1 The Bayesian approach to Statistics

1.1 Bayes' Theorem

The Bayesian approach to statistical inference is based on a particular interpretation of the content of the well-known Theorem of Bayes:

Theorem 1.1 *Let* $\{A_i, i = 1, \ldots, n\}$ *form a partition of the sample space* Ω *such that* $P(A_i) > 0$ *for all* $i = 1, \ldots, n$. *Let* B *be an event such that* $P(B) > 0$. *Then, for all* $i = 1, \ldots, n$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)\ P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)\ P(A_j)}$$

The use of this theorem in a deductive context, that of Probability Theory, is not controversial; $P(B \mid A_i)$ and $P(A_i)$ are assumed known and we want merely to compute $P(A_i \mid B)$

The possible controversy arises in an inductive context, that of Statistics, as an instrument of learning from an experiment:

- A_i denotes an hypothesis or a model that we use to explain a certain phenomenon; a theory to which the researcher attributes *a priori* a degree of credibility given by $P(A_i)$ — prior information
- B represents the result of observing that phenomenon
- $P(B \mid A_i)$ denotes the likelihood of observing B when explanation A_i is assumed correct — sampling information
- The prior probabilities $P(A_i)$ are then updated into posterior probabilities after B has been observed: $P(A_i \mid B)$
- This use of Bayes' theorem raises questions regarding the interpretation of the concept of probability involved in $P(A_i)$ and therefore in $P(A_i \mid B)$
- The frequentist interpretation is not flexible enough; we need to resort to its subjective interpretation

1.2 Bayesian methodology

We need to extend the classical notion of statistical model in order to introduce Bayesian methodology. In (parametric) Statistics, we have $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ as a collection of possible probabilistic models for the observable data $\boldsymbol{X}$; however,

- in frequentist Statistics, θ is unknown but treated as fixed
- in Bayesian statistics, all unknowns are regarded as random quantities because everything that is unknown is uncertain and all uncertainty must be quantified using the language of probability — probability distribution on the parameter space Θ denoted by $\pi(\theta)$ and referred to as prior distribution

$$\pi(\theta) - \text{prior distribution}$$

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) - \text{likelihood function}$$

$$\Downarrow$$

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\ \pi(\theta)}{\int_\Theta f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\ d\theta}, \quad \theta \in \Theta - \text{posterior distribution}$$

Remarks:

- $\pi(\theta)\ f(\boldsymbol{x} \mid \theta) = \pi(\theta, \boldsymbol{x})$ defines a joint distribution on $(\mathcal{X}, \Theta)$
- $m(\boldsymbol{x}) = \int_\Theta f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\ d\theta$ is the so-called prior predictive distribution of the data $\boldsymbol{x}$
- Another way of writing Bayes' theorem is $\pi(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta)\ \pi(\theta)$ where the normalization constant $m(\boldsymbol{x})$ is omitted

Example 1.1 *Suppose* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$ *and that a priori* $\theta \sim Be(a, b)$, $a, b > 0$ *known.*

Beta distribution: if $Y \sim \mathrm{Be}(a, b)$, then

$$f(y \mid a, b) = \frac{1}{B(a, b)} y^{a-1}\ (1-y)^{b-1}\ , \quad 0 < y < 1$$

where $B(a, b) = \Gamma(a)\ \Gamma(b)/\Gamma(a+b)$ is the beta function.

Then, with $t = \sum_{i=1}^n x_i$,

$$f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t(1-\theta)^{n-t}$$

and

$$\pi(\theta) = \frac{1}{B(a, b)} \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0, 1)\ .$$

We can do the calculations to conclude that

$$m(\boldsymbol{x}) = \frac{B(t+a, n-t+b)}{B(a, b)}$$

Hence,

$$\pi(\theta \mid \boldsymbol{x}) = \frac{1}{B(t+a, n-t+b)} \theta^{t+a-1}(1-\theta)^{n-t+b-1}$$

