Chapitre 1

Introduction

1.1 Rôle de l'analyse statistique

L'objective d'une méthode statistique est de tirer, au vu d'observations d'un phénomène, une *inférence* au sujet de la loi générant ces observations, afin, soit d'analyser le passé, soit de prévoir le futur.

La *statistique*, au même titre que la *physique*, est donc à envisager comme *interprétation* du monde et non comme explication de celui-ci. Lorsque la modélisation déterministe d'un phénomène est trop complexe, ou lorsque les observations sont incertain on fait appel à la modélisation probabiliste qui permet de dépasser le stade descriptif des approches déterministes.

L'inférence s'accompagne donc d'une modélisation probabiliste du phénomène observée. Deux approches sont alors en compétition, modélisation *paramétrique* et modélisation *non paramétrique*. Dans ce travail nous nous limiterons à la modélisation *paramétrique* où, on représente la distribution des observations par une loi $p(x|\boldsymbol{\theta})$ où seul le paramètre $\boldsymbol{\theta}$ de dimension finie est inconnue.

Cette perspective nous semble plus pragmatique, dans la mesure où elle considère qu'un nombre fini d'observations ne peut permettre d'estimer efficacemest qu'un nombre fini de paramètres. L'objet de l'inférence statistique est alors de déterminer un ensemble fini de paramètres $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_k\}$ à partir d'un nombre fini d'observations $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ d'une quantité X dont on a modélisé sa génération par la loi $p(x|\boldsymbol{\theta})$.

Définition 1 [Modèle paramétrique] Un modèle statistique paramétrique consiste en l'observation d'une variable X, de loi $p(x|\boldsymbol{\theta})$, où seul le paramètre $\boldsymbol{\theta}$, appartenant à un sous-ensemble Θ d'un espace vectoriel de dimension finie, n'est pas connue.

D'une manière plus générale, une fois le modèle construit, on cherche à établir une inférence sur $\boldsymbol{\theta}$, c'est-à-dire à utiliser les observations $\boldsymbol{x}$ afin d'évaluer $\boldsymbol{\theta}$, en vue d'une décision liée à ces paramétres. L'inférence peut ne concerner qu'une partie des composantes de $\boldsymbol{\theta}$. Elle peut se formuler de manière précise comme "Quelle est la valeur de θ_1 ? ", ou vague comme "θ_1 est-il positive? ", ou encore "θ_2 est-il entre a et $a + \Delta a$? ". Dans le premier cas on parle de l'*estimation* et dans le deuxième cas on parle de *test d'hypothèse*.

Comparée à la théorie des probabilités, la démarche statistique est fondamentalement une démarche d'*inversion* qui est de remonter des "effets" (observations) aux "causes" (paramètres).

1.2 Approche statistique classique

Dans l'approche classique de statistique cett inversion est flagrant dans la notion de *vraisemblance*. En effet, on écrit

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})$$

en considérant $l(\boldsymbol{\theta}|\boldsymbol{x})$ comme une fonction de $\boldsymbol{\theta}$, on la normalise (quand cela est possible) pour en faire une fonction densité sur $\boldsymbol{\theta}$ et on l'utilise soit pour estimation soit pour le teste d'hypothèse sur $\boldsymbol{\theta}$. Par exemple, en estimation on cherche la valeur $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}$ qui maximise $l(\boldsymbol{\theta}|\boldsymbol{x})$, c'est l'*estimation au sense du maximum de vraisemblance*:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MV}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \{l(\boldsymbol{\theta}|\boldsymbol{x})\}$$

En test d'hypthèse on utilise le rapport de deux vraisemblances

$$r = \frac{l(\boldsymbol{\theta} = \boldsymbol{a}|\boldsymbol{x})}{l(\boldsymbol{\theta} \neq \boldsymbol{a}|\boldsymbol{x})}$$

pour décider entre deux hypothèses $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{a}$ et $\mathrm{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{a}$.

