Chapitre 1

Modélisation statistique

1.1 Exemple du jeu de pile ou face

Une pièce a une probabilité inconnue $\theta_0 \in]0,1[$ de tomber sur pile. Sur les 1000 lancers réalisés indépendamment les uns des autres, 520 piles ont été obtenus. Intuitivement, θ_0 est donc proche de 0.52. Mais comment justifier cette approximation ? Par ailleurs, de la même manière qu'il est sans intérêt de donner une valeur approchée d'une intégrale sans préciser l'erreur d'approximation, ce résultat n'a que peu de valeur car il ne nous renseigne pas sur l'erreur commise. Dans cette section, nous allons développer une démarche, qui porte en elle l'architecture du raisonnement en statistique inférentielle, dans le but de formaliser cette intuition et préciser la qualité de l'approximation.

Notons $x_1, \cdots, x_n$ les résultats des $n = 1000$ lancers de pièce, en adoptant la convention $x_i = 1$ si le i-ème lancer a donné pile, 0 dans le cas contraire. L'*observation* $(x_1, \cdots, x_n)$ est une réalisation d'un n-uple de variables aléatoires $(X_1, \cdots, X_n)$ définies sur l'espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$, i.e. pour un $\omega \in \Omega$:

$$\big(X_1(\omega), \cdots, X_n(\omega)\big) = (x_1, \cdots, x_n).$$

Les conditions de l'expérience apportent en outre des informations sur ce n-uple : les variables aléatoires $X_1, \cdots, X_n$ sont indépendantes, car les lancers de pièce le sont, et de même loi de Bernoulli de paramètre θ_0. Cependant, une telle modélisation probabiliste est basée sur la connaissance de θ_0. La valeur de ce paramètre étant inconnue, la seule affirmation raisonable est que

$(x_1,\cdots,x_n)$ est une réalisation d'un n-uple de variables aléatoires indépendantes $(X_1,\cdots,X_n)$ de même loi de Bernoulli $\mathcal{B}(\theta)$, pour un certain paramètre $\theta\in]0,1[$. La loi de ce n-uple est la loi produit sur $\{0,1\}^n$ notée $\mathcal{B}(\theta)^{\otimes n}$, i.e. la probabilité sur $\{0,1\}^n$ définie pour tout $(y_1,\cdots,y_n)\in\{0,1\}^n$ par

$$\begin{aligned}\mathcal{B}(\theta)^{\otimes n}(\{y_1,\cdots,y_n\}) &= \prod_{i=1}^{n}\mathcal{B}(\theta)(\{y_i\})\\ &= \prod_{i=1}^{n}\big(\theta\mathbb{1}_{\{1\}}(y_i)+(1-\theta)\mathbb{1}_{\{0\}}(y_i)\big).\end{aligned}$$

De manière équivalente, $(x_1,\cdots,x_n)$ est une réalisation de l'une des lois de probabilité de l'ensemble $\{\mathcal{B}(\theta)^{\otimes n}\}_{\theta\in]0,1[}$. A partir de cet ensemble de probabilités, appelé *modèle statistique*, l'enjeu est de déduire une valeur du paramètre du modèle qui s'ajuste à l'observation $(x_1,\cdots,x_n)$.

Reprenons le n-uple $(X_1,\cdots,X_n)$ de variables aléatoires indépendantes de même loi $\mathcal{B}(\theta)$. D'après la loi des grands nombres, la variable aléatoire

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

appelée *estimateur* est, avec une probabilité élevée, proche de θ lorsque n est grand. Cette propriété est vraie pour chaque valeur de $\theta\in]0,1[$, donc en particulier pour la valeur inconnue θ_0. Pour la réalisation $\bar{x}_n$ de $\bar{X}_n$, définie par

$$\bar{x}_n=\frac{1}{n}\sum_{i=1}^{n}x_i,$$

$\bar{x}_n\approx\theta_0$ d'où $\theta_0\approx 0.52$ avec un *niveau de confiance* élevé. Nous allons nous appuyer sur cette modélisation pour préciser l'erreur commise, ceci avec deux critères : *risque quadratique* et *intervalle de confiance*.

