Chapitre I-1

Modèles statistiques

L'objectif de ce chapitre est d'introduire la notion de **modèle statistique**. Les concepts principaux abordés incluent les observations, le modèle statistique, l'identifiabilité, la notion de statistique, le modèle statistique induit, le modèle statistique dominé, et le n-échantillon. On construit également de premiers estimateurs élémentaires.
La partie principale du chapitre suit le déroulé du cours. Elle est suivie d'une partie *Outils mathématiques*, qui rassemble les prérequis utilisés, et d'une partie *Compléments*, hors programme d'examen, qui développe des exemples supplémentaires (données censurées, sondage sans remise, régression logistique, modèles non dominés).

Qu'est-ce qu'un modèle ?

Moralement, un *modèle statistique* est une *description mathématique simplifiée d'un phénomène aléatoire réel*. Il sert de pont entre le monde des données, que l'on observe, et celui des lois de probabilité, qui représentent nos hypothèses sur la manière dont ces données sont produites. En choisissant un modèle, on fixe un cadre théorique dans lequel on pourra raisonner, estimer des grandeurs inconnues, prédire de nouvelles observations ou tester des hypothèses. Comme toute simplification, un modèle ne capture pas toute la réalité, mais en retient les aspects jugés essentiels pour répondre à une question précise.

Un *modèle statistique* est donc une *hypothèse mathématiquement formalisée* sur le mécanisme aléatoire ayant généré les données. Mathématiquement, nous verrons qu'il correspond à définir un espace d'observations et une famille de lois de probabilité candidates, servant de cadre à l'inférence statistique.

I-1.1 Les fondamentaux

I-1.1.1 Introduction : la démarche statistique

La statistique est l'étude de la *collecte*, de la *modélisation* et du *traitement* des données. Ses applications couvrent aujourd'hui un champ immense : médecine et santé (épidémiologie, génomique, médecine de précision), sciences de l'environnement (pollution, météorologie, climat), assurance et finance (modélisation d'actifs, calcul de risques), apprentissage statistique et traitement du signal (reconnaissance de la parole, vision par ordinateur, traduction automatique), marketing (systèmes de recommandation, scoring). Dans tous ces exemples avancés, les données observables sont complexes. Pour dégager les idées fondamentales, commençons par des questions plus simples :

1. Quelle est la probabilité qu'une pièce retombe sur « pile » ? Les pièces de 1 centime sont-elles biaisées ?
2. Quelle est la taille moyenne d'une personne dans une salle de cours ? Les élèves de Polytechnique sont-ils plus grands que la moyenne ?
3. Quelle proportion des élèves de deuxième année a déjà utilisé un modèle d'IA pour faire un devoir ?

Apporter une réponse à ces questions demande de les reformuler dans un langage mathématique, puis de collecter, modéliser et analyser des données.

Le **point de départ** est donc un jeu de *données* $(x_1, \dots, x_n)$. Celles-ci peuvent être numériques – scalaires, vectorielles ($x_i \in \mathbb{R}^d$), matricielles ($x_i \in \mathbb{R}^{p\times q}$) – ou symboliques (labels, catégories, par exemple $x_i \in \{0, 1\}$), mais aussi plus structurées : graphes, séries chronologiques, fonctions, textes, images, etc.

I-1.1.2 Modélisation statistique

La démarche de modélisation repose sur deux étapes.

1. Nous modélisons les observations comme des *réalisations de variables aléatoires* :

$$\text{Données} = \text{réalisation d'un élément aléatoire } Z = (X_1, \dots, X_n)\ , \quad Z(\omega) = (x_1, \dots, x_n)\ ,$$

   où $(X_1, \dots, X_n)$ sont des variables aléatoires définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ (toujours omis dans la suite), à valeurs dans un espace mesurable $(\mathsf{Z}, \mathcal{Z})$, dit *espace des observations*.
2. Nous modélisons la *loi* de ces variables aléatoires : la loi du vecteur $Z = (X_1, \dots, X_n)$ est *partiellement connue*, c'est un élément d'une famille $\mathcal{C}$ de lois de probabilité.

Définition I-1.1.1: Modèle statistique

Un *modèle statistique* est la donnée de :

— un espace mesurable $(\mathsf{Z}, \mathcal{Z})$, dit l'*espace des observations*,

— une famille de probabilités $\mathcal{C}$ sur $(\mathsf{Z}, \mathcal{Z})$.

Nous notons $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ le modèle statistique.
Le modèle est dit *paramétrique* lorsque $\mathcal{C}$ correspond à une famille $\{\mathbb{P}_\theta\ ,\ \theta \in \Theta\}$, où Θ est un sous-ensemble de $\mathbb{R}^d$, avec $d \geq 1$; i.e., il existe une fonction associant à chaque $\theta \in \Theta$ un élément $\mathbb{P}_\theta$ de $\mathcal{C}$. Le paramètre θ est alors de dimension finie, pour une paramétrisation naturelle du modèle. Il est dit *non paramétrique* lorsqu'il ne se décrit pas par un nombre fini de paramètres réels (par exemple, l'ensemble de toutes les lois à densité continue sur $\mathbb{R}$).

Remarque. Lorsque l'espace des observations $(\mathsf{Z}, \mathcal{Z})$ est clair d'après le contexte, on pourra, par abus de langage, appeler « modèle statistique » la seule famille de lois $\mathcal{C}$; on parlera ainsi du modèle $\{\mathbb{P}_\theta\ ,\ \theta \in \Theta\}$ sans rappeler l'espace mesurable sur lequel ces lois sont définies.

Exemple I-1.1.1: Pile ou face

On observe le nombre de « piles » issus de n lancers d'une pièce. On observe une unique réalisation d'une variable aléatoire à valeurs dans $\{0,\dots,n\}$, muni de la tribu des parties. En supposant les lancers indépendants, chacun de loi de Bernoulli $\mathcal{B}\mathrm{er}(\theta)$ pour un $\theta \in [0,1]$, l'observation a pour loi $\mathrm{Bin}(n,\theta)$. Le modèle s'écrit

$$(\{0,\dots,n\}, \mathcal{P}(\{0,\dots,n\}), \{\mathrm{Bin}(n,\theta)\,,\ \theta\in[0,1]\})\ .$$

Exemple I-1.1.2: Tailles

On observe la taille de n personnes : l'observation est un vecteur de $\mathbb{R}^n$, muni de la tribu borélienne $\mathcal{B}(\mathbb{R}^n)$. On modélise la taille par une loi gaussienne de moyenne μ et de variance σ^2, et les différentes observations par des variables aléatoires indépendantes. Le modèle est

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{\mathcal{N}(\mu\mathbf{1}_n, \sigma^2\mathrm{I}_n)\,,\ \mu\in\mathbb{R},\ \sigma^2\in\mathbb{R}_+^*\right\}\right)\ .$$

Exemple I-1.1.3: Temps d'attente

On mesure n temps d'attente, en minutes : l'observation est un vecteur de $\mathbb{R}_+^n$, muni de sa tribu borélienne. Deux situations conduisent à deux modèles différents.

