STAT8310 - Bayesian Data Analysis

Chi-Kuang Yeh

2026-01-13

Table of contents

Preface **6**
Description 6
Prerequisites 6
Instructor 6
Office Hour 6
Grade Distribution 7
Assignment 7
Midterm 7
Topics and Corresponding Lectures 7
Recommended Textbooks 7
Side Readings 7

1 Quick Overview **8**
1.1 Why Bayesian? 8
1.2 Some Bayesian Topics and their Computational Focus 8
1.3 Interesting Article: 10

2 Week 1 — Introduction and Bayesian Thinking **11**
2.1 Introduction to Bayesian Inference 11
2.2 Foundational Concepts 11
2.2.1 Why Use Bayesian Methods? 11
2.3 Bayesian vs. Frequentist Comparison 12
2.3.1 Motivating Examples 12
2.3.2 Probability as a Measure of Uncertainty 13
2.3.3 Building Blocks of Bayesian Inference 13
2.3.4 Bayes' Theorem 13
2.3.5 Inference from the Posterior Distribution 14
2.4 One-Parameter Models 14
2.4.1 The Beta-Binomial Model 14
2.4.2 The Normal Model with Known Variance 16
2.4.3 Posterior Predictive Distribution 17
2.4.4 Conjugate Priors 18
2.4.5 Practical Considerations 19
2.4.6 R Examples 19

3 Week 2 — Conjugate Priors and Analytical Posteriors 23
3.1 Overview . . . 23
3.2 Learning Goals . . . 23
3.3 Lecture 1: The Concept of Conjugacy . . . 24
3.3.1 1.1 Definition . . . 24
3.3.2 1.2 Why Conjugacy Matters . . . 24
3.3.3 1.3 Examples of Conjugate Pairs . . . 24
3.4 Lecture 2: Beta–Binomial and Gamma–Poisson Models . . . 24
3.4.1 2.1 Beta–Binomial Model (Review and Generalization) . . . 24
3.4.2 2.2 Gamma–Poisson Model (Counts) . . . 25
3.4.3 2.3 R Example: Gamma–Poisson Updating . . . 25

4 Week 3 — Monte Carlo Integration and Simulation-Based Bayesian Inference 27
4.1 Overview . . . 27
4.2 Learning Goals . . . 27
4.3 Lecture 1: Motivation and Fundamentals of Monte Carlo . . . 28
4.3.1 1.1 The Problem . . . 28
4.3.2 1.2 Monte Carlo Idea . . . 28
4.3.3 1.3 Monte Carlo Error . . . 28
4.3.4 1.4 Simple Example . . . 28

5 Week 4 — Markov Chain Monte Carlo (MCMC) Methods 30
5.1 Overview . . . 30
5.2 Learning Goals . . . 30
5.3 Lecture 1: Introduction to MCMC . . . 31
5.3.1 1.1 Motivation . . . 31
5.3.2 1.2 Markov Chain Basics . . . 31
5.3.3 1.3 Core Idea . . . 31
5.4 Lecture 2: The Metropolis–Hastings Algorithm . . . 31
5.4.1 2.1 Algorithm Steps . . . 31
5.4.2 2.2 Special Case: Symmetric Proposal . . . 32
5.4.3 2.3 Example: Posterior for a Normal Mean (Unknown Mean, Known Variance) . . . 32

6 Week 5 — Model Checking and Comparison 34
6.1 Learning Goals . . . 34
6.2 Lecture 1 — Posterior Predictive Checking . . . 34
6.2.1 Posterior Predictive Distribution . . . 34
6.2.2 Implementation Steps . . . 35
6.2.3 Example A — Binomial Model . . . 35
6.2.4 Example B — Normal Model (Standard Deviation Check) . . . 36
6.2.5 Practical Tips . . . 37

6.3 Lecture 2 — Bayesian Model Comparison 37
6.3.1 Motivation 37
6.3.2 Bayes Factors 38
6.3.3 WAIC and LOO (Predictive Criteria) 38
6.3.4 Example A — Comparing Two Regression Models with `brms` (optional heavy computation) 38
6.3.5 Example B — Quick WAIC Comparison via Frequentist Approximation 39
6.3.6 Visual Predictive Comparison 40
6.3.7 Practical Summary 41
6.4 Lab 5 — Model Checking and Comparison 41
6.5 Homework 5 42
6.6 Key Takeaways 43

7 Week 6 — Hierarchical Bayesian Models 44
7.1 Learning Goals 44
7.2 Lecture 1 — Motivation and Structure of Hierarchical Models 44
7.2.1 1.1 Why Hierarchical Models? 44
7.2.2 1.2 Model Structure 45
7.2.3 1.3 Three Extremes of Pooling 45
7.2.4 1.4 Shrinkage Intuition 45
7.2.5 1.5 Example — Simulated Group Means 46
7.2.6 1.6 Advantages of Hierarchical Models 47
7.3 Lecture 2 — Hierarchical Regression and Implementation 47
7.3.1 2.1 Hierarchical Linear Regression 47
7.3.2 2.2 Example — Hierarchical Regression with `brms` 47
7.3.3 2.3 Interpretation 48
7.3.4 2.4 Practical Considerations 48
7.3.5 2.5 Summary of Hierarchical Modeling Benefits 49
7.4 Homework 6 49
7.5 Key Takeaways 50

8 Week 7 — Bayesian Decision Theory 51
8.1 Learning Goals 51
8.2 Lecture 1 — Principles of Bayesian Decision Theory 51
8.2.1 1.1 Motivation 51
8.2.2 1.2 The Decision-Theoretic Setup 52
8.2.3 1.3 Common Loss Functions and Bayes Rules 52
8.2.4 1.4 Example — Estimation under Quadratic Loss 52
8.2.5 1.5 Decision Rules and Risk 53
8.2.6 1.6 Example — Hypothesis Testing with 0–1 Loss 54
8.3 Lecture 2 — Applications and Extensions 54
8.3.1 2.1 Bayesian Credible Intervals as Decision Regions 54
8.3.2 2.2 Decision Theory for Classification 55

8.3.3 2.3 Loss vs Utility . . . 56
8.3.4 2.4 Connection to Frequentist Estimation . . . 56
8.3.5 2.5 Example — Optimal Cutoff for a Diagnostic Test . . . 56
8.3.6 2.6 Summary of Bayesian Decision Theory . . . 57
8.4 Homework 7 . . . 57
8.5 Key Takeaways . . . 58

9 Week 8 — Advanced Bayesian Computation 59
9.1 Learning Goals . . . 59
9.2 Lecture 1 — Hamiltonian Monte Carlo (HMC) . . . 59
9.2.1 1.1 Motivation . . . 59
9.2.2 1.2 Hamiltonian Dynamics . . . 60
9.2.3 1.3 Leapfrog Integration . . . 60
9.2.4 1.4 Intuition . . . 60
9.2.5 1.5 Example — Logistic Regression with HMC (Stan) . . . 61
9.2.6 1.6 Diagnosing HMC Performance . . . 61
9.2.7 1.7 Advantages of HMC . . . 61
9.3 Lecture 2 — Variational Inference (VI) . . . 62
9.3.1 2.1 Motivation . . . 62
9.3.2 2.2 Objective Function . . . 62
9.3.3 2.3 Mean-Field Approximation . . . 62
9.3.4 2.4 Example — Variational Bayes for a Normal Mean . . . 63
9.3.5 2.5 Automatic VI with `brms` . . . 63
9.3.6 2.6 Comparison: HMC vs VI . . . 64
9.3.7 2.7 Visual Comparison (Conceptual) . . . 64
9.3.8 2.8 Practical Advice . . . 65
9.4 Homework 8 . . . 65
9.5 Key Takeaways . . . 66

References 67

Preface

Description

This course will cover the topics in the theory and practice of *Bayesian statistical inference*, ranging from a review of fundamentals to questions of current research interest. Motivation for the Bayesian approach. Bayesian computation, Monte Carlo methods, asymptotics. Model checking and comparison. A selection of examples and issues in modelling and data analysis. Discussion of advantages and difficulties of the Bayesian approach. This course will be computationally intensive through analysis of data sets using the R statistical computing language.

Prerequisites

MATH 4752/6752 – Mathematical Statistics II or equivalent, and the ability to program in a high-level language.

Instructor

Chi-Kuang Yeh, Assistant Professor in the Department of Mathematics and Statistics, Georgia State University.

- Office: Suite 1407, 25 Park Place.
- Email: cyeh@gsu.edu.

