Bayes' Theorem: Derivation, Interpretation, and a First Worked Example

Bayesian Inference & Computation

Plan for today

- Warm-up recap: independence, conditional probability, exchangeability
- Bayes' theorem: statement + proof from conditional probability
- What each term means: posterior, likelihood, prior, marginal likelihood
- Practical view: $\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta)$
- A first worked example: Normal likelihood with three different priors
- (Optional) Implementation idea: simple R code + plotting intuition

Big message

Bayesian inference is about learning θ from data y by combining:

$$\text{data information (likelihood)} \quad \times \quad \text{belief information (prior).}$$

Warm-up 1: Exchangeability

Definition (exchangeability)

Random variables $Y_1, \ldots, Y_N$ are *exchangeable* if for every permutation σ of $\{1, \ldots, N\}$,

$$\pi(Y_1, \ldots, Y_N) = \pi(Y_{\sigma(1)}, \ldots, Y_{\sigma(N)}).$$

Equivalently, the joint distribution is invariant under re-ordering.

Warm-up 1: Exchangeability

Definition (exchangeability)

Random variables $Y_1, \ldots, Y_N$ are *exchangeable* if for every permutation σ of $\{1, \ldots, N\}$,

$$\pi(Y_1, \ldots, Y_N) = \pi(Y_{\sigma(1)}, \ldots, Y_{\sigma(N)}).$$

Equivalently, the joint distribution is invariant under re-ordering.

- Exchangeability expresses **symmetry** in the model: order does not matter.
- It is **weaker than independence**: you can have exchangeability without factorisation.
- It is often a modelling assumption for “similar” observations.

Independence vs exchangeability

Independence (stronger)

If $Y_1, \ldots, Y_N$ are independent, then

$$\pi(Y_1, \ldots, Y_N) = \prod_{i=1}^{N} \pi(Y_i).$$

Independence vs exchangeability

Independence (stronger)

If $Y_1, \ldots, Y_N$ are independent, then

$$\pi(Y_1, \ldots, Y_N) = \prod_{i=1}^{N} \pi(Y_i).$$

Exchangeability (weaker, symmetry only)

If $Y_1, \ldots, Y_N$ are exchangeable, we only know

$$\pi(Y_1, \ldots, Y_N) = \pi(Y_{\sigma(1)}, \ldots, Y_{\sigma(N)}) \quad \forall \sigma,$$

but we *do not* necessarily get factorisation.

Why did we care (briefly)?

Exchangeability was used as a stepping stone to justify subjective probability modelling (De Finetti-type ideas): it explains why a “probability on parameters” can be coherent.

Warm-up 2: A useful independence manipulation

Let A and B be independent events.

Key property

Independence means:

$$\pi(A \cap B) = \pi(A)\pi(B).$$

Warm-up 2: A useful independence manipulation

Let A and B be independent events.

Key property

Independence means:

$$\pi(A \cap B) = \pi(A)\pi(B).$$

Conditioning on an event C

A common algebraic trick is to expand conditional probabilities like brackets:

$$\pi(A \cap B \mid C) = \frac{\pi(A \cap B \cap C)}{\pi(C)}.$$

Similarly,

$$\pi(B \mid C) = \frac{\pi(B \cap C)}{\pi(C)}.$$

Warm-up 2: A useful independence manipulation

Why this comes up repeatedly

These "expand-and-cancel" manipulations are used all the time when deriving Bayes' theorem, posterior distributions, and conditional independence arguments.

The Bayesian question

Main goal of Bayesian inference

We want the distribution of a model parameter θ *after* observing data y:

$$\pi(\theta \mid y).$$

The Bayesian question

Main goal of Bayesian inference

We want the distribution of a model parameter θ *after* observing data y:

$$\pi(\theta \mid y).$$

- θ can be a single parameter (scalar) or a vector $\boldsymbol{\theta}$.
- y can be a single observation or a dataset $\boldsymbol{y} = (y_1, \ldots, y_n)$.
- The meaning is the same: "what do we know about θ given the observed data?"