1. *Des bus à horaire régulier.* Les bus partent toutes les θ minutes et l'on arrive à l'arrêt à un instant sans lien avec l'horaire : on modélise le temps d'attente par une loi uniforme sur $[0,\theta]$, et les n attentes, observées des jours différents, par des variables aléatoires indépendantes. Le modèle est

$$\left(\mathbb{R}_+^n,\ \mathcal{B}(\mathbb{R}_+^n),\ \left\{\mathcal{U}([0,\theta])^{\otimes n}\,,\ \theta>0\right\}\right).$$

2. *Des clients qui arrivent dans un café.* Les arrivées sont dispersées et sans mémoire : le temps qui sépare deux arrivées successives ne dépend pas du temps déjà écoulé depuis la dernière. Cette propriété caractérise la loi exponentielle (Section I-1.2) ; on modélise les n temps entre arrivées successives par des variables aléatoires indépendantes de loi $\mathcal{E}(\lambda)$, où $\lambda>0$ est le nombre moyen d'arrivées par minute. Le modèle est

$$\left(\mathbb{R}_+^n,\ \mathcal{B}(\mathbb{R}_+^n),\ \left\{\mathcal{E}(\lambda)^{\otimes n}\,,\ \lambda>0\right\}\right).$$

Dans les deux cas, l'espace des observations et la tribu sont les mêmes ; seule la famille de lois change, et avec elle l'hypothèse faite sur le mécanisme des arrivées. Le second modèle est le point de départ de l'Exercice 1 (PC1).

Exemple I-1.1.4: Modèle de sondage

Une élection entre deux candidats A et B a lieu : on effectue un sondage à la sortie des urnes en interrogeant n votants. On appelle x_i la réponse du i-ème individu sondé : $x_i = 1$ si le i-ème individu sondé vote A et $x_i = 0$ sinon. En considérant que l'expérience consiste en un tirage uniforme de n individus avec replacement dans la population, le modèle statistique associé est donné par

$$\left(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \left\{\mathcal{B}\mathrm{er}(\theta)^{\otimes n}, \theta\in\Theta\right\}\right),$$

où $\mathcal{P}(\{0,1\}^n)$ est l'ensemble des parties de $\{0,1\}^n$ et $\mathcal{B}\mathrm{er}(\theta)^{\otimes n}$ est le produit de n lois de

Bernoulli de paramètre θ (voir Section I-1.2), i.e.

$$\mathcal{B}\mathrm{er}(\theta)^{\otimes n}(\{x_1,\dots,x_n\}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \,.$$

Par ce choix de la famille de lois, on exprime que les tirages sont indépendants et le résultat de chaque tirage est une variable de Bernoulli de paramètre $\theta \in \Theta$. Le cas d'un tirage *sans* replacement, qui conduit à la loi hypergéométrique, est traité en Section I-1.4.

Intuition. *Construire un modèle, c'est répondre à trois questions :* qu'est-ce que j'observe ? *(l'espace* $(\mathsf{Z},\mathcal{Z})$*),* quels mécanismes aléatoires ai-je envisagés ? *(la famille* $\mathcal{C}$*), et* que veux-je apprendre ? *(la problématique). Plus la famille* $\mathcal{C}$ *est grande, moins le modèle fait d'hypothèses – mais plus il sera difficile d'en extraire une réponse précise.*

La plupart des modèles rencontrés dans ce cours s'écrivent sous la forme « $p_\theta \cdot \mu$ » (une densité par rapport à une mesure de référence). Cette propriété porte un nom :

Définition I-1.1 (Famille dominée et modèle statistique dominé). *Soient* $(\mathsf{Z},\mathcal{Z})$ *un espace mesurable et* μ *une mesure* σ*-finie sur* $(\mathsf{Z},\mathcal{Z})$*. Une* famille *de lois* $\mathcal{C}$ *est* dominée *par la mesure* μ *si toute loi* $\mathbb{P} \in \mathcal{C}$ *admet une densité de probabilité* p *par rapport à* μ*, i.e.* $\mathbb{P} = p \cdot \mu$*.*

Le modèle statistique $(\mathsf{Z},\mathcal{Z},\mathcal{C})$ *est dit* dominé *lorsque la famille de lois* $\mathcal{C}$ *est dominée.*

Exemple I-1.1.5: Modèles dominés : quelques densités

Une loi et sa densité sont deux objets différents : la densité est une fonction, la loi est la mesure obtenue en l'intégrant contre la mesure de référence, $\mathbb{P}(A) = \int_A p \,\mathrm{d}\mu$. Il faut d'ailleurs toujours préciser par rapport à quelle mesure une loi admet une densité : une densité est définie pour une loi *par rapport à une mesure (*σ*-finie) qui la domine.*

1. **Le modèle gaussien à** n **observations** est dominé par la mesure de Lebesgue $\mathrm{Leb}^{\otimes n}$, avec la densité

$$p_{\mu,\sigma^2}(x_1,\dots,x_n) = (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\Big) = \prod_{i=1}^n (2\pi\sigma^2)^{-1/2}\exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big)\,,$$

   et $\mathcal{N}(\mu,\sigma^2)^{\otimes n}(A) = \int_A p_{\mu,\sigma^2}\,\mathrm{dLeb}^{\otimes n}$ pour $A \in \mathcal{B}(\mathbb{R}^n)$; ici $p_{\mu,\sigma^2}(x)$ n'est pas une probabilité (la loi d'un singleton est nulle). Les densités des autres lois usuelles du cours (exponentielle, Gamma, uniforme) sont rappelées en Section I-1.2.

