Olivier Marchal

Cours et exercices corrigés de statistiques inférentielles

2e édition

ellipses

Chapitre I. Généralités

Introduction

Ce chapitre a pour but de présenter les problématiques générales liées à l'inférence en statistique. Il permet ainsi de comprendre les questions qui sont abordées et traitées dans les différents chapitres à venir. En particulier, il précise la notion d'inférence qui peut être difficile à appréhender pour les étudiants ayant reçu une formation mathématique qui sont plus naturellement habitués à des raisonnements hypothético-déductifs. Néanmoins, une fois le cadre et les hypothèses générales posées, nous verrons que les constructions et résultats développés pour résoudre les différents problèmes respectent la logique, la rigueur ainsi que les raisonnements habituels en mathématiques et fournissent donc des théorèmes et propositions au sens mathématique du terme.

1 Les grands problèmes en statistiques inférentielles

1.1 Modélisation paramétrique

Comme son nom l'indique, le but des statistiques inférentielles consiste à inférer de façon optimale certaines quantités à l'aide de données numériques ou non obtenues dans le cadre d'expériences scientifiques ou de simulations numériques. Comme tout domaine des mathématiques, elle fait appel à une terminologie précise que nous allons introduire :

Définition I.1 (Modèle statistique)**.** *Soit* $(\Omega, \mathcal{F}, \mathbb{P})$ *un espace de probabilité. On appelle* ***modèle statistique*** *ou* ***modèle statistique paramétrique*** *la donnée d'une famille de lois de probabilité* $\mathcal{P} = (P_{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \mathcal{I}}$ *définies*

sur cet espace et indicées par un paramètre $\boldsymbol{\theta}$ appartenant à un ensemble $\mathcal{I}$ non vide. Notons que $\mathcal{I}$ peut être un ensemble fini, dénombrable ou non dénombrable.[1]

En pratique, les cas les plus communs et qui seront développés dans ce document sont les suivants :

- $\mathcal{I}$ est un ensemble fini qui par bijection peut être identifié à $[\![1, M]\!]$.
- $\mathcal{I}$ est un ensemble dénombrable mais non fini. Dans ce cas, par bijection, il peut être identifié à $\mathbb{N}$.
- $\mathcal{I}$ est une réunion finie d'intervalles de $\mathbb{R}$.
- $\mathcal{I}$ est une réunion finie de pavés de $\mathbb{R}^d$ avec $d \geq 2$. Dans ce cas, $\boldsymbol{\theta}$ est un paramètre vectoriel de dimension d : $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$.

En parallèle du modèle théorique, les statistiques inférentielles font appel à des données $(x_i)_{i\leq n}$, c'est-à-dire à n valeurs numériques ou vectorielles recueillies lors d'une expérience ou à l'issue de simulations numériques. L'hypothèse fondamentale des statistiques inférentielles est alors de supposer que ces données peuvent être correctement modélisées par le modèle statistique.

Définition I.2. *Un jeu de données $(x_i)_{i\leq n}$ est dit compatible avec un modèle statistique paramétrique $\mathcal{P} = (P_\theta)_{\theta\in\mathcal{I}}$ s'il existe une valeur $\boldsymbol{\theta}_0 \in \mathcal{I}$ telle que les données $(x_i)_{i\leq n}$ puissent être considérées comme la réalisation d'un tirage aléatoire de n variables aléatoires $(X_i)_{i\leq n}$ définies sur l'espace de probabilité associé au modèle et indépendantes et identiquement distribuées (i.i.d.) suivant la loi de probabilité $P_{\boldsymbol{\theta}_0}$. En d'autres termes, les données $(x_i)_{i\leq n}$ peuvent être vues comme un* ***échantillon*** *de taille n de la loi $P_{\boldsymbol{\theta}_0}$.*

