idée du bootstrap consiste à effectuer des simulations de Monte Carlo *en simulant de la loi empirique* $\hat{F}$ (au lieu de la loi F).

La démarche du bootstrap est la suivante :

1. On détermine la loi empirique $\hat{F}$ associée à l'échantillon observé $\mathbf{x}$. Posons $b = 1$.
2. Ensuite on génère une réalisation, notée $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$, de $\mathbf{X}$ de la loi empirique $\hat{F}$.

3. On évalue l'estimateur $\hat{\theta}$ pour l'échantillon $\mathbf{x}^{*b}$: $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{x}^{*b})$. Posons $b = b + 1$.
4. On répète 2. et 3. B fois pour une valeur de B assez grande, donnant lieu à un échantillon de taille B de l'estimateur $\hat{\theta}^* = \hat{\theta}(\mathbf{X})$ où $\mathbf{X}$ est de la loi empirique $\hat{F}$:

$$(\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*R}) \, .$$

5. On calcule l'écart type empirique associé à cet échantillon d'estimateurs :

$$\hat{\text{se}}_B(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^{*i}\right)^2} \, .$$

Cette démarche est également illustrée par la Figure 8.2. On remarque la grande similitude avec le schéma des simulations de Monte Carlo de la Figure 8.1.

Il est courant d'utiliser des étoiles * pour désigner des objets créés lors de la procédure bootstrap. En particulier, $\mathbf{x}^{*b}$ désigne un échantillon de la loi empirique $\hat{F}$, dit *échantillon bootstrap*, et $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{x}^{*b})$ est appelé une *réplique bootstrap* de l'estimateur $\hat{\theta}$.

Comment obtenir un échantillon bootstrap ? Autrement dit, comment générer des réalisations de la loi $\hat{F}$ associée aux observations $\mathbf{x} = (x_1, \ldots, x_n)$? Notons que $\hat{F}$ est la loi discrète à valeurs dans $\{x_1, \ldots, x_n\}$ qui associe le poids $1/n$ à chaque observation x_i. Plus précisément, soit X^* une variable aléatoire de la loi $\hat{F}$. Si toutes les observations x_i sont deux à deux différentes, on a

$$\mathbb{P}_{\hat{F}}(X^* = x_i) = \frac{1}{n} \, , \quad i = 1, \ldots, n \, .$$

En conséquence, simuler une réalisation de la fonction de répartition empirique $\hat{F}$ est équivalent à tirer un point x_i au hasard dans l'échantillon observé $\mathbf{x}$. Autrement dit, on génère alors un échantillon $x_1^*, \ldots, x_n^*$ de la loi $\hat{F}$ par tirage avec remise de n valeurs dans l'échantillon observé $\mathbf{x}$.

Dans le cas général, si on admet la possibilité que l'échantillon $\mathbf{x}$ contient certaines valeurs plusieurs fois, (ce qui est le cas lorsque F est une loi discrète), on a

$$\mathbb{P}_{\hat{F}}(X^* = x_i) = \frac{\#\{j : x_i = x_j\}}{n} \, , \quad i = 1, \ldots, n \, .$$

Un échantillon bootstrap est alors composé d'autant de valeurs que l'échantillon original à valeurs issues de l'échantillon original $\mathbf{x}$, mais avec des fréquences potentiellement différentes. On parle souvent de *rééchantillonnage* car on reconstruit un ensemble d'échantillons en partant de l'échantillon observé.

Sous `R` on peut tout simplement utiliser la fonction `sample()` pour générer des échantillons bootstrap.

En effet, l'idée du bootstrap repose sur la méthode de substitution du Chapitre 5 en remplaçant F par sa version empirique $\hat{F}$. D'après ce principe, un estimateur de l'erreur standard $\text{se}(\hat{\theta}) = \text{se}_F(\hat{\theta}(\mathbf{X}))$ (l'erreur standard de $\hat{\theta}(\mathbf{X})$ lorsque $\mathbf{X}$ est un échantillon de la loi F) est donné par sa version empirique $\hat{\text{se}}_\infty(\hat{\theta}) = \text{se}_{\hat{F}}(\hat{\theta}(\mathbf{X}))$ (l'erreur standard de $\hat{\theta}(\mathbf{X})$ lorsque $\mathbf{X}$ est un échantillon de la loi $\hat{F}$). Or, en général, l'estimateur $\hat{\text{se}}_\infty(\hat{\theta})$ n'est pas disponible car la loi de $\hat{\theta}(\mathbf{X})$ où $\mathbf{X}$ suit la loi empirique $\hat{F}$ est inconnue. Le bootstrap propose d'approcher $\hat{\text{se}}_\infty(\hat{\theta})$ par l'estimateur bootstrap $\hat{\text{se}}_B(\hat{\theta})$. En fait, pour un échantillon $\mathbf{x}$ fixé, on a

$$\hat{\text{se}}_B(\hat{\theta}) \xrightarrow{P} \hat{\text{se}}_\infty(\hat{\theta}) \, , \quad B \to \infty \, .$$

Parfois, on appelle $\hat{\text{se}}_\infty(\hat{\theta})$ l'*estimateur bootstrap idéal*.

En résumé, en utilisant le bootstrap on effectue deux approximations de l'erreur standard $\mathrm{se}(\hat{\theta})$. Plus précisément,

$$\mathrm{se}(\hat{\theta}) = \mathrm{se}_F(\hat{\theta}(\mathbf{X})) \underbrace{\approx}_{\text{petit si } n \text{ est grand}} \hat{\mathrm{se}}_\infty(\hat{\theta}) = \mathrm{se}_{\hat{F}}(\hat{\theta}(\mathbf{X})) \underbrace{\approx}_{\text{petit si } B \text{ est grand}} \hat{\mathrm{se}}_B(\hat{\theta}) \,.$$

La précision de la première approximation est déterminée par la taille n de l'échantillon. En revanche, la précision de la deuxième approximation dépend de B, le nombre d'échantillons bootstrap générés par l'algorithme, choisi par le statisticien. Il est donc possible de rendre cette erreur d'approximation négligeable. Seul inconvénient, plus B est grand, plus le temps de calcul est long.

En pratique, il est suffisant de choisir le nombre B de répliques bootstrap entre 25 et 200. C'est suffisant dans le contexte de l'approximation de l'erreur standard.

Exemple

Reprenons l'exemple de la Section 8.1. La fonction suivante met en œuvre le bootstrap afin d'obtenir des approximations des erreurs standard de la moyenne empirique, de la médiane et de la moyenne tronquée.

```
se.boot ← function(data, B){
# calcul des approximations des erreurs standard associees a la
   moyenne empirique, la mediane et la moyenne tronquee par le
   bootstrap
# ENTREE: data - echantillon observe ; B - nombre de replique
   bootstrap
# SORTIE : erreurs standard associees a la moyenne empirique, la
   mediane et la moyenne tronquee

   nb.obs ← length(data)
   # pour la moyenne tronquee :
   gamma = .1
   nb.tronq ← round(gamma/2* nb.obs)
   ind.tronq ← (nb.tronq+1):(nb.obs-nb.tronq)

   estim.moy ← rep(0,B)
   estim.med ← rep(0,B)
   estim.tronq ← rep(0,B)
   for (i in 1:B)
   {  # generer un echantillon bootstrap :
      data.boot ← sample(data, nb.obs, replace=TRUE)
      # calcul des 3 estimateurs :
      estim.moy[i] ← mean(data.boot)
      estim.med[i] ← median(data.boot)
      data.boot ← sort(data.boot)
      estim.tronq[i] ← mean(data.boot[ind.tronq])

   }
   # calcul des erreurs standard approchees :
   std.err.moy ← sd(estim.moy)
   std.err.med ← sd(estim.med)
   std.err.tronq ← sd(estim.tronq)
   return(c(std.err.moy,std.err.med,std.err.tronq))
}
```

Les deux fonctions `se.MC()` et `se.boot()` sont très similaires. La différence principale est située à la première ligne de la boucle `for` dans la génération des échantillons `data` resp.