2. **Une observation de Poisson.** On observe un comptage X à valeurs dans $\mathbb{N}$ (nombre de clients arrivés en une heure, par exemple), modélisé par $\big(\mathbb{N},\mathcal{P}(\mathbb{N}),\{\mathrm{Poisson}(\lambda)\,,\ \lambda>0\}\big)$ (voir l'Exercice 1 (PC1)). Le modèle est dominé par la mesure de comptage μ sur $\mathbb{N}$, avec la densité

$$p_\lambda(k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}\,, \qquad \text{de sorte que} \qquad \mathrm{Poisson}(\lambda)(A) = \sum_{k\in A} p_\lambda(k) \quad \text{pour } A \subset \mathbb{N}\,.$$

   Ici $p_\lambda(k)$ coïncide avec la probabilité du singleton $\{k\}$: c'est propre à la mesure de comptage, et cela ne dispense pas de distinguer la fonction p_λ de la loi $\mathrm{Poisson}(\lambda)$.

3. **Le modèle de Bernoulli à** n **observations** est dominé par la mesure de comptage sur $\{0,1\}^n$, avec la densité

$$p_\theta(x_1,\dots,x_n) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}\,, \qquad \mathcal{B}\mathrm{er}(\theta)^{\otimes n}(A) = \sum_{x\in A} p_\theta(x) \quad \text{pour } A \subset \{0,1\}^n\,.$$

La mesure de domination n'est jamais unique : si μ domine le modèle, toute mesure σ-finie μ' telle que $\mu \ll \mu'$ le domine aussi. Ce choix est sans conséquence pour la statistique, car deux densités relatives à deux mesures dominantes diffèrent d'un facteur multiplicatif qui ne dépend pas de θ ; ce point, ainsi que des exemples de modèles non dominés, sont discutés en Section I-1.4. La modélisation d'arrivées successives (temps exponentiels, comptages poissoniens) est l'objet de l'Exercice 1 (PC1).

I-1.1.3 Objectifs de la statistique

À partir des observations, on cherche à *affiner notre connaissance de la loi de l'observation*. Plus spécifiquement, on distingue plusieurs tâches :

- **estimation ponctuelle** : donner une valeur *plausible* du paramètre θ ;
- **estimation par région** : déterminer un sous-ensemble $\Theta_0(Z) \subset \Theta$ auquel le paramètre θ appartient de façon plausible ;
- **test** : décider si le paramètre θ appartient à un sous-ensemble Θ_0, et évaluer la plausibilité de la décision ;
- **prédiction** : donner une valeur plausible pour une quantité non observée, par exemple prédire la valeur d'une sortie Y à partir d'une entrée X. Le modèle de régression linéaire gaussienne, outil central de prédiction, est l'objet de l'Exercice 4 (PC1).

Nous étudierons ces différentes tâches au fur et à mesure du cours : l'estimation aux cours 2 et 3, les régions de confiance au cours 3, les tests aux cours 4 et 5, et les problèmes de régression, que l'on rencontrera aussi bien en statistique qu'en apprentissage.

Statistique et probabilités. En *probabilités*, la loi est supposée *connue* : étant donnée une loi $\mathbb{P}$ et un vecteur aléatoire Z, on évalue des quantités de la forme $\mathbb{P}(f(Z) \geq c)$, $\mathbb{E}[f(Z)]$, etc. En *statistique*, on résout un *problème inverse* : étant donnée une *réalisation* $z = (x_1, \ldots, x_n)$ d'un élément aléatoire $Z = (X_1, \ldots, X_n)$, on cherche à *inférer* certaines caractéristiques de la loi de ce vecteur.

Le modèle comme point de départ. Dans ce cours de *statistique mathématique*, le modèle est le point de départ, et tout ce qui en est déduit est mathématiquement exact. La phrase

> « *on considère le modèle* $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ »

signifie : « on suppose observer une réalisation d'une variable aléatoire Z à valeurs dans Z dont la loi est un élément de $\mathcal{C}$ ». Par exemple, « on considère le modèle $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{q_\lambda \cdot \mathrm{Leb},\ \lambda > 0\})$ » avec $q_\lambda(x) = \lambda \mathrm{e}^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$ signifie que l'on suppose observer X de loi exponentielle de paramètre λ, pour un certain $\lambda > 0$. On souligne cette hypothèse, et la dépendance dans le paramètre, par les notations

$$\mathbb{E}_\theta[\cdot]\,, \qquad \mathbb{P}_\theta(\cdot)\,,$$

qui désignent l'espérance et la probabilité sous l'hypothèse que l'observation suit la loi $\mathbb{P}_\theta$.

Un modèle est une hypothèse. C'est même précisément notre hypothèse de départ pour toute la partie mathématique de la statistique. *« Tous les modèles sont faux, mais certains sont utiles ! »* (George Box). En pratique, la pertinence d'une modélisation dépend du contexte : modéliser un pile ou face par une loi de Bernoulli, des tailles par une gaussienne, les réponses d'un sondage téléphonique par des Bernoulli i.i.d. , ou un temps d'attente par une loi exponentielle, sont des simplifications utiles mais discutables (dépendances mécaniques, queues de distribution, biais de réponse, non-stationnarité...).

Intuition. *Il faut distinguer deux niveaux :* à l'intérieur *du modèle, tous les énoncés sont rigoureux ;* le choix *du modèle, lui, est une hypothèse de modélisation, qui peut être inexacte. La conclusion statistique sur le monde réel doit donc être prise avec la même précaution que la conclusion de tout modèle physique.*

I-1.1.4 Identifiabilité

Définition I-1.2 (Identifiabilité). *Soit un modèle paramétrique* $(\mathsf{Z}, \mathcal{Z}, \{\mathbb{P}_\theta\,,\ \theta \in \Theta\})$. *Le modèle est* identifiable *si la fonction* $\theta \mapsto \mathbb{P}_\theta$ *est injective, i.e. si* $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ *implique* $\theta = \theta'$.

L'identifiabilité signifie que la connaissance complète de la loi $\mathbb{P}_\theta$ permet d'identifier de façon unique θ. C'est donc une propriété *minimale* pour pouvoir espérer retrouver θ à partir des observations.

