I - Introduction

Consider the following simple problem:

A bag contains a ball of unknown colour, which may be either black or white. A white ball is added to the bag, then a ball is drawn at random from it. The drawn ball happens to be white. What is the colour of the ball that was initially in the bag?

A natural way to solve this problem is to proceed as follows:

- Let C_I be the colour of the ball that was initially in the bag and assume that $\mathbb{P}(C_I = w) = \mathbb{P}(C_I = b) = 1/2$.
- Denoting by C_D the colour of the drawn ball, we have

$$\mathbb{P}(C_I = w | C_D = w) = \frac{2}{3}.$$

- Then, we answer to the above question by saying that with probability 2/3 a white ball was initially in the bag.

Lessons from this problem

Most of you would have probably performed the same computations. However, these latter contain the two main ideas underlying Bayesian inference:

1. Prior uncertainty/prior beliefs ("Of which colour is the initial ball?") can be expressed in terms of probabilities:

$$\mathbb{P}(C_I = w) = \mathbb{P}(C_I = b) = 1/2.$$

2. Taking into account any new information that arises from the experiment ("the drawn ball is white") can be done by writing conditional probabilities:

$$\mathbb{P}(C_I = w | C_D = w) = 2/3.$$

The fact that we accept almost unnoticeably these concepts prove they are natural and convenient. Therefore we are all Bayesian!

Frequentist versus Bayesian interpretation of probabilities

From a frequentist perspective the probability of an event is the limit of its relative frequency as the number of trials goes to infinity.

From a Baysian perspective the probability of an event represents our reasonable expectation about it in a single trial.

To illustrate this difference of interpretation let X denotes the output of the roll of a fair dice (with 6 sides numbered from 1 to 6). Then, both schools agree that $\mathbb{P}(X = 1) = 1/6$ but

- For the frequentist school $\mathbb{P}(X = 1) = 1/6$ because, if we roll the dice n times and denote by x_i the outcome of the i-th roll, we have $\frac{1}{6} = \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{1\}}(x_i)$.
- For the Bayesian school $\mathbb{P}(X = 1) = 1/6$ because all the 6 sides of the dice are equally likely to face upwards (since the dice is fair).

The frequentist view of the introductory example

Since C_I is deterministic (i.e. the color of the ball that was initially in the bag is maybe unknown but whether it is white or black is now part of the 'state of the world') the above computations do not make any sense from a frequentist point of view.

Bayesian formulae

Bayesian statistics is named after Reverend Thomas Bayes (1701-1761), who discovered the Bayes formula:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)},$$

where A and B are two events.

This formula allows for deducing the relation $A \to B$ (A gives B) from its opposite $B \to A$ (B gives A) and from the prior information $\mathbb{P}(B)$.

Actually, Thomas Bayes proved a continuous version of the above formula: Given random variables X and Y, with conditional distribution $f(x|y)$ and marginal distribution $g(y)$, the conditional distribution of y given x is

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y')g(y')\mathrm{d}y'}.$$

Historical remark: Pierre Simon de Laplace (1749-1827) rediscovered independently Bayes formula and can be considered as the second Bayesian scientist in History.

The Bayesian viewpoint in Statistical Science

Bayes and Laplace went one step further and consider that the uncertainty about the parameter θ of a parametric model $\{f(\cdot|\theta),\, \theta \in \Theta\}$ can be expressed through a probability distribution $\pi(\theta)$ on Θ.

Then, it is possible to apply the Bayes formula to the couple (θ, x):

$$\boxed{\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{m(x)}}$$

Bayesian terminology:

1. $\pi(\theta)$ is called prior distribution of θ,
2. $f(x|\theta)$ is called the likelihood of the model,
3. $\pi(\theta|x)$ is called the posterior distribution of θ,
4. $m(x) = \int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta$ is called the marginal density of x.

We can now provide a formal definition of Bayesian models:

Definition 1.1 *A Bayesian statistical model consists of a parametric statistical model* $\{f(\cdot|\theta),\, \theta \in \Theta \subseteq \mathbb{R}^d\}$ *and a prior distribution* $\pi(\theta)$ *on* Θ.

