Introduction to Bayesian analysis for medical studies
—
Part I: Bayesian theory

Boris Hejblum
Université de Bordeaux, ISPED, Inserm BPH U1219/Inria SISTM, Bordeaux, France
boris.hejblum@u-bordeaux.fr
https://borishejblum.science

Graduate School of Health and Medical Sciences
at the University of Copenhagen

May 13th, 2019

Nice to meet you

First things first: **a show of hands**

- who has used R before?
- who knows what does Maximum Likelihood Estimator means ?
- who is afraid/uncomfortable with math formulas ?
- who knows what a regression/linear model is ?
- who has ever heard of random-effects before ?

$\Rightarrow$ What do you do, and what are your expectations from this course ?

Bayesian vocabulary

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

Course objectives

1. **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

2. **Study** and **perform** Bayesian analyses in **biomedical applications**:
    1. understand, discuss and reproduce a Bayesian (re-)estimation of a Relative Risk
    2. understand and perform a Bayesian meta-analysis using R
    3. understand and explain an adaptive design for Phase I/II trials and 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.

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

1% of the population is affected by this rare disease. A medical test has the following properties:

- if someone has the disease, its test will come out positive 99% of the time
- if someone does not have the disease, its test will come out negative 95% of the time

Given that someone got a positive result, what is his/her probability to have the disease ?

Bayes theorem: exercise

1% of the population is affected by this rare disease. A medical test has the following properties:

- if someone has the disease, its test will come out positive 99% of the time
- if someone does not have the disease, its test will come out negative 95% of the time

Given that someone got a positive result, what is his/her probability to have the disease ?

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

Bayes theorem: exercise

1% of the population is affected by this rare disease. A medical test has the following properties:

- if someone has the disease, its test will come out positive 99% of the time
- if someone does not have the disease, its test will come out negative 95% of the time

Given that someone got a positive result, what is his/her probability to have the disease ?

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

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

Bayes theorem: exercise

1% of the population is affected by this rare disease. A medical test has the following properties:

- if someone has the disease, its test will come out positive 99% of the time
- if someone does not have the disease, its test will come out negative 95% of the time

Given that someone got a positive result, what is his/her probability to have the disease ?

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

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

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*

$\Rightarrow$ 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 **18-19$^{\text{th}}$ centuries**
2. Galton & Pearson, then Fisher & Neymann ⇒ **frequentist** theory became dominant during the **20$^{\text{th}}$ century**
3. turn of the **21$^{\text{th}}$ 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}

Bayesian vs. Frequentists: an outdated debate

Fisher firmly rejected Bayesian reasoning

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

> *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 $\boldsymbol{Y} = (Y_1, \dots, Y_n)$
- we observe a sample $\boldsymbol{y} = (y_1, \dots, y_n)$

Refresher on frequentist modeling

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

⇒ **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

⇒ (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**

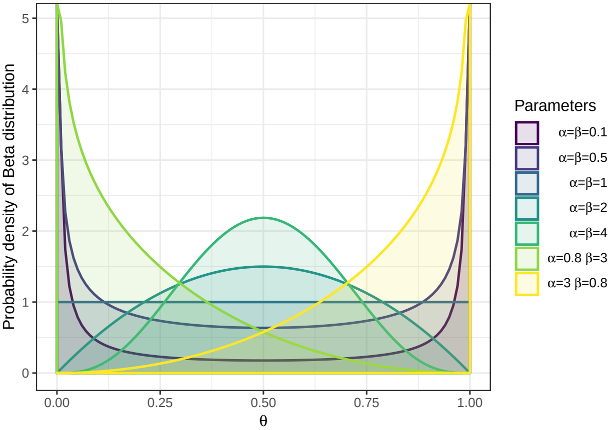
$$p(\theta|\boldsymbol{y}) = \frac{\theta^S(1-\theta)^{n-S}}{f(\boldsymbol{y})} = p(\theta|\boldsymbol{y}) = \binom{n}{S}(n+1)\theta^S(1-\theta)^{n-S}$$