Notons que pour un jeu de données fixé $(x_i)_{i\leq n}$, il n'est pas toujours évident de choisir un modèle statistique approprié et que la théorie qui sera développée ne permet pas de vérifier si le modèle inclus bien la loi de probabilité correspondant réellement aux données. Ce point constitue la faiblesse essentielle des statistiques inférentielles et plus généralement du domaine de

[1] La notation $\boldsymbol{\theta}$ du paramètre indiçant la famille de lois est la notation standard en statistique paramétrique. Elle sera utilisée dans cet ouvrage de façon systématique pour développer la théorie en toute généralité.

la statistique paramétrique par rapport aux statistiques non paramétriques. En effet, **supposer que les données sont correctement modélisées par des variables aléatoires appartenant à une famille de lois de probabilité fixée est une hypothèse majeure qui constitue le postulat de départ de toutes les analyses qui seront effectuées** dans cet ouvrage. Néanmoins, une fois cette hypothèse supposée, la théorie permet alors de démontrer la valeur « optimale » (dans un sens qui sera précisé par la suite) que le paramètre $\boldsymbol{\theta}$ du modèle doit prendre ainsi que d'en préciser ses propriétés importantes.

1.2 Les différentes problématiques à résoudre

Une fois le modèle statistique sélectionné (supposé), la théorie mathématique qui va être développée permettra de résoudre les questions naturelles suivantes :

1. Comment peut-on sélectionner le paramètre $\boldsymbol{\theta}_0$ dans l'ensemble $\mathcal{I}$ qui correspond « au mieux » aux données $(x_i)_{i\leq n}$ recueillies ? Cette question est au centre de la **théorie des estimateurs** qui permet de définir tout d'abord les bons critères mathématiques pour sélectionner le paramètre de façon optimale ainsi que la **théorie du maximum de vraisemblance** qui offre une construction pour le calculer dans un cadre très général.

2. Une fois la détermination optimale du paramètre du modèle réalisée, on applique alors la formule aux données dont on dispose pour calculer de façon effective le paramètre du modèle associé à ce jeu de données. Néanmoins, compte tenu du caractère fini de la taille de l'échantillon, il est évident que l'ajout ou la suppression d'une des données va modifier l'estimation effective du paramètre du modèle. Il est donc nécessaire de pouvoir fournir de façon théorique des bornes contrôlées de la véritable valeur du paramètre du modèle (que l'on obtiendrait si on disposait d'un nombre infini de valeurs ce qui est impossible en pratique). Cette question donne lieu à la démonstration de formules **d'intervalles de confiance et de prédiction autour de la valeur estimée du paramètre du modèle**. Dans certaines applications pratiques, il peut également être intéressant de décider si la valeur véritable du paramètre est supérieure ou inférieure à une valeur donnée, ou appartient à un intervalle donné, etc. Ce type de

questions donne lieu à l'élaboration de **tests paramétriques d'hypothèses** qui sont très utilisés en sciences expérimentales.

3. Parmi les outils statistiques les plus utilisés, la **régression linéaire** fait partie des plus connus. En effet, il s'agit d'un outil enseigné très tôt dans le cadre scolaire pour les besoins des sciences expérimentales. Nous verrons donc en quoi la régression linéaire, en plus de son aspect graphique intuitif, peut être considérée comme un cas particulier de tests paramétriques et nous préciserons à quel modèle ce test est associé.

2 Quelques exemples et modèles standards

Afin d'illustrer les problèmes typiques et la notion de modèle statistique, voici une liste non exhaustive de problèmes statistiques qui pourront être résolus dans cet ouvrage.

