L'approche bayésienne

Contenu

1. La philosophie bayésienne
2. Probabilité subjective
3. Procédures bayésiennes (exemples)
4. Lois a priori et a postériori
5. Densité de Jeffreys
6. Estimation

1 La philosophie bayésienne

La nature du contenu de ce chapitre est conceptuellement différente de ce qui a été vu durant la majeure partie de la session. Les méthodes utilisées lors des séances précédentes sont dites fréquentielles (ou classiques). Elles sont basées sur les postulats suivants :

1. Probabilité = limite de fréquences relatives.
   Probabilité = propriétés objectives du monde réel.
2. Paramètres = constantes fixes mais inconnues.
   Puisque ce sont des constantes, on ne peut pas leur associer des lois de probabilités.
3. Méthodes statistiques : doivent être construites pour avoir des propriétés asymptotiques bien définies. Par exemple, un $IC_{0.95}$ devrait contenir la vraie valeur dans 95% des cas si l'expérience est répétée un grand nombre de fois. Pour ce faire, on choisit un EAS $X_1, \cdots, X_n$ d'une population ayant une loi P_θ ($\theta \in \Theta$, espace des paramètres). Le processus de décision dépend alors de la fonction de vraisemblance
$$L(\theta) = \prod_{i=1}^{n} f(x_i \,|\theta)$$

Une autre approche pour faire de l'inférence est l'inférence bayésienne. Voici ses postulats :

1. Probabilité = degré de croyance (différent de fréquence limite). On peut donc énoncer des affirmations de type probabilistique sur beaucoup de choses, pas seulement sur des données sujettes à des fluctuations aléatoires.
   Exemples. La probabilité que la COVID-19 sera éradiquée avant la fin de l'année est inférieure à 0.01. La probabilité qu'il pleuve aujourd'hui à Bamako est égale à 0.5.
2. On peut attribuer des probabilités au(x) paramètre(s) θ, bien que ce soit une (des) constante(s). Nous interprétions la quantité
$$f(x_i \,|\theta)$$
   comme une loi (densité ou fonction de masse) dépendant du paramètre fixe (mais inconnu) θ. Cette quantité devient une loi conditionnelle :
$$f_{X|\Theta}(x_i \,|\theta)$$
   avec $X = (X_1, \cdots, X_n)$.
   Note. Θ désigne aussi bien l'espace des paramètres qu'une variable aléatoire de support cet espace (pour ne pas alourdir les notations).

3. On fait des inférences sur le paramètre θ en trouvant une loi de probabilité $h(\theta)$ pour θ (dite loi a priori : reflète la croyance subjective du chercheur). Cette loi est généralement adoptée avant la collecte des données (expertise, information annexe, etc.).
4. On peut alors faire toutes les inférences et analyses statistiques usuelles en combinant l'information sur θ donnée par l'échantillon à travers la fonction de vraisemblance et la loi a priori $h(\theta)$ de Θ et ce, par le biais du **théorème de Bayes**. Ceci mène à la loi a posteriori (mise à jour) de Θ. C'est cette loi qui guide toute l'inférence à faire.

Remarques.
- Cette approche est (de moins en moins) controversée parce qu'elle contient une notion subjective de la probabilité.
- Elle ne garantit donc pas une bonne performance à long terme.
- Les statistiques mettent plus d'emphase sur les méthodes fréquentielles (quoique cela commence à changer).
- Le big data, l'apprentissage-machine, etc. sont plutôt bayésiens.
- Pour les curieux. Le théorème de Bayes (publié à titre posthume en 1763 par un de ses compagnons : Richard Price) du pasteur Thomas Bayes est redécouvert (de façon indépendante) par Laplace sous le nom de probabilité inverse (1774). Les fondements théoriques modernes de cette approche sont notamment dus à de Finetti (1937), Savage (1954), Jeffreys (1957), etc. Un opposant notoire à cette approche est Fisher. A titre annexe, l'approche bayésienne peut être approchée par le biais de la théorie de la décision (formalisée par Abraham Wald en 1950 : *Statistical Decision Functions*) où on cherche des règles de décision optimales. Dans cette optique, on peut aussi consulter les livres classiques de Ferguson (1967) et Berger (1985).