Example: Laplace measuring the mass of Saturn

Around 1812-1816 Laplace used Bayes formula to estimate the mass of Saturn given the available astronomical knowledge and data.

More precisely, he estimated the mass of Saturn applying the Bayes formula

$$\mathbb{P}(M|D,I) = \frac{\mathbb{P}(D|M,I)\mathbb{P}(M|I)}{\mathbb{P}(D|I)}$$

where

- M is the mass of Saturn
- D is the data from various measurements of planetary orbits
- I is the existing background information, e.g., Newtons laws of celestial mechanics.

In this example M is the parameter of interest, $\mathbb{P}(D|M,I)$ the likelihood and $\mathbb{P}(M|I)$ the prior distribution of M.

Remarkably, Laplace's estimated value of the mass of Saturn only differs from the modern value measured by the NASA by about 0.5%.

Remark: Probabilistic statements about M, the mass of Saturn, do not have any sense from a frequentist perspective.

General framework

In this course we will often replace distribution with probability density function (pdf), assuming that this latter is well defined with respect to a dominating measure (e.g. Lebesgue or counting measure).

In particular,

- The dominating measure for the model $\{f(\cdot|\theta),\, \theta \in \Theta \subseteq \mathbb{R}^d\}$ will be denoted by $\mathrm{d}x$;
- The dominated measure for the prior distribution will be denoted by $\mathrm{d}\theta$.

We denote by $\mathcal{X} \subseteq \mathbb{R}^k$ the observation space so that, for any $\theta \in \Theta$, $f(\cdot|\theta)\mathrm{d}x$ is a probability distribution on $\mathcal{X}$.

Following the standard convention, we will use capital letters for random variables and small letters for their realizations. For instance, X is a random variable and x is a realization of X.

However, in the Bayesian literature this standard convention is not used for the parameter θ (the notation Θ is already used for the parameter space!). Hence, throughout this course the notation θ is used for both the random variable and its realizations.

Back to the introductory example

Putting the introductory example into our general framework yields

- $\Theta = \{w, b\}$
- $\mathcal{X} = \{w, b\}$ and $X = C_D$
- $\{f(\cdot|\theta),\, \theta \in \Theta\}$ defined by

$$f(x|w) = \begin{cases} 1, & x = w \\ 0, & x = b \end{cases}, \quad f(x|b) = \begin{cases} \frac{1}{2}, & x = w \\ \frac{1}{2}, & x = b \end{cases}$$

- $\pi(\theta) = \frac{1}{2}\mathbf{1}_{\{w\}}(\theta) + \frac{1}{2}\mathbf{1}_{\{b\}}(\theta)$
- $\pi(\theta|x) = \mathbb{P}(C_I = \theta | C_D = x)$
- $\mathrm{d}x = \delta_{\{w\}}(\mathrm{d}x) + \delta_{\{b\}}(\mathrm{d}x)$
- $\mathrm{d}\theta = \delta_{\{w\}}(\mathrm{d}\theta) + \delta_{\{b\}}(\mathrm{d}\theta)$

Direct application of Bayes formula gives

$$\pi(w|x) = \frac{\pi(w) f(x|w)}{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{\frac{1}{2}\mathbf{1}_{\{w\}}(x)}{\frac{1}{2}f(x|b) + \frac{1}{2}f(x|w)} = \frac{\frac{1}{2}\mathbf{1}_{\{w\}}(x)}{\frac{1}{4} + \frac{1}{2}\mathbf{1}_{\{w\}}(x)}$$

and thus, as per above (!),

$$\mathbb{P}(C_I = w | C_D = w) = \pi(w|w) = \frac{\frac{1}{2}}{\frac{1}{4} + \frac{1}{2}} = \frac{2}{3}.$$

Some remarks about the general framework

The notation x for the observation and $f(x|\theta)$ for the likelihood that we will use throughout this course is quite general.