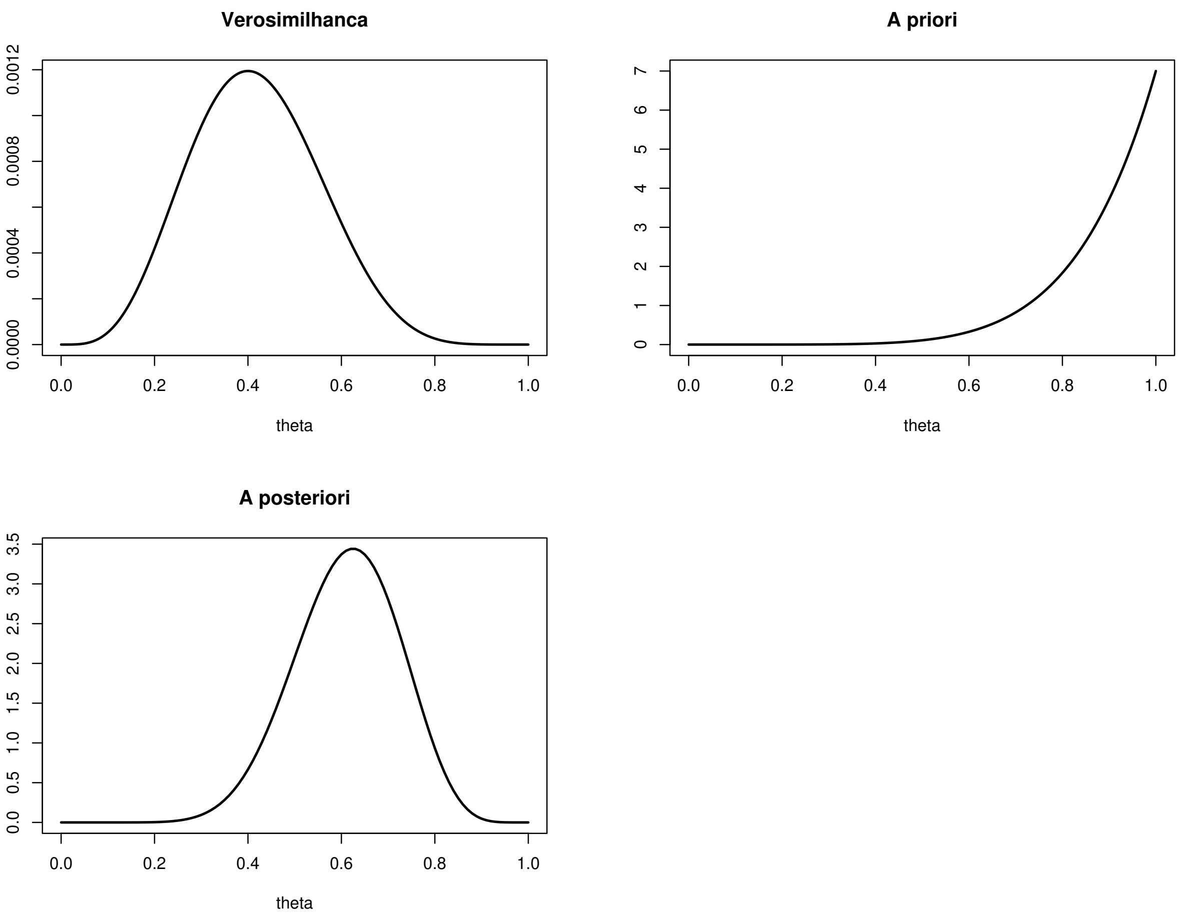
that is,

$$\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(t+a, n-t+b)$$

Example: $n = 10$, $t = 4$, $a = 7$, $b = 1$,

$$\theta \sim \mathrm{Be}(7, 1)$$

$$\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(11, 7)$$


Verosimilhanca
0.0000
0.0004
0.0008
0.0012
0.0
0.2
0.4
0.6
0.8
1.0
theta
A priori
0
1
2
3
4
5
6
7
theta
A posteriori
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
theta

Remarks:

1. If two likelihood functions are proportional, they lead to the same posterior distribution. Implications:
   (a) Bayesian inference only depends on observed data through the observed value of a sufficient statistic
   (b) $\pi(\theta \mid \boldsymbol{x}) = \pi(\theta \mid \boldsymbol{T}(\boldsymbol{x}))$ if $\boldsymbol{T}$ is sufficient for θ
   (c) (Bayesian inference respects the sufficiency principle)
   (d) Bayesian inference only depends on the statistical model through the likelihood function $L(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta)$
   (e) (Bayesian inference respects the likelihood principle)

Remarks (ctd):