2 Probabilité subjective

Les probabilités subjectives sont donc le fondement des méthodes bayésiennes. Donnons un exemple illustratif.

Supposons qu'une personne a attribué $p = P(E) = 2/5$ à un certain évènement E. Alors la cote (odd) contre E est

$$o(E) = \frac{1-p}{p} = \frac{1-2/5}{2/5} = \frac{3}{2} \quad (3 \text{ contre } 2)$$

Si cette personne accepte de parier, elle devrait être prête à accepter les deux faces (duales) du pari :

1. gagner 3\$ si E se réalise et perdre 2\$ si E ne se réalise pas ;
2. gagner 2\$ si E ne se réalise pas et perdre 3\$ si E se réalise.

Si aucune de ces deux situations n'est acceptable pour elle, elle devrait revoir son $P(E)$.

Pour fixer les idées, ceci ressemble à la situation où deux enfants se partagent une barre de chocolat de façon aussi égale que possible : l'un divise la barre et l'autre choisit quel morceau prendre (le plus grand possible). Ceci signifie que le premier enfant essaie de partager la barre aussi fort qu'il peut pour obtenir deux parts égales.

Donc, si vous voulez utiliser les probabilités subjectives, il faut être prêts à accepter les deux possibilités du pari (symétrie).

Supposons que vous acceptez $p = P(E)$ comme étant le prix juste pour l'événement E : vous gagnez 1\$ si E se réalise (et votre profit net est donc égal à $(1-p)$ \$) puisque vous avez déjà payé p comme droit, et vous perdez vos p\$ si E ne se réalise pas.

Il peut être montré que toutes les règles (définitions et théorèmes) des probabilités usuelles restent valides dans les probabilités subjectives.

3 Procédures bayésiennes (exemples)

Exemple 1 *Pour comprendre l'inférence bayésienne, revoyons le théorème de Bayes :*

$$P(E_j|F) = \frac{P(F|E_j)\,P(E_j)}{\sum_i P(F|E_i)}$$

dans un contexte où on veut apprendre quelque chose sur un paramètre d'une loi de probabilité.

Supposons qu'on ait $Poi(\theta)$, $\theta > 0$ *et qu'on sait que* $\theta = 2$ *ou* $\theta = 3$. *Dans l'inférence bayésienne, le paramètre est considéré comme une variable aléatoire* Θ. *Supposons, dans cet exemple, qu'on ait la loi a priori*

Θ	2	3	*Total*
P	1/3	2/3	1

On prend un échantillon aléatoire simple (EAS) de taille $n = 2$ *et voici les résultats obtenus :*

$$x_1 = 2, \quad x_2 = 4$$

Connaissant ces données, quelles sont les probabilités a postériori pour $\Theta = 2$ *et* $\Theta = 4$ *? Bayes :*

$$\begin{aligned} P(\Theta = 2|X_1 = 2, X_2 = 4) &= \frac{P(\Theta = 2, X_1 = 2, X_2 = 4)}{P(X_1 = 2, X_2 = 4)} \\ &= \frac{(1/3)\,\frac{2^2e^{-2}}{2!}\,\frac{2^4e^{-2}}{4!}}{(1/3)\,\frac{2^2e^{-2}}{2!}\,\frac{2^4e^{-2}}{4!} + (2/3)\,\frac{3^2e^{-3}}{2!}\,\frac{3^4e^{-3}}{4!}} \\ &= \frac{\frac{4}{9}e^{-4}}{\frac{4}{9}e^{-4} + \frac{81}{8}e^{-6}} \\ &\approxeq 0.244\,91 \end{aligned}$$

De manière analogue :

$$\begin{aligned} P(\Theta = 3|X_1 = 2, X_2 = 4) &= 1 - P(\Theta = 2|X_1 = 2, X_2 = 4) \\ &= 0.755\,09 \end{aligned}$$

Ainsi, ayant observé $x_1 = 2$ *et* $x_2 = 4$, *la probabilité a postériori de* $\Theta = 2$ *est inférieure à la probabilité a priori de* $\Theta = 2$ *(elle a baissé de* 0.33 *à* 0.25*). On en conclut que les observations* $x_1 = 2$ *et* $x_2 = 4$ *semblent plus en faveur de* $\Theta = 3$ *que de* $\Theta = 2$. *Ceci est conforme avec notre intuition que* $\overline{x} = 3$.