Office Hour

TBA

Grade Distribution

- Homework – 50%
- Exam – 30%
- Final – 20%

Assignment

□ TBA

Midterm

□ TBA

Topics and Corresponding Lectures

Those chapters are based on the lecture notes. This part will be updated frequently.

Topic	Lecture Covered
Introduction to R Programming	1–2

Recommended Textbooks

- Gelman, A., Carlin, J., Stern, H., Rubin, D., Dunson, D., and Vehtari, A. (2021). Bayesian Data Analysis, CRC Press, 3rd Ed.
- Hoff, P.D. (2009). A First Course in Bayesian Statistical Methods, Springer.
- McElreath, R. (2018). Statistical Rethinking: A Bayesian Course with Examples in R and Stan, CRC Press.

Side Readings

- TBA

1 Quick Overview

The posterior distribution is obtained from the prior distribution and sampling model via *Bayes' rule*:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int_{\Theta} p(y \mid \theta')p(\theta')d\theta'}.$$

1.1 Why Bayesian?

- **Intuitive probability interpretation**: Directly quantifies uncertainty about parameters as probability distributions
- **Incorporates prior knowledge**: Systematically combines domain expertise with data through the prior distribution
- **Principled inference**: Bayes' rule provides a coherent framework for updating beliefs based on evidence
- **Natural handling of uncertainty**: Posterior distributions capture full uncertainty, not just point estimates
- **Sequential analysis**: Easily updates beliefs as new data arrives (posterior becomes new prior)
- **Small sample inference**: Performs well with limited data by leveraging prior information
- **Prediction with uncertainty**: Generates predictive distributions that quantify uncertainty in future observations
- **Decision-making**: Naturally incorporates loss functions for optimal decision rules
- **Model comparison**: Bayes factors provide a principled approach to comparing competing models

---

1.2 Some Bayesian Topics and their Computational Focus

Table 1.1: Some of the Bayesian Topics and its computational related focuses.

Topics	**Key Concepts / Readings**	**Computing Focus**
Introduction to Bayesian Thinking	Bayesian vs. Frequentist paradigms; Prior, likelihood, posterior	Review of R basics and reproducible workflows
Bayesian Inference for Simple Models	Conjugate priors, Beta-Binomial, Normal-Normal, Poisson-Gamma	Simulating posteriors, visualization
Prior Elicitation and Sensitivity	Informative vs. noninformative priors, Jeffreys prior	Prior sensitivity plots
Monte Carlo Integration	Law of large numbers, sampling-based inference	Random sampling and Monte Carlo approximation
Markov Chain Monte Carlo (MCMC)	Metropolis-Hastings, Gibbs sampler	Implementing MCMC in R
Convergence Diagnostics	Trace plots, autocorrelation, Gelman–Rubin statistic	`coda`, `rstan`, and `bayesplot` packages
Hierarchical Bayesian Models	Partial pooling, shrinkage, multilevel structures	`rstanarm` / `brms`
Midterm Project: Bayesian Linear Regression	Posterior inference for regression, model selection	`brms`, `rstanarm`, custom Gibbs samplers
Bayesian Model Comparison	Bayes factors, BIC, DIC, WAIC, LOO	Practical comparison via cross-validation
Model Checking and Diagnostics	Posterior predictive checks, residual analysis	`pp_check` in `brms`
Advanced Computation	Hamiltonian Monte Carlo (HMC), Variational Inference	Using `Stan` and `CmdStanR`
Bayesian Decision Theory	Utility functions, decision rules, loss minimization	Simple decision problems in R
Modern Bayesian Methods	Approximate Bayesian computation (ABC), Bayesian neural networks	Examples via `rstan` or `tensorflow-probability`
Student Project Presentations	Applications and case studies	Full workflow demonstration in R

1.3 Interesting Article:

- Goligher, E.C., Harhay, M.O. (2023). What Is the Point of Bayesian Analysis?, American Journal of Respiratory and Critical Care Medicine, 209, 485–487.

2 Week 1 — Introduction and Bayesian Thinking

2.1 Introduction to Bayesian Inference

Bayesian inference is based on a simple principle: the **posterior distribution** (our updated beliefs after observing data) is obtained from the **prior distribution** (our initial beliefs) and the **sampling model** (how data are generated) via **Bayes' rule**:

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

This elegant formula is the foundation of all Bayesian inference. It tells us how to update our beliefs in light of new evidence.

2.2 Foundational Concepts

2.2.1 Why Use Bayesian Methods?

Probability as Uncertainty: The Bayesian framework treats probability as a measure of uncertainty about unknown quantities, not just long-run frequencies. This allows direct probability statements about parameters given observed data.

Incorporates Prior Knowledge: Bayesian methods naturally combine prior information (from expert judgment, previous studies, or domain knowledge) with observed data. This is particularly valuable in: - Medical research where historical trials exist - Engineering where physical constraints are known - Sequential analysis where data arrive over time

Direct Inference: Bayesian inference answers the questions researchers actually ask: - "What is the probability that treatment B is better than A given my data?" - "What is a plausible range for this parameter?" - Rather than "If the null hypothesis were true, what is the probability of observing this data?"

Flexible Modeling: Complex models with multiple parameters, hierarchical structures, or missing data are more naturally expressed in Bayesian frameworks.

Better Small-Sample Performance: With limited data, informative priors can stabilize estimates and provide more stable inference than frequentist methods.

2.3 Bayesian vs. Frequentist Comparison

Both approaches have merits and limitations. The choice depends on the problem context, available prior information, and the questions being asked.

2.3.1 Motivating Examples

2.3.1.1 Example 1.1: Inference for a proportion

Suppose we are interested in estimating the rate at which a disease occurs in a population. We sample $n = 20$ individuals and observe $y = 8$ with the disease.

Questions: - What is our estimate of the disease rate θ? - How certain are we about this estimate? - How would we predict the number with disease in a future sample?

Two approaches:

Frequentist approach: - Point estimate: $\hat{\theta} = y/n = 8/20 = 0.4$ - Confidence interval based on sampling distribution - Does not directly provide $P(\theta \in [a, b] \mid \text{data})$

Bayesian approach: - Treat θ as a random variable with a prior distribution - Update beliefs using data via Bayes' theorem - Obtain posterior distribution: direct probability statements about θ

2.3.1.2 Example 1.2: Comparing two groups

Consider two treatment groups with success rates θ_A and θ_B.

- Group A: 8 successes out of 20 trials
- Group B: 12 successes out of 20 trials

Questions: - Is treatment B better than treatment A? - What is $P(\theta_B > \theta_A \mid \text{data})$? - Bayesian methods provide direct answers to such questions.

2.3.2 Probability as a Measure of Uncertainty

- **Frequentist interpretation:** Probability as long-run frequency of repeated events.
- **Bayesian interpretation:** Probability as a degree of belief or uncertainty about unknown quantities.

The Bayesian view allows us to: - Make probability statements about parameters (not just data) - Incorporate prior information naturally - Update beliefs coherently as new data arrive

2.3.3 Building Blocks of Bayesian Inference

For parameter θ and observed data y, we need three components:

1. **Prior distribution** $p(\theta)$: expresses our beliefs about θ before seeing data
2. **Likelihood** $p(y \mid \theta)$: probability model for the data given θ

3. **Posterior distribution** $p(\theta \mid y)$: updated beliefs after seeing data

2.3.4 Bayes' Theorem

These three components are combined via **Bayes' theorem:**

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

In words:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

Key insight: The posterior is proportional to the likelihood times the prior:

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

The denominator $p(y) = \int p(y \mid \theta)\, p(\theta)\, d\theta$ is a normalizing constant ensuring $\int p(\theta \mid y)\, d\theta = 1$.

2.3.5 Inference from the Posterior Distribution

Once we obtain the posterior $p(\theta \mid y)$, we can:

1. **Point estimation:**
   - Posterior mean: $E[\theta \mid y]$
   - Posterior median or mode
2. **Interval estimation:**
   - Credible intervals: $P(a < \theta < b \mid y) = 0.95$
   - Direct probability statements about parameters
3. **Hypothesis testing:**
   - $P(\theta_B > \theta_A \mid y)$
4. **Prediction:**
   - Posterior predictive distribution for future data $\tilde{y}$:

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

2.4 One-Parameter Models

2.4.1 The Beta-Binomial Model

2.4.1.1 Setup

Consider binary outcome data: $y_1, \dots, y_n$ where each $y_i \in \{0, 1\}$.

Let $y = \sum_{i=1}^{n} y_i$ be the number of successes. We model:

$$y \mid \theta \sim \text{Binomial}(n, \theta)$$

where θ is the probability of success.