`data.boot`.

L'exemple suivant montre qu'avec un échantillon de taille 100 on arrive à estimer les erreurs standard des trois estimateurs avec une très bonne précision sans connaître la loi F. On peut les comparer aux valeurs obtenues par des simulations de Monte Carlo auparavant.

```
> data ← rt(100,1)+5
> se.boot(data,100)
[1] 7.6283988 0.1560726 0.4588739
> data ← rt(100,1)+5
> se.boot(data,100)
[1] 0.8438882 0.1435884 0.3354507
> data ← rt(100,3)+5
> se.boot(data,100)
[1] 0.1666446 0.1918002 0.1335904
> data ← rt(100,3)+5
> se.boot(data,100)
[1] 0.1757404 0.1029560 0.1452283
> se.boot(data,100)
[1] 0.17279203 0.08664147 0.13272875
> data ← rt(100,5)+5
> se.boot(data,100)
[1] 0.1180121 0.1222495 0.1175629
> data ← rt(100,10)+5
> se.boot(data,100)
[1] 0.1171885 0.1462890 0.1151134
> data ← rt(100,50)+5
> se.boot(data,100)
[1] 0.09913085 0.11534663 0.10362818
```

8.2 Intervalle de confiance

Dans le cas de la moyenne empirique l'erreur standard a été utilisée pour établir un intervalle qui contient la vraie valeur du paramètre μ avec une certaine probabilité, voir (8.2) et (8.3). En effet, ce concept d'intervalle qui contient la vraie valeur du paramètre est très intéressant et peut se généraliser. Dans ce paragraphe nous voyons comment trouver deux variables aléatoires A et B telles que la probabilité

$$\mathbb{P}(A \leq \theta \leq B)$$

ait une valeur fixée à l'avance, appelée *niveau de confiance* et notée $1 - \alpha$.

8.2.1 Définition

Le modèle statistique considéré ici est, comme précédemment, $\{\mathbb{P}_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$, et $\mathbf{x} = (x_1, \ldots, x_n)$ est l'échantillon observé, et $\mathbf{X} = (X_1, \ldots, X_n)$ désigne un vecteur aléatoire de loi $\mathbb{P}_\theta$.

Soient $a(\cdot)$ et $b(\cdot)$ des fonctions boréliennes à valeurs dans $\mathbb{R}$, telles que $a(\mathbf{x}) < b(\mathbf{x})$ pour tout $\mathbf{x}$. Soit $0 < \alpha < 1$. L'intervalle $[a(\mathbf{X}), b(\mathbf{X})]$ est dit **intervalle de confiance de niveau** $1 - \alpha$ pour θ si

$$\mathbb{P}_\theta \left(a(\mathbf{X}) \leq \theta \leq b(\mathbf{X})\right) \geq 1 - \alpha \,, \tag{8.4}$$

pour tout $\theta \in \Theta$. On le note $\mathrm{IC}_{1-\alpha}(\theta)$.

On dit que $\mathrm{IC}_{1-\alpha}(\theta)$ est un intervalle de confiance de *taille* $1-\alpha$ pour θ si, pour tout $\theta \in \Theta$, on a égalité en (8.4).

En pratique, on choisit une valeur faible de α, typiquement de l'ordre de 0,1 ou 0,05, et on parle alors d'intervalle de confiance de niveau 90 % ou 95 %.

On doit comprendre un intervalle de confiance de niveau $1-\alpha$ comme un intervalle *aléatoire* qui a une probabilité $1-\alpha$ de contenir le vrai paramètre θ et non comme une région fixée auquel θ aléatoire appartient avec une probabilité $1-\alpha$. Dans la pratique, le statisticien calcule les réalisations numériques $a(\mathbf{x})$ et $b(\mathbf{x})$ de $a(\mathbf{X})$ et $b(\mathbf{X})$ à partir d'observations $\mathbf{x}$, et cela lui fournit une *réalisation* de l'intervalle de confiance. Supposons par exemple que $\alpha = 0{,}05$ et que l'on ait trouvé $a = 2$ et $b = 7$. Même si la tentation est forte, on ne peut pas dire à proprement parler que l'intervalle $[2, 7]$ contient θ avec probabilité 0,95. Soit il contient θ, soit il ne contient pas. Tout ce que l'on peut dire, c'est que la probabilité qu'un intervalle construit de cette manière contienne θ est de 95 %. Ou encore : si l'on construit l'intervalle de confiance de niveau 0,95 pour 100 échantillons $\mathbf{x}$ différents, il est probable que 95 d'entre eux contiennent la vraie valeur de θ (mais on ne sait évidemment pas desquels il s'agit…!).

Bien entendu, l'intervalle $\mathrm{IC}_{1-\alpha}(\theta) = (-\infty, \infty)$ convient toujours, mais n'est guère intéressant. En effet, notre intérêt pratique sera en général de rendre l'intervalle $[a(\mathbf{X}), b(\mathbf{X})]$ le plus petit possible. On notera

$$\ell_{\mathrm{IC}} = b(\mathbf{X}) - a(\mathbf{X})$$

la *longueur de l'intervalle de confiance* $\mathrm{IC}_{1-\alpha}(\theta) = [a(\mathbf{X}), b(\mathbf{X})]$.

Rarement la situation est telle que l'on n'est intéressé qu'à établir une borne inférieure ou une borne supérieure pour θ, $a(\mathbf{X})$ ou $b(\mathbf{X})$ étant rejeté à l'infini. On parle alors d'intervalle de confiance *unilatéral* (par opposition à *bilatéral*).

De façon analogue, on définit l'intervalle de confiance asymptotique.

Un **intervalle de confiance asymptotique de niveau** $1-\alpha$ pour θ est un intervalle aléatoire $[a_n(\mathbf{X}_n), b_n(\mathbf{X}_n)]$ tel que, pour tout $\theta \in \Theta$,

$$\liminf_{n\to\infty} \mathbb{P}_\theta \left(a_n(\mathbf{X}_n) \leq \theta \leq b_n(\mathbf{X}_n)\right) \geq 1-\alpha \,. \tag{8.5}$$

Les intervalles de confiance étant valables pour tout n fini, sont préférables aux intervalles de confiance asymptotiques. En effet, pour les derniers, on ne contrôle pas exactement l'erreur, on ne fait que dire qu'elle est asymptotiquement de l'ordre fixé, sans préciser à partir de quelle taille n de l'échantillon l'approximation devient raisonnable. Cependant, souvent il est bien plus facile de construire des intervalles de confiance asymptotiques.

Remarquons que si le paramètre θ est un vecteur de dimension d avec $d > 1$, on peut généraliser la notion d'intervalle de confiance pour θ. Dans ce cas on cherchera plutôt des *régions de confiance* $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^d$ qui contiennent θ avec probabilité au moins $1-\alpha$.

8.2.2 Construction d'intervalle de confiance

On imagine assez facilement qu'un estimateur ponctuel $\hat{\theta}$ de θ sera un bon point de départ pour construire un intervalle de confiance pour θ. En effet, puisque $\hat{\theta}$ est censé de prendre des valeurs près de θ, il est naturel d'utiliser un voisinage du type $[\hat{\theta} - \hat{\delta}_1, \hat{\theta} + \hat{\delta}_2]$ de $\hat{\theta}$ comme intervalle de confiance. Afin de déterminer la taille exacte de ce voisinage (pour

que l'intervalle de confiance soit de niveau $1-\alpha$), la connaissance de la loi de l'estimateur est indispensable.