Exemple 2 *Reprenons l'exemple de Rice vu dans un chapitre antérieur. Soit donné une pièce de monnaie et on veut déterminer* $p = P(Pile)$. *On lance la pièce* n *fois et on examine la variable* $X =$ *nombre de piles obtenus. Que nous apprend cette expérience ? Posons*

$$\Theta = P(Pile)$$

On résume notre connaissance sur Θ *avant l'expérience par une densité concentrée sur* $[0;1]$ *(on parle de randomisation) : c'est la densité a priori. Si on ne sait rien sur* Θ, *on peut choisir* $\Theta \sim U[0;1]$ *(distribution plate : flat) :*

$$h(\theta) = f_\Theta(\theta) = 1 \text{ si } 0 \leq \theta \leq 1$$

Maintenant, on observe Θ pour obtenir une distribution a posteriori. Sachant θ, on a (loi conditionnelle) $X \mid \theta \sim Bin\,(n, p = \theta)$:

$$f_{X|\Theta}\,(x \mid \theta) = \binom{n}{x}\theta^x\,(1-\theta)^{n-x}\,, \quad x = 0, 1, \ldots, n.$$

La distribution conjointe de X et Θ est alors

$$\begin{aligned} f_{(X,\Theta)}\,(\theta, x) &= f_{X|\Theta}\,(x \mid \theta)\, f_\Theta\,(\theta) \\ &= \binom{n}{x}\theta^x\,(1-\theta)^{n-x} \\ &= \frac{n!}{x!\,(n-x)!}\theta^x\,(1-\theta)^{n-x} \\ &= \frac{\Gamma\,(n+1)}{\Gamma\,(x+1)\,\Gamma\,(n+1-x)}\theta^x\,(1-\theta)^{n-x}\,, \qquad x = 0, 1, \ldots, n \end{aligned}$$

Intégrons par rapport à θ :

$$\begin{aligned} f_X\,(x) &= \frac{\Gamma\,(n+1)}{\Gamma\,(x+1)\,\Gamma\,(n+1-x)}\int_0^1 \theta^x\,(1-\theta)^{n-x}\,d\theta \\ &= \frac{\Gamma\,(n+1)}{\Gamma\,(x+1)\,\Gamma\,(n+1-x)}\frac{\Gamma\,(x+1)\,\Gamma\,(n+1-x)}{\Gamma\,(n+2)} \\ &= \frac{1}{n+1}, \qquad x = 0, 1, \ldots, n. \end{aligned}$$

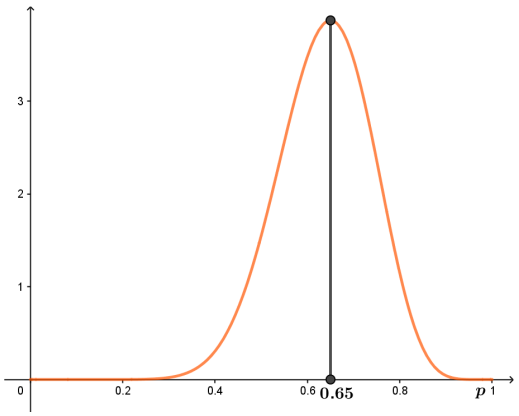
On a utilisé les propriétés de la loi Béta pour aboutir à ce résultat. Ainsi, si la loi de Θ est uniforme, celle de X l'est aussi (discrète). Maintenant, si on observe $X = x$,

$$f_{\Theta|X}\,(\theta \mid x) = \frac{f_{(X,\Theta)}\,(\theta, x)}{f_{\mathbb{X}}\,(x)} = \frac{\Gamma\,(n+2)}{\Gamma\,(x+1)\,\Gamma\,(n+1-x)}\theta^x\,(1-\theta)^{n-x}$$

après un petit calcul. On a donc

$$\Theta \mid X \sim Beta\,(a = x+1, b = n+1-x)$$

Ainsi, si par exemple on a obtenu $x = 13$ piles sur $n = 20$ jets, le maximum de la densité est donné pour la valeur approximative $p \approxeq 0.65$.