Bayes' Theorem (statement)

Theorem (Bayes' theorem)

Given a parameter θ *and observed data* y,

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\,\pi(\theta)}{\pi(y)}.$$

Bayes' Theorem (statement)

Theorem (Bayes' theorem)

Given a parameter θ and observed data y,

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\,\pi(\theta)}{\pi(y)}.$$

Interpretation at a glance

$$\underbrace{\pi(\theta \mid y)}_{\text{posterior}} = \frac{\underbrace{\pi(y \mid \theta)}_{\text{likelihood}}\;\underbrace{\pi(\theta)}_{\text{prior}}}{\underbrace{\pi(y)}_{\text{marginal likelihood}}}.$$

Proof of Bayes' Theorem (just conditional probability)

Start from the definition of conditional probability

$$\pi(\theta \mid y) = \frac{\pi(\theta, y)}{\pi(y)} \qquad \text{(assuming } \pi(y) > 0\text{).}$$

Proof of Bayes' Theorem (just conditional probability)

Start from the definition of conditional probability

$$\pi(\theta \mid y) = \frac{\pi(\theta, y)}{\pi(y)} \qquad \text{(assuming } \pi(y) > 0\text{)}.$$

Also write conditional probability the other way

$$\pi(y \mid \theta) = \frac{\pi(\theta, y)}{\pi(\theta)} \qquad \text{(assuming } \pi(\theta) > 0\text{)}.$$

Proof of Bayes' Theorem (just conditional probability)

Start from the definition of conditional probability

$$\pi(\theta \mid y) = \frac{\pi(\theta, y)}{\pi(y)} \qquad \text{(assuming } \pi(y) > 0\text{).}$$

Also write conditional probability the other way

$$\pi(y \mid \theta) = \frac{\pi(\theta, y)}{\pi(\theta)} \qquad \text{(assuming } \pi(\theta) > 0\text{).}$$

Rearrange the second equation

$$\pi(\theta, y) = \pi(y \mid \theta)\, \pi(\theta).$$

Proof of Bayes' Theorem (just conditional probability)

Start from the definition of conditional probability

$$\pi(\theta \mid y) = \frac{\pi(\theta, y)}{\pi(y)} \qquad \text{(assuming } \pi(y) > 0\text{)}.$$

Also write conditional probability the other way

$$\pi(y \mid \theta) = \frac{\pi(\theta, y)}{\pi(\theta)} \qquad \text{(assuming } \pi(\theta) > 0\text{)}.$$

Rearrange the second equation

$$\pi(\theta, y) = \pi(y \mid \theta)\,\pi(\theta).$$

Substitute into the first equation

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\,\pi(\theta)}{\pi(y)}.$$

Why the proof feels "underwhelming"

- The proof is algebraically simple because Bayes' theorem is essentially a **rearrangement** of conditional probability.
- The power is *not* in the proof.
- The power is in the **interpretation** and in what it lets us do:
  - combine prior beliefs with data evidence,
  - update beliefs coherently,
  - quantify uncertainty in θ,
  - do prediction via $\pi(y^\star \mid y)$.

Labelling the terms in Bayes' theorem

Posterior: $\pi(\theta \mid y)$

- The distribution of interest in Bayesian inference.
- "What can we say about θ after observing y?"

Labelling the terms in Bayes' theorem

Posterior: $\pi(\theta \mid y)$

- The distribution of interest in Bayesian inference.
- "What can we say about θ after observing y?"

Likelihood: $\pi(y \mid \theta)$

- Measures how compatible the data y is with each θ.
- "If θ were true, how likely is the observed data?"

Labelling the terms in Bayes' theorem

Posterior: $\pi(\theta \mid y)$

- The distribution of interest in Bayesian inference.
- "What can we say about θ after observing y?"

Likelihood: $\pi(y \mid \theta)$

- Measures how compatible the data y is with each θ.
- "If θ were true, how likely is the observed data?"

