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

12 novembre 2019

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**

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**

La statistique « fréquentiste » :

- Neyman & Pearson
- paramètres vu comme **déterministes**
- Estimation par le **Maximum de Vraisemblance**
- **théorie des tests** statistiques & **intervalle de confiance**

Théorème de Bayes

Article posthume du révérend Thomas Bayes en 1763

$$\Pr(A|E) = \frac{\Pr(E|A)\Pr(A)}{\Pr(E|A)\Pr(A) + \Pr\left(E|\overline{A}\right)\Pr\left(\overline{A}\right)} = \frac{\Pr(E|A)\Pr(A)}{\Pr(E)}$$

Théorème de Bayes

Article posthume du révérend Thomas Bayes en 1763

$$\Pr(A|E) = \frac{\Pr(E|A)\Pr(A)}{\Pr(E|A)\Pr(A) + \Pr\left(E|\overline{A}\right)\Pr\left(\overline{A}\right)} = \frac{\Pr(E|A)\Pr(A)}{\Pr(E)}$$

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 le probabilité que je sois 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 the population. 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 resultat positif au test, quel est la probabilité qu'il ait la maladie ?

Théorème de Bayes : exercice

1% de la population est affecté par une maladie rare the population. 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 resultat positif au test, quel est la probabilité qu'il ait la maladie ?

$$\mathsf{Pr}(M=+)=0.01 \qquad \mathsf{Pr}(T=+|M=+)=0.99 \qquad \mathsf{Pr}(T=-|M=-)=0.95$$

$$\mathsf{Pr}(M=+|T=+) \quad = \ ?$$

Théorème de Bayes : exercice

1% de la population est affecté par une maladie rare the population. 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 resultat positif au test, quel est la probabilité qu'il ait la maladie ?

$$\mathrm{Pr}(M=+)=0.01 \qquad \mathrm{Pr}(T=+|M=+)=0.99 \qquad \mathrm{Pr}(T=-|M=-)=0.95$$

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

Théorème de Bayes continu

- $f(y|\theta)$: modèle (probabiliste) paramétrique
- θ : paramètres
- π : distribution de probabilité

Théorème de Bayes continu :

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

Théorème de Bayes continu

- $f(y|\theta)$: modèle (probabiliste) paramétrique
- θ : paramètres
- π : distribution de probabilité

Théorème de Bayes continu :

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

Pierre-Simon de Laplace !

Philosophie bayésienne

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

$\Rightarrow$ 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 rejetait fermement le raisonnement bayésien

$\Rightarrow$ communauté divisée en 2 au XX^{e} siècle

Bayésiens vs. Fréquentistes : un débat dépassé

Fisher rejetait fermement le raisonnement bayésien

$\Rightarrow$ 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 necessaire

Gilbert Saporta

Modélisation bayésienne

Rappel sur la modélisation fréquentiste

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

Rappel sur la modélisation fréquentiste

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

Rappel sur la modélisation fréquentiste

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

$\widehat{\theta}$?

Présentation de l'application historique

Laplace

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

$\Rightarrow$ **observations** : naissances observées à Paris entre 1745 et 1770 ($241{,}945$ filles & $251{,}527$ garçons)

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

Trois composants

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 **décrites** ?

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 **décrites** ?

3. **La distribution *a priori***

   Un distribution de probabilité sur les paramètres θ du modèle d'échantillonnage

Le modèle d'échantillonnage

$\boldsymbol{y}$: les observations disponibles

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

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

La distribution *a priori*

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

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

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

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***

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 nait, 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 nait, 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

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

3. **Distribution *a priori***

   ...

Retour à l'exemple historique de Laplace

1. **La question**

   Quand un enfant nait, 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

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

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 $p(\theta|\boldsymbol{Y})$

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 $p(\theta|\boldsymbol{Y})$

Théorème de Bayes :

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

où $f(\boldsymbol{y}) = \int_{\Theta} f(\boldsymbol{y}|\theta)\pi(\theta)\,\mathrm{d}\theta$ 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 $p(\theta|\boldsymbol{Y})$

Théorème de Bayes :

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

où $f(\boldsymbol{y}) = \int_\Theta f(\boldsymbol{y}|\theta)\pi(\theta)\,\mathrm{d}\theta$ 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 $f(\boldsymbol{y}|\theta)$ 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**

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

2. **La loi *a priori***

   ...