Likelihood:

$$p(y \mid \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \propto \theta^y (1 - \theta)^{n-y}$$

2.4.1.2 Prior Distribution

We use a **Beta prior** for θ:

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

with density:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1$$

Prior properties: - $E[\theta] = \frac{\alpha}{\alpha+\beta}$ - $\text{Mode}[\theta] = \frac{\alpha-1}{\alpha+\beta-2}$ (if $\alpha, \beta > 1$) - $\text{Var}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Interpretation: Think of α as prior successes and β as prior failures.

2.4.1.3 Posterior Distribution

By Bayes' theorem:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta) \times p(\theta) && (2.1) \\ &\propto \theta^y (1-\theta)^{n-y} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} && (2.2) \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} && (2.3) \end{aligned}$$

This is the kernel of a $\text{Beta}(\alpha+y, \beta+n-y)$ distribution.

Posterior:

$$\theta \mid y \sim \text{Beta}(\alpha+y, \beta+n-y)$$

Posterior mean:

$$E[\theta \mid y] = \frac{\alpha+y}{\alpha+\beta+n}$$

This is a weighted average of the prior mean $\frac{\alpha}{\alpha+\beta}$ and the sample proportion $\frac{y}{n}$.

2.4.1.4 Example 2.1: Disease Rate

Return to the disease rate example: $n = 20$, $y = 8$.

Suppose we use a weakly informative prior: $\theta \sim \text{Beta}(2, 2)$ (prior mean = 0.5).

Posterior: $\theta \mid y \sim \text{Beta}(10, 14)$

Posterior mean: $E[\theta \mid y] = \frac{10}{24} = 0.417$

95% credible interval: We can compute quantiles of $\text{Beta}(10, 14)$ to get a 95% interval for θ.

2.4.2 The Normal Model with Known Variance

2.4.2.1 Setup

Suppose we observe data $y_1, \ldots, y_n$ that are i.i.d. from:

$$y_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$$

where θ is the unknown mean and σ^2 is a known variance.

Likelihood: For the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$:

$$\bar{y} \mid \theta \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$$

The likelihood is:

$$p(y_1, \ldots, y_n \mid \theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\}$$

2.4.2.2 Prior Distribution

We use a **Normal prior** for θ:

$$\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$$

with density:

$$p(\theta) \propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right\}$$

2.4.2.3 Posterior Distribution

By Bayes' theorem:

$$p(\theta \mid y) \propto p(y \mid \theta) \times p(\theta) \tag{2.4}$$

$$\propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\} \times \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\} \tag{2.5}$$

After completing the square, we obtain:

$$\theta \mid y \sim \mathcal{N}(\mu_n, \tau_n^2)$$

where:

$$\tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1} = \frac{1}{\text{prior precision} + \text{data precision}}$$

$$\mu_n = \tau_n^2 \left(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}\right)$$

Alternative form:

$$\mu_n = w\mu_0 + (1-w)\bar{y}$$

where $w = \frac{\sigma^2/n}{\sigma^2/n+\tau_0^2}$ is the weight on the prior mean.

2.4.2.4 Interpretation

- The posterior mean is a **weighted average** of the prior mean and the sample mean
- As $n \to \infty$, the posterior mean converges to $\bar{y}$ (data dominate)
- With little data, the posterior is pulled toward the prior
- The posterior precision is the sum of the prior and data precisions

---

2.4.3 Posterior Predictive Distribution

After observing data $y = (y_1, \dots, y_n)$, we often want to predict a future observation $\tilde{y}$.

The **posterior predictive distribution** is:

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

This averages the conditional distribution of $\tilde{y}$ given θ over the posterior uncertainty in θ.

2.4.3.1 Example: Beta-Binomial

For the beta-binomial model with $\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y)$:

$$P(\tilde{y} = 1 \mid y) = E[\theta \mid y] = \frac{\alpha + y}{\alpha + \beta + n}$$

2.4.3.2 Example: Normal Model

For the normal model with known variance, if $\theta \mid y \sim \mathcal{N}(\mu_n, \tau_n^2)$ and $\tilde{y} \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$:

$$\tilde{y} \mid y \sim \mathcal{N}(\mu_n, \tau_n^2 + \sigma^2)$$

The predictive variance includes both parameter uncertainty (τ_n^2) and sampling variability (σ^2).

2.4.4 Conjugate Priors

Definition: A prior distribution is **conjugate** to a likelihood if the posterior distribution is in the same family as the prior.

Examples: - Beta prior + Binomial likelihood → Beta posterior - Normal prior + Normal likelihood (known variance) → Normal posterior - Gamma prior + Poisson likelihood → Gamma posterior

Advantages: - Analytical posteriors (no numerical integration needed) - Interpretable parameters - Computationally efficient

Limitations: - May not reflect true prior beliefs - Modern computing makes non-conjugate priors feasible

---

2.4.5 Practical Considerations

2.4.5.1 Prior Elicitation

How do we choose a prior?

1. **Informative priors:** Based on previous studies or expert knowledge
2. **Weakly informative priors:** Provide some regularization without dominating the data
3. **Non-informative priors:** Attempt to be "objective" (e.g., uniform, Jeffreys prior)

2.4.5.2 Sensitivity Analysis

- Try different priors and check if conclusions change substantially
- If posterior is sensitive to prior choice with substantial data, investigate further

2.4.5.3 Comparing Bayesian and Frequentist Inference

Aspect	Bayesian	Frequentist
Parameters	Random variables with distributions	Fixed unknown constants
Probability statements	Direct: $P(\theta \in [a, b] \mid y)$	Indirect: confidence intervals
Prior information	Naturally incorporated	Difficult to include
Small samples	Can be more stable	May have poor properties
Interpretation	Conditional on observed data	Based on repeated sampling

---

2.4.6 R Examples

2.4.6.1 Example 3.1: Beta-Binomial Model

```r
# Data
n <- 20
y <- 8

# Prior: Beta(2, 2)
alpha0 <- 2
```

```r
beta0 <- 2

# Posterior: Beta(10, 14)
alpha1 <- alpha0 + y
beta1 <- beta0 + n - y

# Grid for plotting
theta <- seq(0, 1, length.out = 500)

# Plot prior and posterior
plot(theta, dbeta(theta, alpha0, beta0), type = "l", lwd = 2, col = "blue",
     ylab = "Density", xlab = expression(theta),
     main = "Beta-Binomial Model: Prior and Posterior")
lines(theta, dbeta(theta, alpha1, beta1), col = "red", lwd = 2)
abline(v = y/n, lty = 2, col = "gray")

legend("topright",
       legend = c("Prior Beta(2,2)", "Posterior Beta(10,14)", "MLE"),
       col = c("blue", "red", "gray"), lwd = c(2, 2, 1), lty = c(1, 1, 2))
```

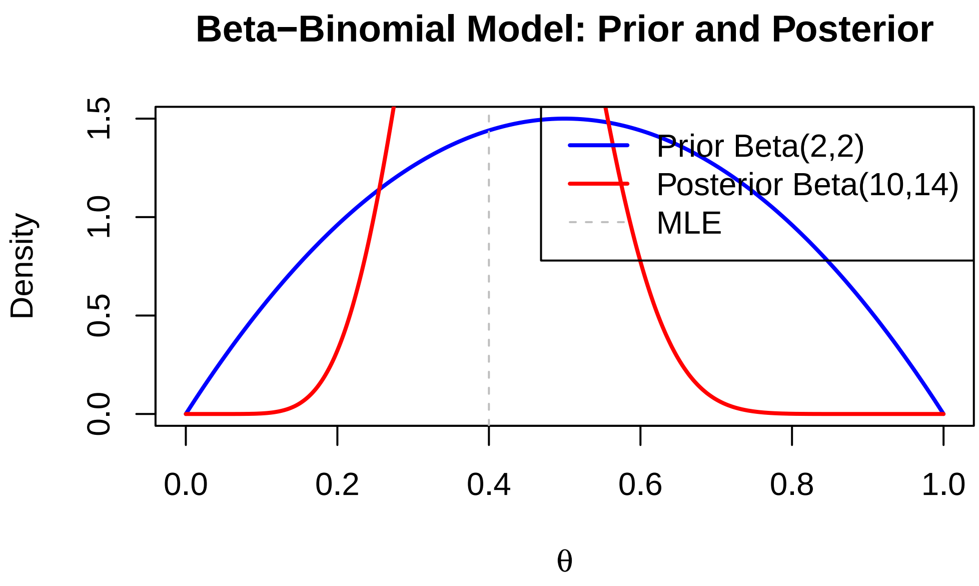


Figure 2.1: Prior, Likelihood, and Posterior for Beta-Binomial Model

```
# Posterior summary
cat("Posterior mean:", alpha1/(alpha1 + beta1), "\n")
```

```
Posterior mean: 0.4166667
```

```
cat("95% credible interval:", qbeta(c(0.025, 0.975), alpha1, beta1), "\n")
```

```
95% credible interval: 0.2319142 0.614581
```

