Cours 1

1 Introduction

Dans de nombreuses situations d'expériences aléatoires, il semble raisonnable d'imaginer que le praticien a une certaine idée du phénomène aléatoire qu'il est en train d'observer. Or, la démarche statistique classique repose essentiellement sur un principe de vraisemblance qui consiste à considérer que ce qui a été observé rend compte de manière exhaustive du phénomène. Mais l'observation ne fournit qu'une image et celle-ci peut être mauvaise. Certes cet inconvénient est en général gommé par les considérations asymptotiques et un certain nombre de théorèmes permettent d'évaluer la bonne qualité des estimateurs si le nombre d'observations est suffisant. L'analyse bayésienne des problèmes statistiques propose d'introduire dans la démarche d'inférence, l'information dont dispose a priori le praticien. Dans le cadre de la statistique paramétrique, ceci se traduira par le choix d'un loi sur le paramètre d'intérêt.

Dans l'approche classique, le modèle statistique est le triplet $(\mathfrak{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$. Ayant un *a priori* sur le paramètre, modélisé par une densité de probabilité que nous noterons $\pi(\theta)$, on "ré-actualise" cet a priori au vu de l'observation en calculant la densité a posteriori $\pi(\theta|x)$, et c'est à partir de cette loi que l'on mène l'inférence.
On peut alors, par exemple, de manière intuitive pour le moment, retenir l'espérance mathématique ou encore le modee de cette densité a posteriori comme estimateur de θ.
Le paramètre θ devient donc en quelque sorte une variable aléatoire, à laquelle on associe une loi de probabilité dite **loi a priori**.
On sent bien d'emblée que les estimateurs bayésiens sont très dépendants du choix de l'a priori. Différentes méthodes existent pour déterminer ces lois a priori. On peut se référer à des techniques bayésiennes empiriques, où l'on construit la loi a priori sur la base d'une expérience passée, usant de méthodes fréquentistes, pour obtenir forme et valeurs de paramètres pour cette loi. Nous verrons que l'on peut aussi modéliser l'absence d'information sur le paramètre au moyen des lois dites *lois non informatives*.
L'approche bayésienne se différencie donc de l'approche classique dans le sens où le paramètre θ n'est plus considéré comme étant totalement inconnu ; il est devenu une v.a. dont le comportement est supposé connu. On fait intervenir dans l'analyse statistique une distribution associée à ce paramètre.
On donne la définition :

Définition 1 – *On appelle* **modèle statistique bayésien**, *la donnée d'un modèle statistique paramétré* $(\mathfrak{X}, \mathcal{A}, P_\theta, \theta \in \Theta)$ *avec* $f(x/\theta)$ *densité de* P_θ **et** *d'une loi* $\pi(\theta)$ *sur le paramètre.*

La démarche de l'analyse bayésienne conduit au calcul d'une **loi a posteriori** $\pi(\theta|x)$; actuali-

sation de la loi a priori $\pi(\theta)$ au vu de l'observation.
Ce calcul repose sur la version continue du théorème de Bayes :

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)}$$

$f(x|\theta)$ désignant la loi de l'observation ou **vraisemblance** et $f(x)$ la loi marginale ou **prédictive** :

$$f(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$$

.

Exemple 1 – Considérons un n-échantillon i.i.d. $\underline{X} = (X_1, \cdots, X_n)$ de loi exponentielle de paramètre $\theta > 0$. (i.i.d. signifie que les v.a. X_i sont**i**ndépendantes et **i**dentiquement **d**istribuées)
La vraisemblance a pour expression :

$$f(\underline{x}|\theta) = (1/\theta)^n \exp\{-\sum_{i=1}^n x_i/\theta\}.$$

On prend une loi a priori de type gamma-inverse sur θ. La densité de cette loi a priori est donnée par :

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/u)^{\alpha+1}\exp\{-\beta/u\}\ ,\quad u \in \mathbb{R}^+\ ,\quad \alpha, \beta > 0.$$

La loi jointe est donc :

$$f(\underline{x}, \theta) = f(\underline{x} \mid \theta)\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/\theta)^{n+\alpha+1}\exp\{-(\beta + \sum_{i=1}^n x_i)/\theta\}$$

et la prédictive s'écrit :

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^{+\infty}(1/\theta)^{n+\alpha+1}\exp\{-(\beta + \sum_{i=1}^n x_i)/\theta\} = \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(n+\alpha)}{(\sum_{i=1}^n x_i + \beta)^{n+\alpha}}$$

Ainsi, la loi a posteriori a pour expression :

$$\pi(\theta \mid \underline{x}) = \frac{(\sum_{i=1}^n x_i + \beta)^{n+\alpha}}{\Gamma(n+\alpha)}(1/\theta)^{n+\alpha+1}e^{-(\beta+\sum_{i=1}^n x_i)/\theta}$$

Il s'agit d'une loi gamma inverse de paramètres $(n + \alpha, \sum_{i=1}^n x_i + \beta)$ □

