Bayesian Statistics

Lectures for the MRes in Big Data and Machine Learning in Physics

Boris Leistedt

Based on course notes by Alan Heavens, Imperial College London

November 9, 2025

Contents

List of Examples

1 Books and Syllabus

1.1 Recommended Books

- D. Silvia & J. Skilling: Data Analysis: a Bayesian Tutorial (CUP). *Nice small book for the basics.*
- P. Saha: Principles of Data Analysis. (Capella Archive)

  `https://www.physik.uzh.ch/ psaha/pda/`

  Similarly, a good, clear, small volume. Free online version as well as a physical book.
- T. Loredo: Bayesian Inference in the Physical Sciences

  `http://www.astro.cornell.edu/staff/loredo/bayes/`
- D. Mackay: Information Theory, Inference and Learning Algorithms. (CUP)

  `http://www.inference.phy.cam`

  More on the information theory basis of the subject.
- A. Gelman et al: Bayesian Data Analysis (CRC Press) *Comprehensive.*

1.2 Course Syllabus

The course begins with an **Introduction** covering the fundamental differences between Bayesian and frequentist interpretations of probability. We discuss how scientific questions are often inverse problems, where we wish to infer properties of a system from observed data rather than predict data from known parameters. The foundation of Bayesian inference, Bayes theorem, is introduced as the key tool for updating our beliefs in light of new evidence.

In the section on **Parameter Inference**, we develop the machinery for inferring model parameters from data. The likelihood function, which quantifies how probable our observed data are given specific parameter values, is central to this discussion. We examine the role of priors in encoding our initial state of knowledge, including the treatment of location parameters (parameters that shift distributions) and scale parameters (parameters that affect the spread of distributions). The concept of 'noninformative' priors is discussed critically, acknowledging both their utility and limitations. We also introduce conjugate priors, which are prior distributions that combine mathematically conveniently with certain likelihood functions to produce posteriors in the same distributional family.

The **Posterior** section focuses on extracting information from the posterior distribution. Marginalisation allows us to focus on parameters of interest by integrating over nuisance parameters—those parameters necessary for the model but not of direct scientific interest. We distinguish between conditional and marginal errors, emphasizing that conditional errors (obtained by fixing other parameters) typically underestimate the true uncertainty. The profile likelihood method, which maximizes the likelihood over nuisance parameters for each value of the parameter of interest, is discussed as an alternative approach with connections to frequentist methods.

Sampling methods are introduced as practical tools for working with complex posterior distributions that cannot be evaluated analytically. We explain how samples from a distribution can be used to represent that distribution numerically, allowing us to compute marginal distributions, moments, and credible intervals. The theoretical foundations include detailed balance, which ensures that a sampling algorithm targets the correct distribution, and Markov processes, which describe how samples are generated sequentially with each new sample depending only on the previous state.

The **MCMC** (Markov Chain Monte Carlo) section presents the Metropolis-Hastings algorithm for low-dimensional problems. This algorithm generates samples by proposing moves in parameter space and accepting or rejecting them according to a criterion that ensures detailed balance. The choice of proposal distribution significantly affects the efficiency of the sampler. We examine sample correlations, quantified by the autocorrelation function, which arise because successive samples in a Markov chain are not independent. The effective sample size accounts for these correlations to give a measure of the equivalent number of independent samples. The burn-in period, during which the chain moves from its arbitrary initial position to the high-probability regions of the posterior, must be discarded. Convergence tests, particularly the Gelman-Rubin diagnostic, help assess whether chains have adequately explored the target distribution.

For **Higher-Dimensional Problems**, standard MCMC methods become inefficient due to the difficulty of proposing moves that are likely to be accepted in high dimensions. Hamiltonian Monte Carlo (HMC) addresses this by using gradient information to propose distant moves that follow the geometry of the posterior distribution. Gibbs sampling offers an alternative strategy: when conditional distributions for individual parameters (or blocks of parameters) are available, we can sample from them in turn, effectively decomposing a high-dimensional problem into a sequence of lower-dimensional ones.

Multi-level Models, or Bayesian Hierarchical Models, extend the basic framework to situations where the data have natural grouping or structure. In these models, latent parameters (also called hyperparameters) govern the distribution of the parameters at lower levels of the hierarchy. This approach allows information to be shared across groups, leading to improved inference compared to analyzing each group independently or pooling all data without accounting for group structure.

Model Comparison addresses the question of which model best explains the data. The Bayesian evidence (or marginal likelihood) provides a natural tool for model comparison, with the ratio of evidences (Bayes factor) quantifying the relative support for competing models. For nested models, the Savage-Dickey density ratio provides a convenient shortcut for computing Bayes factors. We also discuss information criteria such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and DIC (Deviance Information Criterion), which approximate model comparison while accounting for model complexity.

Likelihood-Free Inference (also called Simulation-Based Inference or implicit likelihood methods) tackles problems where the likelihood function cannot be evaluated, but we can simulate data from the forward model. Rejection sampling forms the basis of Approximate Bayesian Computation (ABC), where we accept parameters if they generate data sufficiently similar to our observations. Kernel density estimation (KDE) provides a non-parametric method for estimating probability densities from samples, useful for constructing approximate posteriors.

The topic of **Extreme Data Compression** introduces the MOPED (Multiple Optimized Parameter Estimation and Data compression) algorithm, which shows that high-dimensional data can often be compressed to a number of summary statistics equal to the number of parameters being inferred, without loss of information. This dramatic compression can greatly speed up inference for problems with large datasets.

Finally, we address **Complications** that arise in real data analysis. Selection effects occur when our ability to observe data depends on the values of the data themselves. Truncation occurs when data outside some range are completely unobserved, while censoring occurs when we know that data exist beyond some threshold but do not know their exact values. Both effects must be properly modeled in the likelihood to avoid biased inferences.

1.3 Learning Objectives

By the end of this course, students will be able to apply Bayesian reasoning to scientific problems, construct likelihood functions and priors, derive posterior distributions, implement MCMC sampling, build hierarchical models, perform model comparison, and handle complex data scenarios including missing data and selection effects.

1.4 Mathematical Notation

This course uses standard mathematical notation for probability and statistics. Familiarity with the following conventions will aid understanding:

Random variables and parameters:

- x, y, θ, μ, σ denote scalar quantities (single numbers)
- $\mathbf{x}$, $\mathbf{d}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ denote vectors (sets of numbers) using bold font
- $\mathbf{X}$, $\mathbf{C}$, $\boldsymbol{\Sigma}$ denote matrices using bold capital letters
- $\mathbf{I}$ denotes the identity matrix

Probability notation:

- $p(x)$ denotes the probability mass function (for discrete x) or probability density function (for continuous x)
- $p(x|y)$ denotes conditional probability: the probability of x given that y has occurred or is known
- $p(x, y)$ denotes joint probability: the probability that both x and y occur
- $p(\mathbf{d}|\boldsymbol{\theta}, M)$ denotes the probability of data $\mathbf{d}$ given parameters $\boldsymbol{\theta}$ and model M
- $\mathcal{L}(\boldsymbol{\theta})$ or $\mathcal{L}(\mathbf{d}|\boldsymbol{\theta})$ denotes the likelihood function (data treated as fixed, parameters as variables)
- $\pi(\boldsymbol{\theta})$ or $p(\boldsymbol{\theta}|M)$ denotes the prior distribution on parameters
- $\sim$ means "is distributed as," e.g., $x \sim \mathcal{N}(0, 1)$ means "x is distributed as a standard normal"

Statistical operations:

- $\mathbb{E}[x]$ or $\langle x \rangle$ denotes the expectation (mean) of x
- $\mathrm{Var}[x]$ or σ^2 denotes the variance of x
- $\mathrm{Cov}[x, y]$ denotes the covariance between x and y
- $\int f(x)dx$ denotes integration over all possible values of x
- $\sum_i f(x_i)$ denotes summation over discrete index i
- $\propto$ means "proportional to" (equality up to a normalization constant)

Special symbols:

- $\delta(x)$ denotes the Dirac delta function (infinite at $x = 0$, zero elsewhere, integrates to 1)
- $\delta^D(x - a)$ denotes the Dirac delta function centered at $x = a$
- $\ln x$ or $\log x$ denotes the natural logarithm (base e)
- $\log_{10} x$ denotes the base-10 logarithm (when explicitly needed)
- ∇f denotes the gradient (vector of partial derivatives) of f
- $\partial f/\partial x$ denotes the partial derivative of f with respect to x
- $\arg\max_x f(x)$ denotes the value of x that maximizes $f(x)$

Abbreviations:

- MLE: Maximum Likelihood Estimate
- MAP: Maximum A Posteriori estimate
- MCMC: Markov Chain Monte Carlo
- PDF: Probability Density Function
- CDF: Cumulative Distribution Function
- i.i.d.: independent and identically distributed

Gaussian (Normal) Distribution Notation:

Throughout this course, we use the semicolon notation $\mathcal{N}(x;\mu,\sigma^2)$ to denote the Gaussian probability density function, where x is the random variable, μ is the mean, and σ^2 is the variance. This notation emphasizes that x is the argument of the density function, while μ and σ^2 are parameters.

1D Gaussian:

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

Properties:

- Mean: $\mathbb{E}[x] = \mu$
- Variance: $\text{Var}[x] = \sigma^2$
- Standard deviation: $\sigma = \sqrt{\sigma^2}$
- Precision (inverse variance): $\lambda = 1/\sigma^2$
- The Gaussian is symmetric about μ and integrates to 1: $\int_{-\infty}^{\infty} \mathcal{N}(x;\mu,\sigma^2)dx = 1$
- Linear transformation: If $x \sim \mathcal{N}(\mu,\sigma^2)$, then $ax+b \sim \mathcal{N}(a\mu+b, a^2\sigma^2)$
- Sum of independent Gaussians: If $x_1 \sim \mathcal{N}(\mu_1,\sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu_2,\sigma_2^2)$ are independent, then $x_1+x_2 \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

Multivariate Gaussian:

For a d-dimensional random vector $\mathbf{x} = (x_1,\ldots,x_d)^T$:

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{2}$$

where $\boldsymbol{\mu}$ is the d-dimensional mean vector, $\boldsymbol{\Sigma}$ is the $d \times d$ covariance matrix, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

Properties:

- Mean vector: $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$
- Covariance matrix: $\text{Cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$
- Diagonal elements: $\Sigma_{ii} = \text{Var}[x_i]$
- Off-diagonal elements: $\Sigma_{ij} = \text{Cov}[x_i,x_j]$ for $i \neq j$
- Linear transformation: If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then $\mathbf{A}\mathbf{x}+\mathbf{b} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$
- Marginalization: If $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2)$ is partitioned with $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where

  $$\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix} \tag{3}$$

  then the marginal distribution is $\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11})$
- Conditioning: For the same partitioned Gaussian, the conditional distribution is

  $$\mathbf{x}_1|\mathbf{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \tag{4}$$

- Product of Gaussians: The product of two Gaussian densities (in the same variable) is proportional to a Gaussian:

  $$\mathcal{N}(x;\mu_1,\sigma_1^2)\cdot\mathcal{N}(x;\mu_2,\sigma_2^2) \propto \mathcal{N}(x;\mu_3,\sigma_3^2) \tag{5}$$

  where $\sigma_3^2 = \left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)^{-1}$ and $\mu_3 = \sigma_3^2\left(\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}\right)$

These Gaussian properties are fundamental to Bayesian inference, particularly for conjugate priors, Kalman filtering, and Gaussian processes.