3. **La distribution *a posteriori***

   ...

Application à l'exemple historique

1. **La vraisemblance**

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

2. **La loi *a priori***

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

3. **La distribution *a posteriori***

   . . .

Application à l'exemple historique

1. **La vraisemblance**

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

2. **La loi *a priori***

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

3. **La distribution *a posteriori***

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

Application à l'exemple historique

1. **La vraisemblance**

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

2. **La loi *a priori***

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

3. **La distribution *a posteriori***

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

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

Application à l'exemple historique

1. **La vraisemblance**

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

2. **La loi *a priori***

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

3. **La distribution *a posteriori***

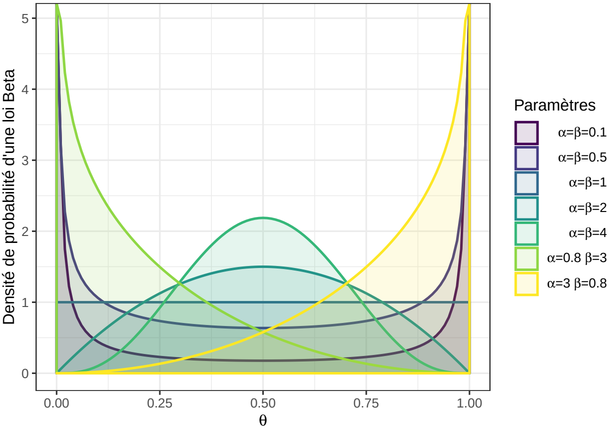
$$p(\theta|\boldsymbol{y}) = \frac{\theta^S(1-\theta)^{n-S}}{f(\boldsymbol{y})} = p(\theta|\boldsymbol{y}) = \binom{n}{S}(n+1)\theta^S(1-\theta)^{n-S}$$

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

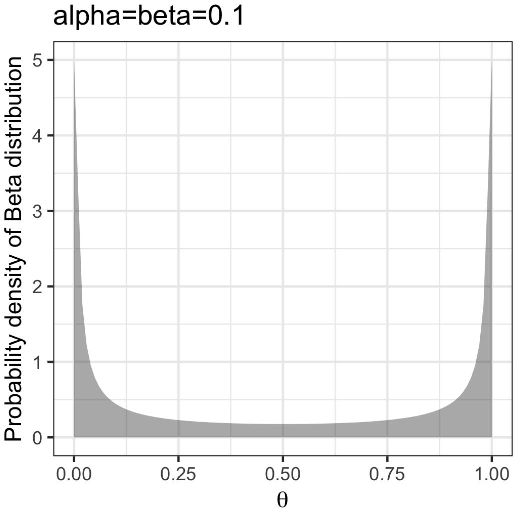
$$P(\theta \geq 0.5|\boldsymbol{y}) = \int_{0.5}^{1} p(\theta|\boldsymbol{y}) = \binom{n}{S}(n+1)\int_{0.5}^{1} \theta^S(1-\theta)^{n-S}\mathrm{d}\theta \approx 1.15\ 10^{-42}$$

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


alpha=beta=0.1
Probability density of Beta distribution
0
1
2
3
4
5
0.00
0.25
0.50
0.75
1.00
θ

Conjugaison de la distribution Beta

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

Loi *a prosteriori* associée : ...

Conjugaison de la distribution Beta

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

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

...

Le signe $\propto$ signifie « proportionel à »

Conjugaison de la distribution Beta

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

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

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

Le signe $\propto$ signifie « proportionel à »

Conjugaison de la distribution Beta

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

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

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

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

Le signe $\propto$ signifie « proportionel à »