2 Estimation Ponctuelle

2.1 Coût et Décision

Le problème très général auquel on s'intéresse ici est celui d'un individu plongé dans un environnement donné (**nature**) et qui, sur la base d'**observations**, est conduit à mener des **actions** et à prendre des **décisions** qui auront un **coût**.
Les espaces intervenant dans l'écriture d'un **modèle de décision** sont :

$\mathfrak{X}$: l'espace des observations,

Θ : l'espace des états de la nature (l'espace des paramètres dans le cas d'un problème statistique),

A : l'espace des actions ou décisions, dont les éléments sont des images de l'observation par une application δ appelée règle de décisions (une statistique (i.e. fonction de observations) dans le cas d'un problème statistique),

$\mathcal{D}$: l'ensemble des *règles de décisions* δ, applications de $\mathfrak{X}$ dans A (les estimateurs possibles).

On note a une action. On a : $a = \delta(x)$.

L'inférence consiste à choisir une règle de décision $\delta \in \mathcal{D}$ concernant $\theta \in \Theta$ sur la base d'une observation $x \in \mathfrak{X}$, x et θ étant liés par la loi $f(x|\theta)$.

En statistique, la règle de décision est un estimateur, l'action est une estimation (valeur de l'estimateur au point d'observation x).

Pour choisir une décision, on construit une relation de préférence en considérant une mesure du coût ou perte encourue lorsqu'on prend la décision $\delta(x)$ et que *l'état de la nature* est θ.

Pour ce faire on introduit la fonction L, appelée **fonction de coût** (ou de perte) définie de la manière suivante :

Définition 2 – *On appelle fonction de coût, toute fonction L de $\Theta \times A$ dans $\mathbb{R}$.*

$L(\theta, a)$ évalue le coût d'une décision a quand le paramètre vaut θ. Elle permet donc, en quelque sorte, de quantifier la perte encourue par une mauvaise décision, une mauvaise évaluation de θ. Il s'agit d'une fonction de θ. Un coût négatif correspond à un gain.

Exemples de construction de fonction de coût

1. (Berger) Pour décider de la commercialisation d'un nouveau médicament antalgique, une entreprise de l'industrie pharmaceutique s'intéresse en particulier à deux facteurs susceptible d'affecter sa décision :

   θ_1 : la proportion d'individus sur laquelle l'antalgique sera efficace,

   θ_2 : la part de marché que le médicament est susceptible de prendre (demande).

   Ces deux paramètres sont inconnues. On pourra faire des expériences pour essayer d'obtenir des informations à leur sujet. On a ici un problème classique de théorie de la décision où le but ultime est de décider de mettre ou non le produit sur le marché, dans quelle proportion, à quel prix, etc.

   Intéressons-nous à θ_2. θ_2 est une proportion.

   On a donc $\Theta = \{\theta_2 : 0 \leq \theta_2 \leq 1\}$ et une décision sera donc ici un nombre compris entre 0 et 1 ; une estimation que les experts veulent faire de la part de marché. L'espace des actions A est l'intervalle $[0, 1]$.

L'information dont on dispose est la suivante. Les experts pensent que le coût d'une surestimation de la demande est 2 fois plus élevé qu'une sous-estimation de celle-ci. Ce qui peut se traduire par une fonction de coût de la forme suivante :

$$L(\theta_2, a) = \begin{cases} \theta_2 - a & \text{si} \quad \theta_2 - a \geq 0 \quad \text{(sous estimation)} \\ 2(a - \theta_2) & \text{si} \quad \theta_2 - a \leq 0 \quad \text{(sur estimation)} \end{cases}$$

2. Une fonction de coût classiquement utilisée est la fonction de **coût quadratique** : $L(\theta, d) = (\theta - d)^2$. C'est le critère des moindres carrés en régression.

3. On retrouve cette notion de coût en théorie des jeux. Dans ce cadre, un jeu est décrit par un triplet $(\Theta, \mathcal{A}, L)$, Θ étant les états possibles de la nature.
   Considérons le jeu suivant à deux joueurs : chaque joueur montre un doigt ou deux. Lorsque la somme est paire, le joueur A gagne. Si la somme est impaire, c'est le joueur B qui gagne. Dans tous les cas, le gagnant reçoit du perdant la somme, en euros, des nombres de doigts apparus. Plaçons nous du point de vue du joueur B. Celui-ci représente le décideur (l'agent économique, le statisticien). Le joueur B représente lui nature. Le joueur A ne sait pas ce que va jouer B, c'est-à-dire il ne connaît pas l'état de la nature. Cet état de la nature est un point θ de $\Theta = \{1, 2\}$ et le joueur A sans être informé de l'état de la nature (ce que va jouer B), va devoir choisir une action a dans $\mathcal{A} = \{1, 2\}$.
   On peut alors décrire une fonction de coût, pour le joueur B de la manière suivante : Si le couple (1,1) apparaît, la somme est paire. Le joueur A l'emporte et gagne 2 euros donc le coût pour B est - 2 euros.
   Si on a (1,2), c'est le joueur B qui gagne et on a un coût de 3 euros pour le joueurs B (gain). En poursuivant ce raisonnement, on complète le tableau ci-dessous.

