One-parameter models

General approach to Bayesian data modelling

Teacher: Matilde Trevisani

DEAMS

A.A. 2025/2026
(aggiornato: 2025-10-03)

Agenda (about 3 lectures)

One-parameter models

- General approach to Bayesian data modelling
- A first example
- Note on accumulation of evidence
- Binomial model
- Note on impact of more evidence
- Summarizing posterior distributions
- Conjugacy
- Interplay between priors and data
- Normal model
- Poisson model
- Other models

General approach to Bayesian data modelling

A ***Bayesianly justifiable*** analysis is one that

> “treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes’ theorem.”
>
> -- *Rubin (1984, p. 1152)*

3-Step General Approach to Bayesian Modeling

1. Set up the **full probability model**: the joint distribution of all entities, including observables (y) and unobservables (θ) in accordance with all that is known about the problem

$$p(y, \theta) \propto p(y \mid \theta) p(\theta)$$

  - $p(y \mid \theta) \propto L(\theta) = (L(\theta, y))$ is the model for the conditional probability of the data, that is (proportional to) the **likelihood**
  - $p(\theta)$ is the **prior** distribution for the unknown parameters, reflecting what is believed about the situation

2. Condition on the observed data (x), calculate the conditional probability distribution for the unobservable entities (θ) of interest given the observed data: the **posterior** distribution

$$p(\theta \mid y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y \mid \theta) p(\theta)}{p(y)} \propto p(y \mid \theta) p(\theta)$$

3. Examine fit of the model, tenability/sensitivity of assumptions, reasonable conclusions?, respecify, summarize results, etc.

See ambiguous notation

Notation

The main characters:

- We denote with Greek letters, typically, θ, the parameter(s), **unobservable** quantities. θ can be a scalar or a vector.
- The **observed** data are denoted by y, if data are gathered on n units:

  $$y = (y_1, \ldots, y_n)$$

  where y_i can be a scalar or a vector (if more than one variable is observed on each unit). y can then be a scalar, a vector, or a matrix.
- We will also use unknown but **potentially observable** quantities, that is, future observations, these will be denoted as $\tilde{y}$.
- If covariates are available, these will be denoted by x.

Model Specification

Specifying a **Bayesian model** means specifying:

- The distribution of y conditional on the parameter θ: $y \mid \theta \sim p(y \mid \theta)$
- The prior distribution on θ: $\theta \sim p(\theta)$

Putting these together, we have specified the **joint distribution of (y, θ)**:

$$p(y, \theta) = p(y \mid \theta) p(\theta)$$

and we can obtain the **marginal distribution of y** as:

$$p(y) = \int_\Theta p(y, \theta) d\theta = \int_\Theta p(y \mid \theta) p(\theta) d\theta$$

Posterior distribution

Inference on θ will be based on the posterior distribution, which is derived through a straightforward application of Bayes theorem

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

The posterior distribution contains all the information on θ we have (from the data and prior to observing the data).

The work will have to do is to understand

- how to summarize the information in $p(\theta|y)$, to obtain for instance point and interval estimates or to perform hypotheses testing;
- how to explore the distribution, but for simple examples $p(y)$ is difficult to derive (impossible to derive analytically), so exploration of the posterior will be based on computational machinery (MCMC and other stuff) whose starting point is

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Predictive Distribution

We are sometimes interested in "unknown but potentially observable quantities" $\tilde{y}$ (e.g., prediction of y on new statistical units).

We assume that they behave like the data y, that is:

$$\tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta)$$

Hence, unconditionally, the distribution of $\tilde{y}$ is:

$$p(\tilde{y}) = \int_\Theta p(\tilde{y} \mid \theta) p(\theta) d\theta$$

which is the same as y. This is also called the ***prior predictive distribution***. After the data y have been observed, we can compute the ***posterior predictive distribution***:

$$p(\tilde{y} \mid y) = \int_\Theta p(\tilde{y}, \theta \mid y) d\theta = \int_\Theta p(\tilde{y} \mid \theta, y) p(\theta \mid y) d\theta = \int_\Theta p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta$$

where we note that the conditional iid assumption implies that:

$$p(\tilde{y} \mid \theta, y) = p(\tilde{y} \mid \theta).$$

Exchangeability

A common hypothesis in statistical inference is that observations are independent and identically distributed (*iid*), meaning we collect $y_1, \ldots, y_n$ and assume these are *iid*.

In Bayesian inference, where the inference process is fully probabilistic. independence of observations would imply that we cannot learn about future observations from past ones (since y_{n+1} would be independent of $y_1, \ldots, y_n$).

Instead, we assume observations are ***exchangeable***, meaning the joint distribution of $(y_1, \ldots, y_n)$ is invariant to index permutations:

$$p(y_1, \ldots, y_n) = p(y_{i_1}, \ldots, y_{i_n})$$

for any permutation $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$.