2. $\pi(\theta \mid \boldsymbol{x})$, $\theta \in \Theta$, contains all the available information about θ, combining the data (through $L(\theta \mid \boldsymbol{x})$) with the prior information (in $\pi(\theta)$)

3. The Bayesian operation of combining knowledge has a sequential nature: Suppose that $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ with $\boldsymbol{X}_1 \amalg \boldsymbol{X}_2 \mid \theta$. Then,

$$\begin{aligned} \pi(\theta \mid \boldsymbol{x}) &= \frac{f(\boldsymbol{x} \mid \theta)\ \pi(\theta)}{\int f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\ d\theta} \\ &= \frac{f(\boldsymbol{x}_2 \mid \theta)\ \pi(\theta \mid \boldsymbol{x}_1)}{\int f(\boldsymbol{x}_2 \mid \theta)\ \pi(\theta \mid \boldsymbol{x}_1)\ d\theta} \end{aligned}$$

   That is: $\pi(\theta \mid \boldsymbol{x})$ can also be viewed as resulting from updating the "prior" $\pi(\theta \mid \boldsymbol{x}_1)$ with the likelihood $f(\boldsymbol{x}_2 \mid \theta)$

Example 1.2 *Suppose* $X_1, \ldots, X_n \mid \lambda \stackrel{iid}{\sim} Po(\lambda)$ *and that* a priori $\lambda \sim G(a, b)$, $a, b > 0$ *known, that is,*

$$\pi(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}, \quad \lambda > 0.$$

Then, with $t = \sum x_i$, we have

$$L(\lambda \mid \boldsymbol{x}) \propto \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \propto e^{-n\lambda} \lambda^t$$

$$\begin{aligned} \pi(\lambda \mid \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \lambda)\, \pi(\lambda) \propto e^{-n\lambda} \lambda^t \times \lambda^{a-1} e^{-b\lambda} \\ &\propto \lambda^{t+a-1}\, e^{-(n+b)\lambda} \\ &\propto \mathrm{G}(\lambda \mid t+a, n+b) \end{aligned}$$

and as a consequence we have that $\lambda \mid \boldsymbol{x} \sim \mathrm{G}(t+a, n+b)$.

Note that (Candidate's formula)

$$m(\boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\ \pi(\theta)}{\pi(\theta \mid \boldsymbol{x})} \quad \forall \theta \in \Theta$$

so that in this case we get that the prior predictive distribution of $\boldsymbol{X}$ is

$$m(\boldsymbol{x}) = b^a \frac{\Gamma(t+a)}{\Gamma(a)} \prod_{i=1}^{n} (x_i!)^{-1}\ (n+b)^{-(t+a)}$$

for $x_i \in \mathbb{N}_0$, $i = 1, \ldots, n$, $t = \sum x_i$.

1.3 Inference

How do we go about addressing inferential questions within the Bayesian framework?

- The complete answer to this question requires the introduction of Statistical Decision Theory ideas: action space, state space, loss function, etc
- However, in practical terms, the posterior distribution contains all the relevant information about θ, it's all a matter of finding its appropriate summary
- If the goal is to find a point estimate of θ, we can use as an estimate
  - the mode of $\pi(\theta \mid \boldsymbol{x})$, the posterior mode
  - the posterior mean $E(\theta \mid \boldsymbol{x})$
  - the posterior median, etc
- If the goal is to estimate θ by an interval, we can obtain $(a(\boldsymbol{x}), b(\boldsymbol{x}))$ such that $P(\theta \in (a(\boldsymbol{x}), b(\boldsymbol{x})) \mid \boldsymbol{x}) = 0.95$
- If the goal is to confront the statistical hypotheses $H_0: \ \theta \in \Theta_0$ and $H_1: \ \theta \in \Theta_1$, we need to compare $P(\Theta_0 \mid \boldsymbol{x})$ and $P(\Theta_1 \mid \boldsymbol{x})$

Prediction

- We observe $X_1, \ldots, X_n$ a random sample from $\{f(\cdot \mid \theta) : \theta \in \Theta\}$
- a prior on θ is set and the posterior $\pi(\theta \mid \boldsymbol{x})$ is computed
- we wish to predict an outcome Y whose probability distribution depends on θ