On utilise $l(\boldsymbol{\theta}|\boldsymbol{x})$ comme si elle était une fonction densité de probabilité de $\boldsymbol{\theta}$ coditionnellemt aux observations $\boldsymbol{x}$. Cette inversion est purement formelle, alors que dans l'approche bayésienne, comme nous le verrons un peu plus tard, cette inversion se fait d'une manière plus satisfaisante par la règle de Bayes.

Exemple 1 [Détection d'un signal constant bruité] Soit $\{x_i = \theta + b_i, i = 1, \cdots, n\}$ les n-échantillons d'un signal constant noyé dans un bruit blanc gaussien, *i.e.*; $b_i \sim \mathbf{N}(b_i|0,1)$ ou d'une manière équivalente $x_i \sim \mathbf{N}(x_i|\theta,1)$. Il s'agit alors d'abord de détecter l'existance d'un signal et s'il existe d'estimer sa valeur θ. Pour cela on note que

$$\begin{aligned} l(\theta|\boldsymbol{x}) &= \prod_{i=1}^{n} \mathbf{N}(x_i|\theta,1) = c_1 \ \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right] \\ &= c_2 \ \exp\left[-\frac{1}{2}[s^2 + n(\bar{x}-\theta)^2]\right] \end{aligned}$$

avec

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{et} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Ainsi, la détection est équivalent au test d'hypothèse $\mathrm{H}_0 : \theta \neq 0$ contre $\mathrm{H}_1 : \theta = 0$. On définie alors le rapport de vraisemblance r

$$r = \frac{l(\theta \neq 0|\boldsymbol{x})}{l(\theta = 0|\boldsymbol{x})} = \frac{c\exp\left[-\frac{1}{2}[s^2 + n(\bar{x}-\theta)^2]\right]}{c\exp\left[-\frac{1}{2}[s^2 + n\bar{x}^2]\right]} = \exp\left[-\frac{1}{2}[n(\bar{x}-\theta)^2 - n\bar{x}^2]\right],$$

et un seuil r_0 et on décide pour H_0 si $r > r_0$, *i.e.*; si

$$(\bar{x}-\theta)^2 - \bar{x}^2 < \frac{2}{n}\log r_0.$$

Notons que le choix du seuil r_0 est arbitraire. Un choix usuel est $r_0 = 1$ ce qui conduit à la condition

$$(\bar{x}-\theta)^2 < \bar{x}^2,$$

pour acceptation de l'hypothèse H_0. Pour estimer θ par maximum de vraisemblance on obtient

$$\widehat{\theta}_{\mathrm{MV}} = \arg\max_{\theta} \{l(\theta|\boldsymbol{x})\} = \bar{x}.$$

Exemple 2 [Estimation de l'amplitude d'un signal sinusoîdale bruité] Soit $\{x_i = \theta \sin(\omega t_i) + b_i, i = 1, \cdots, n\}$ les n-échantillons d'un signal sinusoîdale noyé dans un bruit blanc gaussien, *i.e.*; $b_i \sim \mathbf{N}(b_i|0,1)$ ou d'une manière équivalente $x_i \sim \mathbf{N}(x_i|\theta \sin(\omega t_i), 1)$. Il s'agit alors d'estimer θ par la méthode du maximum de vraisemblance. Pour cela on note

$$\begin{aligned} l(\theta|\boldsymbol{x}) &= \prod_{i=1}^{n} \mathbf{N}(x_i|\theta \sin(\omega t_i), 1) \\ &= c \exp\left[-\frac{1}{2}\sum_{i=1}^{n}[x_i - \theta \sin(\omega t_i)]^2\right] \end{aligned}$$

et ainsi, on obtient

$$\widehat{\theta}_{\mathrm{MV}} = \arg\max_{\theta}\{l(\theta|\boldsymbol{x})\} = \arg\min_{\theta}\left\{\sum_{i=1}^{n}[x_i - \theta \sin(\omega t_i)]^2\right\}.$$

On retrouve la notion de l'estimation au sense des moindres carrés.

