Introduction to the Bayesian framework in Biometrics
—
Part I: Bayesian theory

Boris Hejblum

PhD Training course
Digital Public Health program, University of Bordeaux

Bayesian vocabulary

- **paradigm**
- ***a priori***
- ***a posteriori***
- **elicitation**

Course objectives

I **Familiarize** oneself with the **Bayesian framework**:

1. understand and assess a Bayesian modeling strategy, and discuss its underlying assumptions
2. rigorously describe expert knowledge by a quantitative prior distribution

II **Study** and **perform** Bayesian analyses in **biomedical applications**:

1. understand and discuss assumptions and methodological choices in biomedical literature using Bayesian methods, including "under the hood" estimation machinery
2. understand, discuss and reproduce a Bayesian estimation of a relative risk
3. understand and perform a Bayesian meta-analysis using R
4. understand continuous reassessment method for phase I/II dose trials, along with the associated decision-rule

NB : this course is by no means exhaustive, and the curious reader will be referred to more complete works such as *The Bayesian Choice* by C Robert.

Motivational examples: diagnostic tests

Search The Guardian International

The obscure maths theorem that governs the reliability of Covid testing

There's been much debate about lateral flow tests – their accuracy depends on context and the theories of a 18th-century cleric

[Good, *J GEN INTERN MED* 2020]

Table 1 Estimates for Post-Test Probability of Acute COVID-19 Infection for Simulated Patient Scenarios

Clinical Scenarios	Pre-test probability (%)	PCR assay sensitivity (%)	Post-test probability of acute COVID-19 infection	
			Positive test (%)	Negative test (%)
Patient 1: high pre-test probability	70	70	100	41.2
		90	100	18.9
	90	70	100	73.0
		90	100	47.4
Patient 2: low pre-test probability	5	70	97.4	1.6
		90	97.9	0.5
	10	70	98.7	3.2
		90	99.0	1.1

Original Article

Bayesian analysis of tests with unknown specificity and sensitivity

Andrew Gelman, Bob Carpenter

First published: 13 August 2020 | https://doi.org/10.1111/rssc.12435 | Citations: 6

Motivational examples: clinical trial design

Design

CLINICAL TRIALS

Clinical Trials
1–10

DOI: 10.1177/1740774520943846
journals.sagepub.com/home/ctj
SAGE

Anti-Thrombotic Therapy to Ameliorate Complications of COVID-19 (ATTACC): Study design and methodology for an international, adaptive Bayesian randomized controlled trial

Methods: An international, open-label, adaptive randomized controlled trial. Using a Bayesian framework, the trial will declare results as soon as pre-specified posterior probabilities for superiority, futility, or harm are reached. The trial uses response-adaptive randomization to maximize the probability that patients will receive the more beneficial treatment approach, as treatment effect information accumulates within the trial. By leveraging a common data safety monitoring

[Houston *et al.*, *Clinical Trials*, 17(5):491-500, 2020]

Motivational examples: study/trial analyses

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 APRIL 22, 2021 VOL. 384 NO. 16

Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19

The REMAP-CAP Investigators*

lumab group, and 0 (interquartile range, –1 to 15) in the control group. The median adjusted cumulative odds ratios were 1.64 (95% credible interval, 1.25 to 2.14) for tocilizumab and 1.76 (95% credible interval, 1.17 to 2.91) for sarilumab as compared with control, yielding posterior probabilities of superiority to control of more than 99.9% and of 99.5%, respectively. An analysis of 90-day survival showed improved survival in the pooled interleukin-6 receptor antagonist groups, yielding a hazard ratio for the comparison with the control group of 1.61 (95% credible interval, 1.25 to 2.08) and a posterior probability of superiority of more than 99.9%. All secondary analyses supported efficacy of these interleukin-6 receptor antagonists.

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 31, 2020 VOL. 383 NO. 27

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D., Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D., Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D., Satrajit Roychoudhury, Ph.D., Kenneth Koury, Ph.D., Ping Li, Ph.D., Warren V. Kalina, Ph.D., David Cooper, Ph.D., Robert W. Frenck, Jr., M.D., Laura L. Hammitt, M.D., Özlem Türeci, M.D., Haylene Nell, M.D., Axel Schaefer, M.D., Serhat Ünal, M.D., Dina B. Tresnan, D.V.M., Ph.D., Susan Mather, M.D., Philip R. Dormitzer, M.D., Ph.D., Uğur Şahin, M.D., Kathrin U. Jansen, Ph.D., and William C. Gruber, M.D., for the C4591001 Clinical Trial Group*

Table 2. Vaccine Efficacy against Covid-19 at Least 7 days after the Second Dose.*

Efficacy End Point	BNT162b2		Placebo		Vaccine Efficacy, % (95% Credible Interval)‡	Posterior Probability (Vaccine Efficacy >30%)§
	No. of Cases	Surveillance Time (n)†	No. of Cases	Surveillance Time (n)†		
	(N=18,198)		(N=18,325)			
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.3–97.6)	>0.9999
	(N=19,965)		(N=20,172)			
Covid-19 occurrence at least 7 days after the second dose in participants with and those without evidence of infection	9	2.332 (18,559)	169	2.345 (18,708)	94.6 (89.9–97.3)	>0.9999

* The total population without baseline infection was 36,523; total population including those with and those without prior evidence of infection was 40,137.
† The surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.
‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.
§ Posterior probability was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

Introduction

Statistics:

- a **mathematical** science
- to **describe** what has happened and
- to assess what **may** happen in **the future**
- relies on the **observation** of natural phenomena in order to propose an interpretation, often through **probabilistic models**

Statistics:

- a **mathematical** science
- to **describe** what has happened and
- to assess what **may** happen in **the future**
- relies on the **observation** of natural phenomena in order to propose an interpretation, often through **probabilistic models**

Frequentist statistics:

- Neyman & Pearson
- **deterministic** view of the parameters
- **Maximum Likelihood Estimation**
- statistical **test theory** & **confidence interval**

Bayes' theorem

Reverend Thomas Bayes posthumous article in 1763

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

(conditional probability formula: $Pr(A|E) = \frac{Pr(A \cap E)}{Pr(E)}$)

Bayes' theorem

Reverend Thomas Bayes posthumous article in 1763

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

(conditional probability formula: $Pr(A|E) = \frac{Pr(A \cap E)}{Pr(E)}$)

In practice:

Last time you visited the doctor, you got **tested for a rare disease**. Unluckily, the result was positive...