Un procédé assez général pour la construction d'intervalle de confiance repose sur l'utilisation des fonctions pivotales. Une fonction $\theta \mapsto \mathcal{T}(\hat{\theta}, \theta)$ dont la loi ne dépend pas du paramètre θ (ou d'autres paramètres inconnus du modèle) est dite **fonction pivotale** (ou **pivot**) pour le modèle statistique $\{P_\theta, \theta \in \Theta\}$.

On procède de la manière suivante :

1. On détermine un estimateur ponctuel $\hat{\theta}$ de θ.
2. On détermine la loi de l'estimateur $\hat{\theta}$.
3. On cherche une transformation $\mathcal{T}(\hat{\theta}, \theta)$ de $\hat{\theta}$ dont la loi ne dépend plus de paramètres inconnus. Autrement dit, on cherche une fonction pivotale $\mathcal{T}(\hat{\theta}, \theta)$ dont on détermine la loi.
4. On choisit $\gamma_1 \in [0,1]$ et $\gamma_2 \in [0,1]$ tels que $\gamma_2 - \gamma_1 = 1-\alpha$. On détermine les quantiles q_{γ_1} et q_{γ_2} d'ordre γ_1 et γ_2 de la loi de $\mathcal{T}(\hat{\theta}, \theta)$ à l'aide d'une table statistique ou d'un ordinateur tels que
$$\mathbb{P}_\theta \left(q_{\gamma_1} \leq \mathcal{T}(\hat{\theta}, \theta) \leq q_{\gamma_2} \right) = \gamma_2 - \gamma_1 = 1 - \alpha \, .$$
5. En "inversant" $\mathcal{T}$ (lorsque c'est possible...), on encadre alors θ par deux quantités aléatoires A et B, fonctions *uniquement* de $\hat{\theta}, q_{\gamma_1}, q_{\gamma_1}$ et de paramètres connus, telles que
$$\mathbb{P}_\theta \left(A \leq \theta \leq B \right) = 1 - \alpha \, .$$

Appliquons cette démarche à un exemple : la construction d'un intervalle de confiance pour le paramètre μ à partir d'un échantillon i.i.d. $\mathbf{X}$ de loi normale $\mathcal{N}(\mu, \sigma^2)$. Dans ce modèle, l'estimateur du maximum de vraisemblance de μ est la moyenne empirique $\bar{X}$ (étape 1) . On sait que $\bar{X}$ suit la loi normale $\mathcal{N}(\mu, \sigma^2/n)$ (étape 2). On en déduit que $\sqrt{n}(\bar{X}-\mu)/\sigma$ suit la loi normale standard $\mathcal{N}(0,1)$. Si σ est connu, cette quantité est une fonction pivotale, car sa loi ne dépend pas de paramètres inconnus. En revanche, si σ est inconnu, il faut chercher une autre statistique. Au chapitre 7 nous avons vu que $\mathcal{T} = \sqrt{n}(\bar{X} - \mu)/S$ avec $S = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ suit la loi de Student t_{n-1}. Donc, $\mathcal{T}$ convient pour la construction d'intervalle de confiance (étape 3). Choisissons $\gamma_1 = \alpha/2$ et $\gamma_2 = 1 - \alpha/2$ et notons t_{n-1}^γ le quantile d'ordre γ de la loi de Studente t_{n-1}. On obtient alors

$$\begin{aligned}
1 - \alpha &= \mathbb{P}(t_{n-1}^{\alpha/2} \leq \mathcal{T} \leq t_{n-1}^{1-\alpha/2}) \qquad \text{(étape 4)} \\
&= \mathbb{P}\left(t_{n-1}^{\alpha/2} \leq \sqrt{n}\frac{\bar{X}_n - \mu}{S} \leq t_{n-1}^{1-\alpha/2}) \right) \\
&= \mathbb{P}\left(\bar{X}_n - t_{n-1}^{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}_n - t_{n-1}^{\alpha/2}\frac{S}{\sqrt{n}} \right) \, .
\end{aligned}$$

On en déduit que $\mathrm{IC}_{1-\alpha}(\mu) = \left[\bar{X}_n - t_{n-1}^{1-\alpha/2} S/\sqrt{n}, \bar{X}_n - t_{n-1}^{\alpha/2} S/\sqrt{n} \right]$ est un intervalle de confiance de taille $1-\alpha$ pour μ (étape 5). La longueur de cet intervalle est

$$\ell_{\mathrm{IC}} = 2t_{n-1}^{1-\alpha/2}\frac{S}{\sqrt{n}} \, ,$$

car $-t_{n-1}^{\alpha/2} = t_{n-1}^{1-\alpha/2}$, par symétrie de la loi de Student.

Il est courant de choisir $\gamma_1 = \alpha/2$ et $\gamma_2 = 1 - \alpha/2$. En fait, quand la loi de la fonction pivotale $\mathcal{T}(\hat{\theta}, \theta)$ est symétrique et unimodale, ce choix minimise la longueur d'intervalle parmi tous les γ_1, γ_2 tels que $\gamma_1 + \gamma_2 = 1 - \alpha$. Il permet donc de localiser la vraie valeur du paramètre μ avec plus de précision que tout intervalle non symétrique.

Le procédé décrit ci-dessus pour la construction d'intervalle de confiance n'est pas incontournable. D'autres approches sont possibles. Dans certaines situations on peut par exemple utiliser des inégalités comme l'inégalité de Markov ou de Hoeffding pour obtenir un intervalle de confiance.

Dans des nombreux cas, l'étape 2 et/ou l'étape 3 du procédé ci-dessus ne sont pas évidentes. Il peut s'avérer plus facile de considérer la *loi limite* de l'estimateur $\hat{\theta}_n$ au lieu de la loi pour n fini. Il faut alors trouver une transformation $\mathcal{T}_n(\hat{\theta}_n, \theta)$ telle que sa *loi limite* soit indépendante de tout paramètre inconnu, et puis, on pourra en déduire un intervalle de confiance *asymptotique.*

Exemple : Intervalle de confiance asymptotique

Considérons un échantillon $\mathbf{X}$ i.i.d. de loi exponentielle $\mathcal{E}(\lambda)$ avec $\lambda > 0$ inconnu. Dans ce modèle, l'estimateur du maximum de vraisemblance de λ est $1/\bar{X}_n$. Par le théorème central limite et la delta méthode, on montre que

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda^2) \, .$$

Pour obtenir une statistique dont la loi limite ne dépend pas du paramètre inconnu λ, on multiplie le terme à gauche par $\bar{X}_n$. D'après le théorème de Slutsky, on a

$$\mathcal{T}_n = \sqrt{n}\bar{X}_n\left(\frac{1}{\bar{X}_n} - \lambda\right) = \underbrace{\lambda\bar{X}_n}_{\xrightarrow{P} 1} \underbrace{\sqrt{n}\frac{\frac{1}{\bar{X}_n} - \lambda}{\lambda}}_{\xrightarrow{\mathcal{L}} \mathcal{N}(0,1)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \, .$$

Enfin, on a pour une variable aléatoire $Z \sim \mathcal{N}(0,1)$ et pour tout $\lambda > 0$

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(-q^N_{1-\alpha/2} \leq Z \leq q^N_{1-\alpha/2}) \\ &= \lim_{n\to\infty} \mathbb{P}_\lambda\left(-q^N_{1-\alpha/2} \leq \sqrt{n}\bar{X}_n\left(\frac{1}{\bar{X}_n} - \lambda\right) \leq q^N_{1-\alpha/2}\right) \\ &= \lim_{n\to\infty} \mathbb{P}_\lambda\left(\frac{1}{\bar{X}_n} - \frac{q^N_{1-\alpha/2}}{\sqrt{n}\bar{X}_n} \leq \lambda \leq \frac{1}{\bar{X}_n} + \frac{q^N_{1-\alpha/2}}{\sqrt{n}\bar{X}_n}\right) \, . \end{aligned}$$

Alors, l'intervalle

$$\mathrm{IC}_{1-\alpha}(\lambda) = \left[\frac{1}{\bar{X}_n} - \frac{q^N_{1-\alpha/2}}{\sqrt{n}\bar{X}_n}, \frac{1}{\bar{X}_n} + \frac{q^N_{1-\alpha/2}}{\sqrt{n}\bar{X}_n}\right]$$

est un intervalle de confiance asymptotique de niveau $1 - \alpha$ pour λ.