Exemple I-1.1.6: Les modèles rencontrés jusqu'ici sont identifiables

Les modèles paramétriques rencontrés jusqu'ici sont tous identifiables, à savoir :

1. le modèle du pile ou face $(\{0,\dots,n\}, \mathcal{P}(\{0,\dots,n\}), \{\mathrm{Bin}(n,\theta)\,,\ \theta \in [0,1]\})$, le nombre de lancers $n \geq 1$ étant fixé ;
2. le modèle des tailles $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \big\{\mathcal{N}(\mu\mathbf{1}_n, \sigma^2\mathrm{I}_n)\,,\ (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+\big\}\big)$, où l'on autorise $\sigma^2 = 0$ avec la convention $\mathcal{N}(\mu\mathbf{1}_n, 0) = \delta_{\mu\mathbf{1}_n}$, masse de Dirac au point $\mu\mathbf{1}_n$ (toutes les tailles sont alors égales à μ) ;
3. le modèle de sondage avec remise $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathcal{B}\mathrm{er}(\theta)^{\otimes n}\,,\ \theta \in [0,1]\})$;
4. le modèle exponentiel du temps d'attente $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \{q_\lambda \cdot \mathrm{Leb}\,,\ \lambda > 0\})$, où $q_\lambda(x) = \lambda \mathrm{e}^{-\lambda x}\mathbb{1}_{\mathbb{R}_+}(x)$.

Dans chaque cas, il s'agit de vérifier que la connaissance de la loi $\mathbb{P}_\theta$ détermine la valeur du paramètre, autrement dit que deux valeurs distinctes du paramètre donnent deux lois distinctes.

Démonstration

On montre dans chaque cas que le paramètre s'exprime en fonction de la loi $\mathbb{P}_\theta$, ce qui entraîne l'injectivité de $\theta \mapsto \mathbb{P}_\theta$.

1. *Pile ou face.* Pour $\theta \in [0,1]$, la formule $\mathrm{Bin}(n,\theta)(\{k\}) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ donne, pour $k = n$, la probabilité d'obtenir n piles : $\mathrm{Bin}(n,\theta)(\{n\}) = \theta^n$. Si $\mathrm{Bin}(n,\theta) = \mathrm{Bin}(n,\theta')$, alors $\theta^n = \theta'^n$, donc $\theta = \theta'$, l'application $t \mapsto t^n$ étant strictement croissante, donc injective, sur $[0,1]$.
2. *Tailles.* Une loi de probabilité sur $\mathbb{R}^n$ admettant un moment d'ordre deux détermine son vecteur espérance et sa matrice de covariance. Pour $(\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+$, la loi $\mathcal{N}(\mu\mathbf{1}_n, \sigma^2\mathrm{I}_n)$ a pour vecteur espérance $\mu\mathbf{1}_n$ et pour matrice de covariance $\sigma^2\mathrm{I}_n$: ce sont les paramètres qui la définissent, et c'est encore vrai pour $\sigma^2 = 0$, la masse de Dirac $\delta_{\mu\mathbf{1}_n}$ ayant pour espérance $\mu\mathbf{1}_n$ et pour matrice de covariance la matrice nulle. Si $\mathcal{N}(\mu\mathbf{1}_n, \sigma^2\mathrm{I}_n) = \mathcal{N}(\mu'\mathbf{1}_n, \sigma'^2\mathrm{I}_n)$, alors $\mu\mathbf{1}_n = \mu'\mathbf{1}_n$ et $\sigma^2\mathrm{I}_n = \sigma'^2\mathrm{I}_n$, d'où $\mu = \mu'$ et $\sigma^2 = \sigma'^2$. L'application $(\mu,\sigma^2) \mapsto \mathcal{N}(\mu\mathbf{1}_n, \sigma^2\mathrm{I}_n)$ est donc injective sur $\mathbb{R}\times\mathbb{R}_+$, et a fortiori sur $\mathbb{R}\times\mathbb{R}_+^*$.
3. *Sondage avec remise.* D'après l'expression de $\mathcal{B}\mathrm{er}(\theta)^{\otimes n}$, la probabilité que les n personnes interrogées votent toutes pour A vaut $\mathcal{B}\mathrm{er}(\theta)^{\otimes n}(\{(1,\dots,1)\}) = \theta^n$. Si $\mathcal{B}\mathrm{er}(\theta)^{\otimes n} = \mathcal{B}\mathrm{er}(\theta')^{\otimes n}$, alors $\theta^n = \theta'^n$, et l'on conclut comme au premier point.
4. *Temps d'attente.* Pour $\lambda > 0$, la loi $\mathbb{P}_\lambda = q_\lambda \cdot \mathrm{Leb}$ vérifie $\mathbb{P}_\lambda([1,+\infty[) = \int_1^{+\infty} \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \mathrm{e}^{-\lambda}$. Si $\mathbb{P}_\lambda = \mathbb{P}_{\lambda'}$, alors $\mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda'}$, donc $\lambda = \lambda'$ par injectivité de l'exponentielle.

□

Exemple I-1.1.7: Mélange gaussien

Les modèles de l'exemple précédent sont identifiables ; voici un exemple qui ne l'est pas. Soit $\sigma^2 > 0$ *connu*. Considérons le modèle de mélange gaussien sur $\mathbb{R}$ dont les lois ont pour densité

par rapport à Leb

$$p_{\pi,\mu_1,\mu_2} = \pi\, p_{\mathcal{N}(\mu_1,\sigma^2)} + (1-\pi)\, p_{\mathcal{N}(\mu_2,\sigma^2)}\ , \qquad (\pi,\mu_1,\mu_2) \in [0,1]\times\mathbb{R}^2\ ,$$

où π est la proportion de la première composante et μ_1, μ_2 les moyennes. Ce modèle n'est *pas* identifiable : en échangeant les rôles des deux composantes,

$$p_{\pi,\mu_1,\mu_2} = p_{1-\pi,\mu_2,\mu_1}\ .$$

On peut rendre le modèle identifiable en restreignant l'espace des paramètres, par exemple en imposant $\mu_1 < \mu_2$ (et $\pi \in]0,1[$, $\mu_1 \le \mu_2$ avec conventions au bord à préciser). L'identifiabilité des modèles de mélange gaussien et poissonien, une fois cette restriction faite, est démontrée dans l'Exercice 7 (PC1).

I-1.1.5 Statistiques

Étant données des observations, il est possible d'en effectuer des transformations : calculer leur moyenne, ne retenir que la plus grande, les compter, etc. Une *statistique* est une telle transformation, c'est-à-dire une fonction des observations seules, dans laquelle n'interviennent ni le paramètre ni la loi inconnue.

Définition I-1.3 (Statistique). *Soient* $(\mathsf{Z},\mathcal{Z},\mathcal{C})$ *un modèle statistique et* $(\mathsf{T},\mathcal{T})$ *un espace mesurable. On appelle* statistique *sur le modèle statistique* $(\mathsf{Z},\mathcal{Z},\mathcal{C})$ *une application mesurable* T *de* $(\mathsf{Z},\mathcal{Z})$ *à valeurs dans* $(\mathsf{T},\mathcal{T})$.