A more comprehensive notation reference is provided in Appendix A.

1.5 Graphical Models for Probabilistic Reasoning

Note: This material is non-examinable, but provides a very useful tool for visualizing probabilistic models and will be used throughout the course.

Directed Acyclic Graphs (DAGs) provide a visual language for representing probabilistic models. In these diagrams, nodes represent variables and directed edges (arrows) represent probabilistic dependencies. The graph is acyclic, meaning you cannot follow arrows in a loop back to where you started.

The notation conventions for DAGs are:

- **Circles (or unfilled nodes)** represent *unknown/latent variables* or *parameters* that we wish to infer. These are unobserved quantities with uncertainty.
- **Shaded circles (or filled nodes)** represent *observed variables* or *data*. These are quantities we have measured or know.
- **Points or small dots** (or sometimes double circles) represent *fixed/deterministic quantities* or *hyperparameters* whose values are set and not inferred.
- **Plates** (rectangular boxes around groups of nodes) represent *repetition* or *replication*. A plate labeled with $i = 1, \ldots, N$ (or simply N) indicates that the contents are repeated N times, such as N independent observations.
- **Arrows** indicate *direct probabilistic dependence*: an arrow from A to B means that B depends on A, i.e., B is sampled from a distribution conditioned on A.

The graph structure encodes the factorization of the joint probability distribution: each variable is conditionally independent of all non-descendants given its parents (the nodes with arrows pointing to it).

These diagrams will help us visualize model structure, understand the flow of information from data to parameters, and communicate assumptions clearly. Throughout the course, we will use DAGs to represent hierarchical models, parameter inference problems, and complex dependencies between variables. Examples of DAGs that appear in this course include:

- **Virus testing** (Section 2, Figure 1): simple diagnostic test with latent disease status
- **Coin toss** (Section 4, Figure 4): parameter inference with Beta-Binomial conjugacy
- **Supernova distance** (Section 5, Figure 6): nuisance parameter marginalization
- **Line fitting with errors** (Section 6, Figure 8): Gibbs sampling with latent variables
- **Hierarchical models** (Section 7, Figures 11, 10): multi-level structure with hyperparameters
- **Model comparison** (Section 8, Figure 12): nested model structures
- **Censored data** (Section 9, Figure 13): selection effects and missing data
- **Variational inference** (Section 10, Figure 14): Bayesian linear regression

2 Bayesian Inference

Bayesian and frequentist views of probability differ in fundamental ways. The frequentist view is usually expressed in terms of relative occurrences of events in multiply-repeated experiments, such as the fraction of heads thrown in the repeated toss of a coin. In Bayesian statistics, probability is sometimes interpreted as a state of knowledge. It is better matched to answering scientific questions, since this notion of probability encapsulates what is often desired when doing science. To see this, one must first recognize that most scientific questions are inverse problems - what can be learned about the world from the data that has been collected?

2.1 Inverse Problems

Data analysis problems in physics are typically inverse problems: given some observed data, the goal is to infer something about the underlying physical process that generated those data. This stands in contrast to forward modeling (also called generative modeling), where observational outcomes are predicted given a known physical process and its parameters. Inverse problems are generally much harder than forward problems because they require working backwards from effects to causes, and this mapping is often not unique.

There are two principal classes of inverse problems in Bayesian statistics. First, **parameter inference** addresses questions such as: given a model with unknown parameters, what do the data tell us about the values of those parameters? Second, **model comparison** asks: given multiple competing models that could explain the data, which model is best supported by the observations? Both types of questions are naturally framed in the Bayesian framework.

2.2 What Questions Can Bayesian Inference Answer?

Bayesian inference provides a systematic framework for answering scientific questions in the presence of uncertainty. Consider some concrete examples of the types of questions that might be asked.

For **parameter inference**, a model structure is assumed and the goal is to determine the values of its parameters given observed data. For instance: if there is a set of (x, y) pairs with measurement errors and a linear relationship $y = mx+c$ is assumed, what are the slope m and intercept c, and what are the uncertainties in these parameters? Or, if 5 X-ray photons have been detected from a source at a known distance in the laboratory, what is the power output of the source and its uncertainty? As a more complex example, given LIGO gravitational wave observations, what are the masses and other properties of the inspiralling compact objects that generated the signal? In each case, not just point estimates but full probability distributions over the parameters are desired, properly accounting for all sources of uncertainty.

For **model comparison**, the goal is to determine which of several competing models is best supported by the data. Do the observed planetary motions support General Relativity or Newtonian gravity? Is the standard cosmological model (ΛCDM) more probable than specified alternative models given the current observational data? Do Large Hadron Collider (LHC) data support the existence of the Higgs boson, or do they favor a model without it? These questions cannot be answered by simply fitting parameters—a principled way to compare models of different complexity and structure is needed. Bayesian model comparison provides this through the calculation of model evidences and Bayes factors.

2.3 Probability: Frequency vs. Degree of Belief

The interpretation of probability is central to understanding Bayesian statistics. In the frequentist view, probability describes the relative frequency of outcomes in infinitely long sequences of repeated trials—for example, the limiting fraction of heads in an infinite sequence of coin tosses. This interpretation works well for repeatable random experiments but becomes problematic when making statements about unique events or fixed parameters.

The Bayesian view, by contrast, interprets probability as expressing a degree of belief or state of knowledge about a proposition. In this framework, probabilities can be assigned to any logical proposition—a statement that could be true or false. The conditional probability $p(A|B)$ represents the degree to which the truth of logical proposition B implies that proposition A is also true. This interpretation allows discussion of the probability that a scientific hypothesis is correct given the data that has been observed.

The Bayesian interpretation naturally expresses what is often desired in science. For example, one might ask: given the Planck satellite's observations of the cosmic microwave background (CMB), what is the probability that the density parameter of cold dark matter lies between 0.3 and 0.4? This question makes sense in the Bayesian framework, where probabilities can be assigned to parameter values. In the frequentist framework, by contrast, the density parameter is either in this range or it is not—it is a fixed (though unknown) value, not a random variable, and so cannot be assigned a probability.

2.4 Fundamental Probability Rules

Bayesian inference rests on a small number of fundamental probability rules, which are used throughout this course.

The **sum rule** states that probabilities of mutually exclusive, exhaustive outcomes must sum to unity: $p(x) + p(\sim x) = 1$, where $\sim x$ means "not x".

The **product rule** relates joint and conditional probabilities: $p(x, y) = p(x|y)p(y)$, where $p(x|y)$ is the conditional probability of x given y (read as "the probability of x given y"), and $p(x, y)$ is the joint probability that both x and y occur.

The **marginalisation rule** allows us to obtain the probability of one variable by summing or integrating over all possible values of other variables. For discrete variables: $p(x) = \sum_k p(x, y_k)$, summing over all possible discrete values y_k. For continuous variables: $p(x) = \int p(x, y)dy$. Here $p(x, y)$ is a probability density function (pdf), where $p(x, y) \geq 0$ and $p(x, y)dxdy$ represents the probability that x and y occur in an infinitesimal interval $dxdy$ around the values x, y. Note that a probability density can be greater than 1—it is a density, not a probability itself.

Since the joint probability is symmetric, $p(x, y) = p(y, x)$, we can write $p(x|y)p(y) = p(y|x)p(x)$. Rearranging this fundamental equality gives us **Bayes' theorem**:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}. \tag{6}$$

This simple equation is the foundation of all Bayesian inference. It tells us how to update our beliefs about y (the prior $p(y)$) in light of observing x, using the likelihood $p(x|y)$ and the marginal probability $p(x)$ (which serves as a normalization constant).

2.5 The Asymmetry of Conditional Probability

A fundamental error in reasoning with probabilities is to confuse $p(x|y)$ with $p(y|x)$. These conditional probabilities are generally very different, and confusing them leads to serious mistakes in inference.

The rain and umbrella example perfectly illustrates why conditional probabilities aren't symmetric. While about 75% of people carry umbrellas when it's raining ($P(\text{Umbrella}|\text{Raining}) \approx 0.75$), only about 30% of people carrying umbrellas are doing so because it's currently raining ($P(\text{Raining}|\text{Umbrella}) \approx 0.30$). This dramatic difference occurs because umbrellas serve multiple purposes—people carry them for potential rain later, sun protection, or out of habit—not just for current rainfall. In contrast, when it is actually raining, the response is more predictable: most people grab an umbrella. This asymmetry shows why seeing rain is a strong predictor of umbrellas, but seeing an umbrella is only a weak predictor of rain. You can't look out your window, spot someone with an umbrella, and confidently conclude it's raining—but if you see rain, you can reasonably assume most people outside have umbrellas.

One might think that no one would make such an elementary mistake, but confusion between $p(x|y)$ and $p(y|x)$ is surprisingly common, with serious consequences in fields from medical diagnosis to legal reasoning. The next section provides a medical testing example that illustrates the practical importance of getting this distinction right.