8.3 Intervalles de confiance par le bootstrap

Pour la construction d'intervalles de confiance par la méthode pivotale il est indispensable de connaître la loi de l'estimateur $\hat{\theta}$ (ou sa loi limite). Que fait-on si elle est inconnue ? Comme dans le cas de l'erreur standard, on peut utiliser le bootstrap pour s'en sortir.

Nous présentons plusieurs approches pour construire des intervalles de confiances par le bootstrap.

Référence bibliographique : *An Introduction to the Bootstrap*, B. Efron et R. J. Tibshirani, Chapman & Hall, 1993.

8.3.1 Intervalle bootstrap standard

La première méthode bootstrap est inspirée par l'intervalle de confiance asymptotique pour la moyenne empirique. Rappelons que pour un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. de loi F avec moyenne μ et variance finie, on a par le théorème central limite,

$$\mathcal{T} = \frac{\bar{X}_n - \mu}{\mathrm{se}(\bar{X}_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \ , \quad n \to \infty \ .$$

Considérons l'estimateur consistant de l'erreur standard donné par $\hat{\mathrm{se}} = s_{\mathbf{X}}/\sqrt{n}$ où $s_{\mathbf{X}}$ désigne l'écart type empirique associé aux données $\mathbf{X}$. Par le lemme de Slutsky, l'intervalle

$$\mathrm{IC}_{1-\alpha}(\mu) = \left[\bar{X} - q^N_{1-\alpha/2}\hat{\mathrm{se}}, \bar{X} - q^N_{\alpha/2}\hat{\mathrm{se}}\right]$$

est un intervalle de confiance asymptotique de niveau $1-\alpha$ pour la moyenne μ.

Or, considérons un contexte plus général, où $\hat{\theta}$ est un estimateur de θ. L'équivalent à la statistique $\mathcal{T}$ ci-dessus est la statistique

$$\mathcal{U} = \frac{\hat{\theta} - \theta}{\mathrm{se}(\hat{\theta})} \ .$$

Dans des nombreux cas (autres que la moyenne empirique) la loi limite de $\mathcal{U}$ est toujours la loi normale standard. C'est p.ex. le cas de la médiane et de la moyenne tronquée. Si, en plus, $\hat{\mathrm{se}}(\hat{\theta})$ est un estimateur consistant de l'erreur standard de $\hat{\theta}$, alors la statistique

$$\mathcal{Z} = \frac{\hat{\theta} - \theta}{\hat{\mathrm{se}}(\hat{\theta})} \tag{8.6}$$

est également asymptotiquement normale. Par conséquent,

$$\mathrm{IC}_{1-\alpha}(\theta) = \left[\hat{\theta} - q^N_{1-\alpha/2}\hat{\mathrm{se}}(\hat{\theta}), \hat{\theta} - q^N_{\alpha/2}\hat{\mathrm{se}}(\hat{\theta})\right] \ ,$$

est un intervalle de confiance asymptotique de niveau $1-\alpha$.

Cependant, lorsque $\hat{\theta}$ n'est pas la moyenne empirique, l'estimation de l'erreur standard $\hat{\mathrm{se}}(\hat{\theta})$ doit se faire par le bootstrap (comme on l'a vu à la Section 8.1). En pratique, on considère alors l'intervalle

$$\mathcal{I}^{\mathrm{stand}}_{1-\alpha} = \left[\hat{\theta} - q^N_{1-\alpha/2}\hat{\mathrm{se}}_B(\hat{\theta}), \hat{\theta} - q^N_{\alpha/2}\hat{\mathrm{se}}_B(\hat{\theta})\right] \ ,$$

où $\hat{\mathrm{se}}_B(\hat{\theta})$ désigne l'estimation bootstrap de l'erreur standard de $\hat{\theta}$. Cet intervalle est appelé **intervalle bootstrap standard**.

8.3.2 Intervalle bootstrap studentisé

L'approximation de la loi de la statistique $\mathcal{Z}$ en (8.6) par la loi normale standard est une approximation asymptotique. En général, on a le problème de ne pas savoir à partir de quelle taille n de l'échantillon cette approximation est valable. En effet, cette approximation peut être très mauvaise, lorsque n n'est pas suffisamment grand et, par suite, conduire à des intervalles bootstrap standard bien erratiques, ce qui veut dire que la couverture de l'intervalle est loin du niveau $1-\alpha$ requis. Plus précisément, on appelle **couverture** d'un intervalle $\mathcal{I}$ la probabilité

$$\mathbb{P}_\theta(\theta \in \mathcal{I}) \ .$$

Rappelons que pour que $\mathcal{I}$ soit un intervalle de confiance de niveau $1-\alpha$ pour le paramètre θ, la couverture de $\mathcal{I}$ doit être de $1-\alpha$.

Afin d'améliorer l'intervalle bootstrap standard, on peut envisager d'estimer la 'vraie' loi de la statistique $\mathcal{Z}$ et utiliser les quantiles correspondants pour construire un meilleur intervalle. Notons $q_\gamma^{\mathcal{Z}}$ le quantile d'ordre γ de la loi de $\mathcal{Z}$. Un intervalle de confiance de niveau $1-\alpha$ pour θ est donné par

$$\mathrm{IC}_{1-\alpha}(\theta) = \left[\hat{\theta} - q_{1-\alpha/2}^{\mathcal{Z}}\hat{\mathrm{se}}(\hat{\theta}), \hat{\theta} - q_{\alpha/2}^{\mathcal{Z}}\hat{\mathrm{se}}(\hat{\theta})\right] ,$$

car

$$\mathbb{P}\left(\theta \in \left[\hat{\theta} - q_{1-\alpha/2}^{\mathcal{Z}}\hat{\mathrm{se}}(\hat{\theta}), \hat{\theta} - q_{\alpha/2}^{\mathcal{Z}}\hat{\mathrm{se}}(\hat{\theta})\right]\right) = \mathbb{P}\left(\frac{\hat{\theta}-\theta}{\hat{\mathrm{se}}(\hat{\theta})} \in \left[q_{\alpha/2}^{\mathcal{Z}}, q_{1-\alpha/2}^{\mathcal{Z}}\right]\right) = 1-\alpha .$$

Reste à savoir comment déterminer des quantiles $q_\gamma^{\mathcal{Z}}$ de la loi de $\mathcal{Z}$. Dans ce but, il est utile de réécrire (8.6) comme

$$\mathcal{Z} = \mathcal{Z}_{F,\theta} = \frac{\hat{\theta}(\mathbf{X}) - \theta}{\hat{\mathrm{se}}(\hat{\theta}(\mathbf{X}))} ,$$

pour souligner la dépendance de la loi de la statistique $\mathcal{Z}$ du paramètre θ inconnu et de la loi F de l'échantillon $\mathbf{X}$. Il est naturel d'approcher la loi de $\mathcal{Z}_{F,\theta}$ par la loi de la statistique $\mathcal{Z}_{\hat{F},\hat{\theta}(\mathbf{x})}$, c'est-à-dire la statistique $\mathcal{Z}$ où $\mathbf{X}$ est un échantillon de la loi empirique $\hat{F}$ associée à l'observation $\mathbf{x}$ et θ est remplacée par son estimation $\hat{\theta}(\mathbf{x})$. On notera

$$\mathcal{Z}^* = \frac{\hat{\theta}(\mathbf{X}^*) - \hat{\theta}(\mathbf{x})}{\hat{\mathrm{se}}(\hat{\theta}(\mathbf{X}^*))} ,$$

où $\mathbf{X}^*$ désigne un échantillon de la loi empirique $\hat{F}$ associée à l'observation $\mathbf{x}$.

