

STA305 : Partie I

Approche bayésienne

Boris Hejblum

*ISPED M2 Biostatistique, Université de Bordeaux
Inserm BPH U1219 / Inria BSO, équipe SISTM*

boris.hejblum@u-bordeaux.fr
<https://borishejblum.science>

1^{er} décembre 2020



Vocabulaire Bayésien

- **paradigme**
- ***a priori***
- ***a posteriori***
- **élicitation**

Objectifs du cours

Se familiariser avec l'approche **bayésienne** :

- 1 être capable de proposer une modélisation bayésienne adéquate face à un problème concret
- 2 savoir calculer la distribution *a posteriori* dans le cas de relations de conjugaison
- 3 comprendre l'impact de la loi *a priori* et la notion d'*a priori* faiblement-informatif
- 4 comprendre la notion de MAP et de moyenne *a posteriori*, d'intervalle de crédibilité ainsi que la différence avec un intervalle de confiance
- 5 comprendre les notions de risques et de coûts, et leurs implications dans la théorie de la décision

Objectifs du cours

Se familiariser avec l'approche **bayésienne** :

- 1 être capable de proposer une modélisation bayésienne adéquate face à un problème concret
- 2 savoir calculer la distribution *a posteriori* dans le cas de relations de conjugaison
- 3 comprendre l'impact de la loi *a priori* et la notion d'*a priori* faiblement-informatif
- 4 comprendre la notion de MAP et de moyenne *a posteriori*, d'intervalle de crédibilité ainsi que la différence avec un intervalle de confiance
- 5 comprendre les notions de risques et de coûts, et leurs implications dans la théorie de la décision

NB : ces notes ne se veulent en aucun cas exhaustives, et l'on renverra le lecteur curieux aux ouvrages bien plus complets que sont *Le choix bayésien* de C. Robert et *Le raisonnement bayésien* de E. Parent & J. Bernier..

Introduction

La statistique :

- Une science **mathématique**
- **décrire** ce qui s'est produit
- faire des **projections** quant à ce qu'il **peut** advenir dans le **futur**
- s'appuie sur l'**observation** de phénomènes naturels pour en proposer une interprétation, souvent à travers des **modèles probabilistes**

Théorème de Bayes

Article posthume du révérend Thomas Bayes en 1763

$$\Pr(A|E) = \frac{\Pr(E|A) \Pr(A)}{\Pr(E|A) \Pr(A) + \Pr(E|\bar{A}) \Pr(\bar{A})} = \frac{\Pr(E|A) \Pr(A)}{\Pr(E)}$$



Exemple pratique :

La dernière fois que vous êtes allez chez le docteur, on vous a fait un **test pour une maladie rare**. Malheureusement, le résultat était positif. . .

Sachant le résultat du test, quelle la probabilité que je suis réellement malade ?

(les tests médicaux n'étant, après tout, pas parfaits.)

→ *Seeing Theory*, Brown University

<http://students.brown.edu/seeing-theory/bayesian-inference/index.html#section1>

Théorème de Bayes : exercice

1% de la population est affecté par une maladie rare. Un test médical pour cette maladie possède les propriétés suivantes :

- si quelqu'un a cette maladie, son test sera positif dans 99% des cas
- si quelqu'un n'a pas cette maladie, son test sera négatif dans 95% des cas

Sachant que quelqu'un a eu un résultat positif au test, quelle est la probabilité qu'il ait la maladie ?

$$\Pr(M = +) = 0.01 \quad \Pr(T = + | M = +) = 0.99 \quad \Pr(T = - | M = -) = 0.95$$

$$\Pr(M = + | T = +) = ?$$

Théorème de Bayes : exercice

1% de la population est affecté par une maladie rare. Un test médical pour cette maladie possède les propriétés suivantes :

- si quelqu'un a cette maladie, son test sera positif dans 99% des cas
- si quelqu'un n'a pas cette maladie, son test sera négatif dans 95% des cas

Sachant que quelqu'un a eu un résultat positif au test, quelle est la probabilité qu'il ait la maladie ?

$$\Pr(M = +) = 0.01 \quad \Pr(T = + | M = +) = 0.99 \quad \Pr(T = - | M = -) = 0.95$$

$$\begin{aligned}\Pr(M = + | T = +) &= \frac{\Pr(T = + | M = +) \Pr(M = +)}{\Pr(T = +)} \\ &= \frac{\Pr(T = + | M = +) \Pr(M = +)}{\Pr(T = + | M = +) \Pr(M = +) + \Pr(T = + | M = -) \Pr(M = -)} \\ &= \frac{\Pr(T = + | M = +) \Pr(M = +)}{\Pr(T = + | M = +) \Pr(M = +) + (1 - \Pr(T = - | M = -)) (1 - \Pr(M = +))} \\ &= 0.17\end{aligned}$$

Philosophie bayésienne

Les **Paramètres sont des variables aléatoires !** – pas de "vraie" valeur

⇒ induit une distribution de probabilité marginale $\pi(\theta)$ sur les paramètres :
la distribution *a priori*

😊 permet de **formaliser** les hypothèses dans la modélisation

☹️ introduit nécessairement de la **subjectivité** dans l'analyse

Bayésiens vs. Fréquentistes : point historique

- ① **Bayes + Laplace** ⇒ développement de la Statistique aux **XVIII-XIX^e siècles**
- ② Galton & Pearson, puis Fisher & Neymann ⇒ théorie **fréquentiste** devenue dominante au cours du **XX^e siècle**
- ③ au tournant du **XXI^e siècle** : avènement de l'ordinateur moderne
⇒ **comeback du bayésien**



Bayésiens vs. Fréquentistes : un débat dépassé

Fisher rejettait fermement le raisonnement bayésien

⇒ communauté divisée en 2 au XX^e siècle

Bayésiens vs. Fréquentistes : un débat dépassé

Fisher rejettait fermement le raisonnement bayésien

⇒ communauté divisée en 2 au XX^e siècle

Être ou ne pas être bayésien, là n'est plus la question : il s'agit d'utiliser à bon escient les outils adaptés quand cela est nécessaire