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

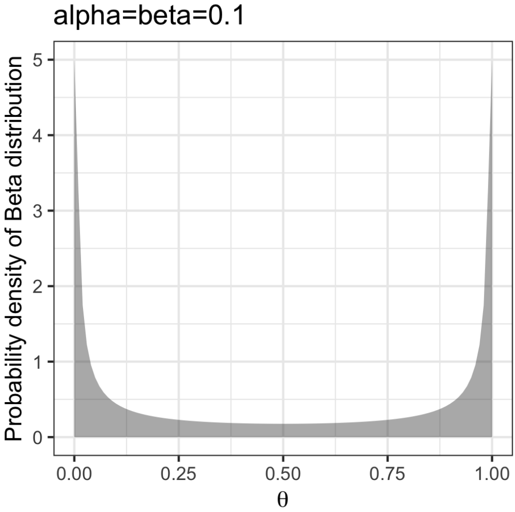
$$P(\theta \geq 0.5|\boldsymbol{y}) = \int_{0.5}^{1} p(\theta|\boldsymbol{y}) = \binom{n}{S}(n+1)\int_{0.5}^{1}\theta^S(1-\theta)^{n-S}\mathrm{d}\theta \approx 1.15\ 10^{-42}$$

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
5
4
3
2
1
0
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 *prosterior*: $p(\theta|\mathbf{y}) \propto \theta^{\alpha+S-1}(1-\theta)^{\beta+(n-S)-1}$

...

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

Conjugacy of the Beta distribution

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

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

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

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

Conjugacy of the Beta distribution

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

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

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

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	$P(\theta \geq 0.5 \mid \boldsymbol{y})$
#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

Impact of the *prior* choice

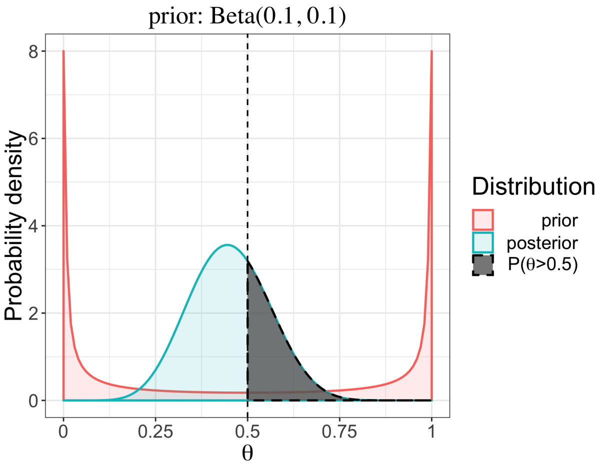
Interpretation of the *prior*	Parameters of the Beta distribution	$P(\theta \geq 0.5 \mid \boldsymbol{y})$
#boys > #girls	$\alpha = 0.1, \beta = 3$	$1.08\ 10^{-42}$
#boys < #girls	$\alpha = 3, \quad \beta = 0.1$	$1.19\ 10^{-42}$
#boys = #girls	$\alpha = 4, \quad \beta = 4$	$1.15\ 10^{-42}$
#boys ≠ #girls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

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

⇒ can **ease modeling** and **elicitation** of the *prior*...

Hyperprior in the historical example

Historical example of birth sex with a Beta *prior*
⇒ two Gamma hyper-*priors* for α and β (conjugated):

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

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*: $p(\theta|\mathbf{y}, \hat{\eta})$

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

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:

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

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

Bayesian inference

Bayesian Inference

Bayesian modeling ⇒ *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**

Sumary of this *posterior* distribution ?

- center
- spread
- ...

Decision theory

Context: estimating an unknown parameterθ

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

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

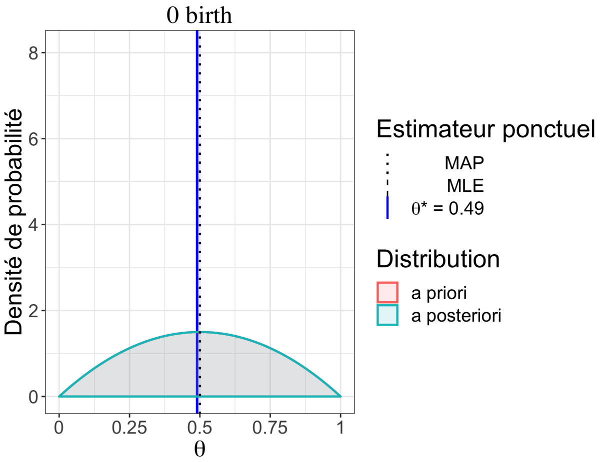
$\Rightarrow$ favored support for either hypothesis from the observed data $\boldsymbol{y}$

Posterior odds

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

Concentration of the posterior

Doob's convergence



→ *Seeing Theory*, Brown University

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

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 ?