1.3 Approche bayésienne

L'idée de base dans cette approche est dans le fait que l'on suppose connaître une information *a priori* supplémentaire sur les paramètres $\boldsymbol{\theta}$ que l'on peut la traduire sous la forme d'une loi *a priori* $p(\boldsymbol{\theta})$. L'objective étant donc d'utiliser cette information supplémentaire. Sachant que l'information contenue dans les observations $\boldsymbol{x}$ est contenue dans $p(\boldsymbol{x}|\boldsymbol{\theta})$ et l'information *a priori* sur $\boldsymbol{\theta}$ dans $p(\boldsymbol{\theta})$, on peut utiliser la règle de Bayes pour combiner ces deux types d'informations en définissant la la loi *a posteriori*

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\displaystyle\int p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}$$

qui contiendra donc toute information sur $\boldsymbol{\theta}$.

On remarque que l'inversion de cause à effet est ici beaucoup plus naturelle. Elle se fait d'une manière cohérente, car l'état de connaissance *a priori* sur $\boldsymbol{\theta}$ traduite par la loi *a priori* $p(\boldsymbol{\theta})$ est transformé, après les observations $\boldsymbol{x}$, en état de connaissance *a posteriori* par la loi *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{x})$.

Remarquons aussi que si les observations x_i sont indépendantes on a

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\boldsymbol{x}),$$

et la loi *a posteriori* peut s'écrire

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto l(\boldsymbol{\theta}|\boldsymbol{x})\, p(\boldsymbol{\theta}).$$

Définition 2 [Modèle bayésien] On appelle *modèle bayésien* la donnée d'un modèle paramétrique, $p(\boldsymbol{x}|\boldsymbol{\theta})$, et d'une loi *a priori* $p(\boldsymbol{\theta})$ sur les paramètres.
Étant donnée la loi des observations $p(\boldsymbol{x}|\boldsymbol{\theta})$ et la loi *a priori* $p(\boldsymbol{\theta})$ on peut construire
(a) la loi jointe de $(\boldsymbol{\theta}, \boldsymbol{x})$,

$$\phi(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta});$$

(b) la loi marginale de $\boldsymbol{x}$,

$$m(\boldsymbol{x}) = \int \phi(\boldsymbol{\theta}, \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta} = \int p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta};$$

et
(c) la loi *a posteriori* de $\boldsymbol{\theta}$, obtenue par la formule de Bayes,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{m(\boldsymbol{x})}.$$

Exemple 3 Si on suppose que $x \sim \mathbf{Bin}(x|\theta, n)$ et $\theta \sim \mathbf{Beta}(\theta|\alpha, \beta)$, alors on a

$$p(x|\theta) = \mathbf{Bin}(x|\theta, n) = C_n^x \theta^x (1-\theta)^{n-x},$$

$$p(\theta) = \mathbf{Beta}(\theta|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

et on peut calculer la loi jointe

$$\phi(\theta, x) = \frac{C_n^x}{B(\alpha, \beta)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1},$$

la loi marginale

$$\begin{aligned} m(x) &= \frac{C_n^x}{B(\alpha, \beta)} B(\alpha + x, n + \beta - x) \\ &= C_n^x \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\Gamma(n+\beta-x)}{\Gamma(\alpha+\beta+n)}, \end{aligned}$$

et la loi *a posteriori*

$$p(\theta|x) = \frac{1}{B(\alpha+x, \beta+n-x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} = \mathbf{Beta}(\alpha+x, n+\beta-x).$$

Le tableau 1.3 montre la relation entre ces différentes lois pour quelque cas classiques des lois usuelles.

L'introduction d'une loi *a priori* sur les paramètres $\boldsymbol{\theta}$ était véritablement révolutionnaire, et aujourd'hui encore, divise les statisticiens. Les orthodoxes considèrent que les paramètres sont *inconnus* mais pas *aléatoire*, et que seule une grandeur aléatoire a droit à avoir une loi de probabilité. C'est cette notion de *quantité ou variable aléatoire* qui est considérée comme une *réalité* qui les conduit à considérer une loi de probabilité comme la représentation de la *réalité* et non pas une représentation de notre état de connaissance sur cette quantité.