Prior: $\pi(\theta)$

- Encodes beliefs/knowledge about θ *before seeing the data*.
- A modelling choice: may be vague or informative.

The marginal likelihood $\pi(y)$ and why we often ignore it

Marginal likelihood / evidence

$$\pi(y) = \int \pi(y \mid \theta)\, \pi(\theta)\, d\theta \qquad \text{(or a sum if } \theta \text{ is discrete).}$$

The marginal likelihood $\pi(y)$ and why we often ignore it

Marginal likelihood / evidence

$$\pi(y) = \int \pi(y \mid \theta)\, \pi(\theta)\, d\theta \qquad \text{(or a sum if } \theta \text{ is discrete).}$$

- $\pi(y)$ is often hard to compute (integrals can be messy).
- It does **not depend on** θ.
- It is just a **normalising constant** to make the posterior integrate to 1.

The marginal likelihood $\pi(y)$ and why we often ignore it

Marginal likelihood / evidence

$$\pi(y) = \int \pi(y \mid \theta)\, \pi(\theta)\, d\theta \qquad \text{(or a sum if } \theta \text{ is discrete).}$$

- $\pi(y)$ is often hard to compute (integrals can be messy).
- It does **not depend on** θ.
- It is just a **normalising constant** to make the posterior integrate to 1.

Practical form we will use most of the time

$$\pi(\theta \mid y) \;\propto\; \pi(y \mid \theta)\, \pi(\theta).$$

A modelling rule: "don't use the data twice"

Important modelling principle

The prior $\pi(\theta)$ should be chosen *before* looking at the current dataset.

A modelling rule: "don't use the data twice"

Important modelling principle

The prior $\pi(\theta)$ should be chosen *before* looking at the current dataset.

- You may use:
  - domain knowledge,
  - previous studies,
  - expert opinion,
  - physical constraints (e.g. $\theta > 0$).
- But you should not set the prior by inspecting the same data y you then plug into the likelihood.

A modelling rule: "don't use the data twice"

Important modelling principle

The prior $\pi(\theta)$ should be chosen *before* looking at the current dataset.

- You may use:
  - domain knowledge,
  - previous studies,
  - expert opinion,
  - physical constraints (e.g. $\theta > 0$).
- But you should not set the prior by inspecting the same data y you then plug into the likelihood.

Simple example

If μ is the population mean height, you cannot:

$$\text{look at the sample } y \quad \Rightarrow \quad \text{choose prior mean equal to the sample mean.}$$

That would "double count" the data.

Historical intermission: Thomas Bayes (1700s)

- Thomas Bayes was a UK minister (Tunbridge Wells) with strong mathematical ability.
- Bayesian inference is named after him largely by historical accident.
- His famous "idea" is often described using a **ball/beanbag** thought experiment:
  - throw objects onto a table (unknown landing distribution),
  - observe outcomes,
  - update beliefs about future outcomes given past outcomes.

Historical intermission: Thomas Bayes (1700s)

- Thomas Bayes was a UK minister (Tunbridge Wells) with strong mathematical ability.
- Bayesian inference is named after him largely by historical accident.
- His famous "idea" is often described using a **ball/beanbag** thought experiment:
  - throw objects onto a table (unknown landing distribution),
  - observe outcomes,
  - update beliefs about future outcomes given past outcomes.

Core theme

Beliefs should be *updated* when data arrives. That is the philosophical backbone of Bayesian thinking.

A (very) short note on the historical motivation

A famous quote attributed to this early line of work is about fixed laws of nature:

> *"...to show what reason we have for believing that there are, in the constitution of things, fixed laws according to which events happen..."*

A (very) short note on the historical motivation

A famous quote attributed to this early line of work is about fixed laws of nature:

> *"...to show what reason we have for believing that there are, in the constitution of things, fixed laws according to which events happen..."*

- Historically, these ideas were tied to philosophical/theological arguments.
- Modern Bayesian statistics is not about theology: it is a practical mathematical framework for uncertainty and learning from data.