Determine the probability distribution of $Y \mid \boldsymbol{x}$, the posterior predictive distribution of Y

$$
\begin{aligned}
f(y \mid \boldsymbol{x}) &= \int_{\Theta} f(y, \theta \mid \boldsymbol{x}) \; d\theta \\
&= \int_{\Theta} f(y \mid \theta, \boldsymbol{x}) \; \pi(\theta \mid \boldsymbol{x}) \; d\theta \\
&= \int_{\Theta} f(y \mid \theta) \; \pi(\theta \mid \boldsymbol{x}) \; d\theta \;\; \text{if } Y \amalg \boldsymbol{X} \mid \theta
\end{aligned}
$$

Example 1.3 $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$; *a priori* $\theta \sim Be(a, b)$, $a, b > 0$ *known.*

We know that $\theta \mid \boldsymbol{x} \sim \mathrm{Be}(a + t, b + n - t)$. Suppose we want to predict the outcome of the next observation, independent of the previous, X_{n+1}. Then,

$$\begin{aligned} f(x_{n+1} \mid \boldsymbol{x}) &= \int_0^1 f(x_{n+1} \mid \theta)\ \pi(\theta \mid \boldsymbol{x})\ d\theta \\ &= \frac{B(a + t + x_{n+1}, b + n - t + 1 - x_{n+1})}{B(a + t, b + n - t)}, \quad x_{n+1} = 0, 1\ . \end{aligned}$$

It would be simpler to use the formula of the iterated expectation:

$$P(X_{n+1} = 1 \mid \boldsymbol{x}) = E[E_\theta[X_{n+1} \mid \theta, \boldsymbol{x}]\boldsymbol{x}] = E[\theta \mid \boldsymbol{x}] = \frac{a + t}{a + b + n}$$

1.4 The prior distribution

Bayesian inference is conceptually very simple and particularly intuitive. However, its practical implementation is often considered difficult:

- $\pi(\theta)$ should reflect information about θ available before the data $\boldsymbol{x}$ are observed. To summarize information that in general will exist in an non-organized fashion in a probability distribution is not trivial
- What should one do when said information is vague or diffuse?
- What if the goal is to produce a statistical analysis which is as "objective" as possible, e.g. one that uses little prior information about θ?
- Calculations: Very rarely will $\pi(\theta \mid \boldsymbol{x})$ exist in closed form, as $m(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\ d\theta$ will not be computable analytically
- The answer to many inferential questions will involve the calculation of $E[\psi(\theta) \mid \boldsymbol{x}]$ for different $\psi(\theta)$

“Solutions”:

- prior distributions which allow analytical calculations
- “non-informative” prior distributions
- Simulation, analytic approximations, numerical calculations

1.4.1 Conjugate prior distributions

Families of prior distributions which allow for analytical calculations.

Example 1.4 *Suppose* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$*; a priori* $\theta \sim Be(a, b)$*,* $a, b > 0$ *known.*

We saw that

$$\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(t + a, n - t + b)$$

that is, the updating is done within the same family of distributions:

$$(a, b) \longrightarrow (t + a, n - t + b)$$

Definition 1.1 *The family* $\Pi = \{\pi(\cdot \mid \tau) : \tau \in \Gamma\}$ *is said to be natural conjugate of the statistical model* $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ *if*

1. $\forall \tau_0, \tau_1 \in \Gamma \; \exists \tau_2 \in \Gamma$*:*

$$\pi(\theta \mid \tau_0)\; \pi(\theta \mid \tau_1) \propto \pi(\theta \mid \tau_2)$$

2. $\exists \tau_0 \in \Gamma : f(\boldsymbol{x} \mid \theta) \propto \pi(\theta \mid \tau_0)$

Consequence:

$$\begin{aligned}
\pi(\theta \mid \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \theta)\; \pi(\theta \mid \tau_1) \\
&\propto \pi(\theta \mid \tau_0)\; \pi(\theta \mid \tau_1) \\
&\propto \pi(\theta \mid \tau_2) \in \Pi
\end{aligned}$$

Example 1.5 *Suppose* $X_i,\ i = 1, \ldots, n \overset{iid}{\sim} Po(\lambda)$.

Then, with $t = \sum x_i$,

$$\begin{aligned} f(\boldsymbol{x} \mid \lambda) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &\propto \lambda^t e^{-n\lambda} \\ &\propto \mathrm{G}(\lambda \mid t+1, n) \end{aligned}$$

Also, $\mathrm{G}(\lambda \mid a, b) \times \mathrm{G}(\lambda \mid c, d) \propto \mathrm{G}(\lambda \mid a+c-1, b+d)$. Hence, the gamma family is the natural conjugate of the Poisson model. The prior-to-posterior update is $(a, b) \to (a+t, b+n)$.