Given the test result, what is the probability that I actually have this disease?

(Medical tests are, after all, not perfectly accurate.)

→ *Seeing Theory*, Brown University

Bayes theorem: exercise

As of May 11th, about 7% of the French population was estimated to have had COVID-19. A medical test has the following properties:

- if someone has COVID-19, its test will come out positive 71% of the time
- if someone does not have the disease, its test will come out negative 98% of the time

Given that someone got a positive result, what is his/her probability to truly have COVID-19 ?

Bayes theorem: exercise

As of May 11$^{\text{th}}$, about 7% of the French population was estimated to have had COVID-19. A medical test has the following properties:

- if someone has COVID-19, its test will come out positive 71% of the time
- if someone does not have the disease, its test will come out negative 98% of the time

Given that someone got a positive result, what is his/her probability to truly have COVID-19 ?

$$\Pr(D=+)=0.07 \qquad \Pr(T=+|D=+)=0.71 \qquad \Pr(T=-|D=-)=0.98$$

Bayes theorem: exercise

As of May 11$^{\text{th}}$, about 7% of the French population was estimated to have had COVID-19. A medical test has the following properties:

- if someone has COVID-19, its test will come out positive 71% of the time
- if someone does not have the disease, its test will come out negative 98% of the time

Given that someone got a positive result, what is his/her probability to truly have COVID-19 ?

$$\Pr(D=+)=0.07 \qquad \Pr(T=+|D=+)=0.71 \qquad \Pr(T=-|D=-)=0.98$$

$$\Pr(D=+|T=+) \quad = \quad ?$$

Bayes theorem: exercise

As of May 11th, about 7% of the French population was estimated to have had COVID-19. A medical test has the following properties:

- if someone has COVID-19, its test will come out positive 71% of the time
- if someone does not have the disease, its test will come out negative 98% of the time

Given that someone got a positive result, what is his/her probability to truly have COVID-19 ?

$$\Pr(D=+)=0.07 \qquad \Pr(T=+|D=+)=0.71 \qquad \Pr(T=-|D=-)=0.98$$

$$\begin{aligned}
\Pr(D=+|T=+) &= \frac{\Pr(T=+|D=+)\Pr(D=+)}{\Pr(T=+)} \\
&= \frac{\Pr(T=+|D=+)\Pr(M=+)}{\Pr(T=+|D=+)\Pr(D=+)+\Pr(T=+|D=-)\Pr(D=-)} \\
&= \frac{\Pr(T=+|D=+)\Pr(M=+)}{\Pr(T=+|D=+)\Pr(D=+)+(1-\Pr(T=-|D=-))(1-\Pr(D=+))} \\
&= (0.71\times 0.07)/(0.71\times 0.07+(1-0.98)\times(1-0.07))=0.73
\end{aligned}$$

Continuous Bayes' theorem

- parametric (probabilistic) model $f(y|\theta)$
- parameters θ
- probability distribution π

Continuous Bayes' theorem:

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

Continuous Bayes' theorem

- parametric (probabilistic) model $f(y|\theta)$
- parameters θ
- probability distribution π

Continuous Bayes' theorem:

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

remember Pierre-Simon de Laplace !

Bayes philosophy

Parameters are random variables ! – *no "true" value*

⇒ induces a marginal probability distribution $\pi(\theta)$ on the parameters: the ***prior*** distribution

- 😁 allows to **formally** take into account hypotheses in the modeling
- ☹ necessarily introduces **subjectivity** into the analysis

Bayesian vs. Frequentists: a historical note

1. **Bayes** + **Laplace** ⇒ development of statistics in the **18-19th centuries**
2. Galton & Pearson, then Fisher & Neymann ⇒ **frequentist** theory became dominant during the **20th century**
3. turn of the **21th century**: rise of the computer ⇒ **Bayes' comeback**

Bayesian vs. Frequentists: an outdated debate

Fisher firmly rejected Bayesian reasoning

$\Rightarrow$ community split in 2 in the 20^{th} century

Bayesian vs. Frequentists: an outdated debate

Fisher firmly rejected Bayesian reasoning

$\Rightarrow$ community split in 2 in the 20^{th} century

> *To be, or not to be, Bayesian, that is no longer the question: it is a matter of wisely using the right tools when necessary*
>
> Gilbert Saporta

Bayesian modeling

Refresher on frequentist modeling

- a series of *iid* (independent and identically distributed) random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$

Refresher on frequentist modeling

- a series of *iid* (independent and identically distributed) random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$
- we observe a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$

Refresher on frequentist modeling

- a series of *iid* (independent and identically distributed) random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$
- we observe a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$
- model their probability distribution as $f(y|\theta)$, $\theta \in \Theta$

Refresher on frequentist modeling

- a series of *iid* (independent and identically distributed) random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$
- we observe a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$
- model their probability distribution as $f(y|\theta)$, $\theta \in \Theta$

This model assumes there is a "true" distribution of Y characterized by the "true" value of the parameter θ^*

$\widehat{\theta}$?