Remarquons, et cela est très important, que T *ne dépend pas* de la loi $\mathbb{P} \in \mathcal{C}$ et donc ne dépend pas du paramètre si le modèle est paramétrique.

Exemple I-1.1.8

Si $(\mathsf{Z},\mathcal{Z},\mathcal{C}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}_\theta\,,\ \theta\in\Theta\})$, en notant $Z = (X_1,\dots,X_n)$:

1. la i-*ème observation* est la statistique X_i définie pour $z = (x_1,\dots,x_n) \in \mathbb{R}^n$ par $X_i(z) = x_i$;
2. $S_1(Z) = \frac{X_1+\dots+X_n}{n}$, la *moyenne empirique*, notée canoniquement $\bar{X}_n$;
3. $S_2(Z) = \min\{X_i,\ i = 1,\dots,n\}$, et plus généralement les *statistiques d'ordre* $(X_{(i)})_{i=1,\dots,n}$;
4. $S_3(Z) = \mathrm{med}(X_1,\dots,X_n)$, la *médiane empirique*.

Remarque. Une statistique est, d'après la définition ci-dessus, une application mesurable $S : (\mathsf{Z},\mathcal{Z}) \to (\mathsf{T},\mathcal{T})$: c'est une fonction définie sur l'espace des observations, dans laquelle n'intervient aucun aléa. En la composant avec l'observation canonique $Z : (\Omega,\mathcal{A},\mathbb{P}) \to (\mathsf{Z},\mathcal{Z})$, on obtient la variable aléatoire $S\circ Z$, que l'on note $S(Z)$, à valeurs dans $(\mathsf{T},\mathcal{T})$; sa loi est la loi image de celle de Z par S (voir la Section I-1.1.6). Les deux points de vue, fonction mesurable sur l'espace des observations d'une part, variable aléatoire dérivée de $(X_1,\dots,X_n)$ d'autre part, décrivent le même objet, et l'on emploiera le terme de *statistique* pour l'un comme pour l'autre. La notation, elle, les distingue : S_1 désigne l'application $z = (x_1,\dots,x_n) \mapsto \frac{1}{n}\sum_{i=1}^n x_i$, tandis que $\bar{X}_n = S_1(Z) = \frac{1}{n}\sum_{i=1}^n X_i$ désigne la variable aléatoire, et non l'application.
Il reste à vérifier que les exemples ci-dessus sont bien des statistiques, c'est-à-dire des applications mesurables de $(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n))$ dans $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. Toute application continue de $\mathbb{R}^n$ dans $\mathbb{R}$ étant borélienne, il suffit de montrer qu'elles sont continues.

— La i-ème observation $z \mapsto x_i$ et la moyenne empirique $z \mapsto \frac{1}{n}\sum_{i=1}^n x_i$ sont des applications linéaires sur $\mathbb{R}^n$, donc continues.

— Pour $k \in \{1, \dots, n\}$, notons $x_{(1)} \leq \cdots \leq x_{(n)}$ le réarrangement croissant de $(x_1, \dots, x_n)$; la k-ième statistique d'ordre est l'application $z \mapsto x_{(k)}$. Elle est 1-lipschitzienne pour la norme $\|z\|_\infty = \max_{1 \leq i \leq n} |x_i|$. En effet, soient $z, z' \in \mathbb{R}^n$ et $\varepsilon = \|z - z'\|_\infty$. Par définition du réarrangement croissant, au moins k coordonnées de z sont inférieures ou égales à $x_{(k)}$; comme $x'_i \leq x_i + \varepsilon$ pour tout i, au moins k coordonnées de z' sont inférieures ou égales à $x_{(k)} + \varepsilon$, ce qui donne $x'_{(k)} \leq x_{(k)} + \varepsilon$. En échangeant les rôles de z et z', on obtient $|x'_{(k)} - x_{(k)}| \leq \|z - z'\|_\infty$. Les statistiques d'ordre sont donc continues, en particulier le minimum $x_{(1)}$ et le maximum $x_{(n)}$.

— La médiane empirique est une fonction continue des statistiques d'ordre : avec la convention usuelle, elle vaut $x_{(m+1)}$ si $n = 2m+1$ est impair et $\frac{1}{2}\big(x_{(m)} + x_{(m+1)}\big)$ si $n = 2m$ est pair. Elle est donc continue.

Toutes ces applications sont continues, donc boréliennes : ce sont bien des statistiques.

Pour servir de support à l'intuition, il est souvent pratique d'introduire des statistiques associées aux observations individuelles. On parlera *d'observations individuelles* $X_1, \cdots, X_n$; on dit aussi que X_i est la *statistique associée* à la i-ème donnée collectée.

Définition I-1.4 (Statistiques indépendantes). *Nous dirons que les statistiques X et Y sur $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ sont indépendantes si pour toute loi $\mathbb{P} \in \mathcal{C}$, les éléments aléatoires X et Y sont indépendants sous $\mathbb{P}$:*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \, .$$

Plus généralement, des statistiques $X_1, \dots, X_n$ sont indépendantes si, pour toute loi $\mathbb{P} \in \mathcal{C}$ et tous $A_1, \dots, A_n$ mesurables, $\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i)$. Reprenons les modèles de Bernoulli et gaussien à n observations rencontrés plus haut (Section I-1.1.2).

Exemple I-1.1.9: Modèle de Bernoulli à n observations

Dans le modèle $\big(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathcal{B}\mathrm{er}(\theta)^{\otimes n}\, ,\ \theta \in [0,1]\}\big)$, les coordonnées $X_1, \dots, X_n$ sont des statistiques indépendantes, chacune de loi $\mathcal{B}\mathrm{er}(\theta)$ sous $\mathbb{P}_\theta$: c'est la définition même de la loi produit $\mathcal{B}\mathrm{er}(\theta)^{\otimes n}$. On dira communément que les observations sont indépendantes et identiquement distribuées (i.i.d.) de loi $\mathcal{B}\mathrm{er}(\theta)$.

Exemple I-1.1.10: Modèle gaussien à n observations

Dans le modèle $(\mathsf{Z}, \mathcal{Z}, \mathcal{C}) = \big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}_\theta\, ,\ \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\}\big)$, avec $\mathbb{P}_\theta = p_\theta \cdot \mathrm{Leb}^{\otimes n}$ et

$$p_\theta(x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \, ,$$

les coordonnées $X_1, \dots, X_n$ sont des statistiques indépendantes, chacune de loi $\mathcal{N}(\mu, \sigma^2)$ sous $\mathbb{P}_\theta$. On dira communément que les observations $(X_1, \dots, X_n)$ sont i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$.