For instance, consider n observations $x_1, \ldots, x_n$. Then, we have $x := (x_1, \ldots, x_n)$ and

- If the observations $x_1, \ldots, x_n$ are modelled as n i.i.d. random variables taking values in $\mathcal{X}_1$ and if $\{\tilde{f}(\cdot|\theta),\, \theta \in \Theta \subseteq \mathbb{R}^d\}$ is the parametric model for observation x_1, then we have

$$\mathcal{X} = \mathcal{X}_1^n, \quad f(x|\theta) = \prod_{k=1}^{n} \tilde{f}(x_k|\theta).$$

- More generally, if the parametric model for x is defined by

$$X_1|\theta \sim \tilde{f}_1(x_1|\theta)\mathrm{d}x_1$$
$$X_k|\big(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, \theta\big) \sim \tilde{f}_k(x_k|x_1, \ldots, x_{k-1}, \theta)\mathrm{d}x_k$$

with $\theta \in \Theta$ and $x_k \in \mathcal{X}_k$, then we have

$$\mathcal{X} = \prod_{k=1}^{n} \mathcal{X}_k, \quad f(x|\theta) = \tilde{f}_1(x_1|\theta) \prod_{k=2}^{n} \tilde{f}_k(x_k|x_1, \ldots, x_{k-1}, \theta).$$

Example: Gaussian model with unknown mean

Let $\phi(\cdot)$ be the density of the $\mathcal{N}_1(0,1)$ distribution.

Proposition 1.1 *Let $(\sigma, \sigma_0, \mu_0) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}$ and consider the Bayesian statistical model defined by*

$$\pi(\theta) = \frac{1}{\sigma_0}\phi\left(\frac{\theta - \mu_0}{\sigma_0}\right), \quad f(x|\theta) = \prod_{k=1}^{n} \frac{1}{\sigma}\phi\left(\frac{x_k - \theta}{\sigma}\right),$$

where $n \in \mathbb{N}$ and $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Then,

$$\pi(\theta|x) = \frac{1}{\sigma_n}\phi\left(\frac{\theta - \mu_n}{\sigma_n}\right)$$

with

$$\mu_n = \lambda_n \mu_0 + (1 - \lambda_n)\bar{x}_n, \quad \sigma_n^2 = \frac{\sigma^2}{n_0 + n}$$

where

$$\bar{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k, \quad \lambda_n = \frac{n_0}{n_0 + n} \in (0,1), \quad n_0 = \frac{\sigma^2}{\sigma_0^2}.$$

Proof: Done in class

Comments:

- The posterior mean μ_n is a weighted average between the empirical mean $\bar{x}_n$ and the prior mean μ_0. The posterior distribution is therefore a compromise between the prior information and the information carried by the observations.
- When n is large, $\mu_n \approx \bar{x}_n$ and $\sigma_n^2 \approx \sigma^2/n$ so that the impact of the prior distribution becomes less important as n increases.