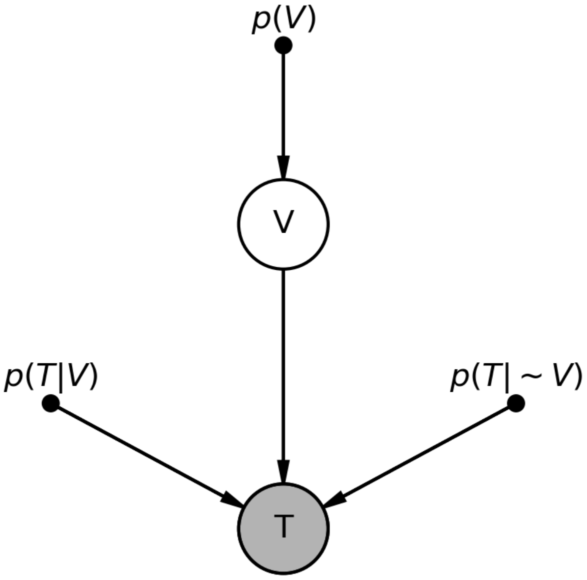

Figure 1: DAG for virus diagnostic testing model. The latent virus status V (prevalence 1%) determines test results T with sensitivity 80% and false positive rate 10%.

Example: Virus Testing: Conditional Probability and Base Rate Fallacy

Consider a medical screening scenario that illustrates the crucial distinction between $p(T|V)$ and $p(V|T)$. A diagnostic test for a virus gives a positive result (T) in infected patients (where V denotes "has the virus") with probability 0.8—this is the test's *sensitivity* or true positive rate. The test also has a false positive rate of 0.1, meaning that $p(T| \sim V) = 0.1$ (the test incorrectly returns positive for 10% of uninfected people). Suppose the prevalence of the virus in the population is 1%, so $p(V) = 0.01$.

Now imagine you take the test and receive a positive result. What is the probability that you actually have the virus? Many people intuitively think the answer is 80% (the sensitivity of the test), but this confuses $p(T|V)$ with $p(V|T)$. We want $p(V|T)$—the probability of having the virus given a positive test—which we can calculate using Bayes' theorem:

$$p(V|T) = \frac{p(T|V)p(V)}{p(T)} = \frac{p(T|V)p(V)}{p(T|V)p(V) + p(T| \sim V)p(\sim V)}. \tag{7}$$

In the denominator, we have used marginalisation to write $p(T) = p(T, V) + p(T, \sim V) = p(T|V)p(V) + p(T| \sim V)p(\sim V)$ (applying the product rule to both terms).

Substituting the numerical values:

$$p(V|T) = \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.1 \times 0.99} = \frac{0.008}{0.008 + 0.099} = \frac{0.008}{0.107} \approx 0.075. \tag{8}$$

Thus, even with a positive test result, there is only a 7.5% probability that you have the virus—and still a 92.5% chance that you do not! This counterintuitive result arises because the disease is rare: false positives among the large uninfected population (0.1 × 0.99 = 0.099) outnumber true positives from the small infected population (0.8 × 0.01 = 0.008) by more than 10 to 1. This example shows why the base rate (prior probability) matters enormously in diagnostic testing, and why medical professionals must be trained to reason correctly with conditional probabilities.

2.6 Bayes' Theorem for Parameter Inference

In parameter inference, we work with three key quantities: the observed data $\mathbf{d}$, a model M that describes how those data arise, and the model parameters $\boldsymbol{\theta}$ whose values we wish to determine. The first and most important rule when approaching any inference problem is to write down precisely what we want to know.

In most cases, what we seek is the probability distribution for the parameters given the data and assuming the model, denoted $p(\boldsymbol{\theta}|\mathbf{d}, M)$. This distribution, called the **posterior**, encapsulates everything the data tell us about the parameter values within the framework of the chosen model. The posterior is computed using Bayes' theorem:

$$p(\boldsymbol{\theta}|\mathbf{d}, M) = \frac{p(\mathbf{d}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}{p(\mathbf{d}|M)}. \tag{9}$$

The three terms on the right-hand side have specific names and interpretations. The term $p(\mathbf{d}|\boldsymbol{\theta}, M)$ is the **likelihood**, commonly denoted $\mathcal{L}(\mathbf{d}|\boldsymbol{\theta})$, which quantifies how probable the observed data are if the parameters take specific values $\boldsymbol{\theta}$. The term $p(\boldsymbol{\theta}|M)$ is the **prior**, denoted $\pi(\boldsymbol{\theta})$, which encodes our state of knowledge about the parameters before observing the data. Finally, $p(\mathbf{d}|M)$ is the **Bayesian evidence** (or marginal likelihood), which plays a crucial role in model comparison but serves only to normalize the posterior in parameter inference, ensuring that it integrates to unity over all possible parameter values.

For clarity, when focusing solely on parameter inference with a fixed model, we often suppress the explicit dependence on M and write Bayes' theorem in the simplified form:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{\mathcal{L}(\mathbf{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{d})}. \tag{10}$$

This formula will be used throughout the remainder of this section, with the understanding that all probabilities are implicitly conditioned on the chosen model. We will restore the model dependence explicitly when we turn to the problem of model comparison.

In the context of parameter inference (i.e. for a given fixed model M), the Evidence serves only to make the posterior a properly normalised probability distribution as a function of the parameters $\boldsymbol{\theta}$. For continuous parameters (re-introducing M),

$$p(\mathbf{d}|M) = \int p(\mathbf{d}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{11}$$

where the integral may be multidimensional (multiple parameters).

2.7 Practical Problem Analysis

Setup is the most important and useful step when tackling almost any problem. It is about comprehensively answering the following questions:

- What are the data, $\mathbf{d}$?
- What is the model for the data? In other words, what is the likelihood function $\mathcal{L}(\boldsymbol{\theta})$?
- What is the model considered? How is it structured? What are the parameters of interest? What are the relationships between the parameters (known and unknown)?
- What is the prior on the parameters $\pi(\boldsymbol{\theta})$?
- Are there any other pieces of information available? It can be useful to write down all of the probability distributions and relationships known or accessible. Typically this involves writing down a list of conditional probability distributions.

3 The Likelihood

3.1 The Likelihood Function and the Sampling Distribution

It is important to pause here to think about $\mathcal{L}$. We can view this distribution two ways. If we fix $\boldsymbol{\theta}$ (as is rather implied by the expression), then we have the distribution of the data for given $\boldsymbol{\theta}$. This is a proper probability distribution that integrates to unity when integrated over all possible data $\mathbf{d}$. Used this way it is properly called the Sampling Distribution.

In Bayesian inference, though, the data are fixed (that is what we have), and this term is treated as a function of $\boldsymbol{\theta}$. In this context, it is called the Likelihood, and is not a proper probability distribution, in the sense that integrating it over $\boldsymbol{\theta}$ at fixed $\mathbf{d}$ does not give unity. Only the full posterior does this.

In practice, real data often contains outliers that can strongly affect inference with standard Gaussian likelihoods. Section 20.10 (Robust Statistics) discusses robust likelihoods using heavy-tailed distributions like the Student-t that down-weight extreme observations.

3.2 Example: Gaussian Likelihood and Maximum Likelihood Estimation

Example: Gaussian Likelihood and Maximum Likelihood Estimation

Problem: A scientist performs repeated measurements of a physical quantity and wants to estimate its true value. Each measurement is subject to random Gaussian noise with known standard deviation. Given a set of independent measurements, how should we combine them to obtain the best estimate of the true value? What is the uncertainty in this estimate? This is one of the most fundamental problems in data analysis, arising in contexts from laboratory experiments to astronomical observations.

Solution:

Suppose we have N independent measurements $\{d_1, d_2, \ldots, d_N\}$ of a quantity whose true value is μ. Each measurement is corrupted by independent Gaussian noise with known variance σ^2:

$$d_i = \mu + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{12}$$

The sampling distribution for each measurement is:

$$p(d_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d_i-\mu)^2}{2\sigma^2}\right) = \mathcal{N}(d_i;\mu,\sigma^2) \tag{13}$$

Since the measurements are independent, the joint sampling distribution (the probability of observing all the data given the parameters) is the product of individual probabilities:

$$p(\mathbf{d}|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d_i-\mu)^2}{2\sigma^2}\right) \tag{14}$$

Taking the logarithm (which is monotonic and thus preserves the location of extrema) gives the log-likelihood:

$$\ln\mathcal{L}(\mu,\sigma) = -\frac{N}{2}\ln(2\pi) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2 \tag{15}$$

Maximum Likelihood Estimate for the Mean:

To find the maximum likelihood estimate (MLE) of μ, we differentiate the log-likelihood with respect to μ and set the derivative to zero:

$$\frac{\partial \ln\mathcal{L}}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(d_i-\mu) = 0 \tag{16}$$

Solving for μ yields the familiar sample mean:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} d_i = \bar{d} \tag{17}$$

This result confirms the intuitive notion that the best estimate of the true value is simply the average of all measurements.

Maximum Likelihood Estimate for the Variance:

If the variance σ^2 is also unknown, we can find its MLE by differentiating with respect to σ:

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(d_i - \mu)^2 = 0 \tag{18}$$

Solving yields:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N}(d_i - \hat{\mu})^2 \tag{19}$$

This is the sample variance. Note that the MLE of the variance uses N in the denominator, which is known to be biased. The unbiased estimator uses $N-1$ instead, accounting for the fact that one degree of freedom was used to estimate μ. From a Bayesian perspective with uniform priors, the MAP estimate coincides with the MLE, but the full posterior provides proper uncertainty quantification without needing to apply ad-hoc bias corrections.

Geometric Interpretation:

The MLE for μ can be understood geometrically: it minimizes the sum of squared residuals $\sum_{i=1}^{N}(d_i - \mu)^2$. This is equivalent to finding the point μ that is closest to all data points in a least-squares sense. The Gaussian assumption makes this the maximum likelihood solution, but least-squares is optimal only when errors are truly Gaussian. For non-Gaussian noise (e.g., heavy-tailed distributions with outliers), other likelihood functions may be more appropriate.

Connection to Bayesian Inference and Posterior Derivation:

The likelihood function derived here forms the foundation for Bayesian parameter estimation. In the Bayesian framework, we combine this likelihood with a prior $\pi(\mu)$ to obtain the posterior using Bayes' theorem:

$$p(\mu|\mathbf{d}, \sigma) = \frac{p(\mathbf{d}|\mu, \sigma)\pi(\mu)}{p(\mathbf{d}|\sigma)} \tag{20}$$

Let us derive the posterior distribution for the simplest case: a uniform (noninformative) prior on μ. This illustrates the basic mechanics of Bayesian updating when we have no prior knowledge.

Step 1: Write down the likelihood

From our earlier derivation, the likelihood for N independent Gaussian measurements is:

$$p(\mathbf{d}|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i - \mu)^2\right) \tag{21}$$

Step 2: Complete the square in μ

Expanding the sum of squares and completing the square (details omitted), we obtain:

$$\sum_{i=1}^{N}(d_i - \mu)^2 = \sum_{i=1}^{N} d_i^2 - N\bar{d}^2 + N(\mu - \bar{d})^2 \tag{22}$$

where $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$ is the sample mean.