Impact du choix de l'*a priori*

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

Pour 493 472 nouveaux-nés dont 241 945 filles

Impact du choix de l'*a priori*

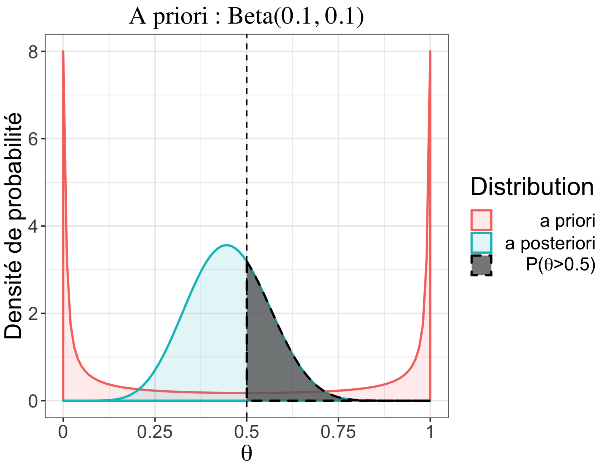
Interprétation de l'*a priori*	Paramètres de la distribution Beta	$P(\theta \geq 0,5 \vert \boldsymbol{y})$
#garçons > #filles	$\alpha = 0,1; \beta = 3$	$1,08\ 10^{-42}$
#garçons < #filles	$\alpha = 3; \quad \beta = 0,1$	$1,19\ 10^{-42}$
#garçons = #filles	$\alpha = 4; \quad \beta = 4$	$1,15\ 10^{-42}$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

Pour 493 472 nouveaux-nés dont 241 945 filles

Interprétation de l'*a priori*	Paramètres de la distribution Beta	$P(\theta \geq 0,5 \vert \boldsymbol{y})$
#garçons > #filles	$\alpha = 0.1, \beta = 3$	0.39
#garçons < #filles	$\alpha = 3, \quad \beta = 0.1$	0.52
#garçons = #filles	$\alpha = 4, \quad \beta = 4$	0.46
#garçons ≠ #filles	$\alpha = 0.1, \beta = 0.1$	0.45
non-informatif	$\alpha = 1, \quad \beta = 1$	0.45

Pour 20 nouveaux-nés dont 9 filles

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



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**

$\Rightarrow$ 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 $p(\theta|\boldsymbol{y}) = 0$

2. Les différents paramètres sont indépendants *a priori*

Élicitation de la loi *a priori*

Stratégies pour communiquer avec des experts non-statisticiens
$\Rightarrow$ 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 plausible 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
- Éliciter 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 ?

La quête des *a priori* non-informatifs

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

$\Rightarrow$ la loi Uniforme, un **a priori non-informatif** ?

La quête des *a priori* non-informatifs

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

$\Rightarrow$ la loi Uniforme, un **a priori non-informatif** ?

2 difficultés majeures :

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

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** $\int_\Theta \pi(\theta)\mathrm{d}\theta = \infty$
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***
$\Rightarrow$ la loi Uniforme, un **a priori non-informatif** ?

2 difficultés majeures :

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

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

- *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*

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

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

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

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

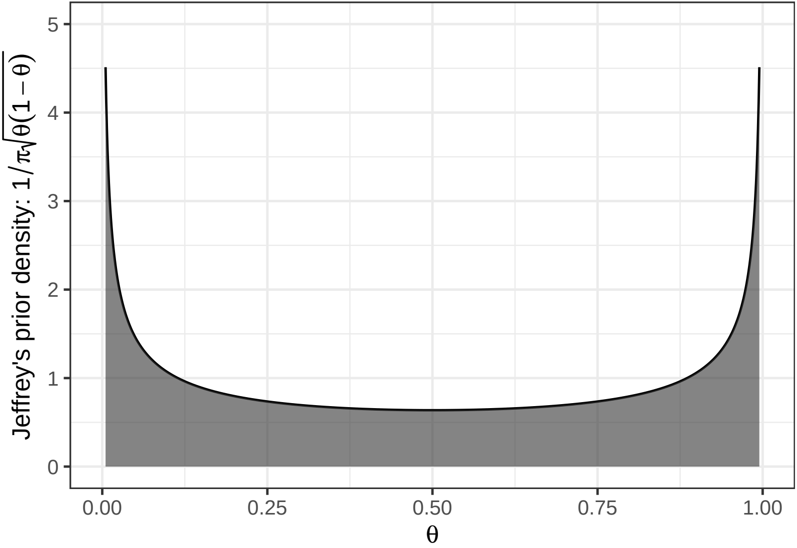
$$\begin{aligned}\pi(\theta) &\propto \sqrt{I(\theta)}\\ &\propto \sqrt{-\mathbb{E}_{Y|\theta}\left[\frac{\mathrm{d}^2\log(f(y|\theta))}{\mathrm{d}\theta^2}\right]}\\ &\propto \dots\end{aligned}$$