Exchangeability and Conditional Independence

We will usually specify the model assuming that

- $y_1, \ldots, y_n$ are iid conditional on θ
- $\theta \sim p(\theta)$

This **implies** that $y_1, \ldots, y_n$ are exchangeable. In fact, consider the unconditional distribution:

$$\begin{aligned} p(y_{i_1}, \ldots, y_{i_n}) &= \int p(y_{i_1}, \ldots, y_{i_n} \mid \theta) p(\theta)\, d\theta \\ &= \int \prod_{j=1}^{n} p(y_{i_j} \mid \theta) p(\theta)\, d\theta \\ &= \int \prod_{i=1}^{n} p(y_i \mid \theta) p(\theta)\, d\theta = p(y_1, \ldots, y_n) \end{aligned}$$

de Finetti's Theorem

For **binary** variables $y_1, \ldots, y_n$, exchangeability is equivalent to conditional *iid*:

Theorem (de Finetti): Let $Y_1, Y_2, \ldots, Y_n$, $n \to \infty$, be a sequence of Bernoulli r.v., then they are exchangeable if and only if there exists a random variable θ valued in $[0, 1]$ such that:

$$p(y_1, \ldots, y_n) = \int_0^1 \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} dP(\theta).$$

An extension of this theorem exists for **general** random variables.

Non independence as θ unknown

The following are equivalent:

- $y_1, \ldots, y_n$ are exchangeable.
- $y_1, \ldots, y_n$ are *iid* conditional on θ.

This means:

> Observations are IID if we know the data-generating mechanism.

Since we do not know it, observations are not independent. Instead:

> y_1 gives information about y_2 ***because*** it provides information about the data-generating mechanism θ.

More on Bayesian prediction interpretation

("The usual Bayesian story":) Bayesian statistics is often described as consisting of assigning a prior on θ and using Bayes rule to compute the posterior distribution. Obtaining the predictive distribution,

$$p(\tilde{y} \mid y) = \int_{\Theta} p(\tilde{y} \mid \theta, y) dp(\theta \mid y)$$

is then just a matter of computations. Bayesian statistics is deeper than that! And a first basic concept we should recall is the interpretation of the **Bayesian predictive distribution**.

Bayesian statistics is about acting under uncertainty, or incomplete *information* (from the data, from domain knowledge, etc.).
If probability is the prescribed formal language to describe this (incomplete) information, then the evolution of information, or *learning*, is expressed through *conditional probabilities*.
In particular, learning on the next observation based on the observed is expressed through the conditional distribution $p(\tilde{y} \mid y)$.
This leads us to the interpretation of the Bayesian predictive distribution:

it is a **learning rule** that formalizes, through conditional probability, how we learn about future events given the available information.

(Thus, it is not meant as the ‘physical mechanism’ generating $\tilde{Y}$ given the past, like in the classic setting).

Exchangeability and Independence

You don’t need to understand the term exchangeability before learning Hierarchical Bayesian Models (Chapter 5).

At this point,

- we consider exchangeable models ***for data,*** $y_1, \ldots, y_n$, in the form of likelihoods in which the n observations are iid, given some parameter vector θ. (Later we will consider exchangeability for parameters.)

- Exchangeability is less strict condition than independence.

  - independence implies exchangeability
  - exchangeability does not imply independence

- exchangeability is related to what information is available (instead of the properties of unknown underlying data generating mechanism. See slide on Bayesian prediction interpretation)

  - Often we may assume that observations are in fact dependent, but if we can’t get information about these dependencies we may assume those observations as exchangeable. "Ignorance implies exchangeability."

A first example

Inference about a discrete quantity

In what follows we consider a real example of the very simplest case of Bayesian calculation.

It is not typical of *statistical* applications of Bayesian inference, as it deals with the **estimation of a single individual's state** (gene carrier or not) - and a very small data sample, rather than with the estimation of a parameter that describes an entire population.

Both the estimand and the observed variable are binary.

Inference about a Genetic Status: Prior

Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent.

Hemophilia is due to a recessive gene in the X-chromosome, that is, if X^* denotes an X-chromosome with the hemophilia gene,

- $X^* X^*$ is a female with the disease
- $X^* X$ is a female without the disease but with the gene
- $X^* Y$ is a male with the disease

Mary has

- an affected brother $\Rightarrow X^* Y$
- an unaffected mother $\Rightarrow XX^*$ or XX
- an unaffected father $\Rightarrow XY$

Overall, the mother must be XX^*.

Let $\theta = 1$ if Mary is a gene carrier (is XX^*) and 0 otherwise (XX), then *based on the above information*, **prior** to any observation,

$$P(\theta = 1) = \frac{1}{2}$$

Inference about a Genetic Status: Data Model and Likelihood

Data consist of the status of Mary's two sons, who are not affected.