Step 3: Apply Bayes' theorem with uniform prior

A uniform prior $\pi(\mu) = \text{const}$ is improper (does not integrate to unity) but is often used as a noninformative prior when we have no prior knowledge about μ. With a uniform prior, the posterior is proportional to the likelihood:

$$p(\mu|\mathbf{d}, \sigma) \propto p(\mathbf{d}|\mu, \sigma) \propto \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right) \tag{23}$$

Step 4: Recognize the Gaussian form

This is the kernel of a Gaussian distribution in μ with mean $\bar{d}$ and variance σ^2/N. The properly normalized posterior is:

$$p(\mu|\mathbf{d}, \sigma) = \mathcal{N}\left(\mu; \bar{d}, \frac{\sigma^2}{N}\right) \tag{24}$$

Interpretation:

This derivation shows several important features of Bayesian inference:

- The posterior mean is the sample mean $\bar{d}$, which coincides with the MLE when using a uniform prior
- The posterior variance is σ^2/N, meaning the posterior uncertainty decreases as $1/\sqrt{N}$ with more measurements—a manifestation of the central limit theorem
- With a uniform prior, the MAP estimate coincides with the MLE, but the Bayesian approach provides the full posterior distribution, which quantifies our uncertainty about μ rather than just a point estimate
- The posterior is properly normalized (integrates to unity over μ), unlike the likelihood which does not integrate to unity over μ for fixed data

For the case of a Gaussian prior on μ, which demonstrates conjugacy and Bayesian learning, see Section 4.4.

4 The Prior

4.1 Specifying Prior Distributions

The prior distribution $\pi(\boldsymbol{\theta})$ encodes our state of knowledge about parameters before observing new data. Priors can be derived from three main sources:

1. **Previous experiments:** When prior experimental data exist, they provide empirical information about parameter values. The posterior from a previous analysis can serve as the prior for a new dataset, creating a sequential updating framework. In fact, probabilities are sometimes explicitly written to include this prior information I, as $p(\boldsymbol{\theta}|I, M)$, though we typically suppress this notation for clarity.

2. **Theoretical models:** Physical theories, conservation laws, or domain knowledge can constrain parameter ranges or suggest functional forms. For example, in physics we might know that a mass must be positive, or that a probability must lie in $[0, 1]$. Theoretical models can also specify relationships between parameters that inform prior construction.

3. **Analysis choices:** In the absence of previous experiments or strong theoretical constraints, we may choose priors based on analytical considerations:

   - **Uniform priors**: $\pi(\theta) = \text{const}$ are typical for *location parameters*, such as the mean of a distribution
   - **Jeffreys priors**: $\pi(\theta) \propto 1/\theta$ are often applied to *scale parameters* where the parameter is positive. Each decade is equally likely, making the prior uniform in $\ln\theta$. For a rigorous information-theoretic foundation of Jeffreys priors, see Section 20.2 (Information Theory) and Section 20.11 (Maximum Entropy Priors).
   - **Conjugate priors**: Priors chosen so that the posterior has the same functional form as the prior, simplifying analytical calculations (see Section 4.4)

The choice of prior matters for two distinct reasons:

- **Impact on results:** The prior directly affects the posterior distribution, particularly when data are sparse or weakly informative. For parameter inference with abundant data, the likelihood dominates and the prior becomes less important. However, for model comparison (see Section 8), the prior remains critically important regardless of data quantity.

- **Analytical simplification:** Certain prior choices (particularly conjugate priors) enable closed-form posterior calculations, avoiding the need for numerical methods like MCMC. This can dramatically reduce computational cost and improve interpretability.

For a Gaussian likelihood, one might reasonably choose a uniform prior for the mean (a location parameter) and a Jeffreys prior for the standard deviation (a scale parameter). Note that we sometimes assume a uniform prior over an infinite range, which is an *improper prior*—it cannot be normalized to integrate to 1. Provided it yields a proper posterior, this is acceptable for parameter inference. (For model comparison, we must use proper priors). Bayesian parameter inference can also be extended to infinite-dimensional parameter spaces using Gaussian Processes (Section 20.3).

4.2 Uninformative Priors and the Curse of Dimensionality

Using previous data to define our state of knowledge is fine, but the very first dataset that was used to determine our state of knowledge will have had to have a prior with no previous data to go on. For such situations, we often try to choose an 'uninformative' prior, which is not as easy as it sounds (and its meaning may not be particularly well-defined).

A uniform prior may seem natural, but it is worth thinking a bit more. Consider this problem: imagine a parameter space with N dimensions, represented by cartesian coordinates $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$, with each coordinate restricted to the range $(-\frac{1}{2}, \frac{1}{2})$. We adopt a uniform prior over this N-dimensional hypercube:

$$\pi(\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } |\theta_i| \leq \frac{1}{2} \text{ for all } i = 1, \ldots, N \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

This prior assigns equal probability density to all points within the hypercube. The total volume of this hypercube is $V_{\text{cube}} = 1^N = 1$.

Now consider the N-sphere of radius $r = \frac{1}{2}$ inscribed within this hypercube (the largest sphere that fits inside). The volume of an N-sphere of radius r is given by:

$$V_{\text{sphere}}(r) = \frac{\pi^{N/2}}{\Gamma(1 + N/2)} r^N \tag{26}$$

For $r = \frac{1}{2}$, this becomes:

$$V_{\text{sphere}}\left(\frac{1}{2}\right) = \frac{\pi^{N/2}}{\Gamma(1 + N/2)} \left(\frac{1}{2}\right)^N = \frac{\pi^{N/2}}{2^N \Gamma(1 + N/2)} \tag{27}$$

Under the uniform prior, the probability of a randomly drawn parameter vector $\boldsymbol{\theta}$ falling inside the inscribed N-sphere is simply the ratio of the sphere volume to the hypercube volume:

$$p(\boldsymbol{\theta} \in \text{sphere}) = \frac{V_{\text{sphere}}}{V_{\text{cube}}} = \frac{\pi^{N/2}}{2^N \Gamma(1 + N/2)} \tag{28}$$

This probability decreases rapidly as N increases, illustrating the curse of dimensionality: in high dimensions, most of the volume of a hypercube is concentrated in its corners, far from the center. The ratio of the volume of the inscribed N-sphere to the volume of the hypercube goes to zero exponentially fast (see Figure 2). For the simplest case of $N = 2$, this reduces to the familiar $\pi/4$. This counterintuitive behavior demonstrates that uniform priors can be highly informative in high-dimensional spaces, as they implicitly favor parameter values at the extremes of the prior range rather than near the center.

An apparently uninformative prior may be highly informative when viewed a different way.

4.3 Sequential Updating of Priors

If we now obtain some more information, perhaps from a new experiment, then we can use Bayes' theorem to update our state of knowledge of the parameters. The posterior of the last experiment becomes the prior for the next one. This is fine, but it begs the question of what prior did the very first experiment use? This is where the 'uninformative' priors come in.

For this to be a consistent process, we should verify that the order and manner in which we analyze data does not affect the final posterior. Given two datasets $\mathbf{d}_1$ and $\mathbf{d}_2$, there are several possible analysis scenarios:

1. **Analyzing $\mathbf{d}_1$ alone:** Update the prior $\pi(\boldsymbol{\theta})$ with only the first dataset to obtain $p(\boldsymbol{\theta}|\mathbf{d}_1)$
2. **Analyzing $\mathbf{d}_2$ alone:** Update the prior $\pi(\boldsymbol{\theta})$ with only the second dataset to obtain $p(\boldsymbol{\theta}|\mathbf{d}_2)$
3. **Analyzing $\mathbf{d}_1$ before $\mathbf{d}_2$:** First obtain $p(\boldsymbol{\theta}|\mathbf{d}_1)$, then use this as the prior for analyzing $\mathbf{d}_2$ to obtain $p(\boldsymbol{\theta}|\mathbf{d}_2, \mathbf{d}_1)$
4. **Analyzing $\mathbf{d}_2$ before $\mathbf{d}_1$:** First obtain $p(\boldsymbol{\theta}|\mathbf{d}_2)$, then use this as the prior for analyzing $\mathbf{d}_1$ to obtain $p(\boldsymbol{\theta}|\mathbf{d}_1, \mathbf{d}_2)$
5. **Analyzing $\mathbf{d}_1$ and $\mathbf{d}_2$ together:** Combine both datasets and update the prior simultaneously to obtain $p(\boldsymbol{\theta}|\mathbf{d}_1, \mathbf{d}_2)$

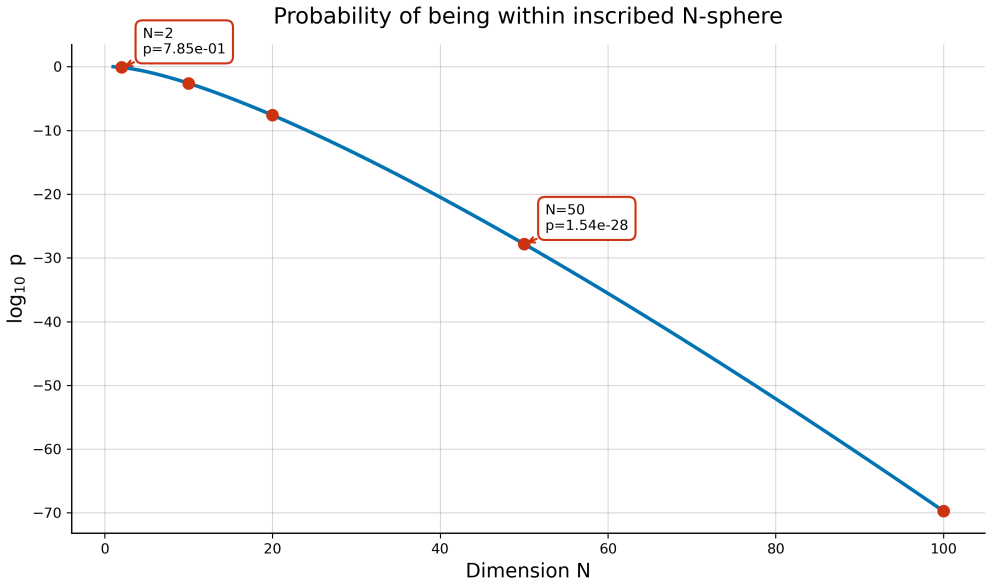

Figure 2: Probability of being within an inscribed N-sphere as a function of dimension N. The plot shows $\log_{10} p$ vs N, demonstrating the curse of dimensionality: as dimensions increase, nearly all the probability mass concentrates in the corners of the hypercube, far from the center.