Le risque quadratique de l'estimateur est le carré de la distance $\mathbb{L}^2(\mathbb{P})$ entre la cible θ et $\bar{X}_n$. Comme $X_1,\cdots,X_n$ sont indépendantes et de même loi $\mathcal{B}(\theta)$,

$$\mathbb{E}\,(\bar{X}_n-\theta)^2=\mathbb{V}(\bar{X}_n)=\frac{1}{n}\mathbb{V}(X_1)=\frac{1}{n}\theta(1-\theta),$$

si $\mathbb{E}$ et $\mathbb{V}$ désignent respectivement l'espérance et la variance pour la probabilité $\mathbb{P}$. Or, $\theta(1-\theta)\leq 1/4$, donc l'erreur quadratique moyenne commise est

majorée par $1/(2\sqrt{n}) \approx 0.02$. Appliqué à l'observation $(x_1, \cdots, x_n)$ issue de la loi $\mathcal{B}(\theta_0)^{\otimes n}$, ce résultat nous donne une information sur la qualité de l'approximation de θ_0 par $\bar{x}_n$.

Un intervalle de confiance par excès de niveau 95% construit avec les variables aléatoires $X_1, \cdots, X_n$ est un intervalle $I(X_1, \cdots, X_n)$ tel que

$$\mathbb{P}\big(\theta \in I(X_1, \cdots, X_n)\big) \geq 0.95.$$

D'après l'inégalité de Bienaymé-Tchebytchev, pour tout $\varepsilon > 0$:

$$\mathbb{P}\big(|\bar{X}_n - \theta| \geq \varepsilon\big) \leq \frac{\mathbb{V}(\bar{X}_n)}{\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

De ce fait, $\mathbb{P}(|\bar{X}_n - \theta| \geq \varepsilon) \leq 0.05$ au moins si $1/(4n\varepsilon^2) \leq 0.05$, et cette contrainte est vérifiée pour $\varepsilon = 0.08$. Par suite,

$$\mathbb{P}\big(\theta \in [\bar{X}_n - 0.08, \bar{X}_n + 0.08]\big) \geq 0.95.$$

En particulier, pour l'observation $(x_1, \cdots, x_n)$ issue de la loi $\mathcal{B}(\theta_0)^{\otimes n}$, $\bar{x}_n = 0.52$ d'où $\theta_0 \in [0.44, 0.60]$ avec un niveau de confiance au moins égal à 95%.

Par le biais de cet exemple simple, nous venons d'exposer quelques uns des concepts de la statistique. Cependant, cette brève étude laisse des zones d'ombre. Par exemple, l'utilisation de la moyenne empirique $\bar{X}_n$ est-elle vraiment la plus judicieuse ? Cet estimateur est-il le meilleur au sens du risque quadratique ? Eu égard au nombre élevé d'observations, la longueur de l'intervalle de confiance est grande : peut-on donner un intervalle de meilleure qualité ? Par ailleurs, le problème peut être envisagé, non plus sous l'angle numérique en cherchant une approximation de θ_0, mais plutôt sous l'angle d'une question posée, par exemple « La pièce est-elle équilibrée ? ». Ni l'approche par calcul de l'erreur quadratique, ni l'intervalle de confiance, qui contient la valeur 0.5, ne permettent de répondre à cette question ; dès lors, quelle stratégie de décision envisager ? L'objectif du livre est de construire des outils permettant de donner quelques éléments de réponse à ces questions.

1.2 Principe fondamental de la statistique

La démarche statistique vise à reconstruire, en se basant seulement sur un n-uple d'observations, la loi de probabilité dont il est issu. Dans l'exemple

précédent, la méthode basée sur la moyenne $\bar{x}_n$ de l'observation $(x_1, \cdots, x_n)$ est très spécifique à ce cas d'école, et ne peut en tout état de cause être exportée à des situations plus complexes. En règle générale, cette démarche de la statistique inférentielle est-elle fondée ? La réponse à cette question fondamentale fait l'objet de cette section.