		B (décideur)	
		1	**2**
A (nature)	**1**	-2	3
	2	3	-4

4. Le paradoxe de Saint-Petersbourg – Un mendiant possède un billet de loterie qui peut lui permettre de gagner 20 000 ducats. Il croise un marchand qui lui propose de lui acheter son billet 9 000 ducats. Le mendiant est donc confronté à un problème de décision qui consiste à choisir entre 2 loteries en quelques sortes. L'une lui offre la possibilté de gagner 20 000 ducats avec une certaine probabilité p ou ne rien gagner du tout avec une probabilité $1 - p$. L'autre est une loterie certaine, il gagne avec certitude 9 000. On peut définir le coût de la décision comme étant l'espérance de gain des loteries. Ainsi, si $p > 0,5$, le mendiant a tout intérêt à jouer puisque l'espérance de gain de la loterie est de $p \times 20000 = 10000 > 9000$ ducats !

2.2 Risque Fréquentiste

On dira qu'une décision est une *bonne décision* si elle conduit à un coût nul.
Autrement dit, une bonne décision est solution de l'équation :

$$L(\theta, \delta(x)) = 0.$$

θ étant inconnu, on ne peut évidemment pas résoudre cette équation. Classer les décisions par la seule considération du coût est donc impossible. Celui-ci ne prend pas en compte l'information apportée par le modèle $f(x \mid \theta)$. Ces remarques conduisent à considérer la moyenne de la perte, c'est le *risque fréquentiste.*

Définition 3 – *On appelle risque fréquentiste le coût moyen (l'espérance mathématique) du coût d'une règle de décision :*

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X))] = \int_{\mathfrak{X}} L[\theta, \delta(x)] dP_\theta(x) \tag{1}$$

On peut alors donner la définition suivante :

Définition 4 – *On dira que δ_1 est préférable à δ_2 et on note $\delta_1 \prec \delta d_2$ si :*

$$R(\theta, \delta_1) \leq R(\theta, \delta_2) \ , \quad \forall \theta \in \Theta.$$

Cette définition permet d'établir un préordre sur l'ensemble $\mathcal{D}$ des décisions.
Cependant, ce préordre est partiel puisqu'il ne permet pas de comparer deux règles de décision telles que :

$$R(\theta_1, \delta_1) < R(\theta_1, \delta_2) \ \text{ et } \ R(\theta_2, \delta_1) > R(\theta_2, \delta_2,).$$

2.3 Risque de Bayes

Puisque l'approche Bayésienne met à la disposition du statisticien une loi a priori $\pi(\theta)$, on peut considérer la moyenne du risque fréquentiste i.e. la moyenne du coût moyen suivant la loi a priori : $E^\pi[R(\theta, \delta(X))]$. Il s'agit du **risque bayésien** ou **risque de Bayes** que l'on note $r(\pi, \delta)$. On a :

$$\begin{aligned} r(\pi, \delta) &= E^\pi[R(\theta, \delta)] \\ &= \int_\Theta R(\theta, \delta)\pi(\theta) d\theta \\ &= \int_\Theta \int_{\mathfrak{X}} L(\theta, \delta(x)) f(x|\theta) dx \pi(\theta) d\theta \\ &= \int_\Theta \int_{\mathfrak{X}} L(\theta, \delta(x)) \pi(\theta|x) f(x) dx d\theta \end{aligned}$$

On définit alors le **coût a posteriori** $\rho(\pi, \delta(x))$ comme étant la moyenne du coût par rapport à la loi a posteriori :

$$\rho(\pi, \delta(x)) = E^{\pi(.|x)}[L(\theta, \delta(x))] = \int_{\Theta} L[\theta, \delta(x)]\pi(\theta \mid x)d\theta$$

Il s'agit d'une fonction de x.
On a le résultat suivant

Proposition 1 – *Le risque de Bayes $r(\pi, \delta)$ est la moyenne du coût a posteriori $\rho(\pi, \delta(x))$ suivant la loi marginale $f(x)$.*

Preuve : $r(\pi, \delta) = \int_{\Theta} \int_{\mathfrak{X}} L(\theta, \delta(x)) f(x \mid \theta)\pi(\theta)dxd\theta$
or $f(x \mid \theta)\pi(\theta) = \pi(\theta \mid x)f(x)$.
On a donc :

$$\begin{aligned} r(\pi, \delta) &= \int_{\mathfrak{X}} \int_{\Theta} L(\theta, \delta(x))\pi(\theta|x)d\theta f(x)dx \\ &= \int_{\mathfrak{X}} \rho(\pi, \delta(x))f(x)dx \end{aligned}$$

□