2.1 Lancers d'une pièce de monnaie

On se propose de déterminer si une pièce de monnaie est équilibrée (i.e. si la probabilité de chacune des deux faces est égale à $\frac{1}{2}$). Afin d'étudier ce problème, on lance successivement la pièce n fois et on note les résultats obtenus : $(x_i)_{i\leq n} \in \{\text{Pile}, \text{Face}\}^n$. Un exemple de résultats possibles pour $n = 5$ lancers est ainsi :

$$x_1 = \text{Pile}\ ,\ x_2 = \text{Pile}\ ,\ x_3 = \text{Face}\ ,\ x_4 = \text{Pile}\ ,\ x_5 = \text{Face}$$

Le modèle statistique correspondant est un modèle de Bernoulli. En effet, si l'on note p la probabilité de faire Pile avec la pièce alors les données $(x_i)_{i\leq n}$ peuvent être modélisées comme la réalisation de n variables aléatoires $(X_i)_{i\leq n}$ indépendantes et identiquement distribuées suivant une loi de Bernoulli de paramètre p. Le paramètre du modèle est ainsi le paramètre $p \in \mathcal{I} = (0, 1)$.[2] Il est réel et continu (dans le sens que ses valeurs potentielles sont un intervalle de $\mathbb{R}$ non réduit à un point). Il est alors naturel de se demander comment estimer de façon optimale cette valeur inconnue du paramètre p du modèle à partir des données recueillies. Par ailleurs,

[2]Dans cet ouvrage, les intervalles seront notés avec les conventions internationales, c'est-à-dire $[a, b) = \{x \in \mathbb{R} \text{ tel que } a \leq x < b\}$.

comme le nombre de lancers est fini, l'information sur ce paramètre ne sera jamais totale (i.e. on ne pourra jamais déterminer avec certitude à partir d'un échantillon fini la vraie valeur du paramètre p) et il sera donc nécessaire d'affiner l'estimation en fournissant des bornes autour de l'estimation obtenue. Enfin, le problème de départ : « Savoir si la pièce est équilibrée ou non » peut se traduire en termes mathématiques sous la forme suivante : $p \stackrel{?}{=} \frac{1}{2}$ qui sera typique d'un test paramétrique d'hypothèses.

2.2 Lancers d'un dé

Le problème précédent peut être généralisé au cas de lancers d'un dé. Dans ce cas, on réalise n lancers indépendants d'un dé et on s'interroge sur son caractère équilibré (i.e. si la probabilité de chaque face est égale à $\frac{1}{6}$). Les résultats obtenus sont notés $(x_i)_{i\leq n} \in [\![1,6]\!]^n$. L'ordre des lancers n'ayant aucune importance (puisque les lancers sont supposés indépendants), on peut résumer les résultats sous la forme d'un tableau :

Numéro de la face	1	2	3	4	5	6
Occurence	12	10	13	11	10	14
Probabilités empiriques	$\frac{12}{70}$	$\frac{10}{70}$	$\frac{13}{70}$	$\frac{11}{70}$	$\frac{10}{70}$	$\frac{14}{70}$

Le modèle sous-jacent est décrit par le fait que les données $(x_i)_{i\leq n}$ peuvent être vues comme la réalisation de variables aléatoires $(X_i)_{i\leq n}$ indépendantes et identiquement distribuées sur l'espace $\Omega = [\![1,6]\!]$ avec les probabilités :

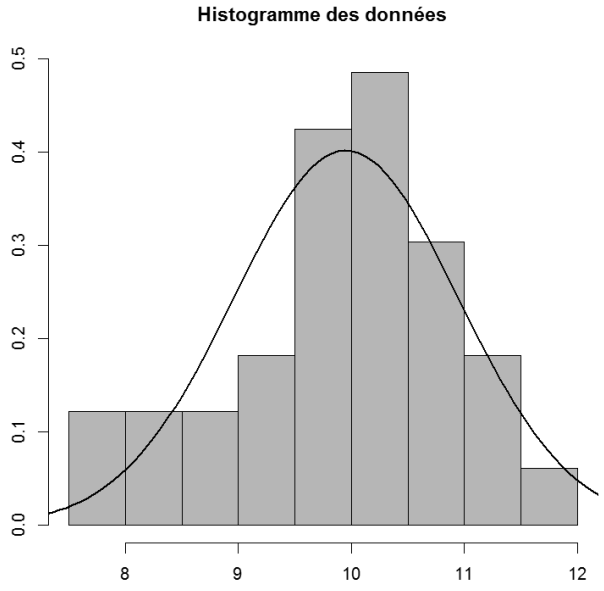
$$\mathbb{P}(X_i = j) = p_j \ , \ \forall \ j \in [\![1,6]\!]$$