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

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

$$\begin{aligned}
\pi(\theta) &\propto \sqrt{I(\theta)} \\
&\propto \sqrt{-\mathbb{E}_{Y|\theta}\left[\frac{\mathrm{d}^2 \log(f(y|\theta))}{\mathrm{d}\theta^2}\right]} \\
&\propto \sqrt{\mathbb{E}_{Y|\theta}\left[\frac{y}{\theta^2} + \frac{1-y}{(1-\theta)^2}\right]} \\
&\propto \sqrt{\theta\left(\frac{1}{\theta^2}\right) + (1-\theta)\left(\frac{1}{(1-\theta)^2}\right)} \\
&\propto \sqrt{\frac{1}{\theta} + \frac{1}{1-\theta}} \\
&\propto \frac{1}{\sqrt{\theta(1-\theta)}}
\end{aligned}$$

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



Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

1. $\pi(\theta)$
2. $f(\boldsymbol{y}|\theta)$

Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

1. $\eta \sim h(\eta)$
2. $\pi(\theta|\eta)$
3. $f(\boldsymbol{y}|\theta)$

Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

1. $\eta \sim h(\eta)$
2. $\pi(\theta|\eta)$
3. $f(\boldsymbol{y}|\theta)$

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

Hyper-priors & modèles hiérarchiques

Niveaux hiérarchiques :

1. $\eta \sim h(\eta)$
2. $\pi(\theta|\eta)$
3. $f(\boldsymbol{y}|\theta)$

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

NB : 3 niveaux hiérarchiques ⇔ 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. $f(\boldsymbol{y}|\theta)$

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

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

⇒ 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 :

$$\begin{aligned}
\alpha &\sim \text{Gamma}(4;0,5)\\
\beta &\sim \text{Gamma}(4;0,5)\\
\theta|\alpha,\beta &\sim \text{Beta}(\alpha;\beta)\\
Y_i|\theta &\overset{iid}{\sim} \text{Bernoulli}(\theta)
\end{aligned}$$

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

$\Rightarrow$ 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

$\Rightarrow$ nécessite d'estimer cet *a priori* à partir des données

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

Bayésien empirique

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

$\Rightarrow$ nécessite d'estimer cet *a priori* à partir des données

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

- Combine les approches bayésienne et fréquentiste
- Distribution *a posteriori* concentrée (variance $\searrow$), mais biais $\nearrow$ (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 : $\widehat{\alpha}_M = 0,020$ et $\widehat{\beta}_M = 0,021$

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

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

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

Bayes séquentiel

Le théorème de Bayes peut être utilisé séquentiellement :

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

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

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

Bayes séquentiel

Le théorème de Bayes peut être utilisé séquentiellement :

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

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

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

Bayes séquentiel

Le théorème de Bayes peut être utilisé séquentiellement :

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

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

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

⇒ 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 $\boldsymbol{y}_{1:20}$ début 1745, dont 9 filles, et que l'on ait un *a priori* uniforme sur θ :

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

Bayes séquentiel : application à l'exemple historique

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

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

On observe ensuite $\boldsymbol{y}_{21:493\,472}$ les 493 452 naissances restantes entre 1745 et 1770, dont 241 936 filles, et on utilise alors cet *a priori* $\mathrm{Beta}(10, 12)$ sur θ :

$$\theta|\boldsymbol{y}_{1:20}, \boldsymbol{y}_{21:493\,472} \sim \dots$$

Bayes séquentiel : application à l'exemple historique

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

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

On observe ensuite $\boldsymbol{y}_{21:493\,472}$ 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 :

$$\begin{aligned}\theta|\boldsymbol{y}_{1:20}, \boldsymbol{y}_{21:493\,472} &\sim \text{Beta}(10 + 241\,936, 12 + 251\,516)\\ &\sim \text{Beta}(241\,946, 251\,528)\end{aligned}$$

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
- ...

Théorie de la décision

Contexte : estimation d'un paramètre inconnu θ

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

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

Pour choisir le $\widehat{\theta}$ 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

- **Espérance *a posteriori* :** $\mu_P = \mathbb{E}(\theta|\boldsymbol{y}) = \mathbb{E}_{\theta|\boldsymbol{y}}(\theta)$

  pas toujours facile car nécessite le calcul d'une intégrale...