Gilbert Saporta

Modélisation bayésienne

Rappel sur la modélisation fréquentiste

- une suite de variables aléatoires *iid* (indépendantes et identiquement distribuées) $\mathbf{Y} = (Y_1, \dots, Y_n)$

Rappel sur la modélisation fréquentiste

- une suite de variables aléatoires *iid* (indépendantes et identiquement distribuées) $\mathbf{Y} = (Y_1, \dots, Y_n)$
- on observe un échantillon $\mathbf{y} = (y_1, \dots, y_n)$

Rappel sur la modélisation fréquentiste

- une suite de variables aléatoires *iid* (indépendantes et identiquement distribuées) $\mathbf{Y} = (Y_1, \dots, Y_n)$
- on observe un échantillon $\mathbf{y} = (y_1, \dots, y_n)$
- on modélise leur distribution de probabilité comme $f(y|\theta)$, $\theta \in \Theta$

Rappel sur la modélisation fréquentiste

- une suite de variables aléatoires *iid* (indépendantes et identiquement distribuées) $\mathbf{Y} = (Y_1, \dots, Y_n)$
- on observe un échantillon $\mathbf{y} = (y_1, \dots, y_n)$
- on modélise leur distribution de probabilité comme $f(y|\theta)$, $\theta \in \Theta$

Ce modèle suppose qu'il existe une « vraie » distribution de Y caractérisée par la « vraie » valeur du paramètre θ^*

$$\hat{\theta} ?$$

Présentation de l'application historique

Laplace

Quelle est la probabilité de naissance d'une fille (plutôt que d'un garçon).

⇒ **observations** : naissances observées à Paris entre 1745 et 1770
(241,945 filles & 251,527 garçons)

Quand un enfant naît, est-il plus probable que ce soit une fille
plutôt qu'un garçon ?

Construction d'un modèle bayésien

Trois composants

① La question

② Le modèle d'échantillonage

③ La distribution *a priori*

Construction d'un modèle bayésien

Trois composants

① La question

La première étape dans la construction d'un modèle est toujours d'identifier la question à laquelle on souhaite répondre

② Le modèle d'échantillonage

③ La distribution *a priori*

Trois composants

1 La question

La première étape dans la construction d'un modèle est toujours d'identifier la question à laquelle on souhaite répondre

2 Le modèle d'échantillonage

Quelles **observations** sont disponibles pour répondre à cette question ?
Comment peuvent-elles être **décrivées** ?

3 La distribution *a priori*

Trois composants

① La question

La première étape dans la construction d'un modèle est toujours d'identifier la question à laquelle on souhaite répondre

② Le modèle d'échantillonage

Quelles **observations** sont disponibles pour répondre à cette question ?
Comment peuvent-elles être **décrivées** ?

③ La distribution *a priori*

Une distribution de probabilité sur les paramètres θ du modèle d'échantillonage

Le modèle d'échantillonnage

y : les observations disponibles

⇒ **modèle probabiliste (paramétrique) génératif** :

$$Y_i \stackrel{iid}{\sim} f(y|\theta)$$

La distribution *a priori*

Dans la modélisation bayésienne, par rapport à la modélisation fréquentiste, on ajoute une loi de probabilité sur les paramètres θ :

$$\theta \sim \pi(\theta)$$

$$Y_i|\theta \stackrel{iid}{\sim} f(y|\theta)$$

θ sera ainsi traité comme une variable aléatoire,
mais qui n'est jamais observée !

Retour à l'exemple historique de Laplace

① La question

② Modèle d'échantillonnage

③ Distribution *a priori*

Construction d'un modèle bayésien

Retour à l'exemple historique de Laplace

① La question

...

② Modèle d'échantillonnage

...

③ Distribution *a priori*

...

Retour à l'exemple historique de Laplace

① La question

Quand un enfant naît, est-il plus probable que ce soit une fille plutôt qu'un garçon ?

② Modèle d'échantillonnage

...

③ Distribution *a priori*

...

Retour à l'exemple historique de Laplace

① La question

Quand un enfant naît, est-il plus probable que ce soit une fille plutôt qu'un garçon ?

② Modèle d'échantillonnage

Distribution de Bernoulli : $Y_i = 1$ si le nouveau né i est une fille, 0 si c'est un garçon

$$Y_i \sim \text{Bernoulli}(\theta) \quad \theta \in [0, 1]$$

③ Distribution *a priori*

...

Retour à l'exemple historique de Laplace

① La question

Quand un enfant naît, est-il plus probable que ce soit une fille plutôt qu'un garçon ?

② Modèle d'échantillonnage

Distribution de Bernoulli : $Y_i = 1$ si le nouveau né i est une fille, 0 si c'est un garçon

$$Y_i \sim \text{Bernoulli}(\theta) \quad \theta \in [0, 1]$$

③ Distribution *a priori*

Un *a priori* uniforme sur θ (la probabilité qu'un nouveau né soit une fille plutôt qu'un garçon) :

$$\theta \sim \mathcal{U}_{[0,1]}$$

Distribution *a posteriori*

L'objet de la modélisation bayésienne : **inférer la distribution *a posteriori* des paramètres**

- **Loi *a posteriori*** : la loi de θ conditionnellement aux observations $p(\theta|Y)$

Distribution *a posteriori*

L'objet de la modélisation bayésienne : **inférer la distribution *a posteriori* des paramètres**

- **Loi *a posteriori*** : la loi de θ conditionnellement aux observations $p(\theta|Y)$

Théorème de Bayes :

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

où $f(y) = \int_{\Theta} f(y|\theta)\pi(\theta) d\theta$ est la loi marginale des données

Distribution *a posteriori*

L'objet de la modélisation bayésienne : **inférer la distribution *a posteriori* des paramètres**

- **Loi *a posteriori*** : la loi de θ conditionnellement aux observations $p(\theta|Y)$

Théorème de Bayes :

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

où $f(y) = \int_{\Theta} f(y|\theta)\pi(\theta) d\theta$ est la loi marginale des données

La distribution *a posteriori* est calculée à partir :

- ① du modèle d'échantillonnage $f(y|\theta)$ – qui donne la vraisemblance $f(y|\theta)$ pour l'ensemble des observations
- ② de la loi *a priori* $\pi(\theta)$

Application à l'exemple historique

① La vraisemblance

② La loi *a priori*

③ La distribution *a posteriori*

Application à l'exemple historique

① La vraisemblance

...