Les paramètres du modèle sont ainsi les 6 probabilités $(p_j)_{1\leq j\leq 6}$ caractérisant les probabilités de chaque face du dé. Notons que ces probabilités ne sont pas indépendantes puisqu'elles sont soumises à la relation $p_1 + \cdots + p_6 = 1$. Ainsi, le paramètre vectoriel du modèle est décrit par $\boldsymbol{\theta} = \begin{pmatrix} p_1 \\ \vdots \\ p_6 \end{pmatrix} \in \mathcal{I} = \left\{ \mathbf{x} \in (0,1)^6 \text{ tel que } \sum_{j=1}^{6} x_j = 1 \right\}$. Par ailleurs, tester le caractère équilibré du dé revient à savoir si $\boldsymbol{\theta} \stackrel{?}{=} \left(\frac{1}{6}, \ldots, \frac{1}{6}\right)$.

2.3 Le modèle gaussien

Le modèle gaussien, aussi appelé modèle normal, est sans nul doute le modèle le plus couramment utilisé dans les applications pratiques. Il correspond à supposer que les données peuvent être décrites comme la réalisation de variables aléatoires $(X_i)_{i\leq n}$ i.i.d. suivant une loi normale $\mathcal{N}(\mu, \sigma^2)$[3]. Ce type de situations est typique de données expérimentales dont l'histogramme présente un profil en forme de « cloche ». D'un point de vue théorique, le choix de la loi normale est naturel lorsque le théorème central limite peut s'appliquer. Notons que les paramètres du modèle gaussien sont $\boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \in \mathcal{I} = \mathbb{R} \times \mathbb{R}_+^*$. Il s'agit donc d'un modèle présentant deux paramètres réels : μ et σ^2. Afin de rendre ce modèle concret, voici un exemple de jeu de données compatibles avec le modèle gaussien :

10	11.5	12	8	8.4	9.2	7.8	8.2	9.1	9.5	10
11.2	10.5	10.6	10.9	8.9	9.6	9.6	8.9	9.8	9.7	10.3
11.2	10.2	10.3	10.4	10.4	10.2	10.1	9.9	10.8	10.6	10.6



[3]La convention utilisée dans ce document concernant la notation de la loi normale est la convention statistique qui consiste à choisir la variance σ^2 comme second paramètre et non l'écart-type σ.

Fig. 1 : Histogramme des données et représentation d'une loi normale semblant modéliser correctement les données.

Notons que l'histogramme permet de vérifier visuellement si les données semblent compatibles avec un modèle gaussien. Néanmoins, surtout dans le cas d'un échantillon de faible taille, il ne faut pas s'attendre à une correspondance parfaite entre l'histogramme et une courbe gaussienne même correctement choisie.

2.4 Un modèle discret

Il arrive parfois que le modèle étudié soit décrit par un paramètre appartenant à un ensemble $\mathcal{I}$ fini ou dénombrable. Par exemple, on peut imaginer le cas suivant : dans une espèce de fleurs, il existe trois sous-espèces numérotées 1, 2 et 3, dont la taille des pétales n'est pas identique. Ainsi les densités de probabilité de chacune des espèces sont bien décrites par des lois normales $\mathcal{N}(\mu_i, \sigma_i^2)$ dont les paramètres sont donnés par le tableau suivant :

	Sous-espèce 1	Sous-espèce 2	Sous-espèce 3
Moyenne	$\mu_1 = 10$	$\mu_2 = 8$	$\mu_3 = 12$
Variance	$\sigma_1^2 = 1$	$\sigma_2^2 = 2.25$	$\sigma_3^2 = 4$