L'objectif est de reconstruire une probabilité Q sur $(\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k))$, inconnue et appelée *mesure théorique*, à partir d'une suite de réalisations de variables aléatoires $X_1, \cdots, X_n$ indépendantes et identiquement distribuées de loi commune Q, définies sur l'espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$. La *mesure empirique*

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

constitue un candidat prometteur d'après la loi des grands nombres. En effet, pour tout élément $A \in \mathscr{B}(\mathbb{R}^k)$, on a avec probabilité 1 :

$$\lim_{n\to\infty} Q_n(A) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i) = Q(A).$$

Mais il ne s'agit pas d'un résultat de convergence de Q_n vers Q, car l'ensemble de probabilité 1 sur lequel cette convergence a lieu dépend de A.

En quel sens Q_n peut-elle tendre vers Q ? Il est tentant de s'orienter vers la *distance en variation totale* entre Q_n et Q. Cette distance est définie, pour deux probabilités ν_1 et ν_2 sur $(\mathbb{R}^k, \mathscr{B}(\mathbb{R}^k))$, par

$$d_V(\nu_1, \nu_2) = \sup_{A \in \mathscr{B}(\mathbb{R}^k)} |\nu_1(A) - \nu_2(A)|.$$

Cependant, si la probabilité Q est sans atomes i.e. $Q(\{x\}) = 0$ pour tout $x \in \mathbb{R}^k$, alors

$$d_V(Q_n, Q) \geq |Q_n(\{X_1, \cdots, X_n\}) - Q(\{X_1, \cdots, X_n\})| = 1.$$

La topologie induite par cette distance est donc en général trop forte.

En revanche, la *topologie de la convergence étroite* [1] offre un cadre d'étude adapté à notre problème.

1. Consulter la section consacrée à la théorie de la mesure dans le chapitre de compléments.

Théorème 1.2.1. [VARADARAJAN] *Avec probabilité 1, la suite de probabilités $(Q_n)_{n\geq 1}$ converge étroitement vers Q.*

Preuve [2]**.** Soient $\mathscr{D}$ un ensemble dénombrable dense de $\mathbb{R}^k$ et $\mathscr{B}$ l'ensemble dénombrable constitué des boules ouvertes de $\mathbb{R}^k$ de rayon rationnel et dont le centre est dans $\mathscr{D}$. La famille $\mathscr{V}$ des intersections finies d'éléments de $\mathscr{B}$ étant dénombrable, la loi forte des grands nombres montre que, avec probabilité 1 :

$$\lim_{n\to\infty} Q_n(A) = Q(A)\ \forall A \in \mathscr{V}.$$

Soit alors $O \subset \mathbb{R}^k$ un ouvert non vide. Il existe une suite $(B_i)_{i\geq 1}$ d'éléments de $\mathscr{B}$ telle que $O = \cup_{i\geq 1} B_i$. D'après le théorème de convergence monotone, pour tout $\varepsilon > 0$, il existe un rang $L \geq 1$ tel que si $\ell \geq L$,

$$Q\Big(\bigcup_{i=1}^{\ell} B_i\Big) \geq Q(O) - \varepsilon.$$

Par ailleurs, l'égalité de Poincaré nous donne

$$Q_n\Big(\bigcup_{i=1}^{\ell} B_i\Big) = \sum_{k=1}^{\ell} (-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq \ell} Q_n\big(B_{i_1}\bigcap\cdots\bigcap B_{i_k}\big).$$

Comme les intersections $B_{i_1}\cap\cdots\cap B_{i_k}$ sont dans $\mathscr{V}$:

$$\lim_{n\to\infty} Q_n\Big(\bigcup_{i=1}^{\ell} B_i\Big) = Q\Big(\bigcup_{i=1}^{\ell} B_i\Big),$$

sur un événement de probabilité 1 indépendant de l'ouvert O. En conséquence, pour tout $\ell \geq L$ et $\mathbb{P}$-p.s. :

$$Q(O) - \varepsilon \leq Q\Big(\bigcup_{i=1}^{\ell} B_i\Big) = \lim_{n\to\infty} Q_n\Big(\bigcup_{i=1}^{\ell} B_i\Big) \leq \liminf_{n\to\infty} Q_n(O),$$

car $\cup_{i=1}^{\ell} B_i \subset O$, l'événement de probabilité 1 sur lequel cette inégalité a lieu ne dépendant pas de l'ouvert O. En faisant ensuite tendre ε vers 0 via les rationnels, on en déduit que, avec probabilité 1, pour tout ouvert O de $\mathbb{R}^k$:

$$Q(O) \leq \liminf_{n\to\infty} Q_n(O).$$

2. Cette preuve peut être omise en première lecture.

Il reste à appliquer le théorème de Portmanteau, d'où le résultat annoncé. □

Pour illustrer la portée de ce théorème, considérons le cas d'une expérience répétée n fois ; le résultat de chaque expérience étant un élément de $\mathbb{R}^k$, l'observation est un n-uple $(x_1,\cdots,x_n) \in (\mathbb{R}^k)^n$. Si les expériences sont menées de manière indépendantes, $(x_1,\cdots,x_n)$ est une réalisation d'un n-uple de variables aléatoires $(X_1,\cdots,X_n)$ indépendantes et de même loi inconnue Q sur $\mathbb{R}^k$. Le théorème de Varadarajan montre que, avec une probabilité élevée, la mesure empirique

$$Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

est proche de la mesure théorique Q, lorsque n est assez grand. Ainsi, en multipliant les expériences, la mesure discrète

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

est proche de la mesure Q avec un niveau de confiance élevé, c'est-à-dire que l'on peut reconstruire la loi Q avec l'observation $(x_1,\cdots,x_n)$. De ce fait, l'approche de base de la statistique inférentielle est fondée, et le théorème de Varadarajan mérite son titre de *principe fondamental de la statistique*.

1.3 Modèle statistique

Le théorème de Varadarajan ne donne pas d'information sur la proximité entre la mesure empirique et la mesure théorique, mais la finalité pratique de la statistique impose de s'y intéresser tout particulièrement. Cependant, une telle quête est hors de portée sans informations complémentaires sur la mesure théorique. Une alternative est alors d'exploiter l'information issue de l'expérience afin de préciser la forme de la mesure théorique, ce qui mène au concept de *modèle statistique*, véritable pilier de toute la démarche statistique. Dans la suite, l'espace des observations est $\mathscr{H}^n = \mathscr{H} \times \cdots \times \mathscr{H}$ (n fois), avec $\mathscr{H} \subset \mathbb{R}^k$.

Définition. *Un modèle statistique est un couple $(\mathscr{H}^n, \mathscr{P})$, où $\mathscr{P}$ est une famille de probabilités sur $(\mathscr{H}^n, \mathscr{B}(\mathscr{H}^n))$.*

Remarque. Un cas important, que nous rencontrerons fréquemment, est celui où l'observation $(x_1, \cdots, x_n) \in \mathscr{H}^n$ est issue de n répétitions indépendantes d'une même expérience. L'observation est alors une réalisation d'une loi produit $Q_0^{\otimes n}$ [3] avec Q_0 une probabilité inconnue sur $(\mathscr{H}, \mathscr{B}(\mathscr{H}))$. Le modèle statistique s'écrit $(\mathscr{H}^n, \{Q^{\otimes n}\}_{Q \in \mathscr{Q}})$, où $\mathscr{Q}$ est un ensemble de probabilités sur $\mathscr{H}$ qui contient Q_0.

Exemple. Dans l'étude statistique du jeu de pile ou face de la section 1.1, x_i vaut 1 si le i-ème lancer a donné pile, et 0 dans le cas contraire. L'espace des observations est $\{0,1\}^n$, et le n-uple $(x_1, \cdots, x_n)$ est une réalisation de l'une des lois de l'ensemble $\{\mathcal{B}(\theta)^{\otimes n}\}_{\theta \in]0,1[}$: le modèle statistique décrivant cette expérience est donc le *modèle de Bernoulli* $(\{0,1\}^n, \{\mathcal{B}(\theta)^{\otimes n}\}_{\theta \in]0,1[})$.

Un même modèle statistique peut décrire plusieurs expériences. Illustrons cette affirmation avec le modèle de Bernoulli de l'exemple précédent.

Exemple. Un lac contient une proportion inconnue de poissons rouges. Muni d'une cane à pêche, on effectue n expériences indépendantes de tirages de poissons dans ce lac.