A key requirement for consistency is that scenarios 3, 4, and 5 must yield the same final posterior distribution. We now demonstrate this equivalence, showing that sequential updating (scenarios 3 and 4) produces identical results to joint analysis (scenario 5), and that the order of sequential updates does not matter.

Let's do the analysis in two stages, firstly analysing $\mathbf{d}_1$. Let's be explicit about the prior information I (defined to be the state of knowledge before the first experiment is done) in Bayes' theorem applied to the first dataset:

$$p(\boldsymbol{\theta}|\mathbf{d}_1, I) = \frac{p(\mathbf{d}_1|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\mathbf{d}_1|I)}. \tag{29}$$

Now we analyse the second data set. It's similar, but the data are different, $\mathbf{d}_1 \to \mathbf{d}_2$ of course, and we have some extra information from the first experiment, so we should update I to include $\mathbf{d}_1$:

$$I \to \mathbf{d}_1, I. \tag{30}$$

So, Bayes' theorem applied to the second dataset gives a posterior

$$p(\boldsymbol{\theta}|\mathbf{d}_2, \mathbf{d}_1, I) = \frac{p(\mathbf{d}_2|\boldsymbol{\theta}, \mathbf{d}_1, I)p(\boldsymbol{\theta}|\mathbf{d}_1, I)}{p(\mathbf{d}_2|\mathbf{d}_1, I)}. \tag{31}$$

We now notice that the new prior in this expression is just the old posterior probability from equation (4.1), i.e. we have updated our prior state of knowledge from the original prior, instead using the posterior from the first dataset.

We can also the rules of probability to write the new likelihood as

$$p(\mathbf{d}_2|\mathbf{d}_1, \boldsymbol{\theta}, I) = \frac{p(\mathbf{d}_2, \mathbf{d}_1|\boldsymbol{\theta}, I)}{p(\mathbf{d}_1|\boldsymbol{\theta}, I)}. \tag{32}$$

Substituting this into equation (4.3) in the old posterior probability along with the expression for the posterior after analysing $\mathbf{d}_1$ (equation 4.1 gives

$$p(\boldsymbol{\theta}|\mathbf{d}_1, \mathbf{d}_2, \boldsymbol{\theta}, I) = \frac{1}{p(\mathbf{d}_2|\mathbf{d}_1, I)} \times \frac{p(\mathbf{d}_2, \mathbf{d}_1|\boldsymbol{\theta}, I)}{\cancel{p(\mathbf{d}_1|\boldsymbol{\theta}, I)}} \times \frac{\cancel{p(\mathbf{d}_1|\boldsymbol{\theta}, I)}p(\boldsymbol{\theta}|I)}{p(\mathbf{d}_1|I)} \tag{33}$$

So we see that the posterior after the second dataset is

$$p(\boldsymbol{\theta}|\mathbf{d}_2, \mathbf{d}_1, I) = \frac{p(\mathbf{d}_2, \mathbf{d}_1|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\mathbf{d}_2, \mathbf{d}_1|I)}, \tag{34}$$

where we have used the product rule in the denominator. This has the same form as equation (4.1), the outcome from the initial experiment, but now with the new data incorporated, i.e. the result we would write down if we analysed the data together ($\mathbf{d} \to \{\mathbf{d}_1, \mathbf{d}_2\}$). So, analysing separately and updating the prior after the first dataset gives the same answer as analysing the combined dataset with the original prior.

Bayes' theorem gives us a natural and self-consistent way of improving our statistical inferences as our state of knowledge increases.

4.4 Conjugate Priors

Sometimes a prior is chosen for mathematical convenience, where, when combined with a given form for the likelihood, the posterior can be calculated analytically and has the same mathematical form as the prior. This property—called conjugacy—offers both computational advantages (avoiding numerical integration or sampling) and interpretive benefits (the prior and posterior have the same functional form, making sequential updating transparent).

Examples of conjugate pairs include:

- **Gaussian likelihood with Gaussian prior:** If the likelihood is Gaussian with known variance, and the mean μ is the parameter of interest, then a conjugate prior is also a Gaussian. The conjugate prior can have any mean and variance, so it is flexible enough to encode various degrees of prior knowledge.
- **Gaussian likelihood with known mean:** If the mean μ is known but the variance σ^2 is unknown, an inverse gamma distribution is a conjugate prior for σ^2.
- **Binomial likelihood:** For coin flips or Bernoulli trials, the Beta distribution is conjugate to the Binomial likelihood (demonstrated in the coin toss example below).

Note that there is nothing inherently special about conjugate priors; they are just convenient mathematically, but they may be flexible enough to specify sensible location and scale constraints. When they are not flexible enough or inappropriate for the problem, non-conjugate priors should be used, accepting the need for numerical methods.

We now present a detailed worked example of Gaussian-Gaussian conjugacy, extending the Gaussian likelihood example from Section 3.2 by incorporating an informative prior.

Example: Gaussian Likelihood with Gaussian Prior: Complete Conjugacy Derivation

Problem: Building on the Gaussian likelihood example from Section 3.2, suppose we now have prior knowledge about the mean parameter μ from previous experiments or theoretical considerations. Rather than using a uniform (noninformative) prior, we encode this prior knowledge as a Gaussian distribution with mean μ_0 and variance σ_0^2. How do we combine this informative prior with new data to obtain the posterior distribution? This scenario is fundamental in Bayesian inference and demonstrates the power of conjugacy: when both the prior and likelihood are Gaussian, the posterior is also Gaussian, and we can derive its parameters analytically.

Solution:

As in Section 3.2, we have N independent measurements $\{d_1, d_2, \ldots, d_N\}$ with known measurement variance σ^2. The likelihood is:

$$p(\mathbf{d}|\mu, \sigma) = \prod_{i=1}^{N} \mathcal{N}(d_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i - \mu)^2\right) \tag{35}$$

We now specify a Gaussian prior on μ:

$$\pi(\mu) = \mathcal{N}(\mu; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \tag{36}$$

Step 1: Apply Bayes' theorem
The posterior is proportional to the product of likelihood and prior:

$$p(\mu|\mathbf{d}, \sigma) \propto p(\mathbf{d}|\mu, \sigma)\pi(\mu) \tag{37}$$

Taking logarithms (to work with the exponents more easily):

$$\ln p(\mu|\mathbf{d}, \sigma) = \text{const} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i - \mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} \tag{38}$$

where "const" absorbs all terms that do not depend on μ.
Step 2: Express likelihood as Gaussian in μ
From Section 3.2, completing the square in the likelihood gives:

$$p(\mathbf{d}|\mu, \sigma) \propto \mathcal{N}\left(\mu; \bar{d}, \frac{\sigma^2}{N}\right) \tag{39}$$

where $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$ is the sample mean. The likelihood is a Gaussian in μ with mean $\bar{d}$ and variance σ^2/N.
Step 3: Apply product of Gaussians identity
The posterior is proportional to the product of two Gaussians (likelihood and prior), both in μ:

$$p(\mu|\mathbf{d}, \sigma) \propto \mathcal{N}\left(\mu; \bar{d}, \frac{\sigma^2}{N}\right) \cdot \mathcal{N}(\mu; \mu_0, \sigma_0^2) \tag{40}$$

Using the product of Gaussians identity from Section 1.4, this product is proportional to a Gaussian:

$$p(\mu|\mathbf{d}, \sigma) \propto \mathcal{N}(\mu; \mu_N, \sigma_N^2) \tag{41}$$

where the posterior precision (inverse variance) is the sum of the individual precisions:

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{42}$$

and the posterior mean is the precision-weighted average:

$$\mu_N = \sigma_N^2\left(\frac{N\bar{d}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) = \frac{\frac{N}{\sigma^2}\bar{d} + \frac{1}{\sigma_0^2}\mu_0}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \tag{43}$$

Step 4: Write the posterior
The posterior distribution is:

$$p(\mu|\mathbf{d}, \sigma) = \mathcal{N}(\mu; \mu_N, \sigma_N^2) \tag{44}$$

with parameters given above.
Interpretation:
This result reveals several profound insights:

- **Precision addition:** The posterior precision is the sum of the data precision (N/σ^2) and the prior precision ($1/\sigma_0^2$). Information combines additively when expressed as precisions.

- **Weighted average:** The posterior mean μ_N is a precision-weighted average of the sample mean $\bar{d}$ and the prior mean μ_0. Sources with higher precision (lower variance) receive more weight.
- **Data vs. prior dominance:**
  - If $N\sigma_0^2 \gg \sigma^2$ (many data or tight prior), then $\mu_N \approx \bar{d}$ (data dominates)
  - If $N\sigma_0^2 \ll \sigma^2$ (few data or loose prior), then $\mu_N \approx \mu_0$ (prior dominates)
- **Variance reduction:** The posterior variance σ_N^2 is always smaller than both the prior variance σ_0^2 and the data variance σ^2/N, reflecting increased certainty from combining information.
- **Conjugacy:** The Gaussian prior and Gaussian likelihood combine to yield a Gaussian posterior with updated parameters. This makes sequential updating trivial: the posterior from one dataset becomes the prior for the next.
- **Limiting cases:**
  - Uniform prior limit: As $\sigma_0^2 \to \infty$ (infinitely diffuse prior), we recover the result from Section 3.2: $\mu_N \to \bar{d}$ and $\sigma_N^2 \to \sigma^2/N$
  - Strong prior limit: As $\sigma_0^2 \to 0$ (infinitely precise prior), $\mu_N \to \mu_0$ and $\sigma_N^2 \to 0$ (data cannot overcome an infinitely strong prior belief)

Gaussian likelihood with known mean: If on the other hand the mean μ is known, but the variance σ^2 is unknown, an inverse gamma distribution is a conjugate prior for $x = \sigma^2$:

$$f(x; \alpha, \beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x} \tag{45}$$

Exercise: show that the posterior is also an inverse gamma distribution, with parameters updated as follows $\alpha \to \alpha + n/2$, and $\beta \to \beta + \sum_{i=1}^{n}(x_i - \mu)^2/2$ for n data.

5 The Posterior

The posterior is the natural outcome of a Bayesian inference problem. It encapsulates our current state of knowledge of the model parameters. It may be very high dimensional, if there are many parameters, and we may want to put it into a more digestible form. It is common to marginalise over all but two parameters, and plot marginal posteriors as a function of each pair of parameters. These are often plotted in 'corner plots'.