Alors que les bayésiens considèrent que les paramétres et les observations, sont deux quantités liées, l'une observable mais incertain (mesurable avec une précision finie), l'autre non observable, mais sur laquelle nous avons une information trés incomplète. On utilise alors la théorie des probabilités pour modéliser cette incertitude (manque de précision) des observations par l'attribution des lois $p(\boldsymbol{x}|\boldsymbol{\theta})$ et $p(\boldsymbol{\theta})$, la régle de Bayes pour le calcul de $p(\boldsymbol{\theta}|\boldsymbol{x})$ et la Statistiques pour tirer une inférence et des conclusions (décision : estimation ou test) sur $\boldsymbol{\theta}$ à partir de cette loi *a posteriori* .

Exemple 4 [Estimation de l'amplitude d'un signal sinusoîdale bruité] Prenons le cas de l'exemple 2 où $\{x_i = \theta \sin(\omega t_i) + b_i, i = 1, \cdots, n\}$ sont les n-échantillons d'un signal sinusoîdale noyé dans un bruit blanc gaussien, *i.e.*; $x_i \sim \mathbf{N}(x_i|\theta \sin(\omega t_i), 1)$ et supposons que notre savons que l'amplitude du signal en moyenne est μ et sa variation autour de μ est en moyenne égale à λ. Il n'est pas alors irrationnelle d'attribuer une loi *a priori* gaussienne $p(\theta) = \mathbf{N}(\theta|\mu, \lambda)$ à θ pour traduire cette connaissance. (Nous verrons plus târd que ce choix correspond à une loi à entropie maximale sous les contraintes de la connaissance de

Relation entre la loi des observations, la loi *a priori*, la loi marginale et la loi *a posteriori*

loi des observation $f(x\|\theta)$	loi *a priori* $\pi(\theta)$	loi marginale $m(x) = \int f(x\|\theta)\,\pi(\theta)\,\mathrm{d}\theta$	loi *a posteriori* $\pi(\theta\|x) = \frac{f(x\|\theta)\,\pi(\theta)}{m(x)}$
Variables discrètes			
Binomiale $\mathbf{Bin}(x\|n,\theta)$	Bêta $\mathbf{Bet}(\theta\|\alpha,\beta)$	Binomiale-Bêta $\mathbf{BinBet}(x\|\alpha,\beta,n)$	Bêta $\mathbf{Bet}(\theta\|\alpha+x,\beta+n-x)$
Binomiale négative $\mathbf{NegBin}(x\|n,\theta)$	Bêta $\mathbf{Bet}(\theta\|\alpha,\beta)$	Binomiale-Bêta négative $\mathbf{NegBinBet}(x\|\alpha,\beta,\theta)$	Bêta $\mathbf{Bet}(\theta\|\alpha+n,\beta+x)$
Poisson $\mathbf{Pn}(x\|\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Poisson-Gamma $\mathbf{PnGam}(x\|\alpha,\beta,1)$	Gamma $\mathbf{Gam}(\theta\|\alpha+x,\beta+1)$
Variables continues			
Gamma $\mathbf{Gam}(x\|\nu,\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Gamma-Gamma $\mathbf{GamGam}(x\|\alpha,\beta,\nu)$	Gamma $\mathbf{Gam}(\theta\|\alpha+\nu,\beta+x)$
Exponentielle $\mathbf{Ex}(x\|\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Pareto $\mathbf{Par}(x\|\alpha,\beta)$	Gamma $\mathbf{Gam}(\theta\|\alpha+1,\beta+x)$
Bêta $\mathbf{Bet}(x\|\alpha,\theta)$	Exponentielle $\mathbf{Ex}(\theta\|\lambda)$	? $\mathbf{?}(x\|\alpha,\lambda)$	Exponentielle $\mathbf{Ex}\,(\theta\|\lambda-\log(1-x))$
Normale $\mathbf{N}(x\|\theta,\sigma^2)$	Normale $\mathbf{N}(\theta\|\mu,\tau^2)$	Normale $\mathbf{N}(x\|\mu+\theta,\tau^2)$?	Normale $\mathbf{N}\left(\mu\|\frac{\mu\sigma^2+\tau^2x}{\sigma^2+\tau^2},\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$
Normale $\mathbf{N}(x\|\mu,\lambda\theta)$	Gamma $\mathbf{Gam}(\theta\|\frac{\alpha}{2},\frac{\alpha}{2})$	Student (t) $\mathbf{St}(x\|\mu,\lambda,\alpha)$	Gamma $\mathbf{Gam}\left(\theta\|\frac{\alpha+1}{2},\frac{\alpha}{2}+\frac{1}{2}(\mu-x)^2\right)$
Normale $\mathbf{N}(x\|\mu,1/\theta)$	Gamma $\mathbf{Gam}(\theta\|\alpha,\beta)$	Student (t)?? $\mathbf{St}(x\|\alpha,\beta)$	Gamma $\mathbf{Gam}\left(\theta\|\alpha+\frac{1}{2},\beta+\frac{1}{2}(\mu-x)^2\right)$