Rappel

$$\int_0^1 x^{a-1}\left(1-x\right)^{b-1}dx = B\left(a,b\right) = \frac{\Gamma\left(a\right)\Gamma\left(b\right)}{\Gamma\left(a+b\right)}, \quad a,b>0.$$

4 Loi a priori et loi a posteriori

Étapes

1. On choisit une densité $h(\theta) = f_\Theta(\theta)$ (densité a priori qui exprime l'état de notre croyance (ou connaissance) sur un paramètre θ avant qu'on ne voie les données.
2. On choisit un modèle statistique
$$f(x|\theta)$$
qui reflète notre croyance sur x sachant θ.
3. On observe les données $x_1, \cdots, x_n$ puis on met à jour notre croyance sur θ en calculant la loi a postériori
$$f(\theta|(x_1, \cdots, x_n))$$

Pour voir comment on effectue l'étape 3, supposons que Θ est discrète et qu'il y a une seule observation $X = x$. On a

$$P(\Theta = \theta|X = x) = \frac{P(\Theta = \theta, X = x)}{P(X = x)} \overset{Bayes}{=} \frac{P(X = x|\Theta = \theta) P(\Theta = \theta)}{\sum_{\theta\in\Theta} P(X = x|\Theta = \theta) P(\Theta = \theta)}$$

si X est discrète.

Dans le cas continu, on passe par les densités :

$$f(\theta|x) = \frac{f(x|\theta) h(\theta)}{\int_{\mathbb{R}} f(x|\theta) h(\theta) d\theta}$$

Si on dispose de n observations, on remplace $f(x|\theta)$ par

$$f((x_1, \cdots, x_n)|\theta) \overset{ind.}{=} \prod_{i=1}^{n} f(x_i|\theta) = L(\theta)$$

Maintenant,

$$f(\theta|(x_1, \cdots, x_n)) = \frac{f((x_1, \cdots, x_n)|\theta) h(\theta)}{\int_{\mathbb{R}} f((x_1, \cdots, x_n)|\theta) h(\theta) d\theta} = \frac{L(\theta) h(\theta)}{c_n}$$

où

$$c_n = \int_{\mathbb{R}} L(\theta) h(\theta) d\theta = \text{ constante de normalisation}$$

On a ainsi le

Théorème 1 *La loi a posteriori est proportionnelle à la vraisemblance multipliée par la loi a priori :*
$$f(\theta|(x_1, \cdots, x_n)) \propto L(\theta) h(\theta) d\theta$$

Remarque 1 *On a ignoré la constante de proportionnalité c_n. On peut la retrouver si besoin est.*

Exemple 3 *Rice (page 286). Poi* $(\theta = \lambda)$*. Le paramètre inconnu est* $\lambda > 0$*. Sa loi a priori est*

$$h(\lambda) = f_{\Lambda}(\lambda)$$

Les données sont des observations iid $X = (X_1, \cdots, X_n)$

$$f_{X_i|\Lambda}(x|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, \cdots$$

et donc leur loi conjointe sachant $\Lambda = \lambda$ *est :*

$$f_{X|\Lambda}(x|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

La loi a posteriori de Λ *sachant* $X = x = (x_1, \cdots, x_n)$ *est :*

$$f_{\Lambda|X}(\lambda|x) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i} h(\lambda)}{\int e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i} h(\lambda)\, d\lambda}$$

Ainsi, pour évaluer la distribution a postériori, il faut faire deux choses :

