CHAPITRE 1

Introduction

Dans ce chapitre, nous introduisons les notions de base de statistiques utiles pour la suite, parmi lesquelles les notions d'expérience, de modèle, d'estimateur et de régions de confiance. Nous illustrons ces notions par des exemples et faisons quelques rappels de probabilité. Enfin nous définissons la notion de loi conditionnelle qui joue un rôle central dans la suite.

1.1 Modèles statistiques

L'objet de départ en statistique est une suite d'observations, appelée *données*, typiquement sous la forme d'une suite numérique $x_1, \dots, x_n$.

La modélisation statistique consiste à écrire $x_i = X_i(\omega)$: les données sont vues comme des réalisations de *variables aléatoires* $X_1, \dots, X_n$, voir un cours basique de probabilités pour les notions élémentaires d'aléatoire.

Definition 1. Une expérience statistique est la donnée de

- un objet aléatoire X à valeurs dans un espace E muni d'une tribu d'événements $\mathcal{E}$.
- une famille de mesures de probabilité sur $(E, \mathcal{E})$ appelée *modèle*

$$\mathcal{P} = \{P_\theta,\ \theta \in \Theta\},$$

où Θ est un ensemble appelé *espace des paramètres*.

Loi. Une 'mesure de probabilité' s'appelle aussi 'loi'.

Souvent, X consiste en un n-uplet $X = X^{(n)} = (X_1, \dots, X_n)$. Dans ce cas, les quantités E et $\mathcal{P}$ de la définition précédente dépendent de n.

Modèle du n-échantillon. Lorsque $X = X^{(n)} = (X_1, \dots, X_n)$, on prendra souvent $P_\theta^{(n)} = P_\theta \otimes \cdots \otimes P_\theta = P_\theta^{\otimes n}$. Si un n-uplet $(Y_1, \dots, Y_n)$ est de loi $P_\theta^{\otimes n}$, on dira que les variables $Y_1, \dots, Y_n$ sont indépendantes et identiquement distribuées (en abbrégé iid).

Definition 2. Un modèle statistique $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ est identifiable si, pour tous $\theta, \theta' \in \Theta$,

$$P_\theta = P_{\theta'} \quad \Rightarrow \quad \theta = \theta'.$$

L'identifiabilité d'un modèle implique que pour une loi donnée Q dans $\mathcal{P}$, il y a un *unique* paramètre tel que $Q = P_\theta$. C'est une propriété très importante, qui assure que le modèle est *bien* paramétré. Nous donnons quelques exemples ci-dessous.

Definition 3. Un modèle statistique $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ est dominé s'il existe une mesure positive μ sur E telle que, pour tous $\theta \in \Theta$, P_θ admet une densité p_θ par rapport à μ, soit

$$dP_\theta(x) = p_\theta(x) d\mu(x).$$

Notons qu'il faut que la mesure μ soit la même pour tous les $\theta \in \Theta$. On parle de mesure *dominante*. Dans la suite, nous travaillerons toujours avec des modèles dominés.

Notations. Si X est une variable aléatoire de loi Q, on note $X \sim Q$. Cela signifie que pour toute g intégrable par rapport à Q, soit $g \in L^1(Q)$,

$$E_{X \sim Q}[g(X)] = E_Q[g(X)] = \int_E g(x) dQ(x).$$

Si $Y \sim P_\theta$, on abrège souvent $E_{Y \sim P_\theta}$ en E_θ. Egalement, dans le cas du modèle du n-échantillon ci-dessus, on note simplement E_θ en lieu et place de $E_{P_\theta^{\otimes n}}$.

Exemples primordiaux

1 Le modèle fondamental est

$$\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}.$$

C'est un modèle dominé, pour μ la mesure de Lebesgue sur $\mathbb{R}$,

$$dP_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} dx.$$

Il s'agit aussi d'un modèle identifiable. En effet, ici $P_\theta = \mathcal{N}(\theta, 1)$, et si $P_\theta = P_{\theta'}$,

(a) *méthode 1.* Si $\mathcal{N}(\theta, 1) = \mathcal{N}(\theta', 1)$, l'espérance pour les deux lois est la même : $E_{\mathcal{N}(\theta,1)}[X] = E_{\mathcal{N}(\theta',1)}[X]$. Or $E_\theta[X] = E_{\mathcal{N}(\theta,1)}[X] = \int x e^{-(x-\theta)^2/2} dx/\sqrt{2\pi} = \theta$. Donc $\theta = \theta'$.

(b) *méthode 2.* Si deux lois à densité par rapport à μ sont égales, alors leurs densité sont égales μ-presque partout. Or $\theta \neq \theta'$ implique que $p_\theta(x) \neq p_{\theta'}(x)$ pour tout $x \in \mathbb{R}$. Ainsi $P_\theta \neq P_{\theta'}$, donc le modèle est identifiable (c'est la contraposée de la définition).

2 Le modèle de tirage de pile ou face est

$$\mathcal{P} = \{\mathrm{Be}(\theta),\ \theta \in [0,1]\},$$

où $\mathrm{Be}(\theta)$ est la loi de Bernoulli. C'est la loi discrète définie par : si $X \sim \mathrm{Be}(\theta)$,

$$P[X = 1] = \theta, \qquad P[X = 0] = 1 - \theta,$$

ce que l'on note aussi $\mathrm{Be}(\theta) = (1-\theta)\delta_0 + \theta\delta_1$. C'est un modèle dominé par $\mu = \delta_0 + \delta_1$. Le modèle est identifiable, par exemple parce que $E_\theta X = \theta$ donc $P_\theta = P_{\theta'}$ implique que $\theta = E_\theta X = E_{\theta'} X = \theta'$.

Definition 4. Un estimateur ponctuel $\hat{\theta}(X)$ dans une expérience statistique $(X, \mathcal{P})$ est une fonction mesurable de X, à valeurs dans l'espace des paramètres Θ.

Une statistique $S(X)$ est une fonction mesurable quelconque de X.

Exemple. Dans le modèle fondamental $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$, si l'on dispose d'observations $X = (X_1, \ldots, X_n)$, alors $\hat{\theta}_1(X) = 1$, $\hat{\theta}_2(X) = \overline{X} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ sont des estimateurs (ponctuels) de θ, et aussi des statistiques.

1.2 Approches statistiques

Nous introduisons deux points de vue principaux, l'approche *fréquentiste* et l'approche *bayésienne*. Ces deux approches ont le même point de départ : l'expérience statistique définie plus haut, et en particulier le modèle $\mathcal{P}$. La principale différence réside dans l'hypothèse que l'on fait sur la loi suivie par les données X.