TAB. 1.1 – *Dans ce tableau, lorsque la loi des observations dépend de plusieurs paramètres on suppose que seul un paramètres noté θ est inconnu. Par exemple dans la septièmes ligne $f(x|\theta) = \mathbf{N}(x|\theta,\sigma^2)$ on suppose que σ^2 est connu alors que dans la huitième ligne $f(x|\theta) = \mathbf{N}(x|\mu,\lambda\theta)$ on suppose que μ et λ sont connus.*

la moyenne et de la variance.) Avec ce choix et en utilisant les mêmes notations que celles de l'exemple 2 on a

$$\begin{aligned} p(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \mathbf{N}(x_i|\theta\sin(\omega t_i), 1) \\ &= c_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{n}[x_i - \theta\sin(\omega t_i)]^2\right], \end{aligned}$$

$$p(\theta) = c_2 \exp\left[-\frac{1}{2\lambda}(\theta-\mu)^2\right]$$

et

$$\begin{aligned} p(\theta|\boldsymbol{x}) &= c_3 p(\boldsymbol{x}|\theta)p(\theta) \\ &= c\exp\left[-\frac{1}{2}\left[\sum_{i=1}^{n}[x_i - \theta\sin(\omega t_i)]^2 + \frac{1}{\lambda}(\theta-\mu)^2\right]\right]. \end{aligned}$$

On remarque que la loi *a posteriori* est une loi gaussienne, par conséquent elle est entièrement définie par sa moyenne et sa variance. Dans ce cas, les estimateurs MAP ou la moyenne *a posteriori* sont équivalents. On a

$$\widehat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta}\{p(\theta|\boldsymbol{x})\} = \arg\min_{\theta}\{J(\theta)\},$$

avec

$$J(\theta) = \sum_{i=1}^{n}[x_i - \theta\sin(\omega t_i)]^2 + \frac{1}{\lambda}(\theta-\mu)^2.$$

On retrouve la notion de l'estimation au sense des moindres carrés.

On a beaucoup critiqué (et cela continue toujour) le choix de la loi *a priori* $\pi(x)$, mais il en va de même pour celui des observations $p(x|\theta)$. Ce qui est étonnant, car ils admettent que X, une quantité physique, suive une loi, par exemple gaussienne de moyen θ, mais ils n'admettent pas que θ qui représente une autre quantité physique qui peut être de la même nature que X (car c'est sa moyenne par exemple), puisse suivre une autre loi gaussienne avec une moyenne μ !