1. *Spécifier la loi a priori* $h(\lambda)$
2. *Calculer l'intégrale.*
   (Noter la présence de la statistique exhaustive pour λ *:* $T = \sum_{i=1}^{n} X_i$*).*

Illustrons ceci. Supposons que $\Lambda \sim Gamma(\alpha, \nu)$ *avec* α, ν *connus. Donc la densité a priori est :*

$$h(\lambda) = \frac{\nu^{\alpha}\lambda^{\alpha-1}e^{-\nu\lambda}}{\Gamma(\alpha)}, \quad \lambda > 0$$

Après calculs, la densité a postériori est

$$\begin{aligned} f_{\Lambda|X}(\lambda|x) &= \frac{\lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-(n+\nu)\lambda}}{\int_0^{\infty} \lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-(n+\nu)\lambda} d\lambda} \\ &= \frac{\lambda^{\alpha'-1}e^{-\nu'\lambda}}{c_n} \end{aligned}$$

On voit ainsi que

$$\Lambda|X \sim Gamma(\alpha', \nu')$$

et

$$c_n = \Gamma(\alpha') = \Gamma\left(\sum_{i=1}^{n} x_i + \alpha\right)$$

5 Lois a priori conjuguées et règle de Jeffreys

Comment trouver la distribution a priori $h(\theta) = f_\Theta(\theta)$? Il arrive souvent que cette loi ne puisse pas être déterminée de manière analytique (par une formule). Ceci empêche alors de trouver une loi a posteriori utilisable. Dans différente situations, cependant, on peut postuler une forme particulière pour $h(\theta)$. On obtient alors les lois a priori conjuguées. Ce postulat dit essentiellement la chose suivante.

Définition 1 *Si $h(\theta)$ appartient à une famille particulière de lois P_θ, alors on dit que $h(\theta)$ est une loi a priori conjuguée pour Θ si et seulement si la loi a posteriori $f(\theta|(x_1, \cdots, x_n))$ appartient à la même famille. On a vu des exemples ci-dessus. Ce point ne sera pas développé ici.*

Pour une une autre approche, parmi plusieurs écoles de pensée (que nous ignorons ici), Jeffreys a suggéré la règle

$$h(\theta) = f_\Theta(\theta) \propto \sqrt{I(\theta)}$$

Cette règle a l'avantage d'être invariante sous une translation.

Exemple 4 *Bernoulli $(\theta = p)$. Rappelons que*

$$I(\theta) = \frac{1}{p(1-p)}$$

La règle de Jeffreys nous dit

$$h(p) \propto \sqrt{I(\theta)} = \sqrt{\frac{1}{p(1-p)}} = p^{-1/2}(1-p)^{-1/2}$$

c-à-d une Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$ (proche de la loi uniforme).

6 Estimation (théorie de la décision)

6.1 Introduction

Une fois qu'on dispose d'une distribution a posteriori, on peut faire une estimation ponctuelle en prenant la moyenne (ou le mode) :

$$\overline{\theta} = \int \theta f\left(\theta \left| (x_1, \cdots, x_n)\right.\right) d\theta = \frac{\int \theta L\left(\theta\right) d\theta}{\int \theta L\left(\theta\right) h\left(\theta\right) d\theta}$$

On peut aussi trouver une estimation par intervalle. On cherche a et b tels que

$$\int_{-\infty}^{a} f\left(\theta \left| (x_1, \cdots, x_n)\right.\right) d\theta = \int_{b}^{+\infty} f\left(\theta \left| (x_1, \cdots, x_n)\right.\right) d\theta = \frac{\alpha}{2}$$

Posons

$$C =]a; b[$$

Alors

$$P\left(\theta \in C \left| (x_1, \cdots, x_n)\right.\right) = \int_{a}^{b} f\left(\theta \left| (x_1, \cdots, x_n)\right.\right) d\theta = 1 - \alpha$$

C'est l'intervalle a posteriori de niveau $1 - \alpha$ (on dit intervalle de crédibilité). On y revient plus bas.