1.2.1 Approche fréquentiste

Dans l'approche *fréquentiste*, on suppose

$$\exists\, \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

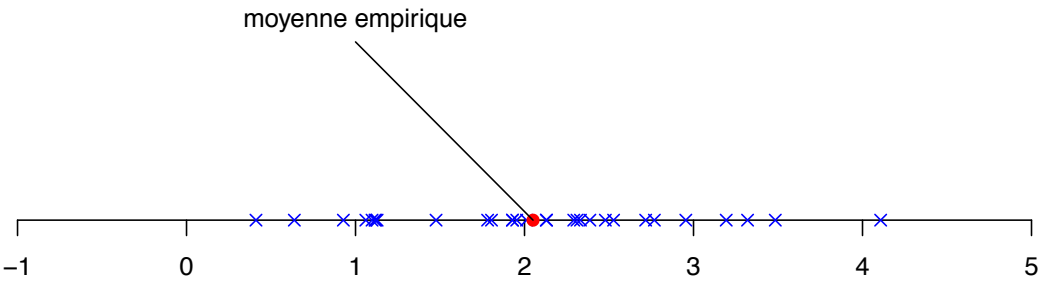
Dans ce cadre, θ_0 s'appelle vraie valeur du paramètre. Typiquement, θ_0 est inconnu et on cherche à l'"estimer" (à l'approcher), à l'aide des données X.

Exemple (modèle fondamental). Supposons l'expérience statistique donnée par $X = (X_1, \dots, X_n)$ et $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$. L'approche fréquentiste consiste à supposer qu'il existe $\theta_0 \in \mathbb{R}$ tel que

$$(X_1, \dots, X_n) \sim \mathcal{N}(\theta_0, 1)^{\otimes n},$$

c'est-à-dire que les données sont i.i.d. de loi commune $\mathcal{N}(\theta_0, 1)$. La Figure 1.1 représente $n = 30$ points tirés aléatoirement de façon indépendante suivant une loi $\mathcal{N}(\theta_0, 1)$. La vraie valeur de θ a été prise égale à $\theta_0 = 2$. On constate que l'échantillon est assez concentré autour de 2, et que la moyenne empirique $\sum_{i=1}^n X_i/n$ est proche de 2.

FIGURE 1.1 – Echantillon de taille $n = 30$ d'une loi $\mathcal{N}(\theta_0, 1)$



Notation. Souvent, lorsque $X = (X_1, \dots, X_n)$ et que l'on travaille avec un modèle d'échantillonage, on notera simplement $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$ au lieu de $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$.

Grandes questions dans le cadre fréquentiste (on peut aussi les poser dans le cadre bayésien)

1. *Estimation.* Il s'agit de construire un estimateur $T(X_1, \dots, X_n)$ qui soit proche, en un sens à préciser, de la vraie valeur θ_0 du paramètre θ.
2. *Intervalles/régions de confiance.* On cherche à construire $\mathcal{C} = \mathcal{C}(X_1, \dots, X_n)$ sous-ensemble (aléatoire) de Θ tel que $\theta_0 \in \mathcal{C}(X_1, \dots, X_n)$ avec grande probabilité.
3. *Tests.* On veut répondre par "vrai" ou "faux" à une propriété donnée de P_θ en construisant $\varphi(X_1, \dots, X_n)$ à valeurs dans $\{0, 1\}$.

1.2.2 Approche bayésienne (intuition)

Thomas Bayes (1702-1761) et Laplace (1749-1827) ont été des pionniers de la méthodologie bayésienne. Dans cette approche, on modélise toutes les quantités inconnues par des variables aléatoires. Ainsi "tout est aléatoire".

Une intuition possible derrière cette approche est que plutôt que de modéliser des quantités par des nombres, il peut être intéressant de les modéliser plutôt par des lois de probabilité.

Prenons comme exemple la température à un endroit précis à un instant du temps. Par exemple, la température dans la pièce où se trouve le lecteur à l'instant présent. On peut répondre en utilisant un thermomètre et en donnant la valeur lue, mettons $T = 18$ degrés. On peut aussi penser qu'il est peut-être raisonnable de tenir compte d'une petite erreur de mesure possible, et donc de répondre plutôt, $\Pi_T = \mathcal{N}(18, v)$, une loi gaussienne centrée en 18, de petite variance, par exemple $v = 0.5$. Si l'on s'intéresse ensuite à la température de-main au même endroit, sachant que soit le temps est resté le même, soit la température a augmenté autour de 3 degrés, au lieu de répondre $T' = 18$ ou $T' = 21$, on peut aussi proposer $\Pi_{T'} = \frac{1}{2}\mathcal{N}(18, v) + \frac{1}{2}\mathcal{N}(21, v)$.

La façon dont on formalise l'approche *bayésienne* est de supposer aléatoire le paramètre inconnu θ du modèle, avec la loi de ce paramètre appelée *loi a priori*. Cette loi reflète notre connaissance a priori (éventuelle) du paramètre. Ainsi, par exemple dans le modèle $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$, si l'on sait à l'avance que le paramètre θ est positif, il est assez naturel de prendre une loi a priori sur θ qui porte sur $\mathbb{R}^+$, par exemple la loi exponentielle $\mathcal{E}(1)$. Ensuite, une fois des données $X_1, \ldots, X_n$ observées, on va mettre à jour la loi a priori en utilisant l'"information" contenue dans les données. Formellement, cette mise à jour se fait par une opération de conditionnement, ce que nous verrons au Chapitre 2. On obtient alors une nouvelle loi, la loi a posteriori, qui est la 'mise à jour de la loi a priori' une fois les données observées. Notons que si l'on n'a pas de connaissance préalable comme la positivité ci-dessus, on pourra choisir plutôt une loi 'qui met un peu de masse partout', comme la loi $\mathcal{N}(0, 1)$, ce qui reflète le fait que potentiellement le paramètre pourrait être partout sur $\mathbb{R}$.

Illustrons les idées ci-dessus dans le cadre du modèle fondamental $\{\mathcal{N}(\theta, 1)^{\otimes n}\}$, avec pour loi a priori sur θ la loi $\mathcal{N}(0, 1)$. Nous verrons au Chapitre 2 qu'après avoir observé n données $X_1, \ldots, X_n$, la loi a posteriori est $\Pi_n = \mathcal{N}\left(\frac{n}{n+1}\overline{X}, \frac{1}{n+1}\right)$. La Figure 1.2 représente la densité de la loi a priori, et de la loi a posteriori Π_3, obtenue après observation des données X_1, X_2, X_3.

1.3 Exemples de modèles

Voici quelques modèles statistiques classiques, décrits par les lois P_θ correspondantes.

→ modèle fondamental

$$P_\theta = \mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}.$$

→ modèle gaussien à moyenne *et* variance inconnues.
Le paramètre du modèle est $\theta = (\mu, \sigma^2)$ et