Historical motivating example

Laplace

What is the probability of birth of girls rather than boys ?

$\Rightarrow$ **observations**: births observed in Paris between 1745 and 1770 (241,945 girls & 251,527 boys)

When a child is born, is it equally likely to be a girl or a boy ?

Three building blocks

1. **the question**
2. **the sampling model**
3. **the prior**

Three building blocks

1. **the question**

   The first step in building a model is always to identify the question you want to answer

2. **the sampling model**

3. **the prior**

Three building blocks

1. **the question**

   The first step in building a model is always to identify the question you want to answer

2. **the sampling model**

   Which **observations** are available to inform our response to this ?
   How can they be **described**?

3. **the prior**

Three building blocks

1. **the question**

   The first step in building a model is always to identify the question you want to answer

2. **the sampling model**

   Which **observations** are available to inform our response to this ?
   How can they be **described**?

3. **the prior**

   A probability distribution on the parameters θ of the sampling model

The sampling model

$\boldsymbol{y}$: the observations available

$\Rightarrow$ (parametric) **probabilistic model** underlying their **generation**:

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

The *prior* distribution

In Bayesian modeling, compared to frequentist modeling, we add a **probability distribution** on the **parameters** θ

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

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

θ will thus be treated like a random variable,
but which is never observed !

Back to Laplace's historical example

1. **The question**
2. **Sampling model**
3. ***prior***

Back to Laplace's historical example

1. **The question**

   ...

2. **Sampling model**

   ...

3. ***prior***

   ...

Back to Laplace's historical example

1. **The question**

   When a child is born, is it equally likely to be a girl or a boy ?

2. **Sampling model**

   ...

3. ***prior***

   ...

Back to Laplace's historical example

1. **The question**

   When a child is born, is it equally likely to be a girl or a boy ?

2. **Sampling model**

   Bernoulli's law for $Y_i = 1$ if the new born i is a girl, 0 if it is a boy:

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

3. ***prior***

   ...

Back to Laplace's historical example

1. **The question**

   When a child is born, is it equally likely to be a girl or a boy ?

2. **Sampling model**

   Bernoulli's law for $Y_i = 1$ if the new born i is a girl, 0 if it is a boy:

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

3. ***prior***

   A uniform prior on θ (the probability that a newborn would be a girl rather than a boy):

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

Posterior distribution

Purpose of a Bayesian modeling: **infer the *posterior*** distribution of the **parameters**

- ***Posterior***: the law of θ conditionally on the observations $p(\theta|\boldsymbol{y})$

Posterior distribution

Purpose of a Bayesian modeling: **infer the *posterior*** distribution of the **parameters**

- ***Posterior***: the law of θ conditionally on the observations $p(\theta|\boldsymbol{y})$

Bayes' theorem:

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

Posterior distribution

Purpose of a Bayesian modeling: **infer the *posterior*** distribution of the **parameters**

- ***Posterior***: the law of θ conditionally on the observations $p(\theta|\boldsymbol{y})$

Bayes' theorem:

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

Posterior is calculated from:

1. the sampling model $f(y|\theta)$ – which yields the likelihood $f(\boldsymbol{y}|\theta)$ for all observations
2. the *prior* $\pi(\theta)$

Application to the historical example

1. **the likelihood**
2. **the prior**
3. **the posterior**

Application to the historical example

1. **the likelihood**

   ...

2. **the prior**

   ...

3. **the posterior**

   ...

Application to the historical example

1. **the likelihood**

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

2. **the prior**

   ...

3. **the posterior**

   ...

Application to the historical example

1. **the likelihood**

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

2. **the prior**

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

3. **the posterior**

   ...

Application to the historical example

1. **the likelihood**

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

2. **the prior**

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

3. **the posterior**

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

Application to the historical example

1. **the likelihood**

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

2. **the prior**

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

3. **the posterior**

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

To answer the question of interest, we can then calculate: ...