Figure 3 shows a corner plot for a three-parameter model with correlated posterior distributions. The diagonal panels display the one-dimensional marginal posterior distributions for each parameter, obtained by integrating over all other parameters. The 68% credible intervals are indicated by the shaded green regions. The off-diagonal panels show the two-dimensional marginal posteriors for each pair of parameters, with contours indicating the 68% and 95% credible regions. These credible regions are highest posterior density (HPD) regions, where the posterior probability density is highest. The correlations between parameters are clearly visible in the elliptical shapes of the 2D contours—for instance, θ_1 and θ_2 are positively correlated, while θ_1 and θ_3 are negatively correlated.

5.1 Example: Beta-Binomial Conjugacy and Prior Dominance

In parameter inference problems, as more data are collected, the likelihood gets progressively more peaked around the true parameter values. For a sufficiently narrow likelihood, the prior is almost constant over the relevant range, and it becomes unimportant (the height of the prior there is irrelevant as it is normalised away by the evidence in the denominator). Let us explore this through a detailed coin-flipping example (adapted from Sivia & Skilling) that demonstrates both the mechanics of Bayesian updating and how data dominates the prior as sample size increases. The probabilistic graphical model is shown in Figure 4.

Example: Coin Toss: Beta-Binomial Conjugacy and Prior Dominance

Problem: Consider a coin-flipping experiment designed to determine whether a coin is fair. The parameter of interest is θ, the probability of obtaining heads on a single toss, which can take any value between 0 and 1. We perform a sequence of tosses and observe the outcomes. We want to: (1) derive the complete posterior distribution using Bayes' theorem, identifying all components; (2) understand how the posterior evolves as more data are collected; and (3) explore how the choice of prior affects inference, particularly comparing uninformative and informative priors.

Solution:

To understand the posterior distribution completely, we derive it step by step and identify all probability distributions involved. We apply Bayes' theorem:

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)\pi(\theta)}{p(\text{data})} \tag{46}$$

The three components are:

1. The Prior: We assume a uniform (flat) prior over the interval $[0, 1]$:

$$\pi(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

This is a special case of the Beta distribution with parameters $\alpha = 1$ and $\beta = 1$, denoted $\text{Beta}(1, 1)$. The general Beta distribution is:

$$\text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{48}$$

For $\alpha = \beta = 1$, this reduces to the uniform distribution.

2. The Likelihood: For n independent coin tosses with k heads observed, the likelihood follows a binomial distribution:

$$p(k|\theta, n) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{49}$$

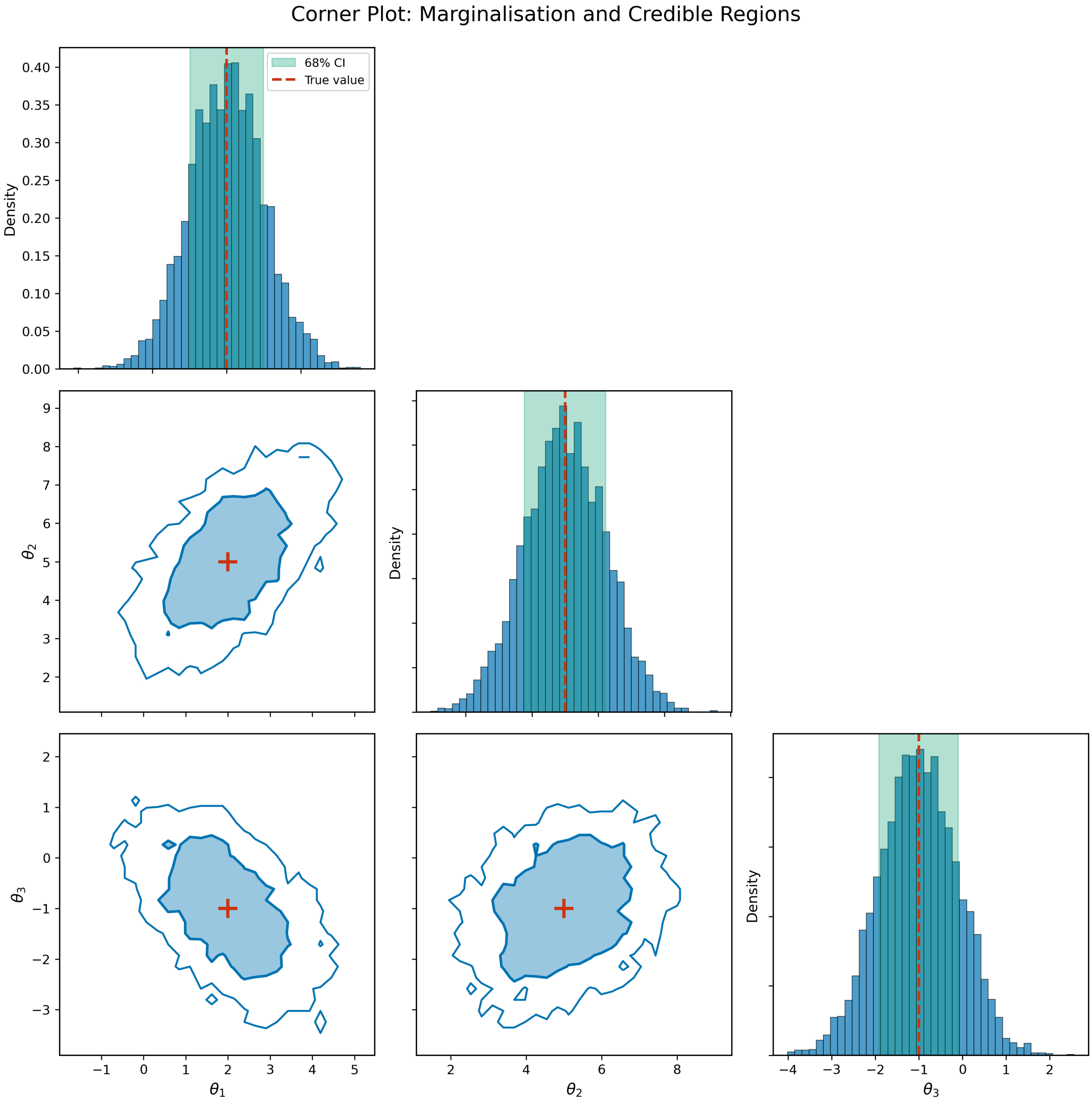

Figure 3: Corner plot illustrating marginalisation and credible regions for a three-parameter model. The diagonal shows 1D marginal posterior distributions with 68% credible intervals (green shaded regions). The off-diagonal panels display 2D marginal posteriors with 68% (dark blue) and 95% (light blue) credible contours. The red cross marks the true parameter values. The elliptical shapes of the contours reveal correlations between parameters.

Figure 4: DAG for coin toss model

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. This represents the probability of observing exactly k heads in n tosses, given that the probability of heads on each toss is θ.
3. The Evidence: The denominator is computed by marginalizing over all possible values of θ:

$$p(k|n) = \int_0^1 p(k|\theta, n)\pi(\theta)d\theta = \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k} \cdot 1\, d\theta \tag{50}$$

This integral can be evaluated using the definition of the Beta function $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. With $\alpha = k+1$ and $\beta = n-k+1$:

$$p(k|n) = \binom{n}{k}B(k+1, n-k+1) = \binom{n}{k}\frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \binom{n}{k}\frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1} \tag{51}$$

4. The Posterior: Combining these elements:

$$\begin{aligned} p(\theta|k, n) &= \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k} \cdot 1}{\frac{1}{n+1}} && (52)\\ &= (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k} && (53)\\ &= \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k} && (54)\\ &= \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)}\theta^k(1-\theta)^{n-k} && (55) \end{aligned}$$

This is precisely the Beta distribution $\text{Beta}(\theta|k+1, n-k+1)$. The posterior has updated the parameters of the Beta prior: $\alpha = 1 \to k+1$ and $\beta = 1 \to n-k+1$.
5. Sequential Updating and Prior Dominance:
After the first toss, suppose we observe a head (H). Using Bayes' theorem, the posterior is:

$$p(\theta|H) \propto p(H|\theta)\pi(\theta) = \theta \times 1 = \theta \tag{56}$$

This posterior is linear in θ, favoring higher values of θ but still allowing for the possibility that the coin is biased toward tails.
As we continue tossing the coin, we accumulate more data. Suppose the sequence begins HHTT... and eventually we observe 40 heads in 64 tosses. For this specific case, the posterior is:

$$p(\theta|k = 40, n = 64) = \text{Beta}(\theta|41, 25) = \frac{\Gamma(66)}{\Gamma(41)\Gamma(25)}\theta^{40}(1-\theta)^{24} \tag{57}$$

The mean of this Beta distribution is $\frac{\alpha}{\alpha+\beta} = \frac{41}{66} \approx 0.621$ and the mode (the peak) is $\frac{\alpha-1}{\alpha+\beta-2} = \frac{40}{64} = 0.625$, which matches the observed frequency of heads. The variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \approx 0.0035$, indicating a relatively narrow distribution concentrated around the observed frequency.

6. Comparison with Informative Priors:
Figure 5 illustrates posteriors for different scenarios comparing uninformative and informative priors with varying amounts of data. The key insights are:

- **Small sample size:** With only a few tosses, the choice of prior matters considerably. An informative prior strongly influences the posterior, and different priors yield substantially different posteriors.
- **Large sample size:** After many tosses, the likelihood dominates, and the posterior is largely determined by the data rather than the prior. If we had started with a different prior—say, one strongly biased toward $\theta = 0.5$ (a highly informative prior favoring a fair coin)—we would need more data to overcome this prior belief. But eventually, with sufficient data, even a highly informative prior would be overwhelmed by the evidence, and the posteriors from different priors would converge to similar distributions centered on the observed frequency.
- **Data dominance principle:** This demonstrates a key principle of Bayesian inference: as more data are collected, the posterior becomes increasingly concentrated around the true parameter value, and the influence of the prior diminishes. The prior is eventually "washed out" by the data.

5.2 Marginalisation and Nuisance Parameters

In many inference problems, we are interested in only a subset of the model parameters. The remaining parameters—called nuisance parameters—must be included in the model because they affect the data, but we do not care about their values for our scientific question. Marginalization is the operation that allows us to "integrate out" these nuisance parameters, obtaining the posterior distribution for only the parameters of interest while properly accounting for uncertainty in the nuisance parameters.