1.4 Vraisemblance et statistique exhaustive

L'analyse statistique (bayésienne ou non bayésienne) est régie par un certain nombre de notions, définitions, régles et principes que j'essaye de les énumérer. Notons que nous connaissons déja les notions de

- la loi de probabilité $p(x|\boldsymbol{\theta})$ d'une grandeur X ;
- les observations $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ que l'on appelle parfois un n-échantillon ;
- la fonction de vraisemblance $l(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})$;
- l'espace des observations $\mathcal{X}$;
- l'espace des paramètres Θ ;
- la loi jointe des observations $p(\boldsymbol{x}|\boldsymbol{\theta})$;
- la loi jointe des observations et des paramètres $p(\boldsymbol{x}, \boldsymbol{\theta})$;
- la loi *a priori* des paramètres $p(\boldsymbol{\theta})$; et
- la loi *a posteriori* des paramètres $p(\boldsymbol{\theta}|\boldsymbol{x})$;

Définition 3 [Statistique exhaustive] Pour un n-échantillon $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ avec $x_i \sim p(x_i|\boldsymbol{\theta})$, une statistique $\boldsymbol{t}(\boldsymbol{x})$, fonctions des observation $\boldsymbol{x}$ à valeur dans Θ, est dite *exhaustive* si la loi $p(\boldsymbol{x}|\boldsymbol{\theta})$ ne dépende de $\boldsymbol{x}$ qu'à travers de $\boldsymbol{t}(\boldsymbol{x})$, *i.e.*;

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x})).$$

La connaissance de $\boldsymbol{\theta}$ et $\boldsymbol{t}(\boldsymbol{x})$ défini entièrement la loi $p(\boldsymbol{x}|\boldsymbol{\theta})$. Notons aussi l'équivalence entre cette condition et les conditions suivantes dans l'approche non bayésienne:

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x}))$$

et

$$l(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{t}(\boldsymbol{x})) = l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x})).$$

Exemple 5 Si on suppose que les observations x_i suivent

$$x_i \sim \mathbf{N}(x_i|\mu, \sigma^2),$$

alors on a

$$\begin{aligned} p(\boldsymbol{x}|\mu, \sigma^2) &= \prod_{i=1}^{n} \mathbf{N}(x_i|\mu, \sigma^2) \\ &= c\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}[s^2 + n(\bar{x} - \mu)^2]\right] \\ &= f(\mu, \sigma^2, \bar{x}, s^2), \end{aligned}$$

avec

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{et} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Ainsi, si on note $\boldsymbol{\theta} = [\mu, \sigma^2]$ et $\boldsymbol{t}(\boldsymbol{x}) = [\bar{x}, s^2]$ on a

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x})),$$

ou en terme de la fonction de vraisemblance on a

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x})),$$

ce qui signifie que $\boldsymbol{t}(\boldsymbol{x}) = [\bar{x}, s^2]$ est une statistique suffisante pour ces observations, et la connaissance de $\boldsymbol{t}$ définit entièrement la loi des observations.

La statistique $\boldsymbol{t}(\boldsymbol{x})$ contient donc "tout" l'information apportée par $\boldsymbol{x}$ sur $\boldsymbol{\theta}$, et sous certaines conditions de régularité, la densité de $\boldsymbol{x}$ se factorise et s'écrit

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = g(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{\theta})\, h(\boldsymbol{x}|\boldsymbol{t}(\boldsymbol{x})),$$

où g est la densité de $\boldsymbol{t}(\boldsymbol{x})$.

Définition 4 [Statistique ancilaire] Pour un n-échantillon $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ avec $x_i \sim p(x_i|\boldsymbol{\theta})$, une statistique $\boldsymbol{a}(\boldsymbol{x})$ est dite ancilaire par rapport aux paramètres $\boldsymbol{\theta}$ si la distribution de $\boldsymbol{x}$ coditionnellement à $[\boldsymbol{a}, \boldsymbol{\theta}]$ ne dépend pas à $\boldsymbol{a}$, i.e;

$$p(\boldsymbol{x}|\boldsymbol{a}(\boldsymbol{x}), \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta}).$$

Notons aussi l'équivalence entre cette condition et les conditions suivantes dans l'approche non bayésienne:

$$l(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{a}(\boldsymbol{x})) = l(\boldsymbol{\theta}|\boldsymbol{x})$$

et

$$l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x}), \boldsymbol{a}(\boldsymbol{x})) = l(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x})).$$

Exemple 6 Dans l'exemple précédent, si on définit $a_1(\boldsymbol{x}) = \min\{x_1, \cdots, x_n\}$, $a_2(\boldsymbol{x}) = \max\{x_1, \cdots, x_n\}$ et $\boldsymbol{a} = [a_1, a_2]$ on a

$$p(\boldsymbol{x}|\mu, \sigma^2, a_1(\boldsymbol{x}), a_2(\boldsymbol{x})) = p(\boldsymbol{x}|\mu, \sigma^2)$$

ou encore

$$p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{a}) = p(\boldsymbol{x}|\boldsymbol{\theta}),$$

autrement dit, la connaissance de $\boldsymbol{a}$ n'apporte rien par rapport à la connaissance de $\boldsymbol{\theta}$.