Let then y_i be an indicator equal to 1 if the i-th son is affected:

$$P(y_i = 1 \mid \theta) = \begin{cases} 0.5 & \text{if } \theta = 1 \\ 0 & \text{otherwise} \end{cases}$$

The outcomes of the two sons are exchangeable and, conditional on the unknown θ, are independent; we assume the sons are not identical twins.

The likelihood function corresponding to Mary's two sons is:

$$L(\theta) = P(y_1 = y_2 = 0 \mid \theta) = \begin{cases} 0.25 & \text{if } \theta = 1 \\ 1 & \text{if } \theta = 0 \end{cases}$$

Inference about a Genetic Status: Prior Predictive

Data consist of the status of Mary's two sons, who are not affected.

We know that

$$P(y_1 = y_2 = 0 \mid \theta) = \begin{cases} 0.25 & \text{if } \theta = 1 \\ 1 & \text{if } \theta = 0 \end{cases}$$

Let $y = (y_1 = y_2 = 0)$, the predictive probability is

$$\begin{aligned} P(y) &= P(y \mid \theta = 1)P(\theta = 1) + P(y \mid \theta = 0)P(\theta = 0) \\ &= 0.25 \times 0.5 + 1 \times 0.5 = 0.625 \end{aligned}$$

Inference about a Genetic Status: Posterior

Prior and likelihood are combined to obtain the posterior, let $y = (y_1 = y_2 = 0)$,

$$\begin{aligned} P(\theta = 1 \,|\, y) &= \frac{P(y \,|\, \theta = 1) P(\theta = 1)}{P(y)} \\ &= \frac{P(y \,|\, \theta = 1) P(\theta = 1)}{P(y \,|\, \theta = 1) P(\theta = 1) + P(y \,|\, \theta = 0) P(\theta = 0)} \\ &= \frac{0.25 \times 0.5}{0.25 \times 0.5 + 1 \times 0.5} = 0.20 \end{aligned}$$

Intuitively it is clear that if a woman has unaffected children, it is less probable that she is a carrier.

When the parameter is discrete, the results can also be effectively described in terms of prior and posterior odds.
The posterior odds are given by the likelihood ratio times the prior odds:

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(y|\theta_1)}{p(y|\theta_2)} \frac{p(\theta_1)}{p(\theta_2)}$$

$$\frac{0.2}{0.8} = \frac{P(\theta = 1 \,|\, y)}{P(\theta = 0 \,|\, y)} = \frac{P(y \,|\, \theta = 1)}{P(y \,|\, \theta = 0)} \frac{P(\theta = 1)}{P(\theta = 0)} = \frac{0.25}{1} \times 1$$

Inference about a Genetic Status: Predictive distributions

Prior to the observations the predictive distribution is

$$\begin{aligned} P(y_1 = 1) &= P(y_1 = 1 \mid \theta = 1)P(\theta = 1) + P(y_1 = 1 \mid \theta = 0)P(\theta = 0) \\ &= 0.5 \times 0.5 + 0 \times 0.5 = 0.25 \end{aligned}$$

Given the data the posterior predictive is

$$\begin{aligned} P(\tilde{y}_3 = 1 \mid y) &= P(\tilde{y}_3 = 1 \mid \theta = 1, y)P(\theta = 1 \mid y) + P(\tilde{y}_3 = 1 \mid \theta = 0, y)P(\theta = 0 \mid y) \\ &= P(\tilde{y}_3 = 1 \mid \theta = 1)P(\theta = 1 \mid y) + P(\tilde{y}_3 = 1 \mid \theta = 1)P(\theta = 0 \mid y) \\ &= 0.5 \times 0.2 + 0 \times 0.8 = 0.1 \end{aligned}$$

Inference about a Genetic Status: Adding More Data

Suppose a third son is born and he is not affected, that is we have a new observation $y_3 = 0$, in order to obtain the new posterior distribution we can use the old posterior $P(\theta = 1 \,|\, y)$ as a prior and update it based on the likelihood $P(y_3 = 0 \,|\, \theta)$

$$\begin{aligned} P(\theta = 1 \,|\, y, y_3 = 0) &= \frac{P(y_3 = 0 \,|\, \theta = 1) P(\theta = 1 \,|\, y)}{P(y_3 = 0 \,|\, \theta = 1) P(\theta = 1 \,|\, y) + P(y_3 = 0 \,|\, \theta = 0) P(\theta = 0 \,|\, y)} \\ &= \frac{0.5 \times 0.2}{0.5 \times 0.2 + 1 \times 0.8} = 0.111 \end{aligned}$$

A similar mechanism works with the odds

$$\begin{array}{ccc} \dfrac{P(\theta = 1 \,|\, y, y_3 = 0)}{P(\theta = 0 \,|\, y, y_3 = 0)} & = & \dfrac{P(y_3 = 0 \,|\, \theta = 1)}{P(y_3 = 0 \,|\, \theta = 0)} \dfrac{P(\theta = 1 \,|\, y)}{P(\theta = 0 \,|\, y)} \\ \dfrac{1}{8} & = & \dfrac{0.5}{1} \qquad\qquad \dfrac{1}{4} \end{array}$$

The same result is obtained by starting from the prior and considering the data $y^{'} = (y_1 = y_2 = y_3 = 0)$.