$$P_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2),$$

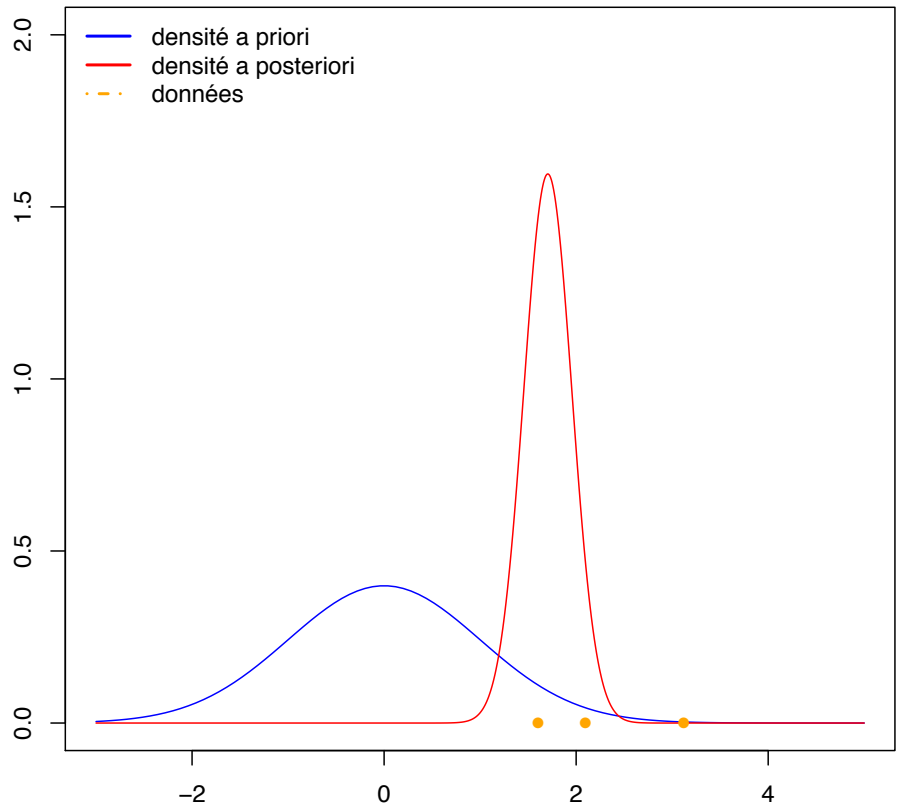
avec ici $\Theta = \{(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$.

→ modèle gaussien en dimension 2.
Il s'agit de l'ensemble des lois

$$\mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \Sigma\right),$$

avec μ_1, μ_2 réels et Σ une matrice 2×2 définie positive, et $\mathcal{N}(v, M)$ désigne la loi d'un vecteur gaussien centré en v et de matrice de variance-covariance M, voir la définition 5 ci-dessous.

FIGURE 1.2 – Densités a priori et a posteriori



→ modèles de translation et changement d'échelle
Il s'agit de la famille de lois de

$$X = \sigma Y + \mu, \quad \text{avec } \sigma > 0,\ \mu \in \mathbb{R},$$

pour Y un variable aléatoire réelle de densité f.
Exercice. Montrer que la densité d'une telle variable X est $\sigma^{-1} f(\frac{\cdot - \mu}{\sigma})$.

→ lois gamma $\Gamma(t, \lambda)$
Il s'agit de lois de densité $f_{t,\lambda}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$,

$$f_{t,\lambda}(x) = \frac{1}{\Gamma(t)} \lambda^t x^{t-1} e^{-\lambda x} \mathbb{1}_{x>0},$$

pour $\Gamma(t) = \int_0^\infty u^{t-1} e^{-u}$ la fonction Gamma. La loi exponentielle $\mathcal{E}(\lambda)$ est la loi $\Gamma(1, \lambda)$.

→ modèle "non-lisse"

$$P_\theta = \text{Unif}[0, \theta],$$

avec pour densité $f_\theta(x) = \theta^{-1} \mathbb{1}_{[0,\theta]}(x)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$.

→ nous verrons d'autres exemples dans la suite du cours et en TD.

1.4 Outils de probabilité

Dans le cadre de ce cours, nous travaillerons essentiellement avec des variables aléatoires à valeurs dans $\mathbb{R}$, $\mathbb{R}^d$, $d \geq 1$. Un cas particulier est celui de variables à valeurs discrètes dans $\mathbb{N}$ ou $\mathbb{Z}$. Dans tout ce qui suit, on pourra prendre les espaces d'arrivée égaux à $\mathbb{R}$ pour fixer les idées.

Types de lois de probabilité sur $\mathbb{R}$, lois discrètes et lois à densité.

La loi de probabilité d'une variable aléatoire X à valeurs dans $\mathbb{R}$ est complètement déterminée par sa fonction de répartition, définie pour t réel par

$$F_X(t) = P[X \leq t].$$

On peut classifier des types de lois suivant la nature de la fonction de répartition F_X. Les deux exemples essentiels sont

→ S'il existe une fonction mesurable positive f telle que

$$F_X(t) = \int_{-\infty}^{t} f(u)du,$$

alors X est dite *à densité*, et f est sa densité par rapport à la mesure de Lebesgue.

→ Si la fonction de répartition F_X est constante par morceaux,

$$F_X(t) = \sum_{i \in \mathcal{D}} p_i \mathbb{1}_{[a_i,\infty)}(t),$$

avec $\mathcal{D}$ un ensemble fini ou dénombrable, $(a_i)_{i\in\mathcal{D}}$ une suite de réels, et $\sum_{i\in\mathcal{D}} p_i = 1$, alors la loi de X est dite *discrète*.

Exemples

→ La loi normale $\mathcal{N}(\mu, \sigma^2)$ est, par définition, une loi à densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$, avec pour densité

$$f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(u-\mu)^2\right\}.$$

→ La loi de Bernoulli $\mathrm{Be}(\theta)$ est la loi discrète de fonction de répartition donnée par $F(t) = p\mathbb{1}_{[0,\infty)} + (1-p)\mathbb{1}_{[1,\infty)}$.

Lois à densité générales.

Soit $(E, \mathcal{E})$ un espace muni d'une tribu $\mathcal{E}$ et soit μ une mesure positive σ-finie sur E.

On rappelle qu'une mesure μ est σ-finie sur E s'il existe une partition dénombrable de E, $E = \cup_{i\geq 1} E_i$, telle que $\mu(E_i) < \infty$. Une telle hypothèse d'ordre technique permet d'invoquer des théorèmes comme celui de Fubini et sera toujours supposée dans la suite.

Par exemple pour $E = \mathbb{R}$ on prend la plupart du temps μ égale à la mesure de Lebesgue sur $\mathbb{R}$. Si la loi P sur E vérifie que pour tout $A \in \mathcal{E}$,

$$P[A] = \int_A p(x)d\mu(x),$$

ce que l'on note aussi $dP(x) = p(x)d\mu(x)$ ou $dP = pd\mu$, on dit que P est à densité p par rapport à μ.

Exemple. On rappelle que δ_x, la masse de Dirac en x, est la mesure positive définie, pour tout A mesurable, par $\delta_x[A] = \mathbb{1}_{x\in A}$.

→ Sur $E = \{0, 1\}$ (ou $E = \mathbb{N}$), la loi de Bernoulli P_θ =Be(θ) admet une densité par rapport à la mesure $\mu = \delta_0 + \delta_1$. En effet, on peut écrire,

$$\begin{aligned} P_\theta[\{0\}] &= 1 - \theta = (1 - \theta)\mu[\{0\}] \\ P_\theta[\{1\}] &= \theta = \theta\mu[\{1\}]. \end{aligned}$$

et donc $dP_\theta = p_\theta d\mu$ avec $p_\theta(x) = (1 - \theta)\mathbb{1}_{x=0} + \theta\mathbb{1}_{x=1}$. On peut par ailleurs remarquer que $P_\theta = (1 - \theta)\delta_0 + \theta\delta_1$.

Lois produits.