Principe d'Exhaustivité : Deux observations pour lesquelles une statistique exhaustive prend la même valeur doivent conduire à la même inférence sur θ.

Principe de Vraisemblance : Toute l'information sur $\boldsymbol{\theta}$ tirée des observations $\boldsymbol{x}$ est contenue dans la vraisemblance $l(\boldsymbol{\theta}|\boldsymbol{x})$, et si $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ sont tels qu'il existe une constante c telle que, pour tout $\boldsymbol{\theta}$,

$$l(\boldsymbol{\theta}|\boldsymbol{x}_1) = c\, l(\boldsymbol{\theta}|\boldsymbol{x}_2),$$

ils apportent la même information sur θ et doivent conduire à la même inférence.

1.5 Lien entre les deux approches

Pensant à l'aspect d'inversion de la statistique, il est tentant de considérer, sous réserve de l'intégrabilité, la fonction de vraisemlance $l(\boldsymbol{\theta}|\boldsymbol{x})$ comme une loi de probabilité sur $\boldsymbol{\theta}$ dont l'estimateur du maximimum de vraisemblance serait la mode. En effet, cell-ci est équivalent à la loi *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{x})$ lorsqu'on choisit une loi *a priori* uniforme pour $p(\boldsymbol{\theta})$. Une telle position était, par exemple, celle de Laplace qui considérait que l'absence d'information *a priori* justifiait le choix de la loi uniforme. De même, Fisher, en introduisant l'analyse fiduciaire, voulait mettre en œuvre le principe de vraisemblance sans passer par une approche bayésienne. Cette position, défendable lorsque $\boldsymbol{\theta}$ est un paramètre de position, conduit cependant à des paradoxes et des contradictions qui montrent clairement la nécessité de la théorie bayésienne plus élaborée incluant les notions des lois *a priori non informatives*, des lois *conjuguées* et des lois *a priori* de *références*.

L'approche à privilégier pour l'inférence bayésienne est celle passant par la loi *a posteriori*. En effet, travaillant conditionnellement aux observations, cette approche suit d'une manière cohérente l'idée d'inversion des causes aux effets, tout en restant fidèle au principe de vraisemblance. En fait, la loi *a posteriori* représente l'actualisation de l'information *a priori*, $p(\boldsymbol{\theta})$, au vu de l'information contenue dans les observations $\boldsymbol{x}$, au travers de la vraisemblance $l(\boldsymbol{\theta}|\boldsymbol{x})$.

Disposant ainsi d'une distribution de probabilité sur $\boldsymbol{\theta}$, le champ de l'inférence est beaucoup plus vaste que dans le cadre classique qui se contentait de $l(\boldsymbol{\theta}|\boldsymbol{x})$. On peut calculer moyenne, médiane, modes ou des régions de confiance sous la forme de régions de plus forte densité *a posteriori*. On peut aussi, en définissant un espace de décisions et d'une fonction de coût définir tout autres estimateurs classiques, ainsi que des estimateurs de la précision des estimateurs employés. On peut également définir rigoureusement la probabilité *a posteriori* d'une hypothèse $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_0$, c'est-à-dire $P(\boldsymbol{\theta} \in \Theta_0|\boldsymbol{x})$, ou comparer deux hypothèses $\mathrm{H}_0 : \boldsymbol{\theta} \in \Theta_0$ et $\mathrm{H}_1 : \boldsymbol{\theta} \in \Theta_1$ en comparant leurs probabilités respectives $\pi_0(\mathrm{H}_0) = P(\boldsymbol{\theta} \in \Theta_0|\boldsymbol{x})$ et $\pi_1(\mathrm{H}_1) = P(\boldsymbol{\theta} \in \Theta_1|\boldsymbol{x})$.