2.4.6.2 Example 3.2: Normal Model

```
# Data
y <- c(1.2, 0.8, 1.5, 1.1, 0.9)
n <- length(y)
ybar <- mean(y)
sigma2 <- 0.25  # known variance

# Prior: N(0, 1)
mu0 <- 0
tau0_sq <- 1

# Posterior
tau_n_sq <- 1 / (1/tau0_sq + n/sigma2)
mu_n <- tau_n_sq * (mu0/tau0_sq + n*ybar/sigma2)

cat("Posterior: N(", round(mu_n, 3), ",", round(tau_n_sq, 3), ")\n")
```

```
Posterior: N( 1.048 , 0.048 )
```

```
# Plot
theta <- seq(-1, 3, length.out = 500)
plot(theta, dnorm(theta, mu0, sqrt(tau0_sq)), type = "l", lwd = 2, col = "blue",
     ylab = "Density", xlab = expression(theta),
     main = "Normal Model: Prior and Posterior")
lines(theta, dnorm(theta, mu_n, sqrt(tau_n_sq)), col = "red", lwd = 2)
abline(v = ybar, lty = 2, col = "gray")

legend("topright",
       legend = c("Prior", "Posterior", "Sample mean"),
       col = c("blue", "red", "gray"), lwd = c(2, 2, 1), lty = c(1, 1, 2))
```

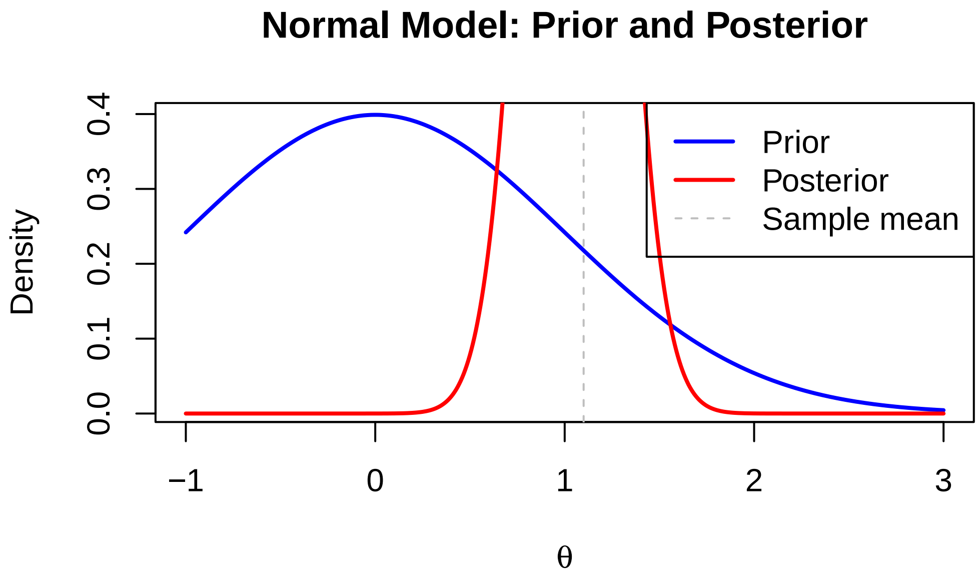


Figure 2.2: Prior, Likelihood, and Posterior for Normal Model

3 Week 2 — Conjugate Priors and Analytical Posteriors

3.1 Overview

This week focuses on **conjugate priors** — special priors that yield posteriors in the same family of distributions as the prior.
Students will learn why conjugacy simplifies Bayesian inference, how to identify conjugate pairs for common likelihoods, and how to perform analytical posterior updates without simulation. We will also introduce the concept of prior sensitivity analysis and noninformative (objective) priors.

3.2 Learning Goals

By the end of Week 2, you should be able to:

- Define and identify conjugate priors for standard likelihood models.
- Derive analytical posteriors for Binomial, Poisson, and Normal models.
- Compute posterior summaries and predictive distributions.
- Discuss the influence of priors on posterior inference.
- Perform prior sensitivity analysis in R.

3.3 Lecture 1: The Concept of Conjugacy

3.3.1 1.1 Definition

A **conjugate prior** for a likelihood $p(y \mid \theta)$ is a prior distribution $p(\theta)$ such that the posterior $p(\theta \mid y)$ belongs to the same family as the prior.

Formally:

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

If $p(\theta \mid y)$ has the same functional form as $p(\theta)$, then $p(\theta)$ is *conjugate* to the likelihood.

3.3.2 1.2 Why Conjugacy Matters

- Provides closed-form expressions for posterior means, variances, and credible intervals.
- Facilitates sequential updating — easy to update priors as new data arrive.
- Useful for educational and analytic illustration before moving to MCMC methods.

3.3.3 1.3 Examples of Conjugate Pairs

Likelihood	Conjugate Prior	Posterior Family
Binomial(n, θ)	Beta(α, β)	Beta$(\alpha + y, \beta + n - y)$
Poisson(λ)	Gamma(a, b)	Gamma$(a + \sum y_i, b + n)$
Normal(μ, σ^2) (known variance)	Normal(μ_0, τ_0^2)	Normal(μ_1, τ_1^2)
Exponential(λ)	Gamma(a, b)	Gamma$(a + n, b + \sum y_i)$
Normal mean/variance (unknown σ^2)	Normal–Inverse-Gamma	Normal–Inverse-Gamma

---

3.4 Lecture 2: Beta–Binomial and Gamma–Poisson Models

3.4.1 2.1 Beta–Binomial Model (Review and Generalization)

Let $y \mid \theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(\alpha_0, \beta_0)$.
Then the posterior is:

$$\theta \mid y \sim \text{Beta}(\alpha_0 + y, \beta_0 + n - y).$$

Posterior Mean:

$$E[\theta \mid y] = \frac{\alpha_0 + y}{\alpha_0 + \beta_0 + n}.$$

Predictive Probability for a Future Success:

$$p(\tilde{y} = 1 \mid y) = E[\theta \mid y].$$

Interpretation:
Each observation updates the Beta prior by adding one success or failure to the corresponding shape parameter.

3.4.2 2.2 Gamma–Poisson Model (Counts)

Suppose we model count data as $y_i \sim \text{Poisson}(\lambda)$, with prior $\lambda \sim \text{Gamma}(a_0, b_0)$ (where the Gamma density is parameterized as $p(\lambda) \propto \lambda^{a_0-1}e^{-b_0\lambda}$).

Posterior:

$$\lambda \mid y_1, \dots, y_n \sim \text{Gamma}\left(a_0 + \sum_{i=1}^{n} y_i,\ b_0 + n\right).$$

Posterior Mean and Variance:

$$E[\lambda \mid y] = \frac{a_0 + \sum y_i}{b_0 + n}, \quad \text{Var}[\lambda \mid y] = \frac{a_0 + \sum y_i}{(b_0 + n)^2}.$$

Posterior Predictive:

$$p(\tilde{y} \mid y) = \int \text{Poisson}(\tilde{y} \mid \lambda)\, p(\lambda \mid y)\, d\lambda,$$

which follows a **Negative Binomial** distribution.

Interpretation:
The Gamma prior acts as if we had observed $a_0 - 1$ pseudo-events over b_0 pseudo-trials.