Sequential analysis

A key aspect of Bayesian analysis is the ease with which **sequential analyses** can be performed.

As new data arrives, we need updating the information.
Considering the whole data (y_1, y_2)

$$p(\theta | y_1, y_2) \propto p(y_1, y_2 | \theta) p(\theta)$$

- Posterior distribution for θ given data y_1 and y_2
- Conditional distribution of y_1 and y_2 given θ
- Prior for θ

Assuming conditional independence, the likelihood can be partitioned:

$$p(y_1, y_2 | \theta) = p(y_2 | \theta) p(y_1 | \theta)$$

$$\text{Then} \quad \begin{aligned} p(\theta | y_1, y_2) \propto p(y_1, y_2 | \theta) p(\theta) &= p(y_2 | \theta) p(y_1 | \theta) p(\theta) \\ &\propto p(y_2 | \theta) p(\theta | y_1) \end{aligned}$$

That is, $p(\theta | y_1, y_2)$ is partitioned into conditional distribution of the sole y_2 given θ and posterior distribution for θ given y_1 (up to a constant of proportionality)

Bayes' Theorem: Accumulation of Evidence

Dataset 1: $p(\theta|y_1) \propto p(y_1|\theta)p(\theta)$

Dataset 2: $p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1)$

Dataset 3: $p(\theta|y_1, y_2, y_3) \propto p(y_3|\theta)p(\theta|y_1, y_2)$

Today's posterior is tomorrow's prior

Bayes' theorem as a mechanism for accumulating evidence

- Update diagnosis as symptoms, test results arrive
- Update beliefs about proficiency as students complete tasks
- Update beliefs about guilty as testimony is heard
- Do a study, use results as basis for prior for next study
- Makes Bayesian approach a natural framework for *meta-analysis* and related approaches that synthesize information from datasets

Note a margine

Modello e verosimiglianza

Il termine $p(y \mid \theta, M)$ ha due nomi diversi a seconda del caso. A causa della notazione concisa utilizzata, si può generare confusione.

1. Il termine $p(y \mid \theta, M)$ è detto **modello** (a volte più specificamente *modello di osservazione* o *modello statistico*) quando è usato per descrivere l'incertezza su y dati θ e M. La notazione più lunga $p_y(y \mid \theta, M)$ mostra esplicitamente che è una funzione di y.
2. Nella regola di Bayes, il termine $p(y \mid \theta, M)$ è chiamato **funzione di verosimiglianza**. La distribuzione a posteriori descrive la probabilità (o densità di probabilità) per diversi valori di θ dato un y fissato, e quindi quando la posteriori è calcolata, i termini sul lato destro (nella regola di Bayes) sono anche valutati come funzione di θ dato un y fissato. La notazione più lunga $p_\theta(y \mid \theta, M)$ mostra esplicitamente che è una funzione di θ.

Il termine ha un proprio nome (verosimiglianza) per differenziarsi rispetto al modello. La funzione di verosimiglianza è distribuzione di probabilità non normalizzata che descrive l'incertezza relativa a θ (ed è per questo che la regola di Bayes ha il termine di normalizzazione per ottenere la distribuzione a posteriori).

Notazione ambigua in statistica

In $p(y|\theta)$

- y può essere variabile o valore,
  - potremmo chiarire usando $p(Y|\theta)$ o $p(y|\theta)$
- θ può essere variabile o valore,
  - potremmo chiarire usando $p(y|\Theta)$ o $p(y|\theta)$
- p può essere una funzione discreta o continua di y o θ
  - potremmo chiarire usando P_Y, P_Θ, p_Y o p_Θ
- $P_Y(Y|\Theta = \theta)$ è una funzione di massa di probabilità, distribuzione campionaria, modello di osservazione
- $P(Y = y|\Theta = \theta)$ è una probabilità
- $P_\Theta(Y = y|\Theta)$ è una funzione di verosimiglianza (può essere discreta o continua)
- $p_Y(Y|\Theta = \theta)$ è una funzione di densità di probabilità, distribuzione campionaria, modello di osservazione
- $p(Y = y|\Theta = \theta)$ è una densità
- $p_\Theta(Y = y|\Theta)$ è una funzione di verosimiglianza (può essere discreta o continua)
- y e θ possono anche essere un misto di continuo e discreto

Back to 3-step general approach