Exemple 5 *Soit $X_1, \cdots, X_n$ iid $\sim N\left(\theta, \sigma^2\right)$ (σ^2 connu). Prenons $\Theta \sim N\left(a, b^2\right)$. On peut vérifier que la loi a posteriori (exercice et cf. lois a priori conjuguées ci-dessus) est*

$$\Theta \left| (X_1, \cdots, X_n) \right. \sim N\left(\overline{\theta}, \tau^2\right)$$

où

$$\overline{\theta} = w\overline{X} + (1 - w)\, a$$

et

$$w = \frac{1/\left(\sigma^2/n\right)}{1/\left(\sigma^2/n\right) + 1/b^2}$$

et

$$\frac{1}{\tau^2} = \frac{1}{\sigma^2/n} + \frac{1}{b^2}$$

Notons que

$$\lim_{n\to\infty} w = 1 \quad et \quad \lim_{n\to\infty} \frac{\tau}{\sigma/\sqrt{n}} = 1$$

Ainsi, pour n assez grand,

$$\Theta \underset{approx.}{\sim} N\left(\widehat{\theta}_{EMV}, \sigma^2/n\right)$$

On a le

Théorème 2 *(Grands échantillons). Soit $\widehat{\theta} = \widehat{\theta}_{EMV}$. Sous certaines conditions de régularité, la densité a posteriori est approximativenment*

$$N\left(\widehat{\theta}, 1/\left(nI\left(\widehat{\theta}\right)\right)\right)$$

Donc

$$\overline{\theta} \approxeq \widehat{\theta}$$

De plus,

$$IC_{1-\alpha}(\theta) \approxeq \left]\widehat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{nI\left(\widehat{\theta}\right)}}; \widehat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{nI\left(\widehat{\theta}\right)}}\right[$$

6.2 Fonction perte

Revoyons le problème de l'estimation ponctuelle sous un autre point de vue (décision).

Soit $X = (X_1, \cdots, X_n)$ un échantillon aléatoire issu d'une loi $f(x|\theta)$, $\theta \in \Theta$ (espace des paramètres). Soit $T = T(X_1, \cdots, X_n)$ une statistique que nous désirons utiliser pour estimer θ. Soit $d(t)$ la fonction de la valeur observée de T (calculée à partir de $T(x_1, \cdots, x_n)$) donnant l'estimation de θ. La fonction d décide quelle est la valeur de notre estimation ponctuelle de θ. On dit que d est une fonction de décision et que $d(t)$ est une décision.

Maintenant, une décision peut être correcte ou incorrecte. On veut introduire une mesure de la différence entre le vrai θ (inconnu) et l'estimation ponctuelle $d(t)$. Ainsi, à chaque couple $(\theta, d(t))$ on associe un nombre non négatif

$$\mathcal{L}(\theta, d(t)) = \text{ fonction perte}$$

(loss function). L'espérance de la fonction perte s'appelle la fonction risque. Soit

$$f_{T|\Theta}(t|\theta), \quad \theta \in \Theta$$

la loi conditionnelle de T. Le risque est (dans le cas continu)

$$R(\theta, d) = E(\mathcal{L}(\theta, d(T))) = \int_{\mathbb{R}} \mathcal{L}(\theta, d(t)) f_{T|\Theta}(t|\theta)\, dt$$

Il est important de choisir une fonction de décision d qui minimise le risque $R(\theta, d)$, $\forall \theta \in \Theta$.

Ceci est généralement impossible car une décision d qui minimise $R(\theta, d)$ pour une valeur particulière de θ peut ne pas le minimiser pour une autre valeur.

On est alors amené soit à restreindre nos fonctions de décision à une certaine famille ou trouver des méthodes pour ordonner les fonctions de risque.

Illustrons ces difficultés sur un exemple.