The Marginalisation Operation:
Marginalising over all n parameters except θ_1 and θ_2 is accomplished by integration:

$$p(\theta_1, \theta_2|\mathbf{d}) = \int p(\theta_1, \ldots, \theta_n|\mathbf{d}) d\theta_3 \ldots d\theta_n \tag{58}$$

More generally, suppose we partition the parameters into those of interest $\boldsymbol{\phi}$ (the target parameters) and nuisance parameters $\boldsymbol{\psi}$, so that $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\psi})$. The joint posterior is:

$$p(\boldsymbol{\phi}, \boldsymbol{\psi}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\phi}, \boldsymbol{\psi})\pi(\boldsymbol{\phi}, \boldsymbol{\psi})}{p(\mathbf{d})} \tag{59}$$

To obtain the marginal posterior for the parameters of interest, we integrate over the nuisance parameters:

$$p(\boldsymbol{\phi}|\mathbf{d}) = \int p(\boldsymbol{\phi}, \boldsymbol{\psi}|\mathbf{d}) d\boldsymbol{\psi} \tag{60}$$

This marginalisation automatically accounts for the uncertainty in the nuisance parameters, producing the correct uncertainties for $\boldsymbol{\phi}$.

Analytical vs Numerical Marginalization:
Marginalization can sometimes be performed analytically, particularly when:

- The posterior has an analytical form (often the case with conjugate priors)
- The integral over nuisance parameters can be evaluated in closed form
- The model structure allows separation of parameters

However, most real-world problems require numerical marginalization through methods like:

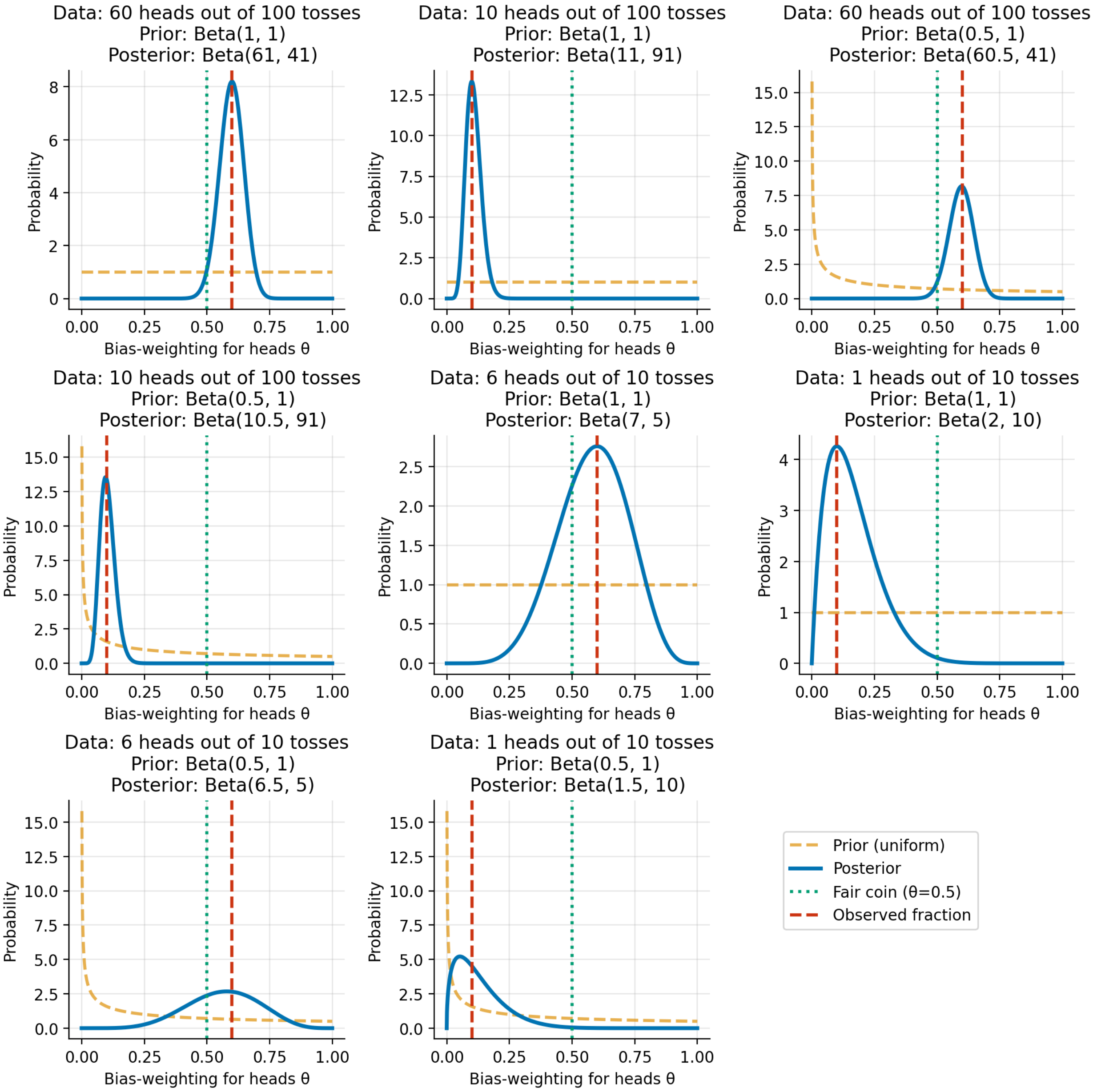

Figure 5: Coin toss analysis showing different scenarios for the data and the prior distributions, yielding different posterior distributions. As the number of observations increases, the posterior becomes increasingly dominated by the data, and the choice of prior becomes less important.

- **Numerical integration** for low-dimensional nuisance parameter spaces
- **Monte Carlo sampling** (MCMC, Section 6) where marginalization is achieved by simply ignoring samples of nuisance parameters when computing summaries
- **Variational inference** (Section 20.4) for approximate marginalization in high dimensions

We now present three examples illustrating analytical marginalization in simple but important scenarios.

Example: Gaussian Measurements: Marginalizing Over Unknown Variance

Problem: In Sections 3.2 and 4.4, we analyzed Gaussian measurements $\{d_1, \ldots, d_N\}$ assuming the measurement variance σ^2 was known. In practice, σ^2 is often unknown and must be inferred alongside the mean μ. If we are primarily interested in μ, how do we properly account for our uncertainty in σ^2? This requires marginalizing over σ^2 as a nuisance parameter.
Solution:
Starting from the Gaussian likelihood (Section 3.2), we now treat both μ and σ^2 as unknown. With a uniform prior on μ and a scale-invariant Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$, the joint posterior is:

$$p(\mu, \sigma^2|\mathbf{d}) \propto (\sigma^2)^{-(N/2+1)} \exp\left(-\frac{S(\mu)}{2\sigma^2}\right) \tag{61}$$

where $S(\mu) = \sum_{i=1}^{N}(d_i - \mu)^2$.
Marginalizing over σ^2:

$$p(\mu|\mathbf{d}) = \int_0^\infty p(\mu, \sigma^2|\mathbf{d})d\sigma^2 \propto \int_0^\infty (\sigma^2)^{-(N/2+1)} \exp\left(-\frac{S(\mu)}{2\sigma^2}\right) d\sigma^2 \tag{62}$$

This integral evaluates to a Student's t-distribution:

$$p(\mu|\mathbf{d}) \propto \left[1 + \frac{(\mu - \bar{d})^2}{s^2/(N-1)}\right]^{-N/2} \tag{63}$$

where $\bar{d} = \frac{1}{N}\sum d_i$ is the sample mean and $s^2 = \frac{1}{N-1}\sum(d_i - \bar{d})^2$ is the sample variance.
Interpretation: Marginalizing over the unknown variance produces heavier tails than the Gaussian posterior from Section 3.2—the t-distribution properly accounts for uncertainty in σ^2. As $N \to \infty$, uncertainty in σ^2 decreases and the t-distribution approaches the Gaussian limit. This demonstrates that ignoring nuisance parameters by fixing them at point estimates underestimates uncertainty.

Example: Coin Toss: Marginalizing Over Hyperparameters

Problem: In Section 5.1, we analyzed coin tosses using a Beta(α, β) prior on the success probability θ. Suppose we are uncertain about the hyperparameters α and β themselves—perhaps we know the coin comes from a manufacturer with variable quality control, but we don't know the exact distribution of bias. How do we account for this higher-level uncertainty when inferring θ?
Solution:
We build a hierarchical model by placing a prior $\pi(\alpha, \beta)$ on the hyperparameters:

$$k|\theta, n \sim \text{Binomial}(n, \theta) \tag{64}$$

$$\theta|\alpha, \beta \sim \text{Beta}(\alpha, \beta) \tag{65}$$

$$\alpha, \beta \sim \pi(\alpha, \beta) \tag{66}$$

The joint posterior is:

$$p(\theta, \alpha, \beta|k, n) \propto \binom{n}{k}\theta^k(1-\theta)^{n-k} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \cdot \pi(\alpha, \beta) \tag{67}$$

To obtain the marginal posterior for θ, we integrate over the hyperparameters:

$$p(\theta|k, n) = \int\int p(\theta, \alpha, \beta|k, n) d\alpha\, d\beta \tag{68}$$

Interpretation: For most choices of $\pi(\alpha, \beta)$, this integral requires numerical evaluation (typically via MCMC, Section 6). Marginalization propagates our uncertainty about the hyperparameters into our inference about θ, yielding wider credible intervals than the fixed-hyperparameter case in Section 5.1. This is the foundation of hierarchical Bayesian modeling (Section 7): uncertainty at each level of the hierarchy is properly accounted for through marginalization.

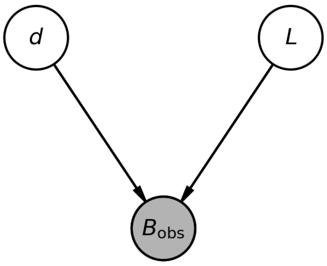

Figure 6: DAG for supernova distance measurement. Distance d (parameter of interest) and luminosity L (nuisance parameter) determine the observed brightness B_{obs}.

Example: Supernova: Nuisance Parameter Marginalization

Problem: Astronomers observe a supernova and measure its apparent brightness. They want to determine the distance to the supernova, which is crucial for cosmological measurements. However, the observed brightness depends on both the distance and the intrinsic luminosity of the supernova. Since each supernova explosion has a different intrinsic luminosity that is not directly observable, we face a classical inference challenge: how to determine the distance when another unknown parameter (the luminosity) affects our measurements. The distance is our parameter of interest, while the luminosity is a nuisance parameter that we must account for but are not primarily interested in. This example demonstrates how to properly marginalize over nuisance parameters to obtain correct uncertainties for the parameter of interest.