Démonstration

La densité se factorise, $p_\theta(x_1, \dots, x_n) = \prod_{i=1}^n p_\theta^{X_i}(x_i)$ avec $p_\theta^{X_i}(x) = (2\pi\sigma^2)^{-1/2} \exp\big(-\frac{1}{2\sigma^2}(x-\mu)^2\big)$, densité de $\mathcal{N}(\mu, \sigma^2)$. Par le théorème de Fubini, pour $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}_\theta(X_1 \in A_1, \dots, X_n \in A_n) = \int_{A_1 \times \cdots \times A_n} \prod_{i=1}^n p_\theta^{X_i}(x_i)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = \prod_{i=1}^n \int_{A_i} p_\theta^{X_i}(x)\, \mathrm{d}x = \prod_{i=1}^n \mathbb{P}_\theta(X_i \in A_i) \, ,$$

c'est-à-dire $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)^{\otimes n}$. □

Exemple I-1.1.11: Moyenne et variance empiriques du modèle gaussien : théorème de Gosset

Dans le modèle gaussien à n observations, avec $n \geq 2$, la moyenne empirique et la variance empirique

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \,, \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

sont des statistiques indépendantes ; de plus, sous $\mathbb{P}_{(\mu,\sigma^2)}$, $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ et $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$. C'est le **théorème de Gosset**. On dira communément que, dans un échantillon gaussien, la moyenne et la variance empiriques sont indépendantes.

Démonstration

Le théorème de Gosset est démontré dans les outils mathématiques (Section I-1.2) à partir du théorème de Cochran, appliqué à la décomposition $\mathbb{R}^n = \mathbb{R}\mathbf{1}_n \oplus (\mathbb{R}\mathbf{1}_n)^\perp$: la projection de $(X_1, \ldots, X_n)$ sur $\mathbb{R}\mathbf{1}_n$ est $\bar{X}_n\mathbf{1}_n$ et le carré de la norme de la projection sur l'orthogonal est $(n-1)S_n^2$. L'Exercice 3 (PC1) reprend ce calcul avec la somme des observations et la somme de leurs carrés. Cette indépendance est propre au modèle gaussien : pour n observations i.i.d. d'une loi à variance finie, l'indépendance de $\bar{X}_n$ et de S_n^2 caractérise la loi gaussienne (théorème de Lukacs, 1942). □

Les statistiques canoniques d'un modèle ne sont pas toujours indépendantes. Dans le même modèle gaussien, la somme des observations $\sum_i X_i = n\bar{X}_n$ et la somme de leurs carrés $\sum_i X_i^2 = (n-1)S_n^2 + n\bar{X}_n^2$ ne sont pas indépendantes, alors que $\bar{X}_n$ et S_n^2 le sont : c'est l'objet de l'Exercice 3 (PC1). Dans le modèle de régression de l'Exercice 4 (PC1), les observations sont indépendantes mais pas identiquement distribuées.

I-1.1.6 Loi image d'une statistique et modèle induit

À partir d'un modèle de référence et d'une statistique, c'est-à-dire d'une transformation des observations, on obtient un nouveau modèle statistique, celui de l'observation transformée : c'est le modèle *induit*.

Intuition. *Un modèle induit est le modèle obtenu en* transformant *nos observations : si j'observe $(X_1, \ldots, X_n)$, je peux choisir de ne regarder que $\sum_{i=1}^n X_i$, ou X_1 seule, ou le maximum. Chaque transformation produit un nouveau modèle statistique, en général plus « pauvre » – on a le droit de jeter de l'information, pas d'en créer.*

Transformer les observations par une statistique T transporte chaque loi du modèle sur l'espace d'arrivée de T ; les lois ainsi obtenues forment un nouveau modèle statistique.

Définition I-1.5 (Modèle induit). *Soient $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ un modèle statistique et T une statistique sur ce modèle, à valeurs dans l'espace mesurable $(\mathsf{T}, \mathcal{T})$. Pour toute loi $\mathbb{P} \in \mathcal{C}$, on appelle* loi image *de $\mathbb{P}$ par T, et l'on note $\mathbb{P}^T$, la loi de probabilité sur $(\mathsf{T}, \mathcal{T})$ définie par*

$$\mathbb{P}^T(A) := \mathbb{P}(T \in A) = \mathbb{P}\left(\{z \in \mathsf{Z} \,:\, T(z) \in A\}\right) \,, \qquad A \in \mathcal{T} \,.$$

On pose $\mathcal{C}^T := \left\{\mathbb{P}^T \,,\, \mathbb{P} \in \mathcal{C}\right\}$. Le modèle statistique $(\mathsf{T}, \mathcal{T}, \mathcal{C}^T)$ est appelé modèle induit *par la statistique T.*

Sous $\mathbb{P}$, l'application T est un élément aléatoire défini sur $(\mathsf{Z}, \mathcal{Z}, \mathbb{P})$, dont $\mathbb{P}^T$ est la loi. Le calcul pratique de la loi image repose sur la méthode de la fonction muette et le changement de variables, rappelés en Section I-1.2 ; c'est l'objet de l'Exercice 2 (PC1).

Exemple I-1.1.12: Modèle induit – Bernoulli

Considérons le modèle $\left(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \{\mathcal{B}\mathrm{er}(\theta)^{\otimes n}\,,\ \theta\in[0,1]\}\right)$ et la statistique $S_n = \sum_{i=1}^n X_i$. Le modèle induit est

$$\left(\{0,\dots,n\},\ \mathcal{P}(\{0,\dots,n\}),\ \{\mathrm{Bin}(n,\theta)\,,\ \theta\in[0,1]\}\right)\ .$$

Exemple I-1.1.13: Modèle induit – loi uniforme

Considérons le modèle $\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{\mathcal{U}([0,\theta])^{\otimes n}\,,\ \theta\in\mathbb{R}_+^*\right\}\right)$: sous $\mathbb{P}_\theta$, chaque observation a pour densité par rapport à Leb la fonction $t\mapsto \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(t)$. Soit la statistique $X_{(n)} = \max_{1\le i\le n} X_i$. Le modèle induit est

$$\left(\mathbb{R},\ \mathcal{B}(\mathbb{R}),\ \left\{q_{n,\theta}\cdot\mathrm{Leb}\,,\ \theta\in\mathbb{R}_+^*\right\}\right) \qquad \text{avec} \qquad q_{n,\theta}(t) = \frac{n\,t^{n-1}}{\theta^n}\,\mathbb{1}_{[0,\theta]}(t)\ .$$

Ce calcul sera repris en PC 4.