Application to the historical example

1. **the likelihood**

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

2. **the prior**

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

3. **the posterior**

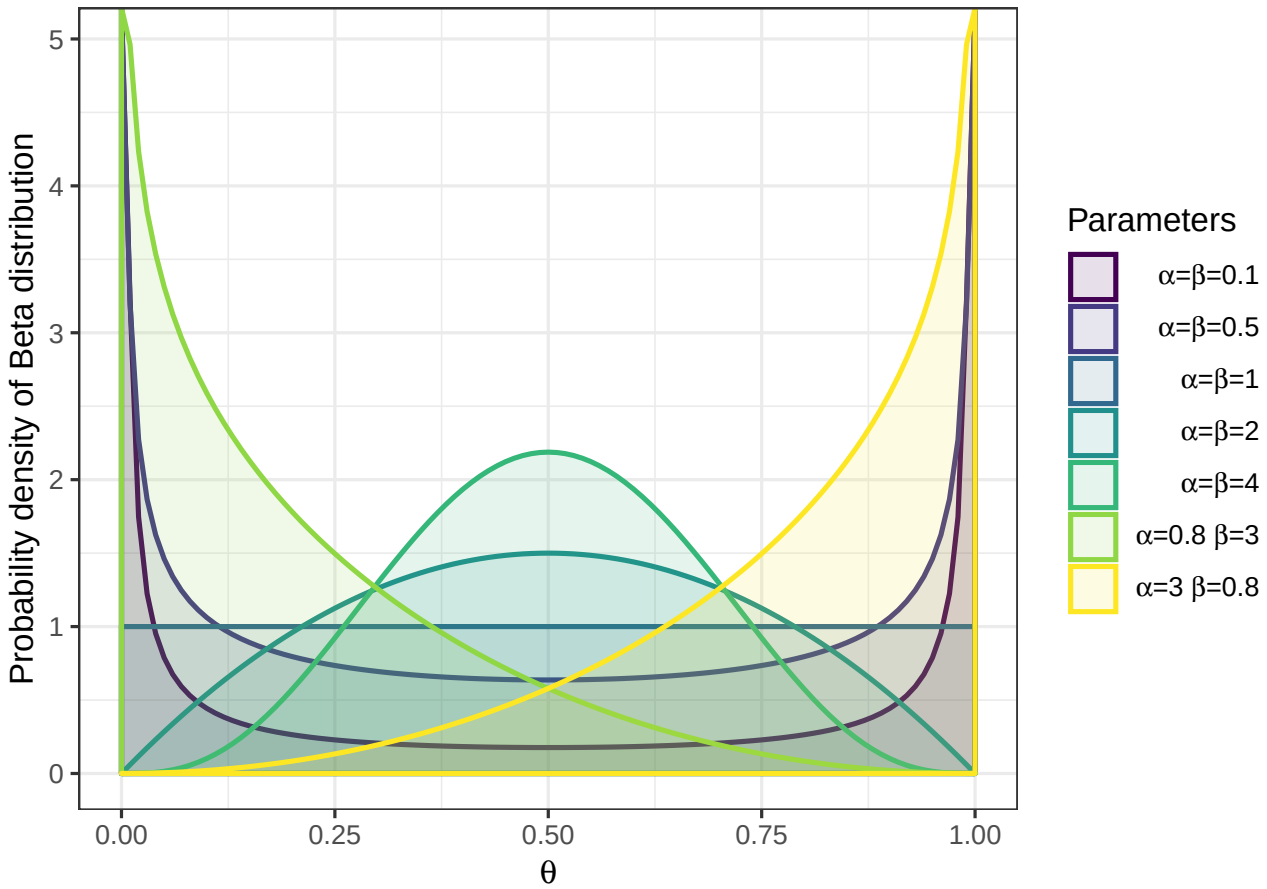
$$p(\theta|\boldsymbol{y}) = \frac{\theta^S(1-\theta)^{n-S}}{f(\boldsymbol{y})} = p(\theta|\boldsymbol{y}) = \binom{n}{S}(n+1)\theta^S(1-\theta)^{n-S}$$

To answer the question of interest, we can then calculate:

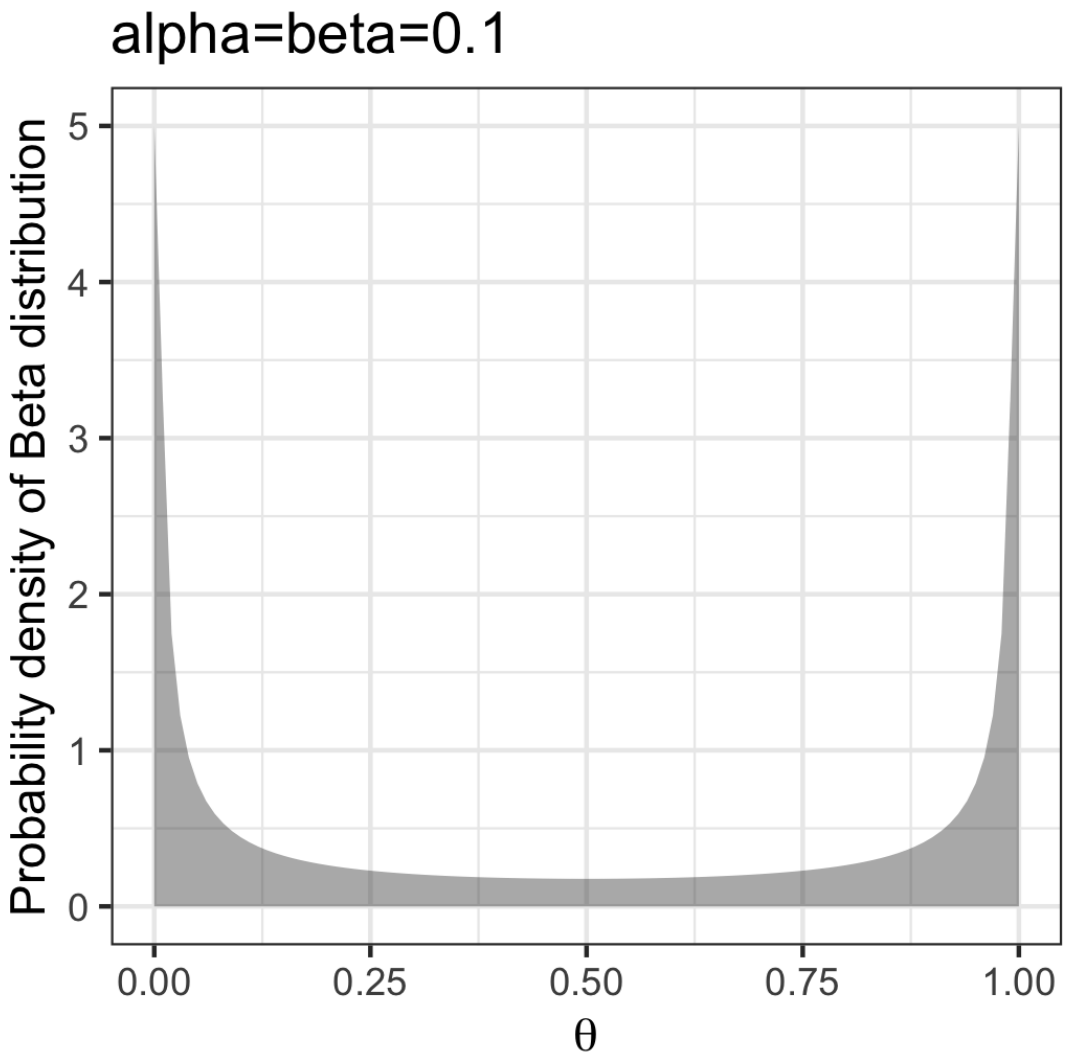
$$P(\theta \geq 0.5|\boldsymbol{y}) = \int_{0.5}^{1} p(\theta|\boldsymbol{y}) = \binom{n}{S}(n+1)\int_{0.5}^{1} \theta^S(1-\theta)^{n-S}\mathrm{d}\theta \approx 1.15\ 10^{-42}$$