Les densités peuvent également être visualisées graphiquement :

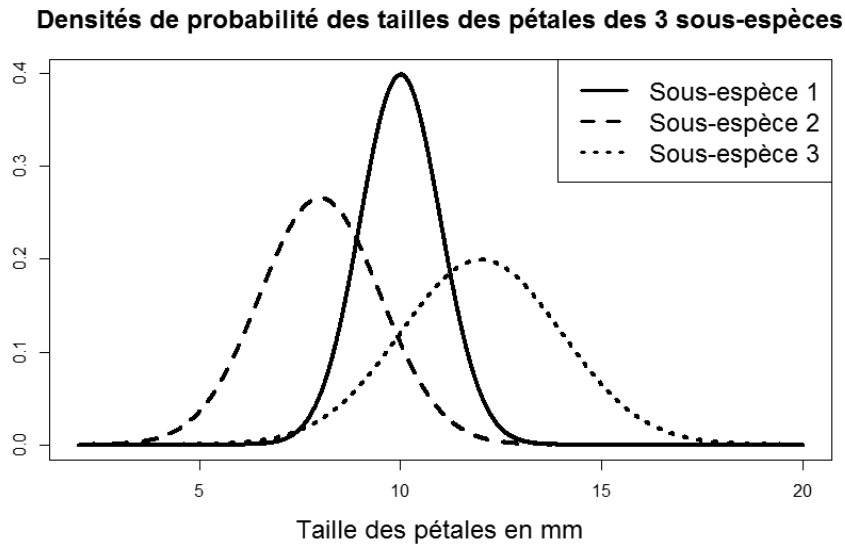


Fig. 2 : Densités de probabilité des tailles des pétales des 3 sous-espèces.

Le problème étudié est alors le suivant : on récupère des pétales d'un plan de l'espèce de fleurs étudiées mais dont la couleur a disparu. On mesure les différents pétales récoltés et on obtient les résultats suivants :

$x_1 = 11.2$	$x_2 = 9.8$	$x_3 = 12.5$	$x_4 = 9.7$	$x_5 = 10.3$
$x_6 = 13.1$	$x_7 = 10.5$	$x_8 = 11.8$	$x_9 = 8.5$	$x_{10} = 9.4$

Peut-on alors inférer à partir de ces données la sous-espèce à laquelle le plan étudié correspond ? Si oui, peut-on quantifier la fiabilité de la prédiction en précisant la probabilité de commettre une erreur ?

Le modèle associé à ce problème correspond à un modèle discret avec $\mathcal{I} = \{1, 2, 3\}$ dont les lois de probabilité sont des lois normales donnant

$$\mathcal{P} = \left\{ P_{\boldsymbol{\theta}} = \mathcal{N}(\mu_{\boldsymbol{\theta}}, \sigma^2_{\boldsymbol{\theta}}) \, , \, \boldsymbol{\theta} \in \{1, 2, 3\} \right\}$$

où les paramètres des lois normales sont donnés dans le tableau ci-dessus. Le problème énoncé correspond donc à savoir si $\boldsymbol{\theta} \stackrel{?}{=} 1$ ou $\boldsymbol{\theta} \stackrel{?}{=} 2$ ou $\boldsymbol{\theta} \stackrel{?}{=} 3$.

3 Conclusion du chapitre

Le but des statistiques inférentielles est donc d'inférer à partir de données expérimentales et d'un modèle (i.e. un ensemble prédéfini de lois de probabilité indicées par un paramètre $\boldsymbol{\theta}$) la valeur « optimale » du paramètre du modèle. Dès lors, il est nécessaire au préalable de définir mathématiquement ce que l'on attend comme critères « d'optimalité » et de développer une théorie permettant de réaliser une estimation du paramètre $\boldsymbol{\theta}$ obéissant à ces critères.