Soit P une mesure de probabilité sur $(E, \mathcal{E})$ et Q une mesure de probabilité sur $(F, \mathcal{F})$. Alors la loi produit $P \otimes Q$ est la loi sur l'espace produit $E \times F$ muni de la tribu produit qui vérifie

$$(P \otimes Q)(A \times B) = P(A) \times Q(B),$$

pour tout $A \in \mathcal{E}$ et $B \in \mathcal{F}$. De plus, si P a une densité p par rapport à une mesure dominante μ sur E et Q une densité q par rapport à une mesure dominante ν sur F, alors $P \otimes Q$ a pour densité $p \times q$ par rapport à $\mu \otimes \nu$

$$\begin{aligned} d(P \otimes Q)(x, y) &= p(x)q(y)d(\mu \otimes \nu)(x, y) \\ &= p(x)q(y)d\mu(x)d\nu(y). \end{aligned}$$

Lois produits et indépendance. Deux variables aléatoires X et Y sont indépendantes si et seulement si la loi du couple $P_{(X,Y)}$ est le produit de la loi P_X de X et P_Y de Y, soit $P_{(X,Y)} = P_X \otimes P_Y$.

Exemple. La loi sur $\mathbb{R}^2$ dont la densité par rapport à la densité produit Leb($\mathbb{R}$) $\otimes$ Leb($\mathbb{R}$) est

$$\frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$$

est une loi produit. En effet, on reconnait le produit de deux lois normales standard $\mathcal{N}(0, 1)$. Donc cette loi est $\mathcal{N}(0, 1) \otimes \mathcal{N}(0, 1)$.

Plus généralement, on peut faire des produits de plusieurs lois, ou de n fois la même loi. Ainsi, $Q = P^{\otimes n}$ est une mesure de probabilité sur l'espace produit E^n. Si P a une densité p par rapport à une mesure dominante μ sur E, soit $dP(x) = p(x)d\mu(x)$, alors $P^{\otimes n}$ a une densité

sur E^n par rapport à $\mu^{\otimes n}$, égale à $q(x_1, \ldots, x_n) = \prod_{i=1}^n p(x_i)$.

VECTEURS GAUSSIENS

Pour $d \geq 1$, soit $\mu \in \mathbb{R}^d$ et V une matrice définie positive, c'est-à-dire telle que $y^T V y > 0$ pour tout y non nul de $\mathbb{R}^d$, où y^T désigne la transposée.

Definition 5. Un vecteur aléatoire X de $\mathbb{R}^d$ suit une loi $\mathcal{N}(\mu, V)$ si sa densité par rapport à la mesure de Lebesgue dans $\mathbb{R}^d$ est, pour $|V| = \det(V)$,

$$x \to \frac{1}{\sqrt{(2\pi)^d |V|}} \exp\left\{-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\right\}.$$

Notons en particulier que si V est une matrice diagonale (et donc également V^{-1}), la densité de la loi $\mathcal{N}(\mu, V)$ s'exprime comme un produit de densités coordonnée par coordonnée. Cela signifie donc d'après ce qui précède que les coordonnées X_i de X sont alors indépendantes. Si en revanche V n'est pas diagonale, V^{-1} non plus et la densité ne s'écrit pas comme un produit : les coordonnées X_i ne sont alors pas indépendantes.

CONVERGENCES.

Pour $x \in \mathbb{R}^d$, $d \geq 1$, on note $\|x\|^2 = \sum_{i=1}^d x_i^2$.

Definition 6. Soit $X_1, \ldots, X_n, \ldots$ et X des variables aléatoires à valeurs dans $\mathbb{R}^d$, $d \geq 1$, définies sur un même espace de probabilité. La suite (X_n) converge en probabilité vers X, ce que l'on note $X_n \xrightarrow{P} X$, si

$$\forall \varepsilon > 0, \qquad P\left[\|X_n - X\| > \varepsilon\right] \to 0 \quad (n \to \infty).$$

Proposition 1. [loi des grands nombres] Soit $(X_n)_{n\geq 1}$ une suite de variables iid à valeurs dans $\mathbb{R}^d$, $d \geq 1$, avec $E[\|X_1\|] < \infty$. Alors (la convergence a aussi lieu presque-sûrement)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} EX_1.$$

Definition 7. Dans une expérience statistique X, $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, l'estimateur $\hat{\theta}(X)$ est consistant si, pour tout $\theta \in \Theta$,

$$\hat{\theta}(X) \overset{P_\theta}{\to} \theta.$$

Definition 8. Soit $(X_n)_{n\geq 1}$ et X des variables aléatoires quelconques à valeurs dans $\mathbb{R}^d$. On dit que X_n converge en loi vers X ce que l'on note $X_n \overset{\mathcal{L}}{\to} X$ si pour toute fonction $f : \mathbb{R}^k \to \mathbb{R}$ continue bornée,

$$E[f(X_n)] \to E[f(X)] \qquad (n \to \infty).$$

De même, on dira que (X_n) converge en loi vers une loi P si $E[f(X_n)] \to E[f(X)]$ pour $X \sim P$, pour toute fontion f continue bornée.

On note la propriété importante suivante de la convergence en loi. On rappelle que pour $A \subset \mathbb{R}^d$, la frontière de A est $\partial A = \overline{A} \setminus \overset{\circ}{A}$.

Proposition 2. Si $X_n \overset{\mathcal{L}}{\to} X$ dans $\mathbb{R}^d$, alors pour tout borélien A de $\mathbb{R}^d$ pour lequel $P[X \in \partial A] = 0$, on a

$$P[X_n \in A] \to P[X \in A] \qquad (n \to \infty).$$

Application. Si $Z_n \overset{\mathcal{L}}{\to} \mathcal{N}(0,1)$, alors pour tout intervalle I de $\mathbb{R}$,

$$P[Z_n \in I] \to P[Z \in I] \qquad (n \to \infty).$$

Proposition 3. [TCL dans $\mathbb{R}^d$] Soit (X_n) une suite de variables aléatoires iid dans $\mathbb{R}^d$, avec $E[\|X_1\|^2] < \infty$. Soit $\mu = EX_1$ et $V = E[(X_1 - E(X_1))(X_1 - E(X_1))^T]$. Alors

$$\sqrt{n}(\overline{X} - \mu) \overset{\mathcal{L}}{\to} \mathcal{N}(0, V),$$

où la $\mathcal{N}(0, V)$ est la loi gaussienne centrée sur $\mathbb{R}^d$ de covariance V.

Definition 9. Un estimateur $\hat{\theta} = \hat{\theta}(X)$ de θ est dit asymptotiquement normal si, pour X de loi $P_\theta^{(n)}$, et pour Σ_θ une matrice positive, quand $n \to \infty$,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_\theta).$$

Proposition 4. [théorème de l'image continue] Soient X_n, X des variables aléatoires à valeurs dans $\mathbb{R}^d$. Soit $g : \mathbb{R}^d \to \mathcal{Y}$ une fonction continue, pour $\mathcal{Y}$ un espace métrique quelconque. Alors $X_n \xrightarrow{\mathcal{L}} X$ implique $g(X_n) \xrightarrow{\mathcal{L}} g(X)$. Egalement, $X_n \xrightarrow{P} X$ implique $g(X_n) \xrightarrow{P} g(X)$.