Exemple 6 *Soit $X_1, \cdots, X_n$ iid $\sim N(\theta, 1)$ avec $n = 25$. Prenons $T = \overline{X}$ et*

$$\mathcal{L}(\theta, d(t)) = (\theta - d(t))^2$$

Nous allons comparer les deux fonctions de décision

$$d_1(t) = t \quad et \quad d_2(t) = 0, \quad \forall t \in \mathbb{R}$$

Les fonctions de risque sont alors

$$R(\theta, d_1(t)) = E\left((\theta - T)^2\right) = \frac{1}{25} \ (= VAR(T))$$

et

$$R(\theta, d_2(t)) = E\left((\theta - 0)^2\right) = \theta^2$$

Si la vraie valeur de θ *est égale à* 0*, alors*

$$d_2(t) = 0$$

est une bonne décision et on a

$$R(0, d_2(t)) = 0$$

Mais si la vraie valeur de θ *est différente de* 0 *(loin de* 0*), alors*

$$d_2(t) = 0$$

est une mauvaise décision. Par exemple, si $\theta = 2$*,*

$$R(2, d_2(t)) = 2^2 = 4 > R(2, d_1(t)) = \frac{1}{25}$$

On voit qu'en général

$$R(\theta, d_2(t)) < R(\theta, d_1(t)) \ \text{ si } \theta \in \left]-\frac{1}{5}; \frac{1}{5}\right[$$

Sinon, c'est le contraire.

Si on avait restreint notre choix uniquement aux fonctions d telles que

$$E(d(T)) = \theta \quad \forall \theta \in \Theta$$

alors la décision $d_2(t) = 0$ n'est plus autorisée.

Remarque 2 *Avec cette restriction,* $T = T(X_1, \cdots, X_n)$*, la fonction risque, avec le* $\mathcal{L}$ *choisi, est l'estimateur sans biais de variance minimum (UMVUE). Nous savons (cf. chapitre 9) que* $\overline{X} = T = d(T)$ *est la solution.*

Supposons maintenant qu'on ne veut pas se restreindre au fonctions de décision d vérifiant

$$E(d(T)) = \theta \quad \forall \theta \in \Theta$$

Cherchons plutôt une fonction de décision qui minimise le maximum de la fonction risque. Dans l'exemple ci-dessus,

$$R(\theta, d_2(t)) = \theta^2$$

est non bornée. Ceci entraîne que $d_2(t) = 0$ n'est pas un bon choix selon ce critère, alors que

$$\max_{\theta \in \mathbb{R}} R(\theta, d_1(t)) = \max_{\theta \in \mathbb{R}} \frac{1}{25} = \frac{1}{25}$$

est une bonne décision car $\frac{1}{25}$ est petit.

Remarque 3 *Il peut être montré que* d_1 *est le meilleur choix selon le critère du minimax avec la fonction de perte*

$$\mathcal{L}(\theta, d(t)) = (\theta - d(t))^2$$

6.3 Estimation ponctuelle bayésienne

Si on veut trouver un estimateur ponctuel de θ (du point de vue bayésien), cela revient à choisir une fonction de décision d de sorte que $d(x) = d((x_1, \cdots, x_n))$ est une valeur prédite de θ (et donc une observation de Θ) quand la valeur calculée $x = (x_1, \cdots, x_n)$ et la loi conditionnelle (a priori) $f_{\Theta|(X_1,\cdots,X_n)}(\theta|x)$ sont connues.

Comment, en pratique, prédire une valeur expérimentale d'une variable aléatoire W si on veut que la prédiction soit "proche" de la valeur à observer ? Comme indiqué plus haut, beaucoup de statisticiens choisissent la moyenne (certains choisissent le mode ou la médiane) de W. Il est raisonnable de faire un choix de décision qui dépend de la fonction perte $\mathcal{L}(\theta, d(x))$. Une façon de construire cette dépendance est de choisir d de façon que l'espérance conditionnelle de la perte soit minimale.

Une estimation bayésienne est une fonction de décision d qui minimise

$$E(\mathcal{L}(\Theta, d(x))|X = x) = \int_{\mathbb{R}} \mathcal{L}(\theta, d(x)) f_{\Theta|X}(\theta|x) d\theta \text{ (cas continu)}$$

La variable aléatoire associée $d(x)$ s'appelle un estimateur bayésien de θ (on ajuste pour le cas discret).