Gaussian model with unknown mean

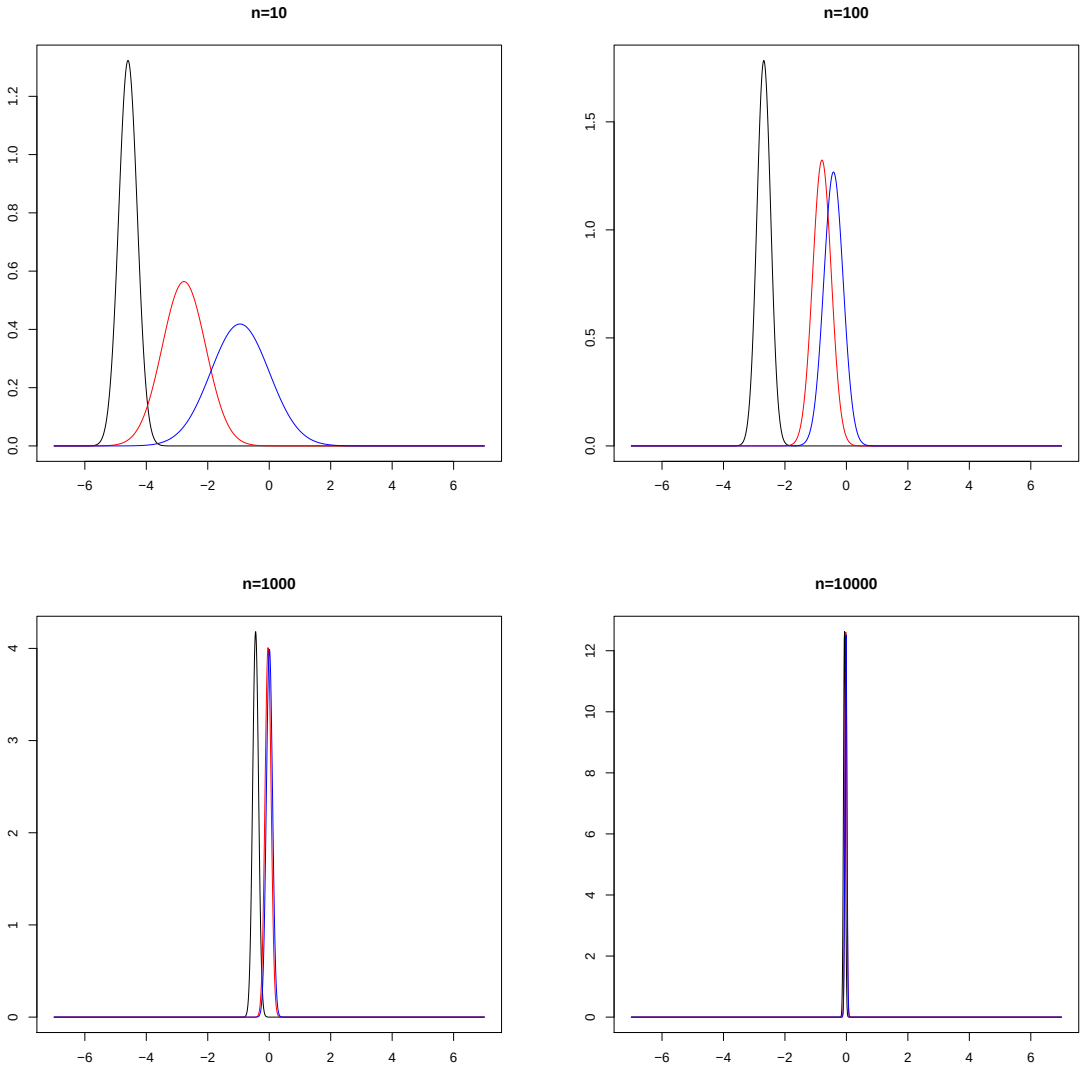

Figure 1: Posterior distribution for $n \in \{10, 100, 1\,000, 10\,000\}$ and for $\sigma_0^2 = 0.1$ (black), $\sigma_0^2 = 1$ (red) and $\sigma_0^2 = 10$. The observations are realizations of i.i.d. $\mathcal{N}_1(0, 10)$ random variables and $\mu_0 = 5$.

General comments about Bayesian inference

1. In the Bayesian framework, all the inference about θ is carried out
   (a) Conditionally on the observation x;
   (b) Uniquely through the posterior distribution $\pi(\theta|x)$.

   In particular,
   - Point estimators of θ are derived from $\pi(\theta|x)$ and are justified only for the current observation x (**Chapter 2**).
   - Hypotheses testing (**Chapter 4**) and model choice (**Chapter 5**) are fully based on $\pi(\theta|x)$ (and not on some asymptotic arguments).
2. As illustrated with the Gaussian example, for a fixed sample size n the choice of the prior distribution can have an important impact on the posterior distribution. The choice of a "good" prior distribution will be the object of **Chapter 3**.
3. As illustrated with the Gaussian example, as the number of observations n increases the impact of a fixed prior distribution becomes less and less important. In particular, as $n \to +\infty$, the posterior distribution concentrates around the true parameter value. Results on Bayesian asymptotics will be presented and derived in **Chapter 6**.

Computational aspects

By definition,

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}, \quad m(x) = \int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta$$

where often (but not always!) $f(x|\theta)$ and $\pi(\theta)$ are known.