- Considérons le cas où chacune de ces expériences consiste à tirer avec remise un poisson dans le lac. L'observation est un n-uple $(x_1, \cdots, x_n)$, avec la convention $x_i = 1$ (resp. 0) si un poisson rouge (resp. un poisson d'une autre variété) a été pêché au i-ème tirage, donc l'espace des observations est $\{0,1\}^n$. De plus, si $\theta \in]0,1[$ est la proportion de poissons rouges, la probabilité d'obtenir la suite $(x_1, \cdots, x_n) \in \{0,1\}^n$ est $\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$ car les n tirages sont indépendants. Il s'agit de la loi $\mathcal{B}(\theta)^{\otimes n}$, c'est-à-dire que le modèle statistique décrivant cette expérience est le modèle de Bernoulli $(\{0,1\}^n, \{\mathcal{B}(\theta)^{\otimes n}\}_{\theta \in]0,1[})$.
- Modifions les modalités, en considérant que chacune des n expériences consiste à tirer avec remise des poissons jusqu'à l'obtention d'un poisson rouge. L'observation est un n-uple $(x_1, \cdots, x_n)$, x_i représentant le nombre de tirages effectués à l'expérience i, donc l'espace des obser-

3. Pour deux probabilités Q et Q' sur $\mathscr{H}$, il existe une unique probabilité $Q \otimes Q'$ sur $(\mathscr{H}^2, \mathscr{B}(\mathscr{H}^2))$ telle que $Q \otimes Q'(A \times A') = Q(A)Q'(A')$ pour tous $A, A' \in \mathscr{B}(\mathscr{H})$. Elle s'appelle probabilité produit, et elle représente la loi du couple de variables aléatoires (X, X'), avec X et X' indépendantes et de lois respectives Q et Q'. Par extension, on définit de même le produit de n probabilités et en particulier, $Q^{\otimes n} = Q \otimes \cdots \otimes Q$ (n fois).

vations est $(\mathbb{N}^\star)^n$. De plus, si la proportion de poissons rouges du lac est $\theta \in]0,1[$, la probabilité d'obtenir la suite $(x_1,\cdots,x_n) \in (\mathbb{N}^\star)^n$ est $\prod_{i=1}^n (1-\theta)^{x_i-1}\theta$ car les n expériences sont indépendantes. Il s'agit du produit n fois de la loi géométrique $\mathcal{G}(\theta)$, donc le modèle statistique décrivant cette expérience est $((\mathbb{N}^\star)^n, \{\mathcal{G}(\theta)^{\otimes n}\}_{\theta\in]0,1[})$.

Plus on dispose d'informations sur l'expérience, plus le modèle statistique peut être décrit de manière précise. La terminologie ci-dessous a pour objectif de classer les modèles selon ce critère.

Définitions.

(i) *Le modèle statistique $(\mathscr{H}^n, \mathscr{P})$ est paramétré par l'ensemble Θ si $\mathscr{P} = \{P_\theta\}_{\theta\in\Theta}$.*

(ii) *Le modèle statistique paramétré $(\mathscr{H}^n, \{P_\theta\}_{\theta\in\Theta})$ est dit paramétrique si Θ est inclus dans un espace vectoriel de dimension finie. Sinon, il est non paramétrique.*

Noter que, dans un soucis de implification des notations, la dépendance de P_θ en n n'est pas mentionnée. L'essentiel de cet ouvrage porte sur l'étude des modèles paramétriques ; ce cadre d'étude est plus restrictif mais, l'espace des paramètres étant moins vaste, les résultats d'approximation sont de meilleure qualité.

En exploitant la nature de l'expérience, on peut imposer des hypothèses mathématiques afin de restreindre l'ensemble des lois de probabilités du modèle statistique. C'est la démarche adoptée dans l'exemple qui suit.

Exemple. L'objectif est de déterminer le modèle statistique associé à l'observation de n durées de vies indépendantes d'ampoules du même type. La variable aléatoire X sur l'espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$ qui représente la durée de vie de ce type d'ampoule vérifie $\mathbb{P}(X>0)>0$ et $\mathbb{P}(X>1)<1$, quitte à changer l'unité de temps. Pour simplifier, considérons qu'une ampoule *ne se souvient pas d'avoir vieilli* ; ceci se traduit par le fait que la loi de X est sans mémoire, i.e. pour tout $s,t \in \mathbb{R}_+$:

$$\mathbb{P}(X > s+t | X > s) = \mathbb{P}(X > t).$$