② La loi *a priori*

...

③ La distribution *a posteriori*

...

Application à l'exemple historique

① La vraisemblance

$$f(\mathbf{y}|\theta) = \prod_{i=1}^n \theta^{y_i} (1-\theta)^{(1-y_i)} = \theta^S (1-\theta)^{n-S} \quad \text{où } S = \sum_{i=1}^n y_i$$

② La loi *a priori*

...

③ La distribution *a posteriori*

...

Application à l'exemple historique

① La vraisemblance

$$f(\mathbf{y}|\theta) = \prod_{i=1}^n \theta^{y_i} (1-\theta)^{(1-y_i)} = \theta^S (1-\theta)^{n-S} \quad \text{où } S = \sum_{i=1}^n y_i$$

② La loi *a priori*

Uniforme : $\pi(\theta) = 1$

③ La distribution *a posteriori*

...

Application à l'exemple historique

① La vraisemblance

$$f(\mathbf{y}|\theta) = \prod_{i=1}^n \theta^{y_i} (1-\theta)^{(1-y_i)} = \theta^S (1-\theta)^{n-S} \quad \text{où } S = \sum_{i=1}^n y_i$$

② La loi *a priori*

Uniforme : $\pi(\theta) = 1$

③ La distribution *a posteriori*

$$p(\theta|\mathbf{y}) = \frac{\theta^S (1-\theta)^{n-S}}{f(\mathbf{y})} = p(\theta|\mathbf{y}) = \binom{n}{S} (n+1) \theta^S (1-\theta)^{n-S}$$

Application à l'exemple historique

① La vraisemblance

$$f(\mathbf{y}|\theta) = \prod_{i=1}^n \theta^{y_i} (1-\theta)^{(1-y_i)} = \theta^S (1-\theta)^{n-S} \quad \text{où } S = \sum_{i=1}^n y_i$$

② La loi *a priori*

Uniforme : $\pi(\theta) = 1$

③ La distribution *a posteriori*

$$p(\theta|\mathbf{y}) = \frac{\theta^S (1-\theta)^{n-S}}{f(\mathbf{y})} = p(\theta|\mathbf{y}) = \binom{n}{S} (n+1) \theta^S (1-\theta)^{n-S}$$

Pour répondre à notre question d'intérêt, on peut alors calculer : ...

Application à l'exemple historique

① La vraisemblance

$$f(\mathbf{y}|\theta) = \prod_{i=1}^n \theta^{y_i} (1-\theta)^{(1-y_i)} = \theta^S (1-\theta)^{n-S} \quad \text{où } S = \sum_{i=1}^n y_i$$

② La loi *a priori*

Uniforme : $\pi(\theta) = 1$

③ La distribution *a posteriori*

$$p(\theta|\mathbf{y}) = \frac{\theta^S (1-\theta)^{n-S}}{f(\mathbf{y})} = p(\theta|\mathbf{y}) = \binom{n}{S} (n+1) \theta^S (1-\theta)^{n-S}$$

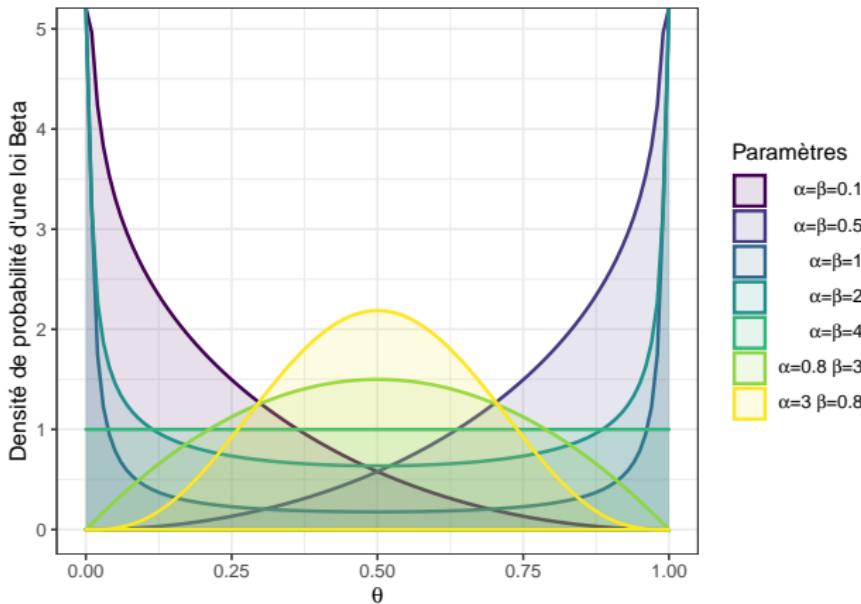
Pour répondre à notre question d'intérêt, on peut alors calculer :

$$P(\theta \geq 0.5|\mathbf{y}) = \int_{0.5}^1 p(\theta|\mathbf{y}) = \binom{n}{S} (n+1) \int_{0.5}^1 \theta^S (1-\theta)^{n-S} d\theta \approx 1.15 \cdot 10^{-42}$$

Construction d'un modèle bayésien

La distribution Beta

$$f(\theta) = \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} \theta^{\alpha-1} (1-\theta)^{\beta-1} \text{ pour } \alpha > 0 \text{ et } \beta > 0$$



Exemples de paramétrisations pour la distribution Beta

Introduction à la l'approche bayésienne
oooooooo

Construction d'un modèle bayésien

Modélisation bayésienne
oooooooo●oooooooooooo

Inférence bayésienne
oooooooooooo

Conclusion
oo

Conjugaison de la distribution Beta

***a priori* Beta** : $\pi = \text{Beta}(\alpha, \beta)$

Loi *a posteriori* associée : ...