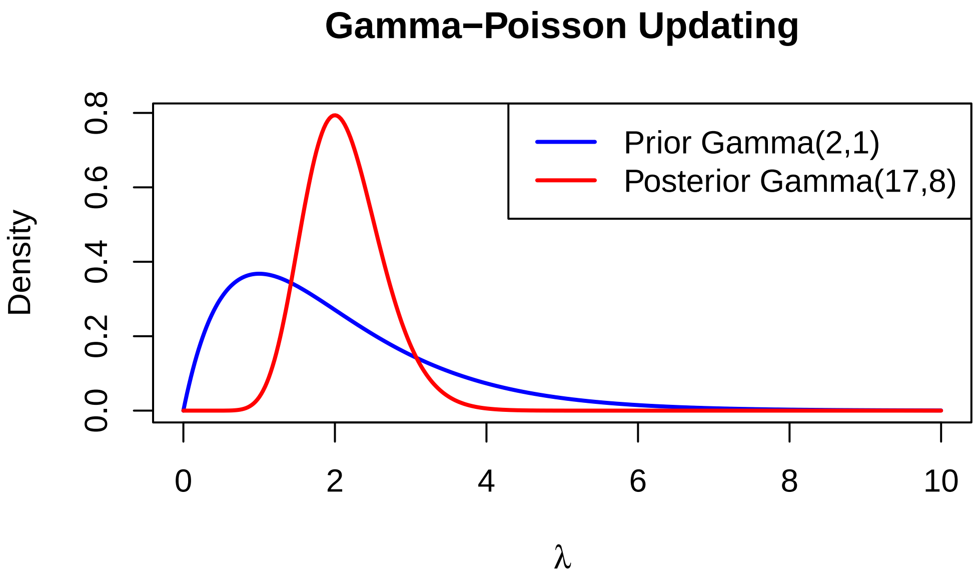
3.4.3 2.3 R Example: Gamma–Poisson Updating

```r
# Posterior update for Gamma-Poisson model
y <- c(3, 2, 4, 1, 0, 2, 3)
a0 <- 2; b0 <- 1   # prior Gamma(2,1)
n <- length(y)

a1 <- a0 + sum(y)
b1 <- b0 + n

lambda <- seq(0, 10, length.out = 400)
prior <- dgamma(lambda, a0, b0)
posterior <- dgamma(lambda, a1, b1)

plot(lambda, prior, type="l", lwd=2, col="blue", ylim=c(0, max(posterior)),
     ylab="Density", xlab=expression(lambda),
     main="Gamma-Poisson Updating")
lines(lambda, posterior, col="red", lwd=2)
legend("topright",
       legend=c("Prior Gamma(2,1)", paste0("Posterior Gamma(", a1, ",", b1, ")")),
       col=c("blue", "red"), lwd=2)
```



4 Week 3 — Monte Carlo Integration and Simulation-Based Bayesian Inference

---

4.1 Overview

This week introduces **Monte Carlo methods**, which allow us to approximate Bayesian quantities when analytical solutions are unavailable.
We explore how random sampling can be used to estimate expectations, posterior summaries, and probabilities.
By the end of this week, students will understand how Monte Carlo simulation forms the foundation for modern Bayesian computation such as MCMC.

4.2 Learning Goals

By the end of Week 3, you should be able to:

- Explain the motivation for Monte Carlo methods in Bayesian inference.
- Approximate expectations, integrals, and posterior summaries using random sampling.
- Implement crude Monte Carlo and importance sampling in R.
- Assess the accuracy and variance of Monte Carlo estimators.
- Interpret Monte Carlo errors and convergence diagnostics.

4.3 Lecture 1: Motivation and Fundamentals of Monte Carlo

4.3.1 1.1 The Problem

Bayesian inference often requires evaluating integrals such as:

$$E[h(\theta) \mid y] = \int h(\theta)\, p(\theta \mid y)\, d\theta,$$

which are rarely available in closed form.

4.3.2 1.2 Monte Carlo Idea

If we can sample $\theta^{(1)}, \dots, \theta^{(M)}$ from the posterior $p(\theta \mid y)$,
then we can approximate the expectation by:

$$\hat{E}[h(\theta)] = \frac{1}{M} \sum_{m=1}^{M} h(\theta^{(m)}).$$

This is called the **Monte Carlo estimator**.

By the **Law of Large Numbers**, $\hat{E}[h(\theta)] \to E[h(\theta)]$ as $M \to \infty$. The **Central Limit Theorem** gives:

$$\sqrt{M}\,(\hat{E} - E[h(\theta)]) \approx N(0, \mathrm{Var}[h(\theta)]).$$

4.3.3 1.3 Monte Carlo Error

We can estimate the simulation error by:

$$\mathrm{SE}(\hat{E}) \approx \sqrt{\frac{\mathrm{Var}(h(\theta))}{M}}.$$

Larger M gives more accurate approximations but increases computation time.

4.3.4 1.4 Simple Example

Compute $E[\theta]$ for $\theta \sim \mathrm{Beta}(2, 5)$ using Monte Carlo.

```r
set.seed(1)
M <- 1e5
theta <- rbeta(M, 2, 5)
mean(theta)          # Monte Carlo estimate
```

```
[1] 0.2861808
```

```r
var(theta) / M       # Monte Carlo variance
```

```
[1] 2.56548e-07
```

5 Week 4 — Markov Chain Monte Carlo (MCMC) Methods

5.1 Overview

This week introduces **Markov Chain Monte Carlo (MCMC)** — a powerful class of algorithms for simulating from complex posterior distributions that are difficult to sample from directly.
You will learn the logic of constructing Markov chains with a desired stationary distribution, how to implement the Metropolis–Hastings (MH) and Gibbs samplers, and how to assess convergence and mixing of MCMC chains.

5.2 Learning Goals

By the end of Week 4, you should be able to:

- Explain the intuition behind MCMC and why it works.
- Implement simple Metropolis–Hastings and Gibbs algorithms in R.
- Diagnose convergence using trace plots and summary statistics.
- Compute posterior means, variances, and credible intervals from MCMC samples.
- Understand practical issues such as burn-in, thinning, and autocorrelation.

5.3 Lecture 1: Introduction to MCMC

5.3.1 1.1 Motivation

For many posteriors, sampling directly is infeasible.
We instead build a *Markov chain* whose stationary distribution is the target posterior $p(\theta \mid y)$.
After sufficient iterations, the draws from the chain behave like samples from the true posterior.

5.3.2 1.2 Markov Chain Basics

A Markov chain $\{\theta^{(t)}\}$ has the **Markov property**:

$$p(\theta^{(t)} \mid \theta^{(t-1)}, \dots, \theta^{(1)}) = p(\theta^{(t)} \mid \theta^{(t-1)}).$$

If the chain is **ergodic**, the distribution of $\theta^{(t)}$ converges to a stationary distribution $\pi(\theta)$.

MCMC constructs such chains so that $\pi(\theta) = p(\theta \mid y)$.

5.3.3 1.3 Core Idea

Repeatedly propose a new value θ^* and decide whether to **accept** or **reject** it
based on how likely it is under the posterior.
This ensures that samples eventually represent the posterior distribution.

5.4 Lecture 2: The Metropolis–Hastings Algorithm

5.4.1 2.1 Algorithm Steps

1. Initialize with $\theta^{(0)}$.

2. For each iteration $t = 1, 2, \dots, T$:

   a. Propose $\theta^* \sim q(\theta^* \mid \theta^{(t-1)})$.

   b. Compute the **acceptance probability**:

   $$\alpha = \min\left(1,\ \frac{p(y \mid \theta^*)\ p(\theta^*)\ q(\theta^{(t-1)} \mid \theta^*)}{p(y \mid \theta^{(t-1)})\, p(\theta^{(t-1)})\, q(\theta^* \mid \theta^{(t-1)})}\right).$$

   c. Accept θ^* with probability α; otherwise, keep $\theta^{(t)} = \theta^{(t-1)}$.

3. After burn-in, the samples $\{\theta^{(t)}\}$ approximate draws from $p(\theta \mid y)$.

5.4.2 2.2 Special Case: Symmetric Proposal

If $q(\theta^* \mid \theta^{(t-1)}) = q(\theta^{(t-1)} \mid \theta^*)$, then

$$\alpha = \min\left(1,\ \frac{p(y \mid \theta^*)\ p(\theta^*)}{p(y \mid \theta^{(t-1)})\, p(\theta^{(t-1)})}\right).$$

This is the **Metropolis algorithm**.

5.4.3 2.3 Example: Posterior for a Normal Mean (Unknown Mean, Known Variance)

Let $y_i \sim N(\mu, 1)$ for $i = 1, \dots, n$ and prior $\mu \sim N(0, 10^2)$.