Choosing (a, b):

- Set $E(\theta) = \mu_0$ and $\mathrm{Var}(\theta) = \sigma_0^2$ subjectively. Then solve $a/b = \mu_0$ and $a/b^2 = \sigma_0^2$.
- $\mathrm{G}(a, b)$ contains the same information as an imaginary sample of “size” b and sample total a:
$$(a, b) \to (a + t, b + n)$$
- Treat a, b as unknown and place a prior on them, $\pi(a, b)$ - hierarchical prior

Drawbacks:

- Conjugate family does not always exist
- Functional form is chosen for convenience and it may have important consequences

1.4.2 Non-informative priors

- Situations where there is no considerable prior information
- Obtain posterior beliefs in situations where the sampling information should overwhelm the prior information
- Obtain a “reference” analysis, an “objective” analysis which can be compared with subjective ones as a way of ascertaining the influence of the prior information
- Research area called “Objective Bayes” — methods or strategies to obtain “objective” priors in various situations which are then evaluated

Bayes-Laplace method

Principle of insufficient reason of Bayes-Laplace: in the absence of any reason to consider that two probabilities are different, they should be considered equal.

Consequences:

- Θ finite, $\Theta = \{\theta_1, \ldots, \theta_k\}$, then $\pi(\theta_i) = 1/k$, $i = 1, \ldots, k$
- If Θ is countable, there is no probability distribution which is compatible with this principle: $\pi(\theta) = c$, $\theta \in \{\theta_1, \ldots, \theta_k, \ldots\}$ implies that $\sum_{\theta \in \Theta} \pi(\theta) = +\infty$: It's an **improper** distribution
- The formal use of Bayes' theorem with an improper prior is controversial; however, it's often utilized as long as the resulting posterior is proper
- Θ not countable: $\pi(\theta) \propto c$, $\theta \in \Theta$ is improper unless Θ is bounded

Most important objection to uniform priors:

Example 1.6 $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$.

The Bayes-Laplace prior would be $\pi(\theta) = 1$, $\theta \in (0, 1)$. An alternative parameterization of the Bernoulli model is in terms of $\psi = \ln[\theta/(1-\theta)]$. The induced distribution in ψ is

$$\pi(\psi) = \frac{e^{\psi}}{(1 + e^{\psi})^2}, \ \psi \in \mathbb{R}$$

Ignorance about θ implies some information about ψ!

In general, with $\theta = g(\psi)$,

$$\pi(\psi) = |g'(\psi)|\pi(g(\psi))$$

Jeffreys Method

Idea: invariance with respect to reparametrizations.

Let $\theta = g(\psi)$ and denote by $I_X(\theta)$ the Fisher information about θ in X. Then, the Fisher information about ψ in X is

$$I_X^*(\psi) = [g'(\psi)]^2 \; I_X(g(\psi)) \; .$$

If *a priori*

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

then the induced prior on ψ is

$$\begin{aligned} \pi(\psi) &= |g'(\psi)| \; \pi(g(\psi)) \\ &= |g'(\psi)| \sqrt{I_X(g(\psi))} \\ &= \sqrt{I_X^*(\psi)} \end{aligned}$$

It does not matter to which parameterization we apply the rule!

Example 1.7 $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$

Recall that $I_X(\theta) = E_\theta[-d^2 \ln f(X \mid \theta)/d\theta^2]$. Hence,

$$I_X(\theta) = E_\theta[X/\theta^2 - (1 - X)/(1 - \theta)^2] = \theta^{-1}(1 - \theta)^{-1}$$

and so

$$\pi^J(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2} \propto \mathrm{Be}(\theta|1/2, 1/2)$$

Example 1.8 $X_1, \ldots, X_n \mid \mu \overset{iid}{\sim} N(\mu, 1)$

Easy calculations show that $I_X(\mu) = 1$, so,

$$\pi^J(\mu) \propto c, \quad \mu \in \mathbb{R}$$

which is an improper distribution. However, the formal use of Bayes' Theorem leads to

$$\mu \mid x_1, \ldots, x_n \sim \mathrm{N}(\bar{x}, 1/n)$$