Grâce au bootstrap, il est possible de générer un échantillon $(\mathcal{Z}^{*1}, \ldots, \mathcal{Z}^{*n})$ de la loi de $\mathcal{Z}^*$ qui peut servir à estimer les quantiles de la loi de $\mathcal{Z}^*$. Plus précisément, on approche les quantiles $q_\gamma^{\mathcal{Z}}$ de la loi de $\mathcal{Z}$ par des quantiles empiriques associés à $(\mathcal{Z}^{*1}, \ldots, \mathcal{Z}^{*n})$ de la loi de $\mathcal{Z}^*$, notés $\hat{q}_\gamma^{\mathcal{Z}^*}$. Cette démarche mène à l'**intervalle bootstrap studentisé** défini par

$$\mathcal{I}_{1-\alpha}^{\mathrm{stud}} = \left[\hat{\theta} - \hat{q}_{1-\alpha/2}^{\mathcal{Z}^*}\hat{\mathrm{se}}(\hat{\theta}), \hat{\theta} - \hat{q}_{\alpha/2}^{\mathcal{Z}^*}\hat{\mathrm{se}}(\hat{\theta})\right] .$$

Remarquons que l'estimateur $\hat{\mathrm{se}}(\hat{\theta})$ de l'erreur standard est typiquement lui-même un estimateur bootstrap.

Algorithme pour le calcul de l' intervalle bootstrap studentisé

`ENTRÉE :`

$\mathbf{x} = (x_1, \ldots, x_n)$	échantillon
$\mathbf{x} \mapsto \hat{\theta}(\mathbf{x})$	estimateur de θ
α	intervalle de confiance de niveau $1-\alpha$
B	nombre d'échantillons bootstrap

1. `for (b = 1:B)`
   (a) On génère un échantillon bootstrap $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$ par tirage avec remise de l'échantillon $\mathbf{x}$.
   (b) On évalue les répliques bootstrap de l'estimateur $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{x}^{*b})$ et de l'erreur standard $\hat{\mathrm{se}}^{*b} = \hat{\mathrm{se}}(\hat{\theta}(X^{*b}))$ où $X^{*b} \sim \hat{F}_{\mathbf{x}^{*b}}$.

(c) On évalue la réplique bootstrap de $\mathcal{Z}$ définie par

$$\mathcal{Z}^{*b} = \frac{\hat{\theta}^{*b} - \hat{\theta}(\mathbf{x})}{\hat{\text{se}}^{*b}} .$$

2. On détermine les quantiles empiriques $\hat{q}^{\mathcal{Z}^*}_{\alpha/2}$ et $\hat{q}^{\mathcal{Z}^*}_{1-\alpha/2}$ d'ordre $\alpha/2$ et $1-\alpha/2$ associés à l'échantillon $(\mathcal{Z}^{*1}, \ldots, \mathcal{Z}^{*B})$:

$$\hat{q}^{\mathcal{Z}^*}_{\alpha/2} = \mathcal{Z}^{*(\lceil \frac{\alpha}{2} B \rceil)} \quad \text{et} \quad \hat{q}^{\mathcal{Z}^*}_{1-\alpha/2} = \mathcal{Z}^{*(\lceil (1-\frac{\alpha}{2}) B \rceil)} ,$$

où $\mathcal{Z}^{*(\lceil b \rceil)}$ désigne la b-ième statistique d'ordre associée à $(\mathcal{Z}^{*1}, \ldots, \mathcal{Z}^{*B})$.

3. On calcule l'estimateur bootstrap de l'erreur standard :

$$\hat{\text{se}}_B(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^{*i}\right)^2} .$$

`SORTIE :` Intervalle bootstrap studentisé :

$$\mathcal{I}^{\text{stud}}_{1-\alpha} = \left[\hat{\theta} - \hat{q}^{\mathcal{Z}^*}_{1-\alpha/2}\hat{\text{se}}_B(\hat{\theta}), \hat{\theta} - \hat{q}^{\mathcal{Z}^*}_{\alpha/2}\hat{\text{se}}_B(\hat{\theta})\right] .$$

Le nombre B d'échantillons bootstrap pour la construction d'un intervalle de confiance doit être bien plus élevé que dans le cas de l'estimation de l'erreur standard d'un estimateur. Il convient de choisir B entre 1000 et 5000.

Le fait d'utiliser l'estimateur bootstrap $\hat{\text{se}}(\hat{\theta})$ de l'erreur standard dans la statistique $\mathcal{Z}$ implique qu'à l'intérieur de l'algorithme ci-dessus il faut effectuer un autre bootstrap pour chaque échantillon bootstrap $\mathbf{x}^{*b}$ afin de calculer $\hat{\text{se}}^{*b}$. On parle de *double bootstrap*. Plus précisément, dans l'étape 1(b) on calcule $\hat{\text{se}}^{*b}$ de la façon suivante : On génère C échantillons bootstrap $\mathbf{x}'^{*1}, \ldots, \mathbf{x}'^{*C}$ par tirage avec remise de l'échantillon bootstrap $\mathbf{x}^{*b}$. Puis, on calcule $\hat{\theta}'^{*c} = \hat{\theta}(\mathbf{x}'^{*c})$ pour $c = 1, \ldots, C$. Enfin, on obtient

$$\hat{\text{se}}^{*b} = \sqrt{\frac{1}{C}\sum_{c=1}^{C}\left(\hat{\theta}'^{*c} - \frac{1}{C}\sum_{i=1}^{C}\hat{\theta}'^{*i}\right)^2} .$$

Au total, il faut alors générer BC échantillons bootstrap de taille n. Si par exemple $B = 1000$ et $C = 100$, cela fait 100 000 échantillons à simuler pour le calcul d'un seul intervalle bootstrap studentisé. Cela peut vite conduire à des longues durées de calcul.

On sait que les intervalles bootstrap studentisés $\mathcal{I}^{\text{stud}}_{1-\alpha}$ sont particulièrement bien lorsque $\hat{\theta}$ est un *estimateur de la tendance centrale* comme par exemple la moyenne empirique, la moyenne tronquée ou la médiane.

8.3.3 Méthode des percentiles centrés

Afin d'éviter de faire du double bootstrap (pour l'estimation de l'erreur standard), on peut construire des intervalles bootstrap basés sur la loi de la statistique $\mathcal{Z}$ sans le dénominateur, c'est-à-dire on considère la statistique

$$\mathcal{D} = \hat{\theta} - \theta .$$

Notons $q_\gamma^{\mathcal{D}}$ le quantile d'ordre γ de la loi de $\mathcal{D}$. Alors un intervalle de confiance de niveau $1-\alpha$ pour θ est donné par

$$\mathrm{IC}_{1-\alpha}(\theta) = \left[\hat{\theta} - q_{1-\alpha/2}^{\mathcal{D}}, \hat{\theta} - q_{\alpha/2}^{\mathcal{D}}\right] ,$$

car

$$\mathbb{P}\left(\theta \in \left[\hat{\theta} - q_{1-\alpha/2}^{\mathcal{D}}, \hat{\theta} - q_{\alpha/2}^{\mathcal{D}}\right]\right) = \mathbb{P}\left(\hat{\theta} - \theta \in \left[q_{\alpha/2}^{\mathcal{D}}, q_{1-\alpha/2}^{\mathcal{D}}\right]\right) = 1-\alpha .$$