The Beta distribution

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



Examples of various parametrizations for the Beta distribution


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

Conjugacy of the Beta distribution

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

Conjugacy of the Beta distribution

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

Corresponding *posterior*: $p(\theta|\mathbf{y}) \propto \theta^{\alpha+S-1}(1-\theta)^{\beta+(n-S)-1}$

...

The $\propto$ symbol means: "proportional to"

Conjugacy of the Beta distribution

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

Corresponding *posterior*: $p(\theta|\mathbf{y}) \propto \theta^{\alpha+S-1}(1-\theta)^{\beta+(n-S)-1}$

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

The $\propto$ symbol means: "proportional to"

Conjugacy of the Beta distribution

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

Corresponding *posterior*: $p(\theta|\mathbf{y}) \propto \theta^{\alpha+S-1}(1-\theta)^{\beta+(n-S)-1}$

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

This is called a **conjugated distribution** because the ***posterior*** and the ***prior*** belong to the **same parametric family**

The $\propto$ symbol means: "proportional to"

Impact of the *prior* choice

Interpretation of the *prior*	Parameters of the Beta distribution	$P(\theta \geq 0.5 \vert \boldsymbol{y})$
#boys > #girls	$\alpha = 0.1, \beta = 3$	$1.08\ 10^{-42}$
#boys < #girls	$\alpha = 3, \quad \beta = 0.1$	$1.19\ 10^{-42}$
#boys = #girls	$\alpha = 4, \quad \beta = 4$	$1.15\ 10^{-42}$
#boys ≠ #girls	$\alpha = 0.1, \beta = 0.1$	$1.15\ 10^{-42}$
non-informative	$\alpha = 1, \quad \beta = 1$	$1.15\ 10^{-42}$

For 493,472 newborns including 241,945 girls

Impact of the *prior* choice

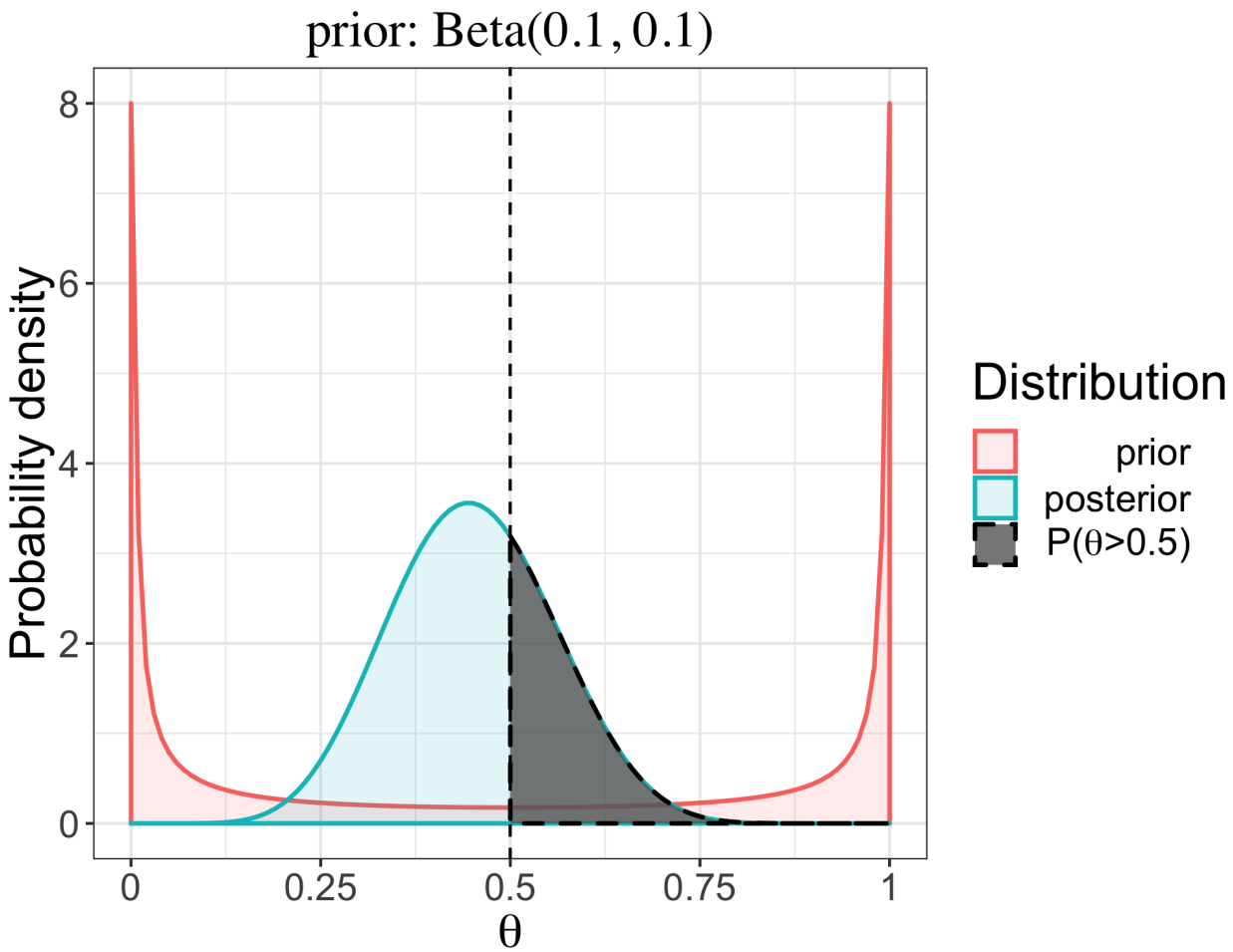
Interpretation of the *prior*	Parameters of the Beta distribution	$P(\theta \geq 0.5 \mid \boldsymbol{y})$
#boys > #girls	$\alpha = 0.1, \beta = 3$	$1.08\ 10^{-42}$
#boys < #girls	$\alpha = 3, \quad \beta = 0.1$	$1.19\ 10^{-42}$
#boys = #girls	$\alpha = 4, \quad \beta = 4$	$1.15\ 10^{-42}$
#boys ≠ #girls	$\alpha = 0.1, \beta = 0.1$	$1.15\ 10^{-42}$
non-informative	$\alpha = 1, \quad \beta = 1$	$1.15\ 10^{-42}$