Construction d'un modèle bayésien

Conjugaison de la distribution Beta

a priori Beta : $\pi = \text{Beta}(\alpha, \beta)$

Loi a posteriori associée : $p(\theta|y) \propto \theta^{\alpha+S-1} (1-\theta)^{\beta+(n-S)-1}$

...

Le signe \propto signifie « proportionnel à »

Conjugaison de la distribution Beta

a priori Beta : $\pi = \text{Beta}(\alpha, \beta)$

Loi a posteriori associée : $p(\theta|y) \propto \theta^{\alpha+S-1} (1-\theta)^{\beta+(n-S)-1}$

$\Rightarrow \theta|y \sim \text{Beta}(\alpha + S, \beta + (n - S))$

Le signe \propto signifie « proportionnel à »

Conjugaison de la distribution Beta

a priori Beta : $\pi = \text{Beta}(\alpha, \beta)$

Loi a posteriori associée : $p(\theta|y) \propto \theta^{\alpha+S-1} (1-\theta)^{\beta+(n-S)-1}$

$\Rightarrow \theta|y \sim \text{Beta}(\alpha + S, \beta + (n - S))$

On parle alors de **distributions conjuguées** car les distributions **a posteriori** et **a priori** appartiennent à la **même famille paramétrique**

Le signe \propto signifie « proportionnel à »

Construction d'un modèle bayésien

Impact du choix de l'*a priori*

Interprétation de l' <i>a priori</i>	Paramètres de la distribution Beta	$P(\theta \geq 0,5 \mathbf{y})$
#garçons > #filles	$\alpha = 0,1; \beta = 3$	$1,08 \cdot 10^{-42}$
#garçons < #filles	$\alpha = 3; \beta = 0,1$	$1,19 \cdot 10^{-42}$
#garçons = #filles	$\alpha = 4; \beta = 4$	$1,15 \cdot 10^{-42}$
#garçons ≠ #filles	$\alpha = 0,1; \beta = 0,1$	$1,15 \cdot 10^{-42}$
non-informatif	$\alpha = 1; \beta = 1$	$1,15 \cdot 10^{-42}$

Pour 493 472 nouveaux-nés dont 241 945 filles

Construction d'un modèle bayésien

Impact du choix de l'*a priori*

Interprétation de l' <i>a priori</i>	Paramètres de la distribution Beta	$P(\theta \geq 0,5 \mathbf{y})$
#garçons > #filles	$\alpha = 0,1; \beta = 3$	$1,08 \cdot 10^{-42}$
#garçons < #filles	$\alpha = 3; \beta = 0,1$	$1,19 \cdot 10^{-42}$
#garçons = #filles	$\alpha = 4; \beta = 4$	$1,15 \cdot 10^{-42}$
#garçons \neq #filles	$\alpha = 0,1; \beta = 0,1$	$1,15 \cdot 10^{-42}$
non-informatif	$\alpha = 1; \beta = 1$	$1,15 \cdot 10^{-42}$

Pour 493 472 nouveaux-nés dont 241 945 filles

Interprétation de l' <i>a priori</i>	Paramètres de la distribution Beta	$P(\theta \geq 0,5 \mathbf{y})$
#garçons > #filles	$\alpha = 0,1, \beta = 3$	0.39
#garçons < #filles	$\alpha = 3, \beta = 0,1$	0.52
#garçons = #filles	$\alpha = 4, \beta = 4$	0.46
#garçons \neq #filles	$\alpha = 0,1, \beta = 0,1$	0.45
non-informatif	$\alpha = 1, \beta = 1$	0.45

Pour 20 nouveaux-nés dont 9 filles

Construction d'un modèle bayésien

Impact de différant *a priori* Beta pour 20 naissances observées

L'épineux choix de la distribution *a priori**a priori* : pour & contre

Avoir une distribution *a priori* :

😊 donne de la **flexibilité**

😊 permet d'incorporer de la **connaissance extérieure**

😢 ajoute nécessairement de la **subjectivité**

⇒ le choix (élicitation) de la distribution *a priori* est un point sensible !

Propriété de la distribution *a priori*

- ① Le support de la loi *a posteriori* doit être inclus dans celui de la distribution *a priori* :
si $\pi(\theta) = 0$, alors $p(\theta|y) = 0$
- ② Les différents paramètres sont indépendants *a priori*

Élicitation de la loi *a priori*

Stratégies pour communiquer avec des experts non-statisticiens

⇒ transformer leurs **connaissances *a priori*** en **distributions *a priori***

- La **méthode des histogrammes** : demander aux experts de donner des poids à des intervalles de valeurs
⚠ peuvent donner une probabilité *a priori* nulle pour des valeurs plausibles des paramètres
- Choisir une **famille de distributions paramétriques** $p(\theta|\eta)$ en **accord avec les experts** (e.g. pour certains quantiles ou moments) – permet de résoudre le problème du support, mais l'impact de la famille paramétrique choisie est important
- Élucider les lois *a priori* à partir de la **littérature** scientifique
- ...

La quête des *a priori* non-informatifs

Parfois, on a **aucune connaissance *a priori***

Quelle loi *a priori* utiliser ?



L'épineux choix de la distribution *a priori*

La quête des *a priori* non-informatifs

Parfois, on a **aucune connaissance *a priori***

⇒ la loi Uniforme, un ***a priori* non-informatif ?**

La quête des *a priori* non-informatifs

Parfois, on a **aucune connaissance *a priori***

⇒ la loi Uniforme, un ***a priori* non-informatif** ?

2 difficultés majeures :

- ➊ **Lois impropres** $\int_{\Theta} \pi(\theta) d\theta = \infty$

La quête des *a priori* non-informatifs

Parfois, on a **aucune connaissance *a priori***

⇒ la loi Uniforme, un ***a priori* non-informatif** ?