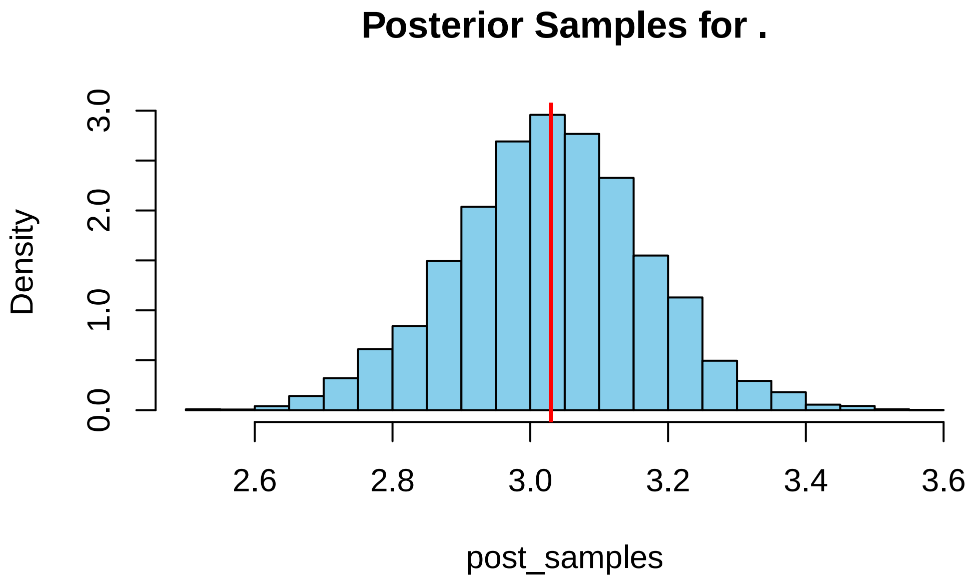
```r
set.seed(123)
# Data
y <- rnorm(50, mean = 3, sd = 1)
n <- length(y)
post_log <- function(mu) {
  sum(dnorm(y, mu, 1, log=TRUE)) + dnorm(mu, 0, 10, log=TRUE)
}

# Metropolis sampler
T <- 10000
mu <- numeric(T)
mu[1] <- 0
proposal_sd <- 0.5

for(t in 2:T) {
  mu_star <- rnorm(1, mu[t-1], proposal_sd)
  log_alpha <- post_log(mu_star) - post_log(mu[t-1])
  if(log(runif(1)) < log_alpha) mu[t] <- mu_star else mu[t] <- mu[t-1]
}

burnin <- 1000
post_samples <- mu[(burnin+1):T]

hist(post_samples, prob=TRUE, col="skyblue", main="Posterior Samples for  ")
abline(v = mean(post_samples), col="red", lwd=2)
```


Posterior Samples for .
Density
0.0
1.0
2.0
3.0
2.6
2.8
3.0
3.2
3.4
3.6
post_samples

6 Week 5 — Model Checking and Comparison

This week introduces methods for evaluating Bayesian model adequacy and comparing models. We focus on **Posterior Predictive Checking (PPC)** and **Bayesian Model Comparison** via Bayes factors, WAIC, and LOO.
Students will diagnose model fit using replicated data and compare predictive accuracy among competing models.

6.1 Learning Goals

By the end of this week, you should be able to:

- Generate and interpret posterior predictive distributions.
- Use posterior predictive checks to detect model misspecification.
- Compute and interpret WAIC, LOO, and Bayes factors.
- Evaluate model adequacy visually and numerically in R.

6.2 Lecture 1 — Posterior Predictive Checking

6.2.1 Posterior Predictive Distribution

For data y and parameters θ,

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

If a model is adequate, the observed data y should look typical among replicated datasets $\tilde{y}$ simulated from this distribution.

6.2.2 Implementation Steps

1. Choose a **discrepancy statistic** $T(y, \theta)$ capturing an aspect of fit.

2. For each posterior draw $\theta^{(m)}$:

   - Simulate $\tilde{y}^{(m)} \sim p(\tilde{y} \mid \theta^{(m)})$.

   - Compute $T(y, \theta^{(m)})$ and $T(\tilde{y}^{(m)}, \theta^{(m)})$.

3. Compute posterior predictive p-value:

$$p_{\text{ppc}} = P(T(\tilde{y}, \theta) > T(y, \theta) \mid y)\,.$$

Values near 0 or 1 suggest lack of fit.

6.2.3 Example A — Binomial Model

```r
set.seed(5)
M <- 5000
y_obs <- 7; n <- 10

theta <- rbeta(M, 2 + y_obs, 2 + n - y_obs)   # posterior draws
y_rep <- rbinom(M, n, theta)

ppc_p <- mean(y_rep >= y_obs)
ppc_p
```

```
[1] 0.509
```

```r
hist(y_rep, breaks=seq(-0.5, n+0.5, by=1),
     col="skyblue", main="Posterior Predictive Distribution", xlab="Replicated ỹ")
abline(v=y_obs, col="red", lwd=2)
legend("topright", legend=c("Observed y"), col="red", lwd=2, bty="n")
```

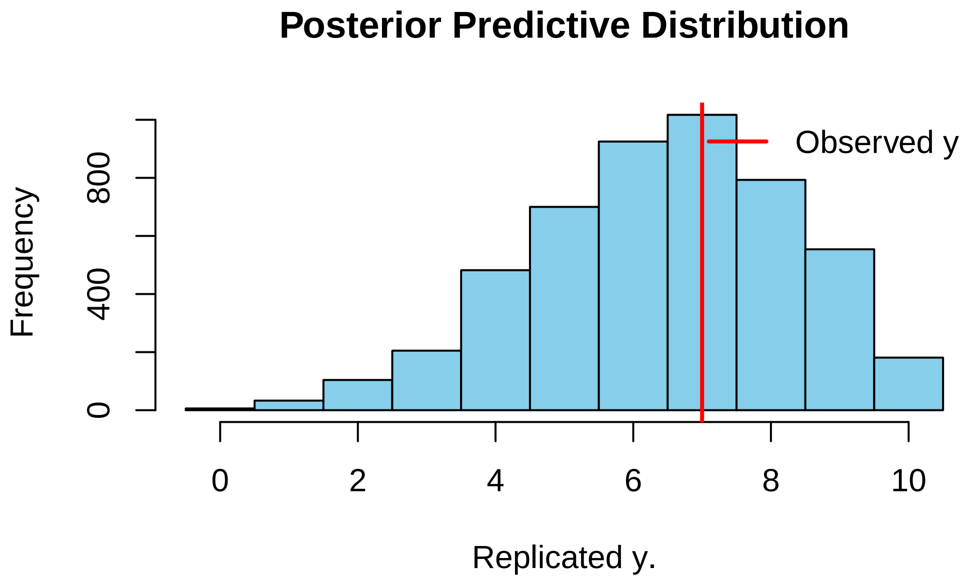


Figure 6.1: Posterior predictive distribution of ỹ

6.2.4 Example B — Normal Model (Standard Deviation Check)

```
set.seed(6)
y <- rnorm(30, mean=0, sd=1)
mu_draw <- rnorm(1000, 0, 1)
y_rep_mat <- replicate(200, rnorm(length(y), sample(mu_draw,1), 1))

T_obs <- sd(y)
T_rep <- apply(y_rep_mat, 2, sd)

mean(T_rep >= T_obs)
```

```
[1] 0.145
```

```
hist(T_rep, col="lightgray", main="Posterior Predictive Check: SD",
     xlab="Replicated sd(ỹ)")
abline(v=T_obs, col="red", lwd=2)
legend("topright", legend=c("Observed sd(y)"), col="red", lwd=2, bty="n")
```

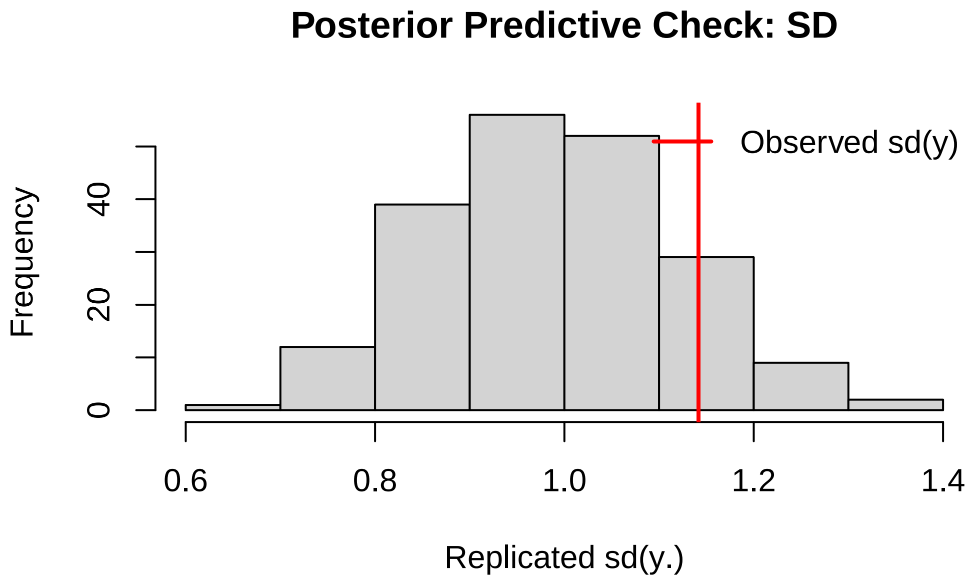


Figure 6.2: PPC for sample standard deviation

6.2.5 Practical Tips

- Use **multiple** statistics $(T_1, T_2, ...)$.
- Visual checks often outperform a single numeric p_{ppc}.
- Integrate PPC with subject-matter knowledge and residual plots.