Worked example: Normal model for an unknown mean

Model

Assume a single observation Y satisfies

$$Y \mid \theta \sim \mathcal{N}(\theta, 1).$$

Worked example: Normal model for an unknown mean

Model

Assume a single observation Y satisfies

$$Y \mid \theta \sim \mathcal{N}(\theta, 1).$$

- The parameter is $\theta \in \mathbb{R}$, the (unknown) mean.
- The variance is known and fixed: 1.
- We will compare different priors $\pi(\theta)$ and see their effect on $\pi(\theta \mid y)$.

Worked example: Normal model for an unknown mean

Model

Assume a single observation Y satisfies

$$Y \mid \theta \sim \mathcal{N}(\theta, 1).$$

- The parameter is $\theta \in \mathbb{R}$, the (unknown) mean.
- The variance is known and fixed: 1.
- We will compare different priors $\pi(\theta)$ and see their effect on $\pi(\theta \mid y)$.

Bayes' rule for this model

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta).$$

Step 1: Write down the likelihood

Likelihood function

Since $Y \mid \theta \sim \mathcal{N}(\theta, 1)$, we have

$$\pi(y \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2}\right).$$

Step 1: Write down the likelihood

Likelihood function

Since $Y \mid \theta \sim \mathcal{N}(\theta, 1)$, we have

$$\pi(y \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2}\right).$$

- Think of this as a function of θ with y fixed.
- It tells us which values of θ make the observed y plausible.

Step 2: Choose a prior (three options)

We now choose $\pi(\theta)$, encoding beliefs about θ *before* observing the data.

Step 2: Choose a prior (three options)

We now choose $\pi(\theta)$, encoding beliefs about θ *before* observing the data.

Prior option 1: Very vague (wide uniform)

$$\theta \sim \mathrm{Unif}(-10{,}000,\ 10{,}000), \qquad \pi(\theta) = \frac{1}{20{,}000}.$$

So $\pi(\theta) \propto 1$ on $[-10{,}000, 10{,}000]$.

Step 2: Choose a prior (three options)

We now choose $\pi(\theta)$, encoding beliefs about θ *before* observing the data.

Prior option 1: Very vague (wide uniform)

$$\theta \sim \mathrm{Unif}(-10{,}000,\ 10{,}000), \qquad \pi(\theta) = \frac{1}{20{,}000}.$$

So $\pi(\theta) \propto 1$ on $[-10{,}000, 10{,}000]$.

Prior option 2: Constraint information (positive only)

$$\theta \sim \mathrm{Unif}(0,\ 10{,}000), \qquad \pi(\theta) = \frac{1}{10{,}000}.$$

So $\pi(\theta) \propto 1$ on $[0, 10{,}000]$ and 0 otherwise.

Step 2 continued: An informative prior

Prior option 3: Informative normal

Assume expert knowledge suggests θ is around 3 with uncertainty 0.7:

$$\theta \sim \mathcal{N}(3,\ 0.7^2), \qquad \pi(\theta) = \frac{1}{\sqrt{2\pi}\,0.7} \exp\left(-\frac{(\theta-3)^2}{2(0.7)^2}\right).$$

Step 2 continued: An informative prior

Prior option 3: Informative normal

Assume expert knowledge suggests θ is around 3 with uncertainty 0.7:

$$\theta \sim \mathcal{N}(3,\ 0.7^2), \qquad \pi(\theta) = \frac{1}{\sqrt{2\pi}\,0.7} \exp\left(-\frac{(\theta-3)^2}{2(0.7)^2}\right).$$

- This prior concentrates probability mass near $\theta = 3$.
- Smaller prior variance 0.7^2 means **stronger prior belief**.