Proposition 5. [Lemme de Slutsky] Soient X_n, Y_n, Z_n des suites de variables aléatoires, X une variable aléatoire fixée, et a, b des constantes. On suppose

$$X_n \xrightarrow{\mathcal{L}} X, \quad Y_n \xrightarrow{\mathcal{L}} b, \quad Z_n \xrightarrow{\mathcal{L}} a.$$

Alors $Z_n X_n + Y_n \xrightarrow{\mathcal{L}} aX + b$.

Remarques. Si $a = 0$, alors $aX = 0X = 0$.
Si a est une constante, alors on peut vérifier que convergence que convergence en loi ou en proba vers a sont équivalentes : $Z_n \xrightarrow{\mathcal{L}} a$ si et seulement si $Z_n \xrightarrow{P} a$.

Exercice. Montrer que si X_n est asymptotiquement normal, alors X_n est consistant. *[On pourra écrire $X_n - \theta = \sqrt{n}(X_n - \theta)/\sqrt{n}$]*

1.5 Outils de statistique

Nous introduisons d'abord une notion de risque qui sera précisée dans la suite du cours. Dans la suite, $\hat{\theta} = \hat{\theta}(X)$ est un estimateur dans une expérience statistique $(X, \mathcal{P} = \{P_\theta,\ \theta \in \Theta\})$.

Point important : estimateurs et statistiques sont des fonctions mesurables de X seulement ; elles ne sont pas autorisées pas à dépendre du paramètre inconnu θ (sinon, on pourrait prendre $\hat{\theta} = \theta$).

Definition 10.

- Cas où $\Theta \subset \mathbb{R}$. Le risque quadratique d'un estimateur $\hat{\theta}(X)$ au point θ est la fonction

$\theta \to R(\theta, \hat{\theta})$ définie par

$$R(\theta, \hat{\theta}) = E_\theta \left[(\hat{\theta} - \theta)^2 \right] = \int (\hat{\theta}(x) - \theta)^2 dP_\theta(x).$$

- Cas où (Θ, d) est un espace métrique. Le risque quadratique est

$$R(\theta, \hat{\theta}) = E_\theta \left[d(\hat{\theta}, \theta)^2 \right].$$

Typiquement, on souhaitera contrôler le risque en donnant des bornes pour celui-ci.

Voyons maintenant quelques inégalités pour contrôler la probabilité de déviation $P_\theta[|\hat{\theta} - \theta| \geq t]$.

Inégalité de Markov. Soit Y une variable aléatoire réelle et $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ une fonction croissante. Alors

$$P[|Y| \geq t] \leq \frac{1}{\psi(t)} E[\psi(|Y|)].$$

En particulier, pour la fonction $x \to x^p$, avec p entier, on obtient

$$P[|Y| \geq t] \leq t^{-p} E[|Y|^p].$$

L'*inégalité de Tchébychev* est un cas particulier, avec $\mathrm{Var}[Y] = E[(Y - EY)^2]$,

$$P[|Y - EY| \geq t] \leq \frac{1}{t^2} \mathrm{Var}[Y].$$

Preuve.

L'inégalité de Markov découle de, en utilisant la croissance de ψ,

$$E\psi(|Y|) \geq E\left[\psi(|Y|) \mathbb{1}_{|Y| \geq t}\right] \geq \psi(t) E[\mathbb{1}_{|Y| \geq t}] = \psi(t) P[|Y| \geq t].$$

On en déduit l'inégalité de Tchébychev en l'appliquant à $Z = E - EY$ et $\psi(x) = x^2$.

Conséquence de l'inégalité de Tchébychev.

$$P_\theta[|\hat{\theta} - \theta| \geq t] \leq \frac{1}{t^2} E_\theta[(\hat{\theta} - \theta)^2] = \frac{R(\theta, \hat{\theta})}{t^2}.$$

Ainsi, un risque quadratique petit implique qu'avec grande probabilité, $|\hat{\theta} - \theta|$ est petit.

Décomposition biais-variance, cas où $\Theta \subset \mathbb{R}$. On peut toujours décomposer

$$\hat{\theta} - \theta = \hat{\theta} - E_\theta \hat{\theta} + E_\theta \hat{\theta} - \theta.$$

En prenant le carré puis l'espérance, et en utilisant la linéarité de l'espérance, qui donne $E_\theta(\{\hat\theta - E_\theta\hat\theta\}\{E_\theta\hat\theta - \theta\}) = \{E_\theta\hat\theta - \theta\}E_\theta(\hat\theta - E_\theta\hat\theta) = 0$, on obtient

$$\begin{aligned} E_\theta(\hat\theta - \theta)^2 &= E\left[(\hat\theta - E_\theta\hat\theta)^2\right] + \left[E_\theta\hat\theta - \theta\right]^2 + 2E_\theta\left[\{\hat\theta - E_\theta\hat\theta\}\{E_\theta\hat\theta - \theta\}\right] \\ &= \quad \left[E_\theta\hat\theta - \theta\right]^2 \quad + \quad E\left[(\hat\theta - E_\theta\hat\theta)^2\right] \\ &= \text{Biais au carré} \quad + \quad \text{Variance} \end{aligned}$$

Pour rendre le risque petit, on cherchera donc à rendre à la fois le biais et la variance petits.

Exemples de calculs de risques. Soient $X_1, \dots, X_n$ iid $\mathcal{N}(\theta, 1)$

→ estimateur constant $\hat\theta = 1$.

> $R(\theta, \hat\theta) = E_\theta(1-\theta)^2 = (\theta-1)^2$. Le risque est imbattable si $\theta = 1$ puisqu'il est nul, mais si $\theta \neq 1$ il est strictement positif et ne tend pas vers 0.

→ estimateur $\hat\theta(X) = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$

> Effectuons le calcul explicite pas à pas
>
> $$\begin{aligned} R(\theta, \hat\theta) = E_\theta(\overline{X} - \theta)^2 &= \frac{1}{n^2}E_\theta\Big[\sum_{i=1}^n (X_i - \theta)\Big]^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n E_\theta(X_i - \theta)^2 + \frac{1}{n^2}\sum_{i\neq j} E_\theta\left[(X_i - \theta)(X_j - \theta)\right] \end{aligned}$$
>
> Pour $i \neq j$, les variables X_i et X_j sont indépendantes donc
>
> $$E_\theta[(X_i - \theta)(X_j - \theta)] = E_\theta[X_i - \theta]E_\theta[X_j - \theta] = 0.$$
>
> Les X_i sont de même loi, donc $E_\theta[(X_i - \theta)^2] = E_\theta[(X_1 - \theta)^2] = \mathrm{Var}_\theta[X_1]$ pour tout i, soit
>
> $$R(\theta, \hat\theta) = \frac{1}{n}\mathrm{Var}_\theta[X_1] = \frac{1}{n}.$$
>
> D'un point de vue global, le risque est bien meilleur que celui de l'estimateur constant. On peut aussi aller plus vite dans le calcul ci-dessus en écrivant
>
> $$R(\theta, \hat\theta) = \frac{1}{n^2}\mathrm{Var}_\theta\Big[\sum_{i=1}^n X_i\Big] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}_\theta(X_i) = \frac{1}{n}\mathrm{Var}_\theta(X_1) = \frac{1}{n},$$
>
> en utilisant que pour des variables indépendantes, la variance de la somme est la somme des variances.