2 difficultés majeures :

① **Lois impropres** $\int_{\Theta} \pi(\theta) d\theta = \infty$

② **Lois non-invariantes**

Non invariance de la loi uniforme : détail

Soit $F_X(x) = P(X < x)$

Non invariance de la loi uniforme : détail

Soit $F_X(x) = P(X < x)$

Si $Y = g(X)$, alors

$$F_Y(y) = P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y))$$

Non invariance de la loi uniforme : détail

Soit $F_X(x) = P(X < x)$

Si $Y = g(X)$, alors

$$F_Y(y) = P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y))$$

En dérivant par rapport à y , on obtient :

$$f_Y(y) = \frac{\partial g^{-1}(y)}{\partial y} f_X(g^{-1}(y))$$

Non invariance de la loi uniforme : détail

Soit $F_X(x) = P(X < x)$

Si $Y = g(X)$, alors

$$F_Y(y) = P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y))$$

En dérivant par rapport à y , on obtient :

$$f_Y(y) = \frac{\partial g^{-1}(y)}{\partial y} f_X(g^{-1}(y)) \neq f_X(g^{-1}(y)) = f_X(x)$$

NB : La formule s'étend au cas multidimensionnel où $|J|$ désigne le déterminant de la matrice jacobienne J (matrice des dérivées partielles)

La quête des *a priori* non-informatifs

Parfois, on a **aucune connaissance *a priori***

⇒ la loi Uniforme, un **a priori non-informatif** ?

2 difficultés majeures :

① **Lois impropres** $\int_{\Theta} \pi(\theta) d\theta = \infty$

② **Lois non-invariantes**

Autres solutions ?

La loi *a priori* de Jeffreys

⇒ Un *a priori* **faiblement informatif**, invariant par re-paramétrisation

- *a priori* unidimensionnel de Jeffreys :

$$\pi(\theta) \propto \sqrt{I(\theta)} \quad \text{où } I \text{ est la matrice d'information de Fisher}$$

- *a priori* multidimensionnel de Jeffreys :

$$\pi(\boldsymbol{\theta}) \propto \sqrt{|I(\boldsymbol{\theta})|}$$

En pratique, il est généralement plus facile (et commun) de considérer les paramètres indépendants *a priori*

L'épineux choix de la distribution a priori

La loi *a priori* de Jeffreys : application à l'exemple historique

$$f(y|\theta) = \theta^y(1-\theta)^{(1-y)}$$

L'épineux choix de la distribution a priori

La loi *a priori* de Jeffreys : application à l'exemple historique

$$f(y|\theta) = \theta^y(1-\theta)^{(1-y)}$$

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

$$\propto \sqrt{-\mathbb{E}_{Y|\theta} \left[\frac{d^2 \log(f(y|\theta))}{d\theta^2} \right]}$$

$\propto \dots$

[Rappel : $I_{(Y_1, \dots, Y_n)}(\theta) = n \times I_Y(\theta)$ si les Y_i sont iid]

L'épineux choix de la distribution a priori

La loi *a priori* de Jeffreys : application à l'exemple historique

$$f(y|\theta) = \theta^y(1-\theta)^{1-y}$$

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

$$\propto \sqrt{-\mathbb{E}_{Y|\theta} \left[\frac{d^2 \log(f(y|\theta))}{d\theta^2} \right]}$$

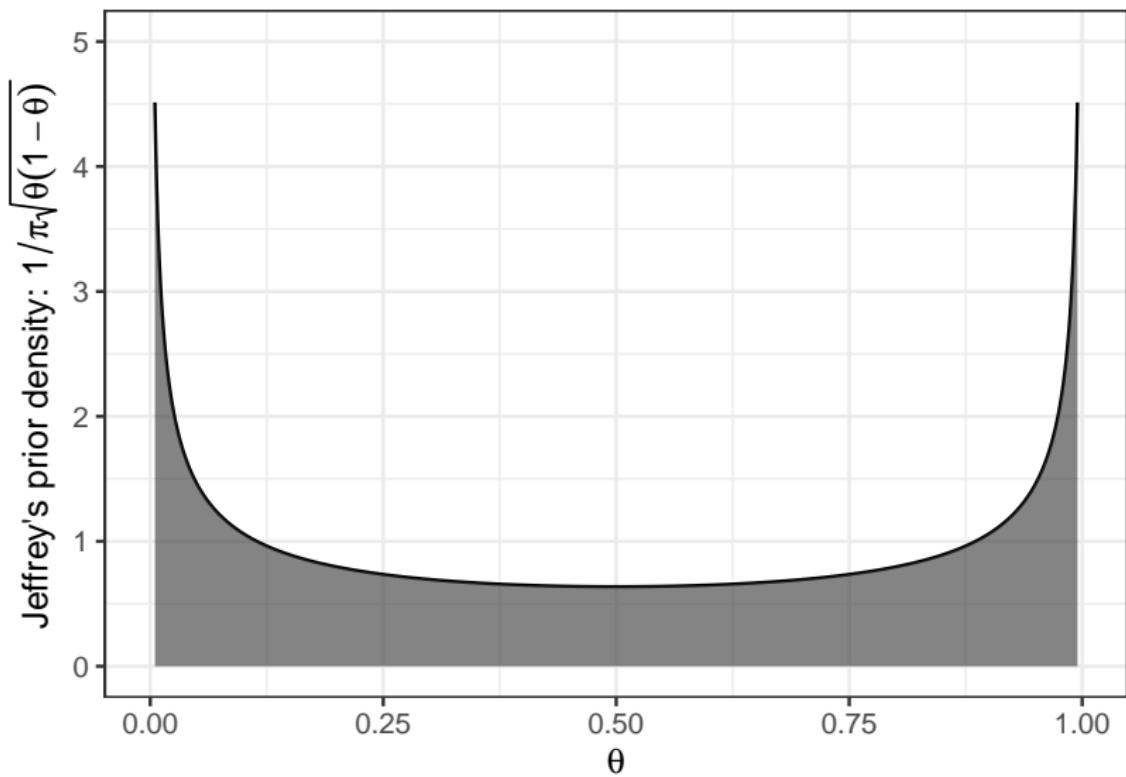
$$\propto \sqrt{\mathbb{E}_{Y|\theta} \left[\frac{y}{\theta^2} + \frac{1-y}{(1-\theta)^2} \right]}$$

$$\propto \sqrt{\theta \left(\frac{1}{\theta^2} \right) + (1-\theta) \left(\frac{1}{(1-\theta)^2} \right)}$$

$$\propto \sqrt{\frac{1}{\theta} + \frac{1}{1-\theta}}$$

$$\propto \frac{1}{\sqrt{\theta(1-\theta)}}$$

[Rappel : $I_{(Y_1, \dots, Y_n)}(\theta) = n \times I_Y(\theta)$ si les Y_i sont iid]