Or, il est naturel d'approcher la loi de $\mathcal{D}$ par la loi de

$$\mathcal{D}^* = \hat{\theta}(\mathbf{X}^*) - \hat{\theta}(\mathbf{x}) ,$$

où $\mathbf{X}^*$ est un échantillon de la loi empirique $\hat{F}$ associée à l'observation $\mathbf{x}$ et $\hat{\theta}(\mathbf{x})$ est l'estimation observée. Par le bootstrap, on calcule des quantiles empiriques $\hat{q}_\gamma^{\mathcal{D}^*}$ de la loi de $\mathcal{D}^*$ pour en déduire l'**intervalle bootstrap des percentiles centrés** $\mathcal{I}_{1-\alpha}^{\mathrm{cent}}$ défini par

$$\mathcal{I}_{1-\alpha}^{\mathrm{cent}} = \left[\hat{\theta} - \hat{q}_{1-\alpha/2}^{\mathcal{D}^*}, \hat{\theta} - \hat{q}_{\alpha/2}^{\mathcal{D}^*}\right] .$$

Notons G^* la fonction de répartition de $\hat{\theta}^* = \hat{\theta}(\mathbf{X}^*)$. On appelle G^* la *distribution bootstrap* de $\hat{\theta}^*$. Remarquons que G^* et la fonction de répartition $F_{\mathcal{D}^*}$ de $\mathcal{D}^*$ vérifient

$$G^*(t) = F_{\mathcal{D}^*}(t - \hat{\theta}(\mathbf{x})) , \quad \text{pour tout } t \in \mathbb{R} ,$$

et, par suite, les quantiles respectifs $q_\gamma^{G^*}$ et $q_\gamma^{\mathcal{D}^*}$ vérifient

$$q_\gamma^{\mathcal{D}^*} = q_\gamma^{G*} - \hat{\theta}(\mathbf{x}) , \quad \text{pour tout } \gamma \in (0,1) .$$

On peut alors réécrire l'intervalle bootstrap des percentiles centrés en utilisant des quantiles empiriques $\hat{q}_\gamma^{G^*}$ de la loi G^* par

$$\mathcal{I}_{1-\alpha}^{\mathrm{cent}} = \left[2\hat{\theta} - \hat{q}_{1-\alpha/2}^{G^*}, 2\hat{\theta} - \hat{q}_{\alpha/2}^{G^*}\right] .$$

Algorithme pour le calcul de l' intervalle bootstrap des percentiles centrés

`ENTRÉE :`

$\mathbf{x} = (x_1, \ldots, x_n)$	échantillon
$\mathbf{x} \mapsto \hat{\theta}(\mathbf{x})$	estimateur de θ
α	intervalle de confiance de niveau $1-\alpha$
B	nombre d'échantillons bootstrap

1. `for (b = 1:B)`
   (a) On génère un échantillon bootstrap $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$ par tirage avec remise de l'échantillon $\mathbf{x}$.
   (b) On évalue les répliques bootstrap de l'estimateur $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{x}^{*b})$.
2. On détermine les quantiles empiriques $\hat{q}_{\alpha/2}^{G^*}$ et $\hat{q}_{1-\alpha/2}^{G^*}$ d'ordre $\alpha/2$ et $1-\alpha/2$ associés à l'échantillon $(\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B})$:

$$\hat{q}_{\alpha/2}^{G^*} = \hat{\theta}^{*(\lceil \frac{\alpha}{2} B \rceil)} \quad \text{et} \quad \hat{q}_{1-\alpha/2}^{G^*} = \hat{\theta}^{*(\lceil (1-\frac{\alpha}{2}) B \rceil)} ,$$

   où $\hat{\theta}^{*(\lceil b \rceil)}$ désigne la b-ième statistique d'ordre associée à $(\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B})$.

`SORTIE :` Intervalle bootstrap des percentiles centrés :

$$\mathcal{I}_{1-\alpha}^{\mathrm{cent}} = \left[2\hat{\theta}(\mathbf{x}) - \hat{q}_{1-\alpha/2}^{G^*}, 2\hat{\theta}(\mathbf{x}) - \hat{q}_{\alpha/2}^{G^*}\right] .$$

8.3.4 Intervalle bootstrap des percentiles

Une autre alternative pour la construction d'intervalles de confiance par le bootstrap repose sur l'utilisation directe de la loi de l'estimateur $\hat{\theta}$. Autrement dit, on cherche à estimer directement la loi de l'estimateur $\hat{\theta}$ et en déduire un intervalle de confiance par les quantiles de cette loi.

En utilisant les notations précédemment introduites, l'intervalle $\left[q^G_{\alpha/2}, q^G_{1-\alpha/2}\right]$ contient $\hat{\theta}$ avec probabilité $1-\alpha$. Comme on suppose que $\hat{\theta} \approx \theta$, l'idée consiste à penser que l'intervalle équivalent avec les quantiles empiriques est susceptible d'être un intervalle de confiance de niveau approximatif $1-\alpha$ pour θ. Comme précédemment, d'après la méthode de substitution, on peut estimer les quantiles q^G_γ par les quantiles de la loi G^*. Par le bootstrap on calcule des quantiles empiriques de la loi G^*, notés $\hat{q}^{G^*}_\gamma$, comme dans le paragraphe précédent, et on en déduit l'**intervalle bootstrap des percentiles** $\mathcal{I}^{\text{perc}}_{1-\alpha}$ définit par

$$\mathcal{I}^{\text{perc}}_{1-\alpha} = \left[\hat{q}^{G^*}_{\alpha/2}, \hat{q}^{G^*}_{1-\alpha/2}\right] .$$

Algorithme pour le calcul de l' intervalle bootstrap des percentiles

`ENTRÉE :`

$\mathbf{x} = (x_1, \ldots, x_n)$	échantillon
$\mathbf{x} \mapsto \hat{\theta}(\mathbf{x})$	estimateur de θ
α	intervalle de confiance de niveau $1-\alpha$
B	nombre d'échantillons bootstrap

1. `for (b = 1:B)`
   (a) On génère un échantillon bootstrap $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$ par tirage avec remise de l'échantillon $\mathbf{x}$.
   (b) On évalue les répliques bootstrap de l'estimateur $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{x}^{*b})$.
2. On détermine les quantiles empiriques $\hat{q}^{G^*}_{\alpha/2}$ et $\hat{q}^{G^*}_{1-\alpha/2}$ d'ordre $\alpha/2$ et $1-\alpha/2$ associés à l'échantillon $(\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B})$:

$$\hat{q}^{G^*}_{\alpha/2} = \hat{\theta}^{*(\lceil \frac{\alpha}{2} B \rceil)} \quad \text{et} \quad \hat{q}^{G^*}_{1-\alpha/2} = \hat{\theta}^{*(\lceil (1-\frac{\alpha}{2}) B \rceil)} ,$$

   où $\hat{\theta}^{*(\lceil b \rceil)}$ désigne la b-ième statistique d'ordre associée à $(\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B})$.

`SORTIE :` Intervalle bootstrap des percentiles :

$$\mathcal{I}^{\text{perc}}_{1-\alpha} = \left[\hat{q}^{G^*}_{\alpha/2}, \hat{q}^{G^*}_{1-\alpha/2}\right] .$$

Comparons cet intervalle à l'intervalle bootstrap des percentiles centrés $\mathcal{I}^{\text{cent}}_{1-\alpha}$ du paragraphe précédent. En fait, quand la loi G^* a par exemple une queue lourdes à droite, alors le quantile $\hat{q}^{G^*}_{1-\alpha/2}$ prend des valeurs très élevées, ce qui implique que l'intervalle $\mathcal{I}^{\text{cent}}_{1-\alpha}$ est très étendu vers la gauche par opposition à l'intervalle $\mathcal{I}^{\text{perc}}_{1-\alpha}$. Qu'est-ce qu'il est mieux ? La réponse est : aucun des deux. Il est vrai que l'intervalle $\mathcal{I}^{\text{perc}}_{1-\alpha}$ donne des meilleurs résultats dans des nombreux cas, mais cet intervalle peut encore être amélioré (voir paragraphe suivant).