Exercice. Soit $X \sim \mathrm{Bin}(n, p)$, la loi binomiale de paramètres n, p. On rappelle qu'il s'agit de la loi de $\sum_{i=1}^n \varepsilon_i$ où ε_i sont iid de loi de Bernoulli $\mathrm{Be}(p)$. Pour $\hat\theta = X/n$,

1. écrire la décomposition biais-variance.
2. montrer que $R(\theta, \hat\theta) \leq 1/(4n)$.

Intervalles et régions de confiance.

On note $[a, b]$ un intervalle d'extrémités a et b.

Definition 11. Soit $\alpha > 0$.

- Cas $\Theta \subset \mathbb{R}$. Un intervalle de confiance de niveau (au moins) $1-\alpha$ est un intervalle aléatoire $I(X) = [a(X), b(X)]$ où $a(X), b(X)$ sont des statistiques à valeurs dans $\mathbb{R}$ vérifiant
$$P_\theta[\,\theta \in I(X)\,] \geq 1-\alpha \qquad \forall\, \theta \in \Theta$$
- Cas général. Une région de confiance de niveau (au moins) $1-\alpha$ est $\mathcal{R}(X) \subset \Theta$ avec
$$P_\theta[\,\theta \in \mathcal{R}(X)\,] \geq 1-\alpha \qquad \forall\, \theta \in \Theta$$

On remarquera que Θ lui-même est toujours une région de confiance, de niveau de confiance égal à 1. Cependant, on souhaite en général trouver une région la plus petite possible (ou proche de la plus petite), telle que le niveau de confiance reste au moins de $1-\alpha$.

Construction d'intervalles de confiance.

1ère technique [utilisation de bornes en probabilité]. On part d'une inégalité comme celle de Tchébychev qui contrôle la probabilité de déviation de $\hat{\theta}$ à θ. Pour $t > 0$,
$$P_\theta[|\hat{\theta} - \theta| > t] \leq t^{-2} R(\hat{\theta}, \theta).$$

→ *Exemple : expérience binomiale.* On observe $X \sim \mathrm{Bin}(n, \theta)$. On pose $\hat{\theta} = X/n$. D'après l'exercice ci-dessus, $R(\hat{\theta}, \theta) \leq 1/(4n)$, donc pour tout $t > 0$
$$P_\theta[|\hat{\theta} - \theta| > t] \leq \frac{1}{4nt^2}$$
soit aussi, en prenant l'événement complémentaire,
$$P_\theta[\theta \in [\hat{\theta} - t, \hat{\theta} + t]] \geq 1 - \frac{1}{4nt^2}.$$
Ainsi pour que $[\hat{\theta} - t, \hat{\theta} + t]$ soit un intervalle de confiance de niveau $1-\alpha$ il suffit de choisir t de sorte que $\alpha = 1/(4nt^2)$ soit $t = 1/\sqrt{4n\alpha}$. On a donc obtenu que
$$I(X) = \left[\hat{\theta} - \frac{1}{\sqrt{4n\alpha}}, \hat{\theta} + \frac{1}{\sqrt{4n\alpha}}\right] = \left[\hat{\theta} \pm \frac{1}{\sqrt{4n\alpha}}\right]$$
est un intervalle de confiance de niveau au moins $1-\alpha$.

Remarque importante : $I(X)$ ne doit pas dépendre de θ ! Or en général $R(\hat{\theta}, \theta)$ dépend de θ. Par exemple dans l'exemple binomial, il vaut $\theta(1-\theta)/n$. C'est pourquoi on peut dans ce cas

le majorer pour obtenir une quantité indépendante de θ.

On peut aussi utiliser d'autres inégalités à la place de celle de Tchébychev, par exemple celles de Markov, Hoeffding ... (voir TDs). On peut parfois aussi utiliser la loi de $\hat{\theta}$ si elle est connue, ce qui n'est pas très fréquent, pour construire l'intervalle de confiance.

Exercice. Dans le modèle fondamental avec observations $X_1, \dots, X_n$ iid de loi $\mathcal{N}(\theta, 1)$, construire un intervalle de confiance pour θ de niveau $1-\alpha$ à partir de $\hat{\theta} = \overline{X}$. On donne l'inégalité $P[|\mathcal{N}(0,1)| \geq t] \leq e^{-t^2/2}$ pour tout $t > 0$.

2ème technique [intervalles de confiance asymptotiques]. On peut utiliser une convergence en loi quand $n \to \infty$ pour construire un intervalle de confiance asymptotique. Supposons, pour $\Theta \subset \mathbb{R}$, que l'on dispose d'un estimateur $\hat{\theta} = \hat{\theta}_n(X)$ asymptotiquement normal, soit

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta)).$$

et que $\theta \to \sigma^2(\theta)$ est continue. Soit z_α tel que $P[|\mathcal{N}(0,1)| \leq z_\alpha] = 1-\alpha$. Alors

$$I(X) = \left[\hat{\theta}_n(X) - z_\alpha \frac{\sigma(\hat{\theta}_n(X))}{\sqrt{n}} \, , \, \hat{\theta}_n(X) + z_\alpha \frac{\sigma(\hat{\theta}_n(X))}{\sqrt{n}} \right]$$

est un intervalle de confiance asymptotique de niveau $1-\alpha$, c'est-à-dire un intervalle tel que

$$\liminf_{n \to \infty} P_\theta[\theta \in I_n(X)] = 1 - \alpha.$$

Preuve.

On constate que

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} = \frac{\sigma(\hat{\theta}_n)}{\sigma(\theta)} \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)}.$$

Comme $\hat{\theta}_n$ est asymptotiquement normal, il est consistant, voir exercice en 1.4, donc $\hat{\theta}_n \xrightarrow{P} \theta$. Par image continue (Proposition 4), on en déduit $\sigma(\hat{\theta}_n) \xrightarrow{P} \sigma(\theta)$. Par ailleurs,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Grâce au lemme de Slutsky (Proposition 5), on en déduit

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)s}{\sigma(\hat{\theta}_n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

La proposition 2 permet d'en déduire

$$P_\theta \left[-z_\alpha \leq \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} \leq z_\alpha \right] \to P\left[|\mathcal{N}(0,1)| \leq z_\alpha\right] = 1 - \alpha,$$

ce qui s'écrit exactement $P_\theta[\theta \in I(X)] \to 1-\alpha$, ce qu'il fallait démontrer.

VRAISEMBLANCE.

Soit $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ un modèle statistique et $X_1, \dots, X_n$ des observations iid de loi P_θ. Supposons le modèle dominé par rapport à une mesure dominante μ, soit $dP_\theta = p_\theta d\mu$. La densité du n-uplet $(X_1, \dots, X_n)$ par rapport à $\mu^{\otimes n}$ est donc $p_\theta(x_1) \dots p_\theta(x_n)$. Cette densité prise calculée en θ et aux points d'observation s'appelle *vraisemblance*.