For 493,472 newborns including 241,945 girls

Interpretation of the *prior*	Parameters of the Beta distribution	$P(\theta \geq 0.5 \mid \boldsymbol{y})$
#boys > #girls	$\alpha = 0.1, \beta = 3$	0.39
#boys < #girls	$\alpha = 3, \quad \beta = 0.1$	0.52
#boys = #girls	$\alpha = 4, \quad \beta = 4$	0.46
#boys ≠ #girls	$\alpha = 0.1, \beta = 0.1$	0.45
non-informative	$\alpha = 1, \quad \beta = 1$	0.45

For 20 newborns including 9 girls

Impact of the *prior* choice for 20 observed births – continued



Priors: pros & cons

Having a *prior* distribution:

- 😁 brings **flexibility**
- 😁 allows to incorporate **external knowledge**
- ☹ adds intrinsic **subjectivity**

⇒ choice (or elicitation) of a *prior* distribution is sensitive !

Prior properties

1. *posterior* support must be included in the support of the *prior*:
   if $\pi(\theta) = 0$, then $p(\theta|\boldsymbol{y}) = 0$

2. independence of the different parameters *a priori*

Prior Elicitation

Strategies to communicate with non-statistical experts

$\Rightarrow$ transform their **knowledge** into *prior* **distribution**

- **histogram method**: experts give weights to ranges of values
  ⚠ might give a zero *prior* for plausible parameter values
- choose a **parametric family** of distributions $p(\theta|\eta)$ in **agreement with what the experts think** (e.g. for quantiles or moments)
  (solves the support problem but the parametric family has a big impact)
- elicit *priors* from the **literature**
- ...

The quest for non-informative *priors*

Sometimes, one has **no prior knowledge whatsoever**

Which *prior* distribution to use ?

The quest for non-informative *priors*

Sometimes, one has **no prior knowledge whatsoever**

$\Rightarrow$ the Uniform distribution, a **non-informative prior** ?

The quest for non-informative *priors*

Sometimes, one has **no prior knowledge whatsoever**

$\Rightarrow$ the Uniform distribution, a **non-informative prior** ?

2 major difficulties:

1. **Improper distributions**
2. **Non-invariant distributions**

The quest for non-informative *priors*

Sometimes, one has **no prior knowledge whatsoever**

$\Rightarrow$ the Uniform distribution, a **non-informative prior** ?

2 major difficulties:

1. **Improper distributions**
2. **Non-invariant distributions**

Other solutions ?

Jeffreys' *priors*

A **weakly informative** *prior* invariant through re-parameterization

- unidimensional Jeffreys' *prior*:

$$\pi(\theta) \propto \sqrt{I(\theta)} \qquad \text{where } I \text{ is Fisher's information matrix}$$

- multidimensional Jeffreys' *prior*:

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

In practice, parameter are considered independent *a priori*

Hyper-*priors* & hierarchical models

Hierarchical levels:

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

Hyper-*priors* & hierarchical models

Hierarchical levels:

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

Hyper-*priors* & hierarchical models

Hierarchical levels:

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* & hierarchical models

Hierarchical levels:

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 hierarchical levels $\Leftrightarrow$ two levels with *prior*: $\pi(\theta) = \int \pi(\theta|\eta)h(\eta)d\eta$

Hyper-*priors* & hierarchical models

Hierarchical levels:

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 hierarchical levels $\Leftrightarrow$ two levels with *prior*: $\pi(\theta) = \int \pi(\theta|\eta)h(\eta)d\eta$

$\Rightarrow$ can **ease modeling** and **elicitation** of the *prior*...