Loi *a priori* de Jeffreys : illustration dans l'exemple historique

Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

① $\pi(\theta)$

② $f(\mathbf{y}|\theta)$

Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

$$① \eta \sim h(\eta)$$

$$② \pi(\theta|\eta)$$

$$③ f(\mathbf{y}|\theta)$$

Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

$$① \eta \sim h(\eta)$$

$$② \pi(\theta|\eta)$$

$$③ f(\mathbf{y}|\theta)$$

$$p(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})} = \frac{\int f(\mathbf{y}|\theta,\eta)\pi(\theta|\eta)h(\eta)d\eta}{f(\mathbf{y})}$$

Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

$$① \eta \sim h(\eta)$$

$$② \pi(\theta|\eta)$$

$$③ f(\mathbf{y}|\theta)$$

$$p(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta) \pi(\theta)}{f(\mathbf{y})} = \frac{\int f(\mathbf{y}|\theta, \eta) \pi(\theta|\eta) h(\eta) d\eta}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\theta) \int \pi(\theta|\eta) h(\eta) d\eta}{f(\mathbf{y})}$$

NB : 3 niveaux hiérarchiques \Leftrightarrow 2 niveaux avec $\pi(\theta) = \int \pi(\theta|\eta) h(\eta) d\eta$

Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

$$① \eta \sim h(\eta)$$

$$② \pi(\theta|\eta)$$

$$③ f(\mathbf{y}|\theta)$$

$$p(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta) \pi(\theta)}{f(\mathbf{y})} = \frac{\int f(\mathbf{y}|\theta, \eta) \pi(\theta|\eta) h(\eta) d\eta}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\theta) \int \pi(\theta|\eta) h(\eta) d\eta}{f(\mathbf{y})}$$

NB : 3 niveaux hiérarchiques \Leftrightarrow 2 niveaux avec $\pi(\theta) = \int \pi(\theta|\eta) h(\eta) d\eta$

\Rightarrow peut faciliter la modélisation & l'élicitation des lois *a priori*

Utilisation d'hyper-priors dans l'exemple historique

Exemple historique du sexe à la naissance avec un *a priori* Beta

...

Utilisation d'hyper-priors dans l'exemple historique

Exemple historique du sexe à la naissance avec un *a priori* Beta

⇒ 2 hyper-priors Gamma pour α et β :

$$\alpha \sim \text{Gamma}(4; 0,5)$$

$$\beta \sim \text{Gamma}(4; 0,5)$$

$$\theta | \alpha, \beta \sim \text{Beta}(\alpha; \beta)$$

$$Y_i | \theta \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$$

Bayésien empirique

Élicitation de la loi *a priori* d'un paramètre d'après sa loi marginale empirique

Bayésien empirique

Élicitation de la loi *a priori* d'un paramètre d'après sa loi marginale empirique

⇒ nécessite d'estimer cet *a priori* à partir des données

Bayésien empirique

Élicitation de la loi *a priori* d'un paramètre d'après sa loi marginale empirique

⇒ nécessite d'estimer cet *a priori* à partir des données

- ① hyper-paramètres η
- ② estimés par $\hat{\eta}$ grâce à des méthodes fréquentistes
(e.g. Max de Vraisemblance ou Méthode des Moments)
- ③ injectés dans la loi *a priori* : $\pi(\theta|\hat{\eta})$
- ④ ⇒ loi *a posteriori* : $p(\theta|y, \hat{\eta})$

Bayésien empirique

Élicitation de la loi *a priori* d'un paramètre d'après sa loi marginale empirique

⇒ nécessite d'estimer cet *a priori* à partir des données

- ① hyper-paramètres η
 - ② estimés par $\hat{\eta}$ grâce à des méthodes fréquentistes
(e.g. Max de Vraisemblance ou Méthode des Moments)
 - ③ injectés dans la loi *a priori* : $\pi(\theta|\hat{\eta})$
 - ④ ⇒ loi *a posteriori* : $p(\theta|y, \hat{\eta})$
-
- Combine les approches bayésienne et fréquentiste
 - Distribution *a posteriori* concentrée (variance ↘), mais biais ↗
(données utilisées 2x !)
 - Approximation d'une approche totalement bayésienne

Bayésien empirique : exemple

Pour une loi Beta(α, β) :

- $\frac{\alpha}{\alpha+\beta}$ est la moyenne
- $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ est la variance

Par la **méthode des moments** on trouve : $\hat{\alpha}_M = 0,020$ et
 $\hat{\beta}_M = 0,021$

puisque $\hat{\theta} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 0,49$ et $\widehat{Var}(\theta) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = 0,24$

$$\theta | \alpha, \beta \sim \text{Beta}(\hat{\alpha}; \hat{\beta})$$

$$Y_i | \theta \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$$

Bayes séquentiel

Le théorème de Bayes peut être utilisé séquentiellement :

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)$$

Si $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, alors :

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}_2|\theta)f(\mathbf{y}_1|\theta)\pi(\theta) \propto f(\mathbf{y}_2|\theta)p(\theta|\mathbf{y}_1)$$

Bayes séquentiel

Le théorème de Bayes peut être utilisé séquentiellement :

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)$$

Si $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, alors :

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}_2|\theta)f(\mathbf{y}_1|\theta)\pi(\theta) \propto f(\mathbf{y}_2|\theta)p(\theta|\mathbf{y}_1)$$

Bayes séquentiel

Le théorème de Bayes peut être utilisé séquentiellement :

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)$$

Si $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, alors :

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}_2|\theta)f(\mathbf{y}_1|\theta)\pi(\theta) \propto f(\mathbf{y}_2|\theta)p(\theta|\mathbf{y}_1)$$