Definition 12. La fonction vraisemblance est la fonction

$$\mathcal{V} : \theta \to \prod_{i=1}^{n} p_\theta(X_i).$$

1.6 Lois conditionnelles

On commence par rappeler, pour A, B des événements avec $P(B) > 0$, la définition de la probabilité de 'A sachant B'. Celle-ci est définie par

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

1.6.1 Le cas discret

Cadre. Soit E un ensemble dénombrable, on peut penser à $\mathbb{N}$ pour fixer les idées. Soient X et Y deux variables aléatoires à valeurs dans E.

On souhaite définir la loi conditionnelle de Y sachant X.

Notons que, s'agissant de variables discrètes, les lois de X et Y sont complètement définies par les données de $P[X = e]$ et $P[Y = e]$ pour tous les éléments possibles e de E. Si Q est la loi $\mathcal{L}(Y \,|\, X = x)$ que l'on cherche à définir, il suffit donc aussi de se donner $Q(\{e\})$ pour tout $e \in E$. On définit tout simplement ces quantités à l'aide de la formule ci-dessus pour la probabilité de A sachant B.

Definition 13. Soit $x \in E$ fixé. La loi conditionnelle de $Y \,|\, X = x$, parfois aussi notée $\mathcal{L}(Y \,|\, X = x)$ est définie par, pour tout $e \in E$, et $x \in E$ tel que $P[X = x] > 0$,

$$P[Y = e \,|\, X = x] = \frac{P[Y = e, X = x]}{P[X = x]}.$$

Exemple. Soient Y, Z deux variables aléatoire *indépendantes* de lois $Y \sim \mathrm{Be}(1/2)$ et $Z \sim \mathrm{Be}(1/2)$. On pose $X = Y + Z$. Quelle est la loi conditionnelle $\mathcal{L}(Y \,|\, X = 1)$?

Notons déjà que $X = 1$ si et seulement $Y = 1$ et $Z = 0$, ou bien $Y = 0$ et $Z = 1$. En utilisant la définition de la loi conditionnelle ainsi que l'indépendance de Y et Z,

$$P[Y = 1 \,|\, X = 1] = \frac{P[X = 1, Y = 1]}{P[X = 1]} = \frac{P[Z = 0, Y = 1]}{P[Y = 1, Z = 0] + P[Y = 0, Z = 1]}$$
$$= \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}} = \frac{1}{2}.$$

Par ailleurs, comme Y ne prend que les valeurs 0 ou 1, on en déduit que $P[Y = 0 \,|\, X = 1] = 1 - P[Y = 1 \,|\, X = 1] = 1 - \frac{1}{2} = \frac{1}{2}$. On en conclut que

$$\mathcal{L}(Y \,|\, X = 1) = \mathrm{Be}\Big(\frac{1}{2}\Big) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

En procédant de la même manière, on obtient (exercice)

$$\mathcal{L}(Y \,|\, X = 0) = \delta_0 \quad \text{et} \quad \mathcal{L}(Y \,|\, X = 2) = \delta_1.$$

Par extension, on définit la loi conditionnelle de Y sachant X, notée $\mathcal{L}(Y \,|\, X)$, comme la loi égale à $\mathcal{L}(Y \,|\, X = x)$ si $X = x$. Dans l'exemple ci-dessus,

$$\mathcal{L}(Y \,|\, X) = \begin{cases} \delta_0 & \text{si } X = 0 \\ \mathrm{Be}\left(\frac{1}{2}\right) & \text{si } X = 1 \\ \delta_1 & \text{si } X = 2, \end{cases}$$

ce qu'on peut aussi écrire de manière un peu plus compacte comme

$$\mathcal{L}(Y \,|\, X) = \left(1 - \frac{X}{2}\right)\delta_0 + \frac{X}{2}\delta_1.$$

1.6.2 Le cas à densité

Cadre.

On se donne

- un espace E muni d'une tribu $\mathcal{E}$ et un espace F muni d'une tribu $\mathcal{F}$
- une mesure α positive σ-finie sur E et une mesure β positive σ-finie sur F
- une variable aléatoire X sur E et une variable aléatoire Y sur F.

On *suppose* que le couple (X, Y) admet une densité notée $f(x, y)$ par rapport à $\alpha \otimes \beta$, ce que l'on écrit aussi, si $P_{X,Y}$ désigne la loi du couple,

$$dP_{X,Y}(x, y) = f(x, y) d\alpha(x) d\beta(y).$$

Loi et densité marginales.

Proposition 6. Dans le cadre ci-dessus, la loi de X seule, appelée loi marginale de X, est la loi P_X de densité f_X donnée par

$$f_X(x) = \int f(x,y)d\beta(y).$$

Preuve.

Pour toute fonction g mesurable bornée, en utilisant le théorème de Fubini,

$$\begin{aligned} E[g(X)] &= \int\int g(x)f(x,y)d\alpha(x)d\beta(y) \\ &= \int g(x)\left[\int f(x,y)d\beta(y)\right]d\alpha(x) = \int g(x)f_X(x)d\alpha(x). \end{aligned}$$

De même, la loi marginale de Y est la loi P_Y dont la densité sur F par rapport à β est donnée par $f_Y(y) = \int f(x,y)d\alpha(x)$.

✍ A partir de la loi du couple (X,Y), on a facilement déduit les lois individuelles de X et Y. Il est important de noter que l'opération inverse n'est pas possible en général sans hypothèse supplémentaire. Il y a en général beaucoup de lois jointes possibles correspondant à deux lois marginales données, voir TDs.

Loi conditionnelle.

Definition 14. La loi conditionnelle de Y sachant $X = x$ est la loi de densité, sur F par rapport à β, donnée par, pour $f_X(x) > 0$,

$$f_{Y\,|\,X=x}(y) = \frac{f(x,y)}{\int f(x,y)d\beta(y)} = \frac{f(x,y)}{f_X(x)}.$$

On notera parfois $f(y\,|\,x)$ au lieu de $f_{Y\,|\,X=x}(y)$ s'il n'y a pas de risque de confusion. Notons que par définition, $y \to f(y\,|\,x)$ est une densité par rapport à β, soit $\int f(y\,|\,x)d\beta(y) = 1$.

Notons que pour avoir une quantité définie pour tous les x de E, on peut étendre la définition de $f_{Y\,|\,X=x}(y)$ au cas où $f_X(x) = 0$ en posant le quotient ci-dessus égal à une valeur quelconque (par exemple 0) lorsque $f_X(x) = 0$. Ces points x n'auront typiquement pas d'incidence dans les calculs car l'ensemble des x tels que $f_X(x) = 0$ est un ensemble de P_X-mesure nulle.

Exercice. Vérifier que le cas discret est un cas particulier de la formule ci-dessous, pour lequel

E et F sont dénombrables, et α, β sont les mesures de comptage sur E et F respectivement, $\alpha = \sum_{e\in E} \delta_e$, $\beta = \sum_{f\in F} \delta_f$.

✍ A partir de la densité conditionnelle de $Y \mid X$ et de la densité marginale de X, on retrouve la densité jointe du couple (X, Y), puisque par définition $f(x, y) = f_{Y\mid X=x}(y) f_X(x)$.