Step 3: Observe data

Data

After effort and cost, we collect one data point:

$$y = 0.$$

Step 3: Observe data

Data

After effort and cost, we collect one data point:

$$y = 0.$$

Likelihood at $y = 0$

Plug $y = 0$ into the likelihood:

$$\pi(y = 0 \mid \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right).$$

Posterior under Prior 1 (wide uniform)

Prior 1

$$\pi(\theta) \propto 1 \quad \text{for } \theta \in [-10{,}000, 10{,}000].$$

Posterior under Prior 1 (wide uniform)

Prior 1

$$\pi(\theta) \propto 1 \quad \text{for } \theta \in [-10{,}000, 10{,}000].$$

Posterior

$$\pi(\theta \mid y = 0) \propto \pi(y = 0 \mid \theta)\pi(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right) \quad \text{for } \theta \in [-10{,}000, 10{,}000].$$

Posterior under Prior 1 (wide uniform)

Prior 1

$$\pi(\theta) \propto 1 \quad \text{for } \theta \in [-10{,}000, 10{,}000].$$

Posterior

$$\pi(\theta \mid y = 0) \propto \pi(y = 0 \mid \theta)\pi(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right) \quad \text{for } \theta \in [-10{,}000, 10{,}000].$$

- Since the prior is (almost) constant, the posterior looks like the likelihood.
- In practice, this behaves like a $\mathcal{N}(0, 1)$ shape (with extremely wide truncation).

Posterior under Prior 2 (positive-only uniform)

Prior 2

$$\pi(\theta) \propto 1 \quad \text{for } \theta \in [0, 10{,}000], \qquad \pi(\theta) = 0 \text{ otherwise.}$$

Posterior under Prior 2 (positive-only uniform)

Prior 2

$$\pi(\theta) \propto 1 \quad \text{for } \theta \in [0, 10{,}000], \qquad \pi(\theta) = 0 \text{ otherwise.}$$

Posterior

$$\pi(\theta \mid y = 0) \propto \exp\left(-\frac{\theta^2}{2}\right) \mathbf{1}_{[0,10{,}000]}(\theta).$$

Posterior under Prior 2 (positive-only uniform)

Prior 2

$$\pi(\theta) \propto 1 \quad \text{for } \theta \in [0, 10{,}000], \qquad \pi(\theta) = 0 \text{ otherwise.}$$

Posterior

$$\pi(\theta \mid y = 0) \propto \exp\left(-\frac{\theta^2}{2}\right) \mathbf{1}_{[0,10{,}000]}(\theta).$$

- Same Gaussian-shaped likelihood, but we **forbid** negative θ.
- This produces a **truncated** normal-like posterior supported on $\theta \geq 0$.

Posterior under Prior 3 (informative normal)

Prior 3

$$\pi(\theta) \propto \exp\left(-\frac{(\theta-3)^2}{2(0.7)^2}\right).$$

Posterior under Prior 3 (informative normal)

Prior 3

$$\pi(\theta) \propto \exp\left(-\frac{(\theta-3)^2}{2(0.7)^2}\right).$$

Posterior (unnormalised)

$$\pi(\theta \mid y=0) \propto \exp\left(-\frac{\theta^2}{2}\right)\exp\left(-\frac{(\theta-3)^2}{2(0.7)^2}\right).$$

Posterior under Prior 3 (informative normal)

Prior 3

$$\pi(\theta) \propto \exp\left(-\frac{(\theta - 3)^2}{2(0.7)^2}\right).$$

Posterior (unnormalised)

$$\pi(\theta \mid y = 0) \propto \exp\left(-\frac{\theta^2}{2}\right) \exp\left(-\frac{(\theta - 3)^2}{2(0.7)^2}\right).$$

- Posterior is the product of two exponentials.
- It will again be proportional to a Gaussian-shaped function in θ.
- Its mean will lie **between** 0 (data) and 3 (prior), typically closer to the more "certain" source.

What the posterior is doing (the intuition)

Key idea

$$\text{Posterior} \;\propto\; \text{Likelihood} \times \text{Prior}.$$

What the posterior is doing (the intuition)

Key idea

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

- Likelihood is "what the data says" about θ.
- Prior is "what we believed before data" about θ.
- Multiplying them **blends** information:
  - A vague prior $\Rightarrow$ posterior mostly follows the likelihood.
  - A strong informative prior $\Rightarrow$ posterior shifts towards the prior.