6.3 Lecture 2 — Bayesian Model Comparison

6.3.1 Motivation

Competing Bayesian models are compared by their fit and complexity. Common approaches include **Bayes factors**, **WAIC**, and **LOO**.

6.3.2 Bayes Factors

Given two models M_1 and M_2,

$$\mathrm{BF}_{12} = \frac{p(y \mid M_1)}{p(y \mid M_2)}, \qquad p(y \mid M) = \int p(y \mid \theta_M, M)\, p(\theta_M \mid M)\, d\theta_M.$$

- $\mathrm{BF}_{12} > 1$ favors M_1.
- Express in $\log_{10}$ scale for interpretation.

$\log_{10} \mathrm{BF}_{12}$	Evidence for M_1
0–0.5	Barely worth mentioning
0.5–1	Substantial
1–2	Strong
>2	Decisive

6.3.3 WAIC and LOO (Predictive Criteria)

When computing marginal likelihoods is infeasible, we use predictive criteria:

- **WAIC (Watanabe–Akaike Information Criterion)**

$$\mathrm{WAIC} = -2(\mathrm{lppd} - p_{\mathrm{WAIC}}),$$

  where $\mathrm{lppd} = \sum_i \log\Big(\frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta^{(s)})\Big)$.

- **LOO (Leave-One-Out Cross-Validation)**
  Approximates the out-of-sample predictive performance:

$$\mathrm{LOO} = \sum_i \log p(y_i \mid y_{-i}),$$

  usually estimated via Pareto-smoothed importance sampling (PSIS-LOO).

Lower WAIC (or higher elpd_loo) indicates better predictive performance.

6.3.4 Example A — Comparing Two Regression Models with `brms` (optional heavy computation)

```r
# Uncomment and install if needed
# install.packages(c("brms", "loo"))

library(brms)
library(loo)

set.seed(7)
dat <- data.frame(x = rnorm(200))
dat$y <- 2 + 3*dat$x + 1.5*dat$x^2 + rnorm(200, sd = 1)  # true quadratic

# Fit linear vs quadratic models
m1 <- brm(y ~ x, data = dat, family = gaussian(), refresh = 0)
m2 <- brm(y ~ x + I(x^2), data = dat, family = gaussian(), refresh = 0)

loo1 <- loo(m1)
loo2 <- loo(m2)
loo_compare(loo1, loo2)

pp_check(m1)
pp_check(m2)
```

6.3.5 Example B — Quick WAIC Comparison via Frequentist Approximation

```r
set.seed(42)
N <- 150
x <- rnorm(N)
y <- 1.5 + 2.2*x + rnorm(N, sd=1.2)

m1 <- lm(y ~ x)
m2 <- lm(y ~ poly(x, 2, raw = TRUE))

sigma1 <- summary(m1)$sigma
sigma2 <- summary(m2)$sigma

# Pseudo-WAIC: approximate lppd - 2p using residual sum of squares
RSS1 <- sum(resid(m1)^2)
RSS2 <- sum(resid(m2)^2)
```

```
npar1 <- length(coef(m1))
npar2 <- length(coef(m2))

WAIC1 <- -2 * (sum(dnorm(y, predict(m1), sigma1, log=TRUE)) - npar1)
WAIC2 <- -2 * (sum(dnorm(y, predict(m2), sigma2, log=TRUE)) - npar2)

data.frame(Model=c("Linear","Quadratic"), WAIC=c(WAIC1,WAIC2))
```

```
      Model     WAIC
1    Linear 473.9530
2 Quadratic 475.7823
```

6.3.6 Visual Predictive Comparison

```
plot(x, y, pch=19, col="#00000055", main="Observed Data with Fitted Models")
xs <- seq(min(x), max(x), length.out=200)
lines(xs, predict(m1, newdata=data.frame(x=xs)), col="steelblue", lwd=2)
lines(xs, predict(m2, newdata=data.frame(x=xs)), col="firebrick", lwd=2)
legend("topleft", lwd=2, col=c("steelblue","firebrick"),
       legend=c("Linear","Quadratic"), bty="n")
```

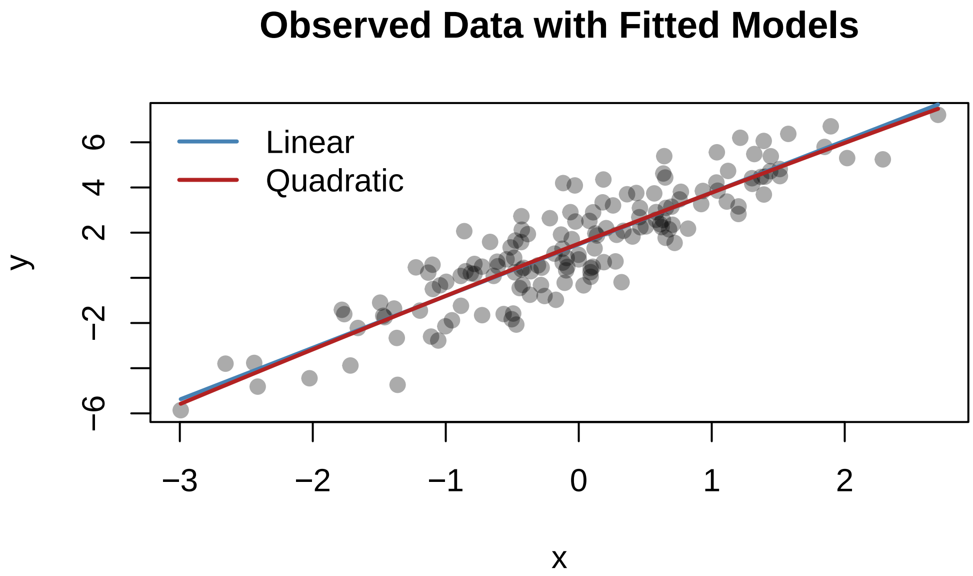


Figure 6.3: Overlay of linear and quadratic model fits

6.3.7 Practical Summary

Method	Strength	Limitation
Posterior Predictive Check	Diagnoses lack of fit to **observed** data	Not inherently comparative
Bayes Factor	Theoretically coherent model evidence	Sensitive to priors; hard integration
WAIC / LOO	Out-of-sample predictive performance	Approximate; needs posterior draws

---

6.4 Lab 5 — Model Checking and Comparison

Objectives

1. Perform PPC using both numerical and visual methods.
2. Compute and interpret WAIC and LOO for model selection.
3. Visualize predictive differences among models.

Packages

`brms`, `loo`, `bayesplot`, `ggplot2`

Tasks

- Fit two Bayesian regression models on the same dataset.
- Conduct posterior predictive checks and compare simulated vs. observed data.
- Compute WAIC and LOO; summarize which model performs better.

6.5 Homework 5

1. **Conceptual**
   - Explain the purpose of posterior predictive checks.
   - Compare WAIC and LOO conceptually.
2. **Computational**
   - Simulate data from a known model. Fit two Bayesian models in R.
   - Use PPC, WAIC, and LOO to assess fit.
   - Discuss how model choice depends on criterion used.
3. **Reflection**
   - Why might visual checks and numerical metrics disagree?
   - Which model would you report and why?

6.6 Key Takeaways

Concept	Summary
Posterior Predictive Check	Compares observed data to replicated draws under the posterior.
Posterior Predictive p-Value	Quantifies fit; extremes suggest model misfit.
WAIC / LOO	Predictive performance measures for Bayesian models.
Bayes Factor	Ratio of marginal likelihoods for model comparison.
Combined Evaluation	Use graphical and numerical criteria together.

Next Week: Hierarchical Bayesian Models — introducing partial pooling and shrinkage.

7 Week 6 — Hierarchical Bayesian Models

This week introduces **hierarchical (multilevel) Bayesian models**, which allow parameters to vary across groups while sharing information through higher-level priors.

We study partial pooling, shrinkage, and their implementation for normal and regression models.

7.1 Learning Goals

By the end of this week, you should be able to:

- Explain the motivation for hierarchical modeling.
- Formulate hierarchical models with group-level parameters.
- Interpret partial pooling and shrinkage.
- Implement a two-level Bayesian model in R using simulation or `brms`.
- Compare complete, no-pooling, and partial-pooling approaches.