Assuming this is the case, three scenarios are possibles:

- The quantity $m(x)$ can be explicitly computed. This happens only when the model $\{f(\cdot|\theta),\, \theta \in \Theta\}$ admits a conjugate prior.
- The quantity $m(x)$ cannot be explicitly computed but can be approximated by a quadrature rule. In this case,

  $$m(x) \approx \sum_{i_1}^{N} w_1^{i_1} \sum_{i_2}^{N} w_2^{i_2} \cdots \sum_{i_d}^{N} w_d^{i_d} f(x|\theta_1^{i_1}, \ldots, \theta_d^{i_d})\pi(\theta_1^{i_1}, \ldots, \theta_d^{i_d})$$

  where the $w_j^{i_j}$'s, the $\theta_j^{i_j}$'s and N depend on the quadrature rule and on the desired level of approximation. This method is only applicable when d is small (say $d \leq 2$) since the cost to reach a fixed approximation error is exponential in d.
- The quantity $m(x)$ cannot be explicitly computed and $d > 2$ (say). In this case, we can use Monte Carlo methods to approximate $\pi(\theta|x)$; that is, we generate a random weighted sample $(W_N^1, \theta_N^1), \ldots, (W_N^N, \theta_N^N)$ such that

  $$\lim_{N\to+\infty} \sum_{i=1}^{N} W_N^i \delta_{\theta_N^i}(\mathrm{d}\theta) = \pi(\theta|x)\mathrm{d}\theta, \quad \text{(almost surely)}.$$

  In **Chapter 7** we will introduce the Metropolis-Hastings algorithm which is arguably the most popular Monte Carlo algorithm to approximate the posterior distribution in this scenario (which is from far the most frequent in practice).

Why do we need Bayesian statistics?

- As Bayesian statistics is based on non-asymptotic arguments it is particularly relevant when the sample size is small (e.g. because collecting observations is costly).
- In some cases (e.g. when the sample size is small) we really want to incorporate some prior information in the analysis.
- Often in science it is easy to predict the outcome given the cause while deducting the cause of a given outcome is much harder. Bayesian theory provides a unified framework to address this problem.
- Frequentist methods are justified as the number of observations goes to infinity and/or according to their average performance over an infinite number of samples. In some cases there exist only a limited number of observations (e.g. there exists only one planet Saturn) so that this kind of justifications does not make any sense.
- Computational reasons: As we will see in Chapters 2 and 6 Bayesian methods have very good frequentist properties while, in some cases, computing Bayesian estimators is easier than computing frequentist estimators.
- In some cases taking a Bayesian approach facilitates the modelling process because specifying the joint distribution $f(x, \theta)$ is simpler than specifying the conditional distribution $f(x|\theta)$. The typical example where this is true is when the observations are similar but not identical (see **Chapters 8 and 9**).
- Philosophical reasons...

A second historical example (Laplace, 1786)

In Paris, $n_m = 251\,527$ *male births and* $n_f = 241\,945$ *female births was recorded in 1786. Using these data, Laplace wanted to test if the probability* $\theta \in \Theta := [0, 1]$ *of a male birth is above* $1/2$.

- Let $x = n_m$ be the observation and $n = n_m + n_f$. Then, the underlying statistical model can be stated as follows:

$$X \sim \text{Bin}(n, \theta).$$

- Laplace had no substantial prior information on the value of θ. Therefore he decided to assign to θ an uniform prior distribution on Θ,

$$\pi(\theta) = \mathbf{1}_{[0,1]}(\theta),$$

  in order to represent the idea that 'a priori all values of θ are equally likely'.

- The corresponding posterior distribution of θ given $X = x$ is the $\text{Beta}(x + 1, n - x + 1)$ distribution; that is

$$\pi(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x}, \quad \theta \in [0, 1].$$

- Using some numerical integration routine, we get

$$\pi(\{\theta : \theta > 1/2\}|x) = 1 - \int_0^{0.5} \pi(\theta|x)\mathrm{d}x \approx 1 - 1.15 \times 10^{-42}.$$

- From this result Laplace deduced that θ is more likely to be above $1/2$.