Hyperprior in the historical example

Historical example of birth sex with a Beta *prior*

$\Rightarrow$ two Gamma hyper-*priors* for α and β (conjugated):

$$\alpha \sim \text{Gamma}(4, 0.5)$$
$$\beta \sim \text{Gamma}(4, 0.5)$$
$$\theta|\alpha, \beta \sim \text{Beta}(\alpha, \beta)$$
$$Y_i|\theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$$

Empirical Bayes

Eliciting the *prior* according to its empirical marginal distribution

$\Rightarrow$ estimate the *prior* from the data

1. hyper-parameters
2. estimate them through frequentist methods (e.g. MLE) by $\hat{\eta}$
3. plug-in estimates into the *prior*
4. $\Rightarrow$ *posterior*: $p(\theta|\boldsymbol{y},\hat{\eta})$

Empirical Bayes

Eliciting the *prior* according to its empirical marginal distribution

$\Rightarrow$ estimate the *prior* from the data

1. hyper-parameters
2. estimate them through frequentist methods (e.g. MLE) by $\hat{\eta}$
3. plug-in estimates into the *prior*
4. $\Rightarrow$ *posterior*: $p(\theta|\boldsymbol{y}, \hat{\eta})$

- Combines Bayesian and frequentist frameworks
- Concentrated *posterior* ($\searrow$ variance) but $\nearrow$ bias (data used twice !)
- Approximate a fully Bayesian approach

Empirical Bayes: example

For a distribution Beta(α, β):

- $\frac{\alpha}{\alpha+\beta}$ is the mean
- $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ is the variance

Thanks to the **method of moments** we get: $\widehat{\alpha}_M = 0,020$ et $\widehat{\beta}_M = 0,021$

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

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

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

Sequential Bayes

Bayes' theorem can be used sequentially:

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

If $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$, then:

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

$\Rightarrow$ *posterior* distribution updates as new observations are aquired/available (*online updates*)

Sequential Bayes in the historical example

Let's imagine that we start by observing 20 births $\boldsymbol{y}_{1:20}$ at the start of 1745, including 9 girls, and that we have a uniform *prior* on θ :

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

Sequential Bayes in the historical example

Let's imagine that we start by observing 20 births $\boldsymbol{y}_{1:20}$ at the start of 1745, including 9 girls, and that we have a uniform *prior* on θ :

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

Then we observe $\boldsymbol{y}_{21:493\,472}$ the remaining 493 452 births between 1745 and 1770, including 241 936 girls, and we then uses this $\text{Beta}(10, 12)$ *prior* for θ :

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

Sequential Bayes in the historical example

Let's imagine that we start by observing 20 births $\boldsymbol{y}_{1:20}$ at the start of 1745, including 9 girls, and that we have a uniform *prior* on θ :

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

Then we observe $\boldsymbol{y}_{21:493\,472}$ the remaining 493 452 births between 1745 and 1770, including 241 936 girls, and we then uses this $\text{Beta}(10, 12)$ *prior* for θ :

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

We get the same *posterior* distribution as with all the observations taken together at once

Bayesian inference

Bayesian Inference

Bayesian modeling $\Rightarrow$ *posterior* distribution:

- all of the information on θ, **conditionally to both the model and the data**

Bayesian Inference

Bayesian modeling $\Rightarrow$ *posterior* distribution:

- all of the information on θ, **conditionally to both the model and the data**

Summary of this *posterior* distribution ?

- center
- spread
- ...

Decision theory

Context: estimating an unknown parameter θ

Decision: choice of an "optimal" point estimator $\widehat{\theta}$

cost function: quantify the penalty associated with the choice of a particular $\widehat{\theta}$

$\Rightarrow$ minimize the cost function to choose the optimal $\widehat{\theta}$

a large number of cost functions are available: each one yields a different point estimator based on its own minimum rule

Point estimates

- ***Posterior* mean:** $\mu_P = \mathbb{E}(\theta|\boldsymbol{y}) = \mathbb{E}_{\theta|\boldsymbol{y}}(\theta)$

  not always easy because it assumes the calculation of an integral...

  $\Rightarrow$ minimize the quadratic error cost

- **Maximum *A Posteriori* (MAP):**
  easy(ier) to compute: just a simple maximization of the *posterior* $f(\boldsymbol{y}|\theta)\pi(\theta)$

- ***Posterior* median:** the median of $p(\theta|(y))$

  $\Rightarrow$ minimize the absolute error cost

⚠ the Bayesian approach gives a full characterization of the *posterior* distribution that goes beyond point estimation

MAP on the historical example

Maximum *A Posteriori* on the historical example of feminine birth in Paris with a uniform prior:

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

with $n = 493,472$ et $S = 241,945$

$$\widehat{\theta}_{MAP} = \frac{S}{n} = 0.4902912$$

Posterior mean on the historical example

***Posterior* mean** on the historical example of feminine birth in Paris with a uniform prior:

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

with $n = 493,472$ et $S = 241,945$

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

$$\tilde{\theta} = \binom{n}{S}(n+1)\frac{S+1}{\binom{n}{S}(n+1)(n+2)} = \frac{S+1}{n+2} = 0.4902913$$

Confidence Interval reminder

What is the interpretation of a frequentist confidence interval at a 95% level ?

...

Confidence Interval reminder

What is the interpretation of a frequentist confidence interval at a 95% level ?

> *95% of the intervals computed on all possible samples (all those that could have been observed) contain the true value θ*

Warning: one cannot interpret a realization of a confidence interval in probabilistic terms ! It is a common mistake...

Credibility interval

The **credibility interval** is interpreted much more naturally than the confidence interval:

> It is an interval that has a 95% chance of containing θ (for a 95% level, obviously)

Defined as an interval with a high *posterior* probability of occurrence.

For example, a **95% credibility interval** is an interval $[t_{inf}, t_{sup}]$ such that $\int_{t_{inf}}^{t_{sup}} p(\theta|\boldsymbol{y})\,\mathrm{d}\theta = 0.95$

NB: usually interested in the shortest possible 95% credibility interval (also called Highest Density Interval).