À ce stade, on peut dire que l'approche statistique classique où toute l'inférence est basée sur la vraisemblance (normalisée et utilisée comme une fonction densité) est un cas particulier de l'approche bayésienne avec une loi *a priori* uniforme. En effet, si $p(\boldsymbol{\theta}) = c$, la loi *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto l(\boldsymbol{\theta}|\boldsymbol{x})$ et, d'aprés le principe de vraisemblance toute inférence

tirée de ces deux approches seron équivalentes. Cependant, seule l'approche bayésienne permet d'introduire des informations complémentaires sur $\boldsymbol{\theta}$ sous forme d'une loi *a priori* non-uniforme.

Une fois la loi *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{x})$ calculée, il reste encore à savoir l'utiliser correctement. Deux avis alors se rencontrent.

Le premier consiste à transmettre cette loi à l'utilisateur comme telle qui en fera ce que lui semble bon. Par exemple, quand il s'agit d'un seule paramètre il peut tout simplement tracer la courbe et en déduire

- la mode :
$$\mathrm{M}\,[\theta] = \sup_{\theta\in\Theta} p(\theta|\boldsymbol{x}),$$
- la moyenne :
$$\mathrm{E}\,[\theta] = \int_{\theta\in\Theta} \theta\, p(\theta|\boldsymbol{x})\,\mathrm{d}\theta,$$
- la variance :
$$\mathrm{Var}\,[\theta] = \int_{\theta\in\Theta} [\theta - \mathrm{E}\,[\theta]]\, p(\theta|\boldsymbol{x})\,\mathrm{d}\theta,$$

etc.

Dans le cas d'un vecteur des paramètres $\boldsymbol{\theta}$ il est plus difficile de résumer la loi *a posteriori* . On peut cependant calculer ces mêmes quantités en utilisant les lois marginales $p(\theta_i|\boldsymbol{x})$:

- la mode :
$$\mathrm{M}\,[\theta_i] = \sup_{\theta_i\in\Theta_i} p(\theta_i|\boldsymbol{x}),$$
- la moyenne :
$$\mathrm{E}\,[\theta_i] = \int_{\theta_i\in\Theta_i} \theta_i\, p(\theta_i|\boldsymbol{x})\,\mathrm{d}\theta_i,$$
- la variance :
$$\mathrm{Var}\,[\theta_i] = \int_{\theta_i\in\Theta_i} [\theta_i - \mathrm{E}\,[\theta_i]]\, p(\theta_i|\boldsymbol{x})\,\mathrm{d}\theta_i,$$

mais aussi

- la covariance :
$$\mathrm{Cov}\,[\theta_i, \theta_j] = \iint_{\boldsymbol{\theta}\in\Theta} [\theta_i - \mathrm{E}\,[\theta_i]]\,[\theta_j - \mathrm{E}\,[\theta_j]]\, p(\boldsymbol{\theta}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta}.$$

Notons aussi que la mode $\mathrm{M}[\theta_i]$ n'est pas forcément égale à $\mathrm{M}[\boldsymbol{\theta}]_i$ où:

$$\mathrm{M}\,[\boldsymbol{\theta}] = \sup_{\boldsymbol{\theta}\in\Theta} p(\boldsymbol{\theta}|\boldsymbol{x})$$

Le deuxième consiste à définir un estimateur $\delta(\theta)$ suivant un critère. En effet, dans l'approche théorie de la décision, on définit une fonction *coût* $C(\theta, \delta)$ (ou une fonction *utilité*) qui mesure, d'une certaine manière, la conformitée de la décision, et choisit comme le meilleur estimateur celui qui minimiserait le coût moyen. Le choix ou la détermination de la fonction coût ne peut se faire qu'une manière subjective.

Dans les chapitres qui vont suivrent nous détaillerons un plus ces deux approches.