  $\Rightarrow$ minimise le coût quadratique

  ...

Estimateurs ponctuels

- **Espérance *a posteriori* :** $\mu_P = \mathbb{E}(\theta|\boldsymbol{y}) = \mathbb{E}_{\theta|\boldsymbol{y}}(\theta)$

  pas toujours facile car nécessite le calcul d'une intégrale. . .

  $\Rightarrow$ minimise le coût quadratique

- **Maximum *A Posteriori* (MAP) :**
  plus facile à calculer : une simple maximisation de $f(\boldsymbol{y}|\theta)\pi(\theta)$

Estimateurs ponctuels

- **Espérance *a posteriori* :** $\mu_P = \mathbb{E}(\theta|\boldsymbol{y}) = \mathbb{E}_{\theta|\boldsymbol{y}}(\theta)$

  pas toujours facile car nécessite le calcul d'une intégrale. . .

  $\Rightarrow$ minimise le coût quadratique

- **Maximum *A Posteriori* (MAP) :**
  plus facile à calculer : une simple maximisation de $f(\boldsymbol{y}|\theta)\pi(\theta)$

- **Median *a posteriori* :** la médiane de $p(\theta|(y))$

Estimateurs ponctuels

- **Espérance *a posteriori* :** $\mu_P = \mathbb{E}(\theta|\boldsymbol{y}) = \mathbb{E}_{\theta|\boldsymbol{y}}(\theta)$

  pas toujours facile car nécessite le calcul d'une intégrale...

  $\Rightarrow$ minimise le coût quadratique

- **Maximum *A Posteriori* (MAP) :**
  plus facile à calculer : une simple maximisation de $f(\boldsymbol{y}|\theta)\pi(\theta)$

- **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*

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 $S = 241\ 945$

$\widehat{\theta}_{MAP} = \ldots$

MAP sur l'exemple historique

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

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

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

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

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 :

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

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

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

$\tilde{\theta} = \ldots$

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 :

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

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

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

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

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

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 $\boldsymbol{y}$, marginalement par rapport à θ :

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

NB : on remarque le lien avec la distribution marginale
$f_Y(y) = \int_\Theta f_Y(y|\theta)\pi(\theta)\,\mathrm{d}\theta$

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 $n \to \infty$:

$$p(\theta|\boldsymbol{y}_n) \xrightarrow{\mathscr{L}} \delta_{\theta^*}$$

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

Bernstein-von Mises : approximation normale

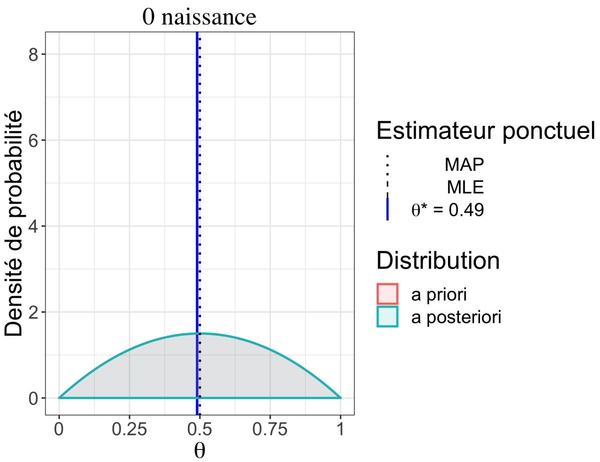
Théorème de Bernstein-von Mises (ou théorème limite central bayésien) : Pour n grand, la distribution *a posteriori* peut être approximée par une loi normale

$$p(\theta|\boldsymbol{y}) \underset{n\to+\infty}{\approx} \mathcal{N}\left(\hat{\theta}, I(\hat{\theta})^{-1}\right)$$

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

Illustration sur l'exemple historique



Conclusion

Concepts essentiels

1. **Modélisation bayésienne** :

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

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

2. **La formule de Bayes** : $p(\theta|\boldsymbol{y}) = \dfrac{f(\boldsymbol{y}|\theta)\pi(\theta)}{f(\boldsymbol{y})}$

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

3. **La distribution *a posteriori* est obtenue par** :

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

4. **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 ?