Si la fonction perte est donnée par

$$\mathcal{L}(\theta, d(x)) = (\theta - d(x))^2$$

alors l'estimation bayésienne est donnée par

$$d(x) = E(\Theta|x)$$

C'est l'espérance de la loi conditionnelle de $\Theta|X = x$ car on sait (cf. cours de proba) que

$$E\left((W - b)^2\right)$$

est minimale si $b = E(W)$ (si cette quantité existe).

Si la fonction perte est

$$\mathcal{L}(\theta, d(x)) = |\theta - d(x)|$$

alors la médiane de la loi conditionnelle de $\Theta|X = x$ est la solution bayésienne.

Ces développements se généralisent à une fonction $g(\theta)$.

Exemple 7 *Considérons le modèle $X_i|\Theta = \theta$ iid $\sim Bin(1, \theta)$ avec $\Theta \sim Beta(a, b)$, a, b connus (> 0). La distribution a priori est donc*

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in]0; 1[$$

On cherche une fonction de décision bayésienne d. La statistique

$$T = \sum_{i=1}^{n} X_i \sim Bin(n, \theta)$$

est exhaustive. On a vu (exemples introductifs) que la loi conditionnelle de $T|\Theta=\theta$ *est*

$$f_{T|\Theta=\theta}(t|\theta)=\binom{n}{t}\theta^t(1-\theta)^{n-t},\quad t=0,1,\cdots,n$$

et elle est donc proportionnelle à

$$\theta^t(1-\theta)^{n-t}\theta^{a-1}(1-\theta)^{b-1}$$

Plus précisément, on a vu que c'est une loi $Beta(a+t,b+n-t)$. *Avec la fonction perte*

$$\mathcal{L}(\theta,d(t))=(\theta-d(t))^2$$

le calcul donne

$$d(t)=\frac{a+t}{a+b+n}\quad(*)$$

Remarque. Réécrivons (*) comme suit :

$$d(t)=\frac{n}{a+b+n}\frac{t}{n}+\frac{a+b}{a+b+n}\frac{a}{a+b}$$

= moyenne pondérée de l'EMV $\frac{T}{n}$ *de* θ. *Si* $n\nearrow$, *alors (estimation asymptotique)*

$$d(T)\approxeq T$$

De plus, $d(T)$ *est convergent. Donc* a *et* b *devraient être choisis tels que, évidemment,*

$$\frac{a}{a+b}$$

est la moyenne a priori, mais $a+b$ *devrait indiquer la valeur de la croyance a priori par rapport à* n. *Ainsi, si on veut que notre croyance a priori ait autant de valeur que la taille de l'échantillon (par ex.* $n=20$*), on devrait prendre* $a+b=20$. *Donc si notre moyenne a priori est* $3/4$, *alors* $a=15$ *et* $b=5$.

Exemple 8 *Considérons le modèle normal* $X_i|\Theta=\theta$ *iid* $\sim N(\theta,\sigma^2)$, σ^2 *connu avec* $\Theta\sim N(\theta_0,\sigma_0^2)$ *(*θ_0,σ_0^2*, connus). Alors* $T=\overline{X}$ *est une statistique exhaustive et une formulation équivalente du modèle est*

$$\begin{aligned}T|\Theta=\theta&\sim N(\theta,\sigma^2/n)\\ \Theta&\sim N(\theta_0,\sigma_0^2)\end{aligned}$$

La distribution à posteriori vérifie

$$f_{T|\Theta}(\theta|t)\propto\exp\left[-\frac{(\sigma_0^2+\sigma^2/n)\theta^2-2(t\sigma_0^2+\theta_0/(\sigma^2/n))\theta}{2(\sigma^2/n)\sigma_0^2}\right]$$

Calculs :

$$\Theta|T\sim N\left(\frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}t+\frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\theta_0,\frac{(\sigma^2/n)\sigma_0^2}{\sigma_0^2+\sigma^2/n}\right)$$

On a encore une moyenne pondérée (en utilisant la même fonction perte quadratique).

Si $n\nearrow$, *on tend de l'EMV* $\overline{X}$.