7.2 Lecture 1 — Motivation and Structure of Hierarchical Models

7.2.1 1.1 Why Hierarchical Models?

Hierarchical models capture **structured variability** among related groups or units:

- Repeated measures within individuals
- Students within classrooms
- Machines within factories

They balance **within-group** and **between-group** information by introducing group-specific parameters drawn from a common population distribution.

7.2.2 1.2 Model Structure

For group $j = 1, \dots, J$ and observations $i = 1, \dots, n_j$:

$$y_{ij} \mid \theta_j, \sigma^2 \sim \mathcal{N}(\theta_j, \sigma^2), \qquad \theta_j \mid \mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2).$$

Top-level priors:

$$\mu \sim \mathcal{N}(0, 10^2), \quad \tau \sim \text{Half-Cauchy}(0, 5).$$

- μ: overall population mean

- τ: between-group standard deviation (pooling strength)
- σ: within-group standard deviation

7.2.3 1.3 Three Extremes of Pooling

Model Type	Description	Behavior
No pooling	Estimate each θ_j separately	Ignores commonality across groups
Complete pooling	Force all groups to share one parameter	Ignores group differences
Partial pooling	Combine information via hierarchical prior	Balances both; default Bayesian choice

---

7.2.4 1.4 Shrinkage Intuition

Posterior for each group mean $ _j$ shrinks toward the global mean μ:

$$E[\theta_j \mid y] = w_j \bar{y}_j + (1 - w_j)\mu,$$

where

$$w_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}.$$

- Large n_j (lots of data): $w_j \to 1 \to$ less shrinkage.
- Small n_j: $w_j \to 0 \to$ stronger shrinkage toward μ.

7.2.5 1.5 Example — Simulated Group Means

```r
set.seed(6)
J <- 8; n_j <- rep(10, J)
mu_true <- 5; tau_true <- 2; sigma_true <- 1

theta_true <- rnorm(J, mu_true, tau_true)
y <- sapply(theta_true, function(tj) rnorm(10, tj, sigma_true))
ybar <- colMeans(y)

# No pooling (group means)
no_pool <- ybar

# Complete pooling (global mean)
complete_pool <- mean(y)

# Partial pooling: simple empirical Bayes shrinkage
tau_hat <- sd(ybar)
sigma_hat <- sd(as.vector(y))
w <- (n_j/sigma_hat^2) / (n_j/sigma_hat^2 + 1/tau_hat^2)
partial_pool <- w*ybar + (1-w)*complete_pool

data.frame(Group=1:J,
           ybar=round(ybar,2),
           NoPool=round(no_pool,2),
           Partial=round(partial_pool,2))
```

```
  Group ybar NoPool Partial
1     1 5.53   5.53    5.53
2     2 4.06   4.06    4.21
3     3 6.90   6.90    6.76
4     4 8.19   8.19    7.91
5     5 4.86   4.86    4.93
6     6 5.42   5.42    5.43
7     7 2.20   2.20    2.55
8     8 6.99   6.99    6.84
```

Observe how partial-pool estimates move small-sample groups toward the global mean.

7.2.6 1.6 Advantages of Hierarchical Models

- Borrow strength across groups.
- Naturally incorporate uncertainty at multiple levels.
- Handle unbalanced data and missingness elegantly.
- Allow group-level predictors and complex dependence structures.

7.3 Lecture 2 — Hierarchical Regression and Implementation

7.3.1 2.1 Hierarchical Linear Regression

General form:

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$
$$\alpha_j \sim \mathcal{N}(\mu_\alpha, \tau_\alpha^2), \quad \beta_j \sim \mathcal{N}(\mu_\beta, \tau_\beta^2).$$

This allows both intercepts and slopes to vary by group.

7.3.2 2.2 Example — Hierarchical Regression with `brms`

```
library(brms)
set.seed(7)

J <- 10
n_j <- 20
group <- rep(1:J, each=n_j)
x <- rnorm(J*n_j, 0, 1)

alpha_true <- rnorm(J, 2, 1)
beta_true  <- rnorm(J, 3, 0.5)
sigma_true <- 0.8
```

```
y <- alpha_true[group] + beta_true[group]*x + rnorm(J*n_j, 0, sigma_true)
dat <- data.frame(y, x, group=factor(group))

# Hierarchical model (random intercept and slope)
m_hier <- brm(y ~ 1 + x + (1 + x | group),
              data=dat, family=gaussian(),
              chains=2, iter=2000, refresh=0)

summary(m_hier)
plot(m_hier)
```

The `(1 + x | group)` formula defines a **varying intercept and slope** for each group.

7.3.3 2.3 Interpretation

Posterior summaries provide:

- Group-level means α_j, β_j.
- Population-level means μ_α, μ_β.
- Variability estimates τ_α, τ_β showing degree of pooling.

Visualize partial pooling by comparing group-specific fits to the global regression line.

```
pp_check(m_hier)
plot(conditional_effects(m_hier), points=TRUE)
```

7.3.4 2.4 Practical Considerations

- Choose weakly informative hyperpriors for scale parameters (e.g., Half-Cauchy or Exponential).

- Inspect group-level posterior intervals to assess pooling.

- Center predictors for numerical stability.

- Use hierarchical models as the default when groups share a common process.

7.3.5 2.5 Summary of Hierarchical Modeling Benefits

Feature	Description
Partial pooling	Shares strength across groups while retaining group differences.
Shrinkage	Stabilizes small-sample estimates toward population mean.
Interpretability	Captures multi-level variation naturally.
Predictive accuracy	Usually superior to separate or fully pooled models.

7.4 Homework 6

1. **Conceptual**
   - Explain why hierarchical modeling is often superior to analyzing groups separately.
   - Distinguish between complete pooling, no pooling, and partial pooling.
2. **Computational**
   - Simulate a small dataset with several groups and fit:
     a. Separate regressions (no pooling).

     b. A single pooled regression.

     c. A hierarchical model (partial pooling).
   - Compare estimates for each group and interpret shrinkage behavior.
3. **Reflection**
   - In what situations would you *not* use a hierarchical model?
   - How does the hierarchical prior act as a regularizer?

7.5 Key Takeaways

Concept	Summary
Hierarchical Model	Combines group-level and population-level inference.
Partial Pooling	Balances within- and between-group information.
Shrinkage	Moves noisy group estimates toward a global mean.
Hierarchical Regression	Extends pooling to both intercepts and slopes.
Practical Insight	Default choice when analyzing grouped or multilevel data.

Next Week: Bayesian Decision Theory — introducing utilities, losses, and optimal decision rules under uncertainty.

8 Week 7 — Bayesian Decision Theory

This week introduces the **decision-theoretic foundation** of Bayesian inference.
We study how posterior distributions lead naturally to optimal decisions when losses or utilities are specified, and apply the theory to point estimation and hypothesis testing.

8.1 Learning Goals

By the end of this week, you should be able to:

- Describe the Bayesian decision-theoretic framework.
- Define loss functions and posterior expected loss.
- Derive Bayes rules for common loss functions.
- Apply Bayesian decision principles to estimation and classification.
- Distinguish between point estimation, interval estimation, and decision-making contexts.

8.2 Lecture 1 — Principles of Bayesian Decision Theory

8.2.1 1.1 Motivation

Statistical inference often involves making decisions under uncertainty:
select an action $a$based on observed data y.
Each action has a **loss** (or **utility**) depending on the true parameter value $ $.

8.2.2 1.2 The Decision-Theoretic Setup

- Parameter: $\theta \in \Theta$
- Data: y
- Action space: $\mathcal{A}$
- Loss function: $L(a, \theta)$

After observing y, the Bayesian chooses an action $a(y)$minimizing **posterior expected loss**:

$$\rho(a \mid y) = E[L(a, \theta) \mid y] = \int L(a, \theta)\, p(\theta \mid y)\, d\theta.$$

Bayes rule:

$$a^*(y) = \arg\min_a \rho(a \mid y).$$

8.2.3 1.3 Common Loss Functions and Bayes Rules

Loss Function	Form	Bayes Action
Squared Error	$L(a, \theta) = (a - \theta)^2$	Posterior mean $E[\theta \mid y]$
Absolute Error	$L(a, \theta) = \|a - \theta\|$	Posterior median
0–1 Loss	$L(a, \theta) = \mathbb{1}\{a \neq \theta\}$	Posterior mode (MAP)

These connect the **posterior mean, median, and mode** to optimal decisions under different losses.

8.2.4 1.4 Example — Estimation under Quadratic Loss