En effet, on peut montrer que l'intervalle des percentiles n'est approprié que quand la distribution G est symétrique par rapport à θ. Pour mieux comprendre, notons que

$$\mathbb{P}(\theta \in \mathcal{I}^{\text{perc}}_{1-\alpha}) = \mathbb{P}\left(\hat{q}^{G^*}_{\alpha/2} - \hat{\theta}(\mathbf{X}) \leq \theta - \hat{\theta}(\mathbf{X}) \leq \hat{q}^{G^*}_{1-\alpha/2} - \hat{\theta}(\mathbf{X})\right) .$$

En fait, le bootstrap utilise les quantiles de $\hat{\theta}(\mathbf{X}^*) - \hat{\theta}(\mathbf{x})$ pour imiter les quantiles de $\theta - \hat{\theta}(\mathbf{X})$. Si la loi de $\theta - \hat{\theta}(\mathbf{X})$ n'est pas symétrique, alors $\hat{\theta}(\mathbf{X}^*) - \hat{\theta}(\mathbf{x})$ n'est pas symétrique non plus, mais dans le sens opposé à celui de $\theta - \hat{\theta}(\mathbf{X})$!

8.3.5 Intervalle bootstrap BC_a

On peut apporter une amélioration à l'intervalle des percentiles $\mathcal{I}_{1-\alpha}^{\text{perc}}$ par un meilleur choix de l'ordre des quantiles de la loi bootstrap G^* afin de corriger certains défauts de l'approche. On définit l'**intervalle** BC_a, où BC_a signifie *bias corrected and accelerated*, noté $\mathcal{I}_{1-\alpha}^{BC_a}$, par

$$\mathcal{I}_{1-\alpha}^{BC_a} = \left[\hat{q}_{\hat{\alpha}_1}^{G^*}, \hat{q}_{\hat{\alpha}_2}^{G^*}\right] ,$$

où les ordres $\hat{\alpha}_1$ et $\hat{\alpha}_2$ de quantiles sont donnés par

$$\hat{\alpha}_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + q_{\alpha/2}^N}{1 - \hat{a}(\hat{z}_0 + q_{\alpha/2}^N)}\right) , \qquad \hat{\alpha}_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + q_{1-\alpha/2}^N}{1 - \hat{a}(\hat{z}_0 + q_{1-\alpha/2}^N)}\right) , \tag{8.7}$$

où Φ est la fonction de répartition de la loi normale standard. De plus, les deux constantes $\hat{z}_0$ et $\hat{a}$ sont données par

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^{*b} < \hat{\theta}(\mathbf{x})\}}{B}\right) , \qquad \hat{a} = \frac{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{3/2}} , \tag{8.8}$$

où $\hat{\theta}_{(i)}$ est l'estimateur $\hat{\theta}$ évalué sur l'échantillon $\mathbf{x}$ où on a supprimé la i-ème observation, et $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_{(i)}$.

Remarquons que si $\hat{z}_0 = 0$ et $\hat{a} = 0$, on retrouve l'intervalle $\mathcal{I}_{1-\alpha}^{\text{perc}}$, car $\hat{\alpha}_1 = \alpha/2$ et $\hat{\alpha}_2 = 1 - \alpha/2$.

Visiblement, $\hat{z}_0$ est un estimateur de la quantité

$$\Phi^{-1}\left(\mathbb{P}(\hat{\theta} \leq \theta)\right) .$$

Or, $\hat{z}_0 \approx 0$ si la loi de $\hat{\theta}$ est symétrique par rapport à θ. En effet, la constante $\hat{z}_0$ est censée de corriger un éventuel biais de l'estimateur $\hat{\theta}$.

Quant à la constante $\hat{a}$, elle est censée d'apporter une accélération. On peut la considérer comme un estimateur du coefficient d'asymétrie (skewness) de la loi de $\hat{\theta}$.

Nous donnons ici quelques éléments de preuve pour un cas idéalisé. Le même raisonnement peut être appliqué dans un contexte beaucoup plus général et avec plus de rigorosité.

Supposons qu'il existe une transformation $m(\cdot)$ monotone croissante, et notons $\varphi = m(\theta)$ et $\hat{\varphi} = m(\hat{\theta})$, telle que

$$\frac{\hat{\varphi} - \varphi}{\sigma_\varphi} \sim \mathcal{N}(-z_0, 1) ,$$

avec $z_0 \in \mathbb{R}$ et $\sigma_\varphi = \text{se}_\varphi(\hat{\varphi})$. En effet, la quantité $-\sigma_\varphi z_0$ représente l'éventuel biais de l'estimateur $\hat{\varphi}$. En plus, on suppose qu'il existe une constante $a \in \mathbb{R}$ telle que σ_φ vérifie la relation suivante

$$\sigma_\varphi = 1 + a\varphi . \tag{8.9}$$

En utilisant des quantiles de la loi normale standard, on construit un intervalle de confiance de niveau $1-\alpha$ pour φ de la façon suivante :

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(q^N_{\alpha/2} \leq \frac{\hat{\varphi}-\varphi}{\sigma_\varphi} + z_0 \leq q^N_{1-\alpha/2}\right) \\
&= \mathbb{P}\left((q^N_{\alpha/2}-z_0)(1+a\varphi) \leq \hat{\varphi}-\varphi \leq (q^N_{1-\alpha/2}-z_0)(1+a\varphi)\right) \\
&= \mathbb{P}\left(\hat{\varphi} - \frac{(\hat{\varphi}a+1)(q^N_{1-\alpha/2}-z_0)}{1+a(q^N_{\alpha/2}-z_0)} \leq \varphi \leq \hat{\varphi} + \frac{(\hat{\varphi}a+1)(q^N_{\alpha/2}-z_0)}{1+a(q^N_{\alpha/2}-z_0)}\right) \\
&= \mathbb{P}\left(\varphi \in \left[\hat{\varphi} - \sigma_{\hat{\varphi}}\frac{q^N_{1-\alpha/2}-z_0}{1+a(q^N_{1-\alpha/2}-z_0)}, \hat{\varphi} - \sigma_{\hat{\varphi}}\frac{q^N_{\alpha/2}-z_0}{1+a(q^N_{\alpha/2}-z_0)}\right]\right) .
\end{aligned}$$

Notons $\tau_1 = -\sigma_{\hat{\varphi}}\frac{q^N_{1-\alpha/2}-z_0}{1+a(q^N_{1-\alpha/2}-z_0)}$ et $\tau_2 = -\sigma_{\hat{\varphi}}\frac{q^N_{\alpha/2}-z_0}{1+a(q^N_{\alpha/2}-z_0)}$. En utilisant que $F_V(V)$ suit la loi uniforme U$[0,1]$ pour toute variable aléatoire V de loi F_V, on a

$$\begin{aligned}
1-\alpha &= \mathbb{P}(\varphi \in [\hat{\varphi}+\tau_1, \hat{\varphi}+\tau_2]) \\
&= \mathbb{P}\left(-\frac{\tau_2}{\sigma_\varphi} + z_0 \leq \frac{\hat{\varphi}-\varphi}{\sigma_\varphi} + z_0 \leq -\frac{\tau_1}{\sigma_\varphi} + z_0\right) \\
&= \mathbb{P}\left(\Phi\left(-\frac{\tau_2}{\sigma_\varphi} + z_0\right) \leq G(\hat{\theta}) \leq \Phi\left(-\frac{\tau_1}{\sigma_\varphi} + z_0\right)\right) \\
&= \mathbb{P}\left(G^{-1}\left(\Phi\left(-\frac{\tau_2}{\sigma_\varphi} + z_0\right)\right) \leq \hat{\theta} \leq G^{-1}\left(\Phi\left(-\frac{\tau_1}{\sigma_\varphi} + z_0\right)\right)\right) \\
&\approx \mathbb{P}\left(G^{-1}\left(\Phi\left(z_0 + \frac{q^N_{\alpha/2}-z_0}{1+a(q^N_{\alpha/2}-z_0)}\right)\right) \leq \hat{\theta} \leq G^{-1}\left(\Phi\left(z_0 + \frac{q^N_{\alpha/2}-z_0}{1+a(q^N_{\alpha/2}-z_0)}\right)\right)\right) ,
\end{aligned}$$