⇒ mise à jour de la distribution *a posteriori* au fur et à mesure qu'arrive les observations (*online*)

Bayes séquentiel : application à l'exemple historique

Imaginons que l'on commence par observer 20 naissances $y_{1:20}$ début 1745, dont 9 filles, et que l'on ait un *a priori* uniforme sur θ :

$$\theta | \mathbf{y}_{1:20} \sim \dots$$

Bayes séquentiel : application à l'exemple historique

Imaginons que l'on commence par observer 20 naissances $\mathbf{y}_{1:20}$ début 1745, dont 9 filles, et que l'on ait un *a priori* uniforme sur θ :

$$\theta | \mathbf{y}_{1:20} \sim \text{Beta}(10, 12)$$

On observe ensuite $\mathbf{y}_{21:493472}$ les 493 452 naissances restantes entre 1745 et 1770, dont 241 936 filles, et on utilise alors cet *a priori* Beta(10,12) sur θ :

$$\theta | \mathbf{y}_{1:20}, \mathbf{y}_{21:493472} \sim \dots$$

Bayes séquentiel : application à l'exemple historique

Imaginons que l'on commence par observer 20 naissances $\mathbf{y}_{1:20}$ début 1745, dont 9 filles, et que l'on ait un *a priori* uniforme sur θ :

$$\theta | \mathbf{y}_{1:20} \sim \text{Beta}(10, 12)$$

On observe ensuite $\mathbf{y}_{21:493472}$ les 493 452 naissances restantes entre 1745 et 1770, dont 241 936 filles, et on utilise alors cet *a priori* Beta(10,12) sur θ :

$$\begin{aligned}\theta | \mathbf{y}_{1:20}, \mathbf{y}_{21:493472} &\sim \text{Beta}(10 + 241\,936, 12 + 251\,516) \\ &\sim \text{Beta}(241\,946, 251\,528)\end{aligned}$$

On retrouve la distribution *a posteriori* avec l'ensemble des observations

Inférence bayésienne

Inférence bayésienne

Modélisation bayésienne \Rightarrow distribution *a posteriori* :

- ensemble de l'information sur θ , **conditionnellement au modèle et aux données**

Inférence bayésienne

Modélisation bayésienne \Rightarrow distribution *a posteriori* :

- ensemble de l'information sur θ , **conditionnellement au modèle et aux données**

Résumé de cette distribution ?

- centre
- incertitude
- ...

Estimateurs ponctuels

Théorie de la décision

Contexte : estimation d'un paramètre inconnu θ

Décision : choix d'un estimateur ponctuel $\hat{\theta}$ « optimal »

fonction de coût : représente la pénalité associée au choix d'un $\hat{\theta}$ particulier

Pour choisir le $\hat{\theta}$ optimal, on minimise la fonction de coût choisie

un grand nombre de fonctions de coût différentes sont possibles : chacune d'entre elle résulte en un estimateur ponctuel optimal différent

Estimateurs ponctuels

Estimateurs ponctuels

- **Espérance *a posteriori*** : $\mu_P = \mathbb{E}(\theta|y) = \mathbb{E}_{\theta|y}(\theta)$
pas toujours facile car nécessite le calcul d'une intégrale...
⇒ minimise le coût quadratique

Estimateurs ponctuels

Estimateurs ponctuels

- **Espérance *a posteriori*** : $\mu_P = \mathbb{E}(\theta|y) = \mathbb{E}_{\theta|y}(\theta)$
pas toujours facile car nécessite le calcul d'une intégrale...
⇒ minimise le coût quadratique
- **Maximum *A Posteriori* (MAP)** :
plus facile à calculer : une simple maximisation de $f(y|\theta)\pi(\theta)$

Estimateurs ponctuels

Estimateurs ponctuels

- **Espérance *a posteriori*** : $\mu_P = \mathbb{E}(\theta|y) = \mathbb{E}_{\theta|y}(\theta)$
pas toujours facile car nécessite le calcul d'une intégrale...
⇒ minimise le coût quadratique
- **Maximum *A Posteriori* (MAP)** :
plus facile à calculer : une simple maximisation de $f(y|\theta)\pi(\theta)$
- **Median *a posteriori*** : la médiane de $p(\theta|(y))$

Estimateurs ponctuels

Estimateurs ponctuels

- **Espérance *a posteriori*** : $\mu_P = \mathbb{E}(\theta|y) = \mathbb{E}_{\theta|y}(\theta)$
pas toujours facile car nécessite le calcul d'une intégrale...
⇒ minimise le coût quadratique
 - **Maximum *A Posteriori* (MAP)** :
plus facile à calculer : une simple maximisation de $f(y|\theta)\pi(\theta)$
 - **Median *a posteriori*** : la médiane de $p(\theta|(y))$
- ⚠ L'approche bayésienne fournit, au delà de l'estimation ponctuelle, une caractérisation complète de la distribution *a posteriori*

Estimateurs ponctuels

MAP sur l'exemple historique

Calcul du Maximum *A Posteriori* dans l'exemple historique des naissances féminines à Paris avec un *a priori* uniforme :

$$p(\theta|\mathbf{y}) = \binom{n}{S} (n+1) \theta^S (1-\theta)^{n-S}$$

avec $n = 493\ 472$ et $S = 241\ 945$

$$\hat{\theta}_{MAP} = \dots$$

Estimateurs ponctuels

MAP sur l'exemple historique

Calcul du Maximum *A Posteriori* dans l'exemple historique des naissances féminines à Paris avec un *a priori* uniforme :

$$p(\theta|\mathbf{y}) = \binom{n}{S} (n+1) \theta^S (1-\theta)^{n-S}$$

avec $n = 493\ 472$ et $S = 241\ 945$

$$\hat{\theta}_{MAP} = \frac{S}{n} = 0,4902912$$

Estimateurs ponctuels

Espérance *a posteriori* sur l'exemple historique

Calcul de l'espérance *a posteriori* dans l'exemple historique des naissances féminines à Paris avec un *a priori* uniforme :