What the posterior is doing (the intuition)

Key idea

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

- Likelihood is "what the data says" about θ.
- Prior is "what we believed before data" about θ.
- Multiplying them **blends** information:
  - A vague prior $\Rightarrow$ posterior mostly follows the likelihood.
  - A strong informative prior $\Rightarrow$ posterior shifts towards the prior.

Important trend with more data

With many observations, the likelihood becomes sharper, and the posterior is dominated more by data.

How to compute and plot in R (conceptual)

We can visualise the relationship:

$$\text{posterior} \;\propto\; \text{likelihood} \times \text{prior}.$$

How to compute and plot in R (conceptual)

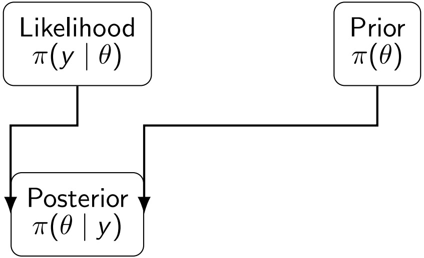
We can visualise the relationship:

$$\text{posterior} \;\propto\; \text{likelihood} \times \text{prior}.$$

Suggested workflow

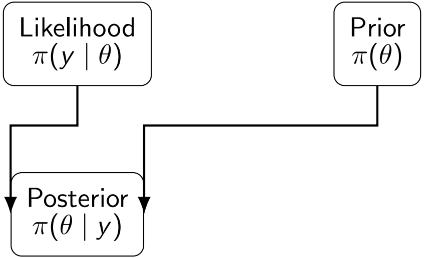
1. Choose a grid of θ values (e.g. from -5 to 5).
2. Evaluate:
$$L(\theta) = \pi(y \mid \theta), \qquad p(\theta) = \pi(\theta).$$
3. Compute unnormalised posterior:
$$\tilde{p}(\theta \mid y) = L(\theta)p(\theta).$$
4. Normalise numerically (optional), or just compare shapes.

A picture you should remember



$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta)$$

A picture you should remember



$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta)$$

Mental model

Bayesian inference is **multiplication of curves** followed by normalisation.

Interpreting the 3-by-3 grid plot (likelihood / prior / posterior)

What the grid is showing

- **Column 1:** the likelihood curve $L(\theta)$ (same for all priors)
- **Column 2:** the prior curve $p(\theta)$ (changes with prior choice)
- **Column 3:** the posterior curve $\tilde{p}(\theta \mid y) = L(\theta)p(\theta)$

Interpreting the 3-by-3 grid plot (likelihood / prior / posterior)

What the grid is showing

- **Column 1:** the likelihood curve $L(\theta)$ (same for all priors)
- **Column 2:** the prior curve $p(\theta)$ (changes with prior choice)
- **Column 3:** the posterior curve $\tilde{p}(\theta \mid y) = L(\theta)p(\theta)$

What you should notice

- Uniform prior $\Rightarrow$ posterior $\approx$ likelihood shape.
- Positive-only prior $\Rightarrow$ posterior is "likelihood chopped in half".
- Normal prior $\Rightarrow$ posterior is a compromise between data-centred and prior-centred beliefs.

Summary

- Bayes' theorem:
$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)}.$$
- Practical version:
$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta).$$
- Posterior combines:
  - data evidence (likelihood),
  - prior belief (prior).
- The marginal likelihood $\pi(y)$ is a normalising constant (often ignored in algebra).
- In the Normal mean example, changing the prior can change the posterior a lot when data are scarce.

Next (preview)

Next lectures: hands-on posterior derivations

We will practise computing posteriors for many common models by repeatedly applying:

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta).$$

Next (preview)

Next lectures: hands-on posterior derivations

We will practise computing posteriors for many common models by repeatedly applying:

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta).$$

- You will see many likelihoods and many priors (and their consequences).
- We will also start discussing prediction and computation.