Exemple I-1.1.14: Modèle induit – moyenne et variance empiriques du modèle gaussien

Reprenons l'Exemple I-1.1.11 (moyenne et variance empiriques du modèle gaussien) : dans le modèle gaussien à n observations, $n\ge 2$, la statistique $T=(\bar X_n, S_n^2)$ est à valeurs dans $\mathbb{R}\times\mathbb{R}_+$. D'après le théorème de Gosset, sous $\mathbb{P}_{(\mu,\sigma^2)}$, $\bar X_n\sim\mathcal{N}(\mu,\sigma^2/n)$ et $(n-1)S_n^2/\sigma^2\sim\chi^2(n-1)$, ces deux variables étant indépendantes. Le modèle induit par T est donc

$$\left(\mathbb{R}\times\mathbb{R}_+,\ \mathcal{B}(\mathbb{R}\times\mathbb{R}_+),\ \left\{\mathcal{N}(\mu,\sigma^2/n)\otimes\tfrac{\sigma^2}{n-1}\,\chi^2(n-1)\,,\ (\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*\right\}\right)\,,$$

où, par un léger abus de notation, $\frac{\sigma^2}{n-1}\,\chi^2(n-1)$ désigne la loi de $\frac{\sigma^2}{n-1}\,Y$ avec $Y\sim\chi^2(n-1)$, c'est-à-dire la loi image de $\chi^2(n-1)$ par $y\mapsto\sigma^2 y/(n-1)$.

Exemple I-1.1.15: Modèle de sondage (suite)

Considérons tout d'abord le modèle statistique donné par

$$(\mathsf{Z},\mathcal{Z},\mathcal{C}) = \left(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \left\{\mathcal{B}\mathrm{er}(\theta)^{\otimes n}\,,\ \theta\in\Theta:=[0,1]\right\}\right)\,.$$

Notons X_i la statistique correspondant au résultat du i-ème sondage, $i\in\{1,\dots,n\}$: pour tout $\mathbf{x}=(x_1,\dots,x_n)\in\{0,1\}^n$, $X_i(\mathbf{x})=x_i$. En notant $\mathbb{P}_\theta=\mathcal{B}\mathrm{er}(\theta)^{\otimes n}$, la loi image $\mathbb{P}_\theta^{X_i}$ est alors une loi de Bernoulli de paramètre $\theta\in\Theta$. On vérifie aussi que les statistiques $X_1,\dots,X_n$ sont indépendantes : pour tout $(x_1,\dots,x_n)\in\{0,1\}^n$, nous avons puisque $\mathbb{P}_\theta$ est $\mathcal{B}\mathrm{er}(\theta)^{\otimes n}$

$$\mathbb{P}_\theta(\{X_1=x_1,\dots,X_n=x_n\}) = \mathbb{P}_\theta^Z(\{x_1,\dots,x_n\}) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \prod_{i=1}^n \mathbb{P}_\theta(\{X_i=x_i\})\,,$$

où $Z=(X_1,\cdots,X_n)$. Pour le modèle $(\mathsf{Z},\mathcal{Z},\mathcal{C})$, les statistiques $(X_1,\dots,X_n)$ sont indépendantes et identiquement distribuées (i.i.d.) : pour tout $\theta\in\Theta$ et $i\in\{1,\dots,n\}$, l'observation X_i est distribuée suivant une loi de Bernoulli de paramètre θ.

I-1.1.7 Modèle statistique du n-échantillon

L'exemple précédent est générique : étudier des systèmes d'observations *indépendantes et de même loi* correspond à la notion de n-échantillon. La construction générale (produit de modèles statistiques, tribu et mesure produit) est donnée en Section I-1.2 ; nous en donnons ici la définition et la propriété fondamentale.

Définition I-1.6 (n-échantillon). *Soient $(\mathsf{X}, \mathcal{X}, \mathcal{C})$ un modèle statistique et $n \in \mathbb{N}^*$. On appelle* n-échantillon *de $(\mathsf{X}, \mathcal{X}, \mathcal{C})$ le modèle statistique*

$$(\mathsf{X}, \mathcal{X}, \mathcal{C})^n := (\mathsf{X}^n,\ \mathcal{X}^{\otimes n}, \{\mathbb{Q}^{\otimes n}\,,\ \mathbb{Q} \in \mathcal{C}\})\,.$$

On appelle i-ème observation (canonique) la statistique X_i définie pour $z = (x_1, \dots, x_n) \in \mathsf{X}^n$ par $X_i(z) = x_i$.

Intuition. *$\mathbb{Q}^{\otimes n}$ est la mesure produit de $\mathbb{Q}$, n fois : c'est la formalisation de « répéter n fois la même expérience, de façon indépendante ». La propriété clé : $(X_1, \dots, X_n) \sim \mathbb{Q}^{\otimes n}$ si et seulement si $(X_1, \dots, X_n)$ sont i.i.d. de loi $\mathbb{Q}$.*

Lemme I-1.1.1

Soit $(\mathsf{X}, \mathcal{X}, \mathcal{C})^n$ un n-échantillon du modèle $(\mathsf{X}, \mathcal{X}, \mathcal{C})$.

(i) Les observations $Z = (X_1, \dots, X_n)$ sont indépendantes.

(ii) Pour tout $i \in \{1, \dots, n\}$, le modèle induit par la statistique X_i est $(\mathsf{X}, \mathcal{X}, \mathcal{C})$.

Démonstration

En effet, pour tout $\mathbb{P} = \mathbb{Q}^{\otimes n} \in \mathcal{C}^{\otimes n}$ et $(A_1, \dots, A_n) \in \mathcal{X}^n$, nous avons

$$\begin{aligned}\mathbb{P}^Z(A_1 \times \cdots \times A_n) &= \mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) \\ &= \mathbb{Q}^{\otimes n}(A_1 \times \cdots \times A_n) = \prod_{i=1}^n \mathbb{Q}(A_i) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i)\,.\end{aligned}$$

□

On adoptera la terminologie : « $(X_1, \dots, X_n)$ *est un n-échantillon du modèle* $(\mathsf{X}, \mathcal{X}, \mathcal{C})$ ».