en utilisant que $\sigma_{\hat{\varphi}}/\sigma_\varphi \approx 1$. Par analogie avec la méthode des percentiles, il faut alors utiliser les quantiles d'ordre

$$\alpha_1 = \Phi\left(z_0 + \frac{q^N_{\alpha/2}-z_0}{1+a(q^N_{\alpha/2}-z_0)}\right) \quad \text{et} \quad \alpha_2 = \Phi\left(z_0 + \frac{q^N_{\alpha/2}-z_0}{1+a(q^N_{\alpha/2}-z_0)}\right) \tag{8.10}$$

de la loi de $\hat{\theta}$ pour encadrer θ. Pour cela il faut encore savoir estimer les constantes z_0 et a. En fait, la constante z_0 vérifie

$$\Phi(z_0) = \mathbb{P}\left(\frac{\hat{\varphi}-\varphi}{\sigma_\varphi} + z_0 \leq z_0\right) = \mathbb{P}\left(\hat{\varphi} \leq \varphi\right) = \mathbb{P}(\hat{\theta} \leq \theta) ,$$

en appliquant m^{-1} pour obtenir la dernière égalité. Cela implique que $\hat{z}_0$ défini par (8.8) est l'estimateur bootstrap de z_0.

Quant à la constante d'accélération a, elle mesure le taux de changement de l'erreur standard (sur une échelle normalisée). On peut montrer que $\hat{a}$ défini par (8.8) est un estimateur de a. Finalement, les constante $\hat{\alpha}_1$ et $\hat{\alpha}_2$ données par (8.7) sont des estimateurs de α_1 et α_2 définis en (8.10).

En fait, il est possible de montrer de façon rigoureuse que l'intervalle BC_a est un intervalle de confiance pour θ de niveau asymptotique $1-\alpha$ sous des conditions assez faibles.

8.3.6 Comparaison de différents intervalles bootstrap

L'approche traditionnelle de la statistique inférentielle repose sur des modèles idéalisés comme des hypothèses fortes sur le type de la distribution des observations. Le bootstrap

n'a pas besoin de modèle statistique bien précis. Bien qu'il existe des méthodes bootstrap dites paramétriques qui s'appliquent aux modèles paramétriques (i.e. la loi des observations est fixée à un paramètre $\theta \in \mathbb{R}^d$ près), le bootstrap est, avant tout, utilisé pour des modèles nonparamétriques. De ce fait, le bootstrap s'avère très utile pour des problèmes pratiques où la définition d'un modèle statistique proche de la réalité est difficile à trouver et/ou un modèle paramétrique est trop contraignant.

L'idée fondamentale des méthodes bootstrap est qu'en absence d'informations précises sur la distribution des données, l'échantillon observé contient toute information disponible sur la distribution sous-jacente, ce qui justifie la méthode de substitution et le rééchantillonnage.

Quelques conditions d'application

La condition pour que l'intervalles bootstrap standard soit un intervalle de confiance asymptotique est que la statistique $\mathcal{U}$ soit asymptotiquement normale. Si la distribution bootstrap de $\hat{\theta}^*$ n'est pas normal et par exemple asymétrique, l'intervalle standard est très erratique. En fait, cet intervalle est peut utilisé dans la pratique, car l'hypothèse gaussienne est très contraignante. En plus, l'intervalles bootstrap standard nécessite que la taille n de l'échantillon est plutôt élevée.

Quant aux intervalles bootstrap studentisés, l'approche conduit à des bons résultats si la loi de la statistique $\mathcal{Z}$ est plus ou moins la même quelque soit la valeur de θ. On dit que $\mathcal{Z}$ est un *pivot approximatif*. En revanche, si la loi de $\mathcal{Z}$ varie beaucoup avec le paramètre θ, les résultats peuvent s'avérer catastrophiques. Dans ce cas, il existe des méthodes pour améliorer l'intervalle bootstrap qui repose sur une transformation $\varphi(\theta)$ du paramètre $\hat{\theta}$ en sorte que la variance de $\varphi(\hat{\theta})$ soit stabilisée, c'est-à-dire qu'elle ne dépend plus de la valeur de θ. Puis, on construit d'abord l'intervalle bootstrap studentisé pour $\varphi(\theta)$, disons $[\ell^*_{\text{down}}, \ell^*_{\text{up}}]$. Ensuite, on en déduit l'intervalle $[\varphi^{-1}(\ell^*_{\text{down}}), \varphi^{-1}(\ell^*_{\text{up}})]$ pour le paramètre θ. La difficulté de cette démarche est de trouver la bonne transformation φ. Un autre élément pénalisant l'approche des intervalles studentisés est lié à l'estimation de l'erreur standard. Le fait de faire du double bootstrap peut représenter un vrai inconvénient de la méthode.

Concernant la méthode des percentiles, elle nécessite la symétrie de la distribution bootstrap G^*. En cas d'asymétrie, il est préférable d'utiliser l'intervalle bootstrap des percentiles centrés, ou encore l'intervalle BC_a.

Justesse des intervalles bootstrap

Pour tous les intervalles bootstrap présentés ici ont peut déduire des conditions sous lesquelles ils sont des intervalles de confiance asymptotique. Plus précisément, on peut montrer que la couverture, c'est-à-dire la probabilité $\mathbb{P}_\theta(\theta \in \mathcal{I}_n)$, tend vers la valeur nominale $1-\alpha$ lorsque la taille n de l'échantillon tend vers l'infini. On dit qu'un intervalle est **juste au premier ordre** si la vitesse de convergence de la couverture vers $1-\alpha$ est de l'ordre $n^{-1/2}$. Autrement dit, si

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in \mathcal{I}_n) = 1 - \alpha + O(n^{-1/2}) \, .$$

En général, les intervalles de confiance asymptotique usuels qui repose sur le théorème central limite sont juste au premier ordre. On peut également montrer que l'intervalle bootstrap des percentiles centrés ainsi que l'intervalle bootstrap des percentiles sont aussi juste au premier ordre.

En revanche, sous des conditions appropriées (mais assez souples), l'intervalle bootstrap studentisé ainsi que l'intervalle BC_a sont **juste au second d'ordre**, autrement dit

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in \mathcal{I}_n) = 1 - \alpha + O(n^{-1}) \ .$$

La convergence de la couverture a lieu plus rapidement que dans les autres cas. Ceci n'est pas seulement un avantage théorique, mais résulte en des meilleurs couvertures sur des échantillons à taille finie. Concernant l'intervalle bootstrap studentisé, cette amélioration s'explique en quelque sorte par le fait de considérer une normalisation de la loi de l'estimateur $\hat{\theta}$, c'est-à-dire de passer à une échelle normalisée par opposition à la méthodes des percentiles. Quant à l'intervalle BC_a, l'amélioration est dûe au choix intelligent des quantiles de la distribution bootstrap.

En fait, même si l'intervalle bootstrap studentisé est juste au second ordre, souvent on peut observer qu'il n'est pas très performant sur des petits échantillons. La raison est que l'approche demande de bootstrapper le rapport de deux variables aléatoires, et seulement quand la taille d'échantillon est suffisamment grande, la variabilité du dénominateur est suffisamment petite pour ne pas jouer négativement sur la justesse de l'intervalle.

En conclusion, l'intervalle BC_a est l'approche récommandée. Il est préférable aux autres méthodes dans une grande majorité de cas.