Exemple. Soit un couple (X, Y) de variables aléatoires sur $E = \mathbb{R}^+ \times \mathbb{R}^+$ de densité

$$f(x, y) = xe^{-x(y+1)}$$

par rapport à la mesure de Lebesgue restreinte à $\mathbb{R}^+ \times \mathbb{R}^+$. Déterminons la loi conditionnelle de Y sachant X. Il suffit de diviser la densité jointe $f(x, y)$ par la densité marginale $f_X(x) = \int_0^\infty xe^{-x(y+1)}dy = e^{-x}$. Ainsi

$$f_{Y\mid X=x}(y) = \frac{xe^{-x(y+1)}}{e^{-x}} = xe^{-xy}$$

On reconnait la densité d'une loi exponentielle de paramètre x. Ainsi, $\mathcal{L}(Y \mid X = x) = \mathcal{E}(x)$. On écrit aussi $\mathcal{L}(Y \mid X) = \mathcal{E}(X)$. Notons que la loi marginale de X a pour densité e^{-x}, ainsi $\mathcal{L}(X) = \mathcal{E}(1)$.

Utilisation du symbole $\propto$='proportionnel à'. Une autre façon de faire pour déterminer la densité conditionnelle est de remarquer qu'il s'agit de reconnaître dans l'expression $f(x, y)/f_X(x)$ une densité en y. En ce sens $f_X(x)$ est simplement une constante de normalisation. De même, tout facteur dépendant seulement de x dans $f(x, y)$ peut se mettre en facteur et intervient seulement dans la normalisation. On écrit ceci à l'aide du symbole proportionnel à '$\propto$'

$$xe^{-x(y+1)} = xe^{-x}e^{-xy} \propto e^{-xy}.$$

La loi dont la densité en y est proportionnelle à e^{-xy} est bien la loi $\mathcal{E}(x)$. Cette méthode évite de devoir calculer la densité marginale $f_X(x)$. Dans cet exemple, ce calcul était immédiat mais ce n'est pas toujours le cas, nous verrons d'autres exemples au prochain chapitre.

Exercice. Déterminer la densité de la loi marginale de Y et montrer que la loi conditionnelle de $X \mid Y$ est une loi Gamma $\Gamma(2, Y + 1)$.

Notion d'espérance conditionnelle

On rappelle l'abbréviation $f(y \mid x) = f_{Y\mid X=x}(y)$.

Definition 15. Si $E[|Y|] < \infty$, on définit l'espérance conditionnelle $E[Y \mid X]$ par

$$E[Y \mid X] = \int yf(y \mid X)d\beta(y).$$

Plus généralement, pour ϕ mesurable avec $\phi(Y)$ intégrable,

$$E[\phi(Y) \mid X] = \int \phi(y)f(y \mid X)d\beta(y).$$

Proposition 7. Pour toute $h : E \times F \to \mathbb{R}$ mesurable, à condition que la variable $h(X,Y)$ soit intégrable,

$$E[h(X,Y)] = E[E[h(X,Y) \,|\, X]] = \int\int h(x,y) dP_{Y \,|\, X=x}(y) dP_X(x).$$

En particulier, sous les mêmes conditions, si $h(X,Y) = \varphi(X)\psi(Y)$, pour φ, ψ mesurables,

$$E[\psi(Y)\varphi(X)] = E[E[\psi(Y) \,|\, X]\varphi(X)].$$

Preuve.

$$\begin{aligned} E[h(X,Y)] &= \int\int h(x,y) f(x,y) d\alpha(x) d\beta(y) \\ &= \int\int h(x,y) \frac{f(x,y)}{f_X(x)} f_X(x) d\alpha(x) d\beta(y) \\ &= \int \left[\int h(x,y) dP_{Y \,|\, X=x}(y) \right] f_X(x) d\alpha(x), \end{aligned}$$

où on a utilisé le théorème de Fubini pour la dernière égalité.

Proposition 8. Dans le cadre précédent, soit (X,Y) un couple de variables aléatoires de densité $f(x,y)$ par rapport à $\alpha \otimes \beta$. Supposons Y de carré intégrable : $E[Y^2] < \infty$. Alors

$$\inf \left\{ E[\,(Y - h(X))^2\,], \quad Eh(X)^2 < \infty \right\} = E[\,(Y - g(X))^2\,],$$

où $g(u) = E[Y \,|\, X = u]$.

Preuve.

On note que pour toute h telle que $E[h(X)^2] < \infty$,

$$E[(Y - h(X))^2] = E[(Y - g(X))^2] + E[(g(X) - h(X))^2].$$

En effet, le double produit est nul puisque, comme $g(X) = E[Y \,|\, X]$,

$$\begin{aligned} E[(Y - g(X))(g(X) - h(X))] &= E[E[Y - g(X) \,|\, X](g(X) - h(X))] \\ &= E[(g(X) - g(X))(g(X) - h(X))] = 0. \end{aligned}$$

On déduit de la première identité ci-dessus que $E[(Y - h(X))^2] \geq E[(Y - g(X))^2]$ pour toute h telle que $E[h(X)^2] < \infty$. Pour conclure il suffit de montrer que $E[g(X)^2] < \infty$. Or

$$Eg(X)^2 = E\left[\int yf(y \mid X)d\beta(y)\right]^2 = \int \left[\int yf(y \mid x)d\beta(y)\right]^2 f_X(x)d\alpha(x)$$
$$\leq \int\int y^2 f(y \mid x)d\beta(y)f_X(x)d\alpha(x) = \int\int y^2 f(x,y)d\beta(y)d\alpha(x),$$

où la dernière ligne résulte de l'inégalité de Cauchy-Schwarz (ou Jensen pour le carré)

$$\left[\int yf(y \mid x)d\beta(y)\right]^2 \leq \int y^2 f(y \mid x)d\beta(y) \int f(y \mid x)d\beta(y) = \int y^2 f(y \mid x)d\beta(y),$$

puisque par définition $y \to f(y \mid x)$ est une densité. En utilisant le théorème de Fubini, nous constatons que $\int\int y^2 f(x,y)d\beta(y)d\alpha(x) = \int\int y^2 f(x,y)d\alpha(x)d\beta(y) = E[Y^2] < \infty$ par hypothèse, ce qui montre $E[g(X)^2] < \infty$.

Interprétation en termes de projection. Soit $\mathcal{H}$ l'espace vectoriel (fermé) de toutes les fonctions $h(X)$ avec h mesurable et $E[h(X)^2] < \infty$. La fonction g, l'espérance conditionnelle de Y sachant $X = \cdot$, est simplement la projection orthogonales de Y sur $\mathcal{H}$.

Remarque culturelle. Le cadre à densité n'est pas le seul cadre où l'on puisse définir des lois conditionnelles. On peut plus généralement proposer une définition de la loi conditionnelle comme opérateur de 'désintégration', dans l'esprit de l'identité de la Proposition 7. Cette notion plus générale est utile notamment en statistique bayésienne pour des modèles complexes, lorsque θ n'est plus un paramètre d'un espace de dimension finie mais par exemple une fonction, mais nous ne la considèrerons pas dans le cadre de ce cours.

1.7 Plan du cours

1. Introduction
2. L'approche bayésienne
3. Bayésien et théorie de la décision
4. Critères de choix de lois a priori
5. Convergences de lois a priori
6. Tests bayésiens
7. Algorithmes de simulation