Solution:

The physical relationship between observed brightness, distance, and luminosity follows the inverse-square law. The model is:

$$B_{\text{obs}} = \frac{L}{4\pi d^2} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{69}$$

where d is the parameter of interest and L is a nuisance parameter. The likelihood is:

$$p(B_{\text{obs}}|d, L) = \mathcal{N}\left(B_{\text{obs}}; \frac{L}{4\pi d^2}, \sigma^2\right) \tag{70}$$

With a uniform prior on d and a prior $\pi(L)$ on the luminosity (perhaps from similar supernovae), the joint posterior is:

$$p(d, L|B_{\text{obs}}) \propto \exp\left(-\frac{1}{2\sigma^2}\left(B_{\text{obs}} - \frac{L}{4\pi d^2}\right)^2\right)\pi(L) \tag{71}$$

The marginal posterior for distance is obtained by integrating over all possible luminosities:

$$p(d|B_{\text{obs}}) = \int_0^\infty p(d, L|B_{\text{obs}})dL \tag{72}$$

This marginal posterior properly accounts for our uncertainty in the intrinsic luminosity L, yielding larger uncertainties in d than would be obtained if we fixed L at a single value. Other examples of nuisance parameters include detector gain (which may have a prior set by laboratory calibration measurements) or calibration parameters that affect the relationship between observed and true quantities.

5.3 Conditional vs Marginal Uncertainties

If all parameters are kept fixed (typically at the maximum posterior values), and the posterior distribution computed as a function of the remaining parameter, this is a conditional distribution, with an associated conditional error (e.g. standard deviation). For example, if we fix $\theta_2, \ldots, \theta_n$ at specific values $\theta_2^*, \ldots, \theta_n^*$ (often the maximum posterior values), the conditional distribution for θ_1 is:

$$p(\theta_1|\theta_2^*, \ldots, \theta_n^*, \mathbf{d}) \propto p(\theta_1, \theta_2^*, \ldots, \theta_n^*|\mathbf{d}) \tag{73}$$

This should be contrasted with the marginal distribution:

$$p(\theta_1|\mathbf{d}) = \int p(\theta_1, \theta_2, \ldots, \theta_n|\mathbf{d})d\theta_2 \ldots d\theta_n \tag{74}$$

The conditional distribution is rarely relevant for reporting parameter uncertainties, since it does not reflect the additional uncertainty that arises from incomplete knowledge of the other parameters. The marginal distribution properly accounts for this uncertainty by integrating over all possible values of the nuisance parameters, weighted by their posterior probability.

To see why conditional errors underestimate the true uncertainty, consider the variance. The marginal variance can be decomposed as:

$$\text{Var}[\theta_1|\mathbf{d}] = \mathbb{E}_{\theta_2,\ldots,\theta_n}[\text{Var}[\theta_1|\theta_2, \ldots, \theta_n, \mathbf{d}]] + \text{Var}_{\theta_2,\ldots,\theta_n}[\mathbb{E}[\theta_1|\theta_2, \ldots, \theta_n, \mathbf{d}]] \tag{75}$$

where the expectation and variance on the right-hand side are taken over the marginal posterior of $\theta_2, \ldots, \theta_n$. This is the law of total variance. The first term represents the average conditional variance, while the second term represents the additional variance due to uncertainty in the other parameters. Since both terms are non-negative, we always have:

$$\text{Var}[\theta_1|\mathbf{d}] \geq \text{Var}[\theta_1|\theta_2^*, \ldots, \theta_n^*, \mathbf{d}] \tag{76}$$

Thus, the marginal standard deviation (uncertainty) is always at least as large as the conditional standard deviation, with equality only when θ_1 is independent of the other parameters in the posterior.

5.4 Summarizing the Posterior Distribution

Once the posterior distribution $p(\boldsymbol{\theta}|\mathbf{d})$ has been computed, we typically need to summarize it for reporting and decision-making. The appropriate summary depends on the shape of the posterior and the intended use. This section describes the main approaches to posterior summarization: point estimates, uncertainty quantification through credible regions, and when simple summaries are sufficient versus when the full distribution must be preserved.

5.4.1 Point Estimates

Several point estimates can be extracted from the posterior, each optimal under different loss functions:

Posterior mean: The expected value of the parameters under the posterior:

$$\bar{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta}|\mathbf{d}] = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{d})d\boldsymbol{\theta} \tag{77}$$

This minimizes the expected squared error and is optimal for symmetric loss functions. It is the natural summary for Gaussian posteriors.

Maximum a posteriori (MAP) estimate: The mode of the posterior:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{d}) \tag{78}$$

With uniform priors, this reduces to the maximum likelihood estimate (MLE). The MAP estimate can be misleading for multimodal or skewed distributions, as it identifies only a single peak and ignores uncertainty.

Posterior median: For each parameter θ_i, the value that divides the marginal posterior into equal probability masses:

$$\int_{-\infty}^{\theta_{i,\mathrm{med}}} p(\theta_i|\mathbf{d}) d\theta_i = 0.5 \tag{79}$$

The median is robust to outliers and skewness, and is optimal for absolute error loss.

5.4.2 Credible Regions and Intervals

Point estimates alone do not convey uncertainty. Credible regions quantify the range of parameter values consistent with the data.

Definition: A $X\%$ credible region is any volume Ω in parameter space such that

$$\int_{\Omega} p(\boldsymbol{\theta}|\mathbf{d}) d\boldsymbol{\theta} = \frac{X}{100}. \tag{80}$$

There is considerable freedom in choosing Ω. Two common choices are:

- **Highest Posterior Density (HPD) regions:** The smallest region containing $X\%$ of the posterior probability. All points inside have higher posterior density than points outside. For unimodal posteriors this is sensible, but for multimodal posteriors the HPD region may consist of several disconnected islands.
- **Central credible intervals:** For a single parameter, the central $(1-\alpha)$ credible interval $[a, b]$ satisfies:

$$\int_{a}^{b} p(\theta|\mathbf{d}) d\theta = 1 - \alpha \tag{81}$$

with equal tail probabilities: $\int_{-\infty}^{a} p(\theta|\mathbf{d}) d\theta = \int_{b}^{\infty} p(\theta|\mathbf{d}) d\theta = \alpha/2$.

In 2D, credible regions are often shown as contour plots for $X = 68.3, 95.5, 99.7$ (corresponding to the probabilities enclosed in a Gaussian distribution by $\pm 1\sigma, 2\sigma, 3\sigma$), though the choice is arbitrary. Always identify what the contours represent when using them.

Credible regions are not confidence regions: The Bayesian credible region is not the same as the frequentist confidence interval. A 95% credible region means there is a 95% probability that the true parameter lies within it, given the data and model. A 95% confidence interval means that in 95% of repeated experiments, the interval would contain the true value—a statement about long-run frequency, not about probability given the observed data. The Bayesian interpretation directly addresses scientific questions about parameter values.

5.4.3 Variance and Covariance

Posterior variance and covariance: For continuous parameters, the posterior covariance matrix is:

$$\mathrm{Cov}[\boldsymbol{\theta}|\mathbf{d}] = \mathbb{E}[(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T|\mathbf{d}] \tag{82}$$

The diagonal elements give the marginal variances, while off-diagonal elements capture correlations between parameters. Strong correlations indicate parameter degeneracies that can only be understood by examining the joint distribution.

5.4.4 Gaussian Posteriors and Chi-Squared Contours

With multivariate Gaussian posteriors, the contour levels that contain $X\%$ of the posterior can be calculated using the chi-squared distribution. For Gaussian likelihoods with independent measurement errors, the chi-squared statistic is:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(d_i - m_i(\boldsymbol{\theta}))^2}{\sigma_i^2} \tag{83}$$

where d_i are observed data, $m_i(\boldsymbol{\theta})$ are model predictions, and σ_i are measurement uncertainties. For correlated errors with covariance matrix $\mathbf{C}$:

$$\chi^2(\boldsymbol{\theta}) = (\mathbf{d} - \mathbf{m}(\boldsymbol{\theta}))^T \mathbf{C}^{-1} (\mathbf{d} - \mathbf{m}(\boldsymbol{\theta})) \tag{84}$$

For a Gaussian likelihood, the log-likelihood is $\ln \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}\chi^2(\boldsymbol{\theta}) + \text{constant}$. The maximum likelihood estimate corresponds to the minimum value χ^2_{min}, and we define $\Delta\chi^2 \equiv \chi^2 - \chi^2_{\text{min}}$ to measure deviations from the best fit.

With uniform priors, the relative posterior probability is $\exp(-\Delta\chi^2/2)$. Standard references like Numerical Recipes provide tables for choosing $\Delta\chi^2$ contour levels that contain specified probabilities for different numbers of parameters. For example, for two parameters jointly, $\Delta\chi^2 = 2.30$ gives a 68.3% credible region, while for a single parameter, $\Delta\chi^2 = 1$ gives 68.3%.

Important: In the non-Gaussian case, it is better to find numerically the HPD regions that contain $X\%$ of the posterior, rather than adopting Gaussian chi-squared contour levels, which may be highly misleading.

5.4.5 When Are Summaries Sufficient?

For approximately Gaussian posteriors, the mean and covariance matrix provide a complete summary. However, simple summaries can be inadequate or misleading for:

- **Multimodal posteriors**: Point estimates can be meaningless, falling between modes or highlighting one mode while ignoring others. The full distribution must be presented, showing all modes and their relative probabilities.
- **Highly skewed posteriors**: The mean may lie in a low-probability region. The median and HPD intervals are more appropriate, but visualizing the full distribution is still recommended.
- **Complex parameter spaces**: Strong correlations and degeneracies require visualization. Corner plots (Figure 3) showing all pairwise marginal distributions are essential for understanding parameter relationships.
- **Decision-making**: Different loss functions require different summaries. For example, the mean is optimal for squared-error loss, the median for absolute-error loss, and the mode for zero-one loss. The appropriate summary depends on the consequences of errors.

In such cases, presenting samples from the posterior (e.g., from MCMC, Section 6) or visualization tools like corner plots are preferable to simple numerical summaries. The full posterior should always be preserved for further analysis, even if summaries are reported for publication. Modern tools make it straightforward to share full MCMC chains or grid-based posterior representations alongside published results.