$$p(\theta|\mathbf{y}) = \binom{n}{S} (n+1) \theta^S (1-\theta)^{n-S}$$

avec $n = 493\ 472$ et $S = 241\ 945$

$$E(\theta|\mathbf{y}) = \int_0^1 \theta p(\theta|\mathbf{y}) d\theta$$

$$\tilde{\theta} = \dots$$

Estimateurs ponctuels

Espérance *a posteriori* sur l'exemple historique

Calcul de l'espérance *a posteriori* dans l'exemple historique des naissances féminines à Paris avec un *a priori* uniforme :

$$p(\theta|\mathbf{y}) = \binom{n}{S} (n+1) \theta^S (1-\theta)^{n-S}$$

avec $n = 493\ 472$ et $S = 241\ 945$

$$E(\theta|\mathbf{y}) = \int_0^1 \theta p(\theta|\mathbf{y}) d\theta$$

$$\tilde{\theta} = \binom{n}{S} (n+1) \frac{S+1}{\binom{n}{S} (n+1)(n+2)} = \frac{S+1}{n+2} = 0,4902913$$

Incertitude

Rappel sur l'Intervalle de confiance

Quelle est l'interprétation d'un intervalle de confiance fréquentiste au niveau 95% ?

...

Rappel sur l'Intervalle de confiance

Quelle est l'interprétation d'un intervalle de confiance fréquentiste au niveau 95% ?

95% des intervalles calculés sur l'ensemble des échantillons possibles (tous ceux qu'il est possible d'observer) contiennent la vraie valeur θ

Attention : on ne peut pas interpréter une réalisation d'un intervalle de confiance en terme probabiliste ! C'est une erreur qui est souvent commise...

Intervalle de crédibilité

L'**intervalle de crédibilité** s'interprète lui bien plus naturellement que l'intervalle de confiance :

C'est un intervalle qui a 95% de chance de contenir θ
(pour un niveau de 95%, évidemment)

Défini comme un intervalle avec une forte probabilité *a posteriori*.

Par exemple, un **intervalle de crédibilité à 95%** est un intervalle $[t_{inf}; t_{sup}]$ tel que $\int_{t_{inf}}^{t_{sup}} p(\theta|y) d\theta = 0.95$

NB : en général on s'intéresse à l'intervalle de crédibilité à 95% le plus étroit possible (*Highest Density Interval*).

Distribution prédictive

distribution prédictive : distribution d'une nouvelle observation Y_{n+1} sachant les observations précédentes \mathbf{y} , marginalement par rapport à θ :

$$\begin{aligned} f_{Y_{n+1}}(y|\mathbf{y}) &= \int_{\Theta} f_{Y_{n+1}}(y, \theta | \mathbf{y}) d\theta \\ &= \int_{\Theta} f_{Y_{n+1}}(y | \theta, \mathbf{y}) p(\theta | \mathbf{y}) d\theta \\ &= \int_{\Theta} f_{Y_{n+1}}(y | \theta) p(\theta | \mathbf{y}) d\theta \end{aligned}$$

NB : on remarque le lien avec la distribution marginale

$$f_Y(y) = \int_{\Theta} f_Y(y | \theta) \pi(\theta) d\theta$$

Concentration de la loi *a posteriori*

Théorème de convergence de Doob

La distribution *a posteriori* se concentre vers la « vraie » valeur du paramètre θ^* lorsque $n \rightarrow \infty$:

$$p(\theta|y_n) \xrightarrow{\mathcal{L}} \delta_{\theta^*}$$

→ *Seeing Theory*, Brown University
<http://students.brown.edu/seeing-theory/bayesian-inference/index.html#section3>

Bernstein-von Mises : approximation normale

Théorème de Bernstein-von Mises (ou théorème limite central bayésien) : Pour n grand, la distribution *a posteriori* peut être approximée par une loi normale

$$p(\theta | \mathbf{y}) \underset{n \rightarrow +\infty}{\approx} \mathcal{N}(\hat{\theta}, I(\hat{\theta})^{-1})$$

Conséquences :

- Méthodes bayésiennes et procédures fréquentistes basées sur le maximum de vraisemblance donnent, pour n suffisamment grand, des résultats très proches
- on peut approximer la loi *a posteriori* par une loi normale dont les paramètres se calcule simplement avec le MAP

Propriétés asymptotiques

Illustration sur l'exemple historique

Conclusion

Concepts essentiels

1 Modélisation bayésienne :

$$\theta \sim \pi(\theta) \quad \text{l'}a priori$$

$$Y_i | \theta \stackrel{iid}{\sim} f(y|\theta) \quad \text{modèle d'échantillonnage}$$

2 La formule de Bayes : $p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}$

avec $p(\theta|y)$ la loi *a posteriori*, $f(y|\theta)$ la vraisemblance (héritée du modèle d'échantillonnage), $\pi(\theta)$ l'*a priori* et $f(y) = \int f(y|\theta)\pi(\theta)$ la distribution marginale des données, i.e. la constante de normalisation (par rapport à θ)

3 La distribution *a posteriori* est obtenue par :

$$p(\theta|y) \propto f(y|\theta)\pi(\theta)$$

4 La loi *a priori* faiblement informative de Jeffreys :

$$\pi(\theta) \propto \sqrt{I(\theta)} \quad \text{en unidimensionnel}$$

possédant la propriété d'invariance.

5 Intervalle de crédibilité, MAP et moyenne *a posteriori*

Usage pratique

L'approche bayésienne est un outil statistique pour l'analyse de données (parmi d'autres)

Usage pratique

L'approche bayésienne est un outil statistique pour l'analyse de données (parmi d'autres)

Particulièrement **utile quand :**

- peu d'observations sont disponibles
- on dispose de connaissances *a priori* importantes

Usage pratique

L'approche bayésienne est un outil statistique pour l'analyse de données (parmi d'autres)

Particulièrement **utile quand :**

- peu d'observations sont disponibles
- on dispose de connaissances *a priori* importantes

Comme tout méthode statistique, l'analyse bayésienne présente des avantages et des inconvénients, qui seront plus ou moins important selon l'application envisagée

Questions ?