Soit $(\mathsf{X}, \mathcal{X}, \{\mathbb{Q}_\theta\,,\ \theta \in \Theta\})$ un modèle paramétrique. Supposons qu'il existe une mesure σ-finie μ sur $(\mathsf{X}, \mathcal{X})$ telle que la famille paramétrique $\{\mathbb{Q}_\theta\,,\ \theta \in \Theta\}$ soit dominée par rapport à μ. Notons, pour tout $\theta \in \Theta$, $q_\theta(\cdot)$ la densité de la loi $\mathbb{Q}_\theta$ par rapport à la mesure de domination μ :

$$\mathbb{Q}_\theta = q_\theta(\cdot) \cdot \mu\,.$$

Pour tout $\theta \in \Theta$, la loi produit $\mathbb{P}_\theta = \mathbb{Q}_\theta^{\otimes n}$ est dominée par $\mu^{\otimes n}$ et admet pour densité

$$(x_1, \dots, x_n) \mapsto p_\theta(x_1, \dots, x_n) := \prod_{i=1}^n q_\theta(x_i)\,. \tag{I-1.1}$$

Les modèles de translation et d'échelle, archétypes de n-échantillons dominés, sont étudiés dans l'Exercice 3 (PC1).

I-1.1.8 Exemples et premiers estimateurs

L'exploration systématique des méthodes de construction d'estimateurs (méthode des moments, maximum de vraisemblance) est l'objet du chapitre suivant. On peut néanmoins, dès maintenant, « proposer » des estimateurs : un *estimateur* est une statistique à valeurs dans Θ (ou dans un espace contenant Θ), destinée à approcher le paramètre inconnu.

Estimation du paramètre d'une Bernoulli. Revenons à la première question de l'introduction (*la pièce est-elle biaisée ?*). On considère un n-échantillon $(X_1, \dots, X_n)$ du modèle

$$\big(\{0,1\},\ \mathcal{P}(\{0,1\}),\ \{\mathcal{B}\mathrm{er}(\theta)\,,\ \theta \in [0,1]\}\big)\,.$$

On souhaite estimer θ. Sous $\mathbb{P}_\theta$, on a $\mathbb{E}_\theta[X_1] = \theta$; on considère donc

$$\hat{\theta}_n := \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \, .$$

On a $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$: l'estimateur est *sans biais*. La qualité de cette approximation (concentration, intervalles de confiance) sera quantifiée dans les chapitres suivants.

Question sensible et réponses randomisées. Considérons la troisième question de l'introduction, en version sensible : *quelle proportion d'élèves a utilisé un modèle d'IA alors que c'était interdit ?* Si l'on pose directement la question, les réponses seront biaisées : on ne peut pas espérer observer $(X_1, \dots, X_n)$. Le protocole des *réponses randomisées* contourne cette difficulté :

1. on pose la question sensible ;
2. chaque participant observe secrètement une variable aléatoire $U_i \sim \mathcal{B}\mathrm{er}(q)$, avec q *connu* (par exemple $q = 0{,}75$), indépendante de X_i ;
3. si $U_i = 1$, il donne la vraie réponse ; si $U_i = 0$, il répond l'opposé.

On observe donc les statistiques $\mathrm{RR}_i = U_i X_i + (1 - U_i)(1 - X_i)$, $i = 1, \dots, n$. Le modèle induit par $(\mathrm{RR}_1, \dots, \mathrm{RR}_n)$ est un n-échantillon du modèle

$$\big(\{0,1\},\ \mathcal{P}(\{0,1\}),\ \{\mathcal{B}\mathrm{er}(\,\theta q + (1-\theta)(1-q)\,),\ \theta \in [0,1]\}\,\big) \, .$$

Intuition. *Chaque répondant est protégé : une réponse « oui » ne révèle pas sa vraie réponse, puisqu'elle peut provenir du retournement aléatoire. Mais* en moyenne*, le signal survit : la proportion de « oui » reste une fonction* connue et inversible *de θ (dès que $q \neq 1/2$), ce qui permet de remonter à θ.*

Sous le modèle induit, $\mathbb{E}_\theta[\mathrm{RR}_i] = \theta q + (1-\theta)(1-q) = \theta(2q-1) + 1 - q$, donc

$$\theta = \frac{\mathbb{E}_\theta[\mathrm{RR}_i] - (1-q)}{2q-1} \, .$$

On considère donc l'estimateur (dit *par la méthode des moments*)

$$\hat{\theta}_n := \frac{\overline{\mathrm{RR}}_n - (1-q)}{2q-1} \, ,$$

que l'on peut de plus contraindre (*clipper*) à l'intervalle $[0,1]$. Pour $q = 0{,}75$: $\hat{\theta}_n = 2(\overline{\mathrm{RR}}_n - 0{,}25)$. On vérifie que $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$ et

$$\mathrm{Var}_\theta(\hat{\theta}_n) = \frac{\mathrm{Var}_\theta(\mathrm{RR}_1)}{n\,(2q-1)^2} \geq \frac{q(1-q)}{n\,(2q-1)^2} \, .$$

Le prix de la confidentialité est une variance accrue (le facteur $(2q-1)^{-2}$ explose quand $q \to 1/2$).

I-1.1.9 En pratique : programme des exercices

Les notions de ce chapitre sont mises en œuvre dans la feuille d'exercices (PC 1), intégrée au chapitre (voir le fascicule d'exercices en fin de chapitre). Exercices traités en petite classe :

Exercice	Notions du cours	Rappels de probabilités
Exo 1 – *Un café pour commencer*	Modélisation, estimation	Lois exponentielle et de Poisson
Exo 2 – *Transformation de v.a.*	Modèle induit	Fonction muette, chgt de variable
Exo 3 – *Translation et échelle*	Modèle induit, th. de Gosset	Manipulation de lois, gaussiennes
Exo 4 – *Régression gaussienne*	Th. de Cochran	Manipulation de lois, gaussiennes

L'exercice 4 sera poursuivi en PC 2, le modèle de régression linéaire gaussienne étant un sujet récurrent du cours. Les exercices bonus (score de football, prix d'un actif financier, identifiabilité des mélanges) sont accompagnés d'un guide de lecture dans le fascicule.

Notions de probabilités clés utilisées cette semaine, à revoir : variable aléatoire, mesure produit, indépendance, densité, théorème de transfert et fonction muette (voir la partie *Outils mathématiques* ci-dessous).