Bayes Factor

Bayes Factor: marginal likelihood ratio between two hypotheses

$$BF_{10} = \frac{f(\boldsymbol{y}|H_1)}{f(\boldsymbol{y}|H_0)}$$

$\Rightarrow$ favored support for either hypothesis from the observed data $\boldsymbol{y}$

Posterior odds

$$\frac{p(H_1|\boldsymbol{y})}{p(H_0|\boldsymbol{y})} = BF_{10} \times \frac{p(H_1)}{p(H_0)}$$

Concentration de la loi *a posteriori*

Doob's convergence theorem

Posteriori distribution concentrate on the "real" valeu of the θ^* parameter when $n \to \infty$:

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

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

Normal approximation

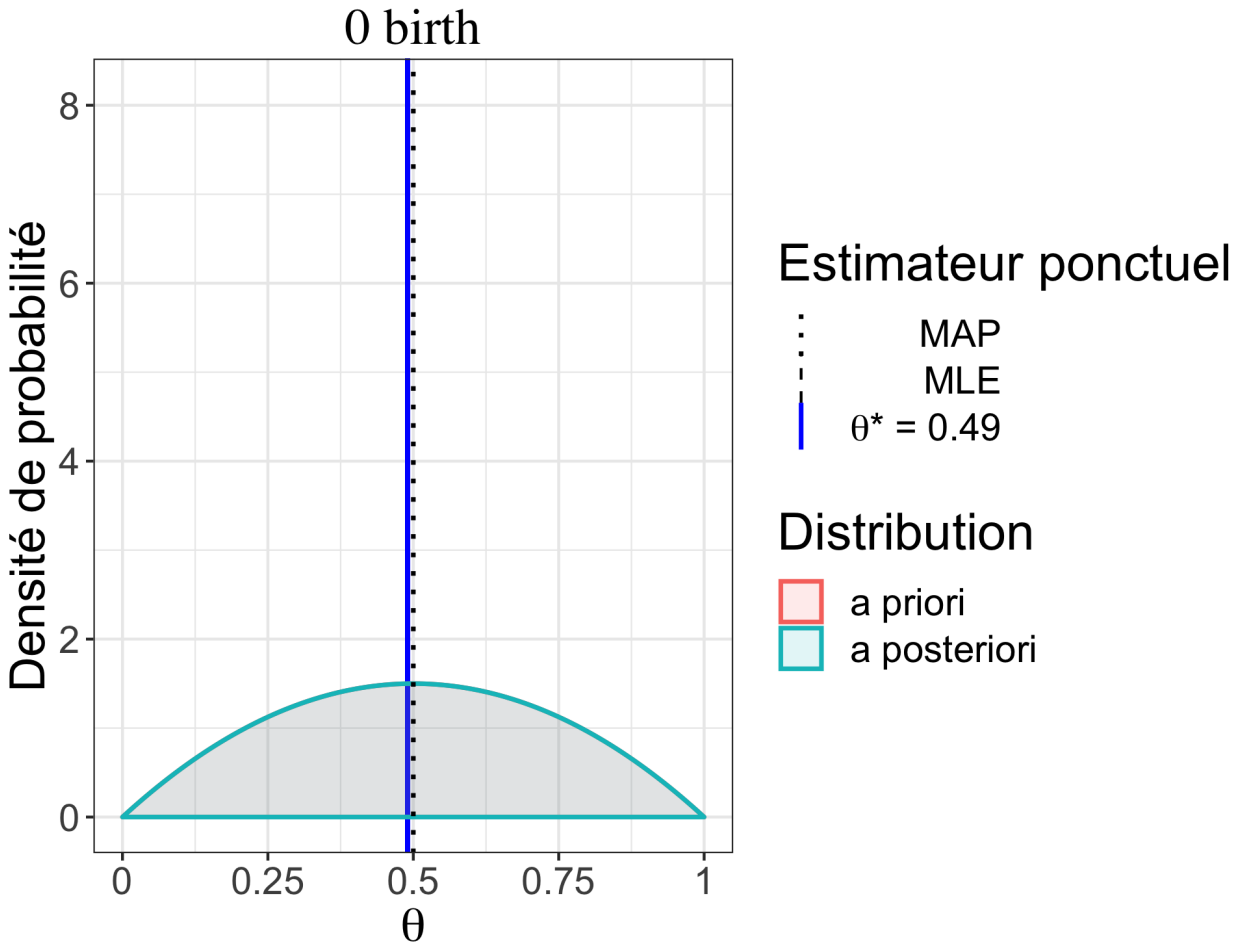
Bernstein-von Mises Theorem (or Bayesian central-limit theorem):
For a large n the *posterior* can be approximated by a normal distribution.

$$p(\theta|\boldsymbol{y}) \approx \mathcal{N}(\hat{\theta}, I(\hat{\theta})^{-1})$$

Consequences:

- Bayesian methods and frequentist procedures based on maximum likelihood give, for large enough n, very close results
- the *posterior* can be computed as a normal whose mean and variance we can calculate simply using the MAP

Illustration on the historical example



→ *Seeing Theory*, Brown University

Conclusion

Essential concepts

1. **Bayesian modeling**:

$$\theta \sim \pi(\theta) \quad \text{the } \textit{prior}$$

$$Y_i|\theta \overset{iid}{\sim} f(y|\theta) \quad \text{sampling model}$$

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

   with $p(\theta|\boldsymbol{y})$ the *posterior*, $f(\boldsymbol{y}|\theta)$ the likelihood (inherited from the sampling model), $\pi(\theta)$ the *prior* and $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\theta)\pi(\theta)$ is the marginal distribution of the data, i.e. the normalizing constant (with respect to θ)

3. **The *posterior* distribution is given by**:

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

4. ***Posterior* mean, MAP, and credibility intervals**

Practical use

The Bayesian framework is (just) another statistical tool for data analysis

Particularly **useful when:**

- few observations only are available
- there is important knowledge *a priori*

Like any statistical method, Bayesian analysis has advantages and disadvantages that will be more or less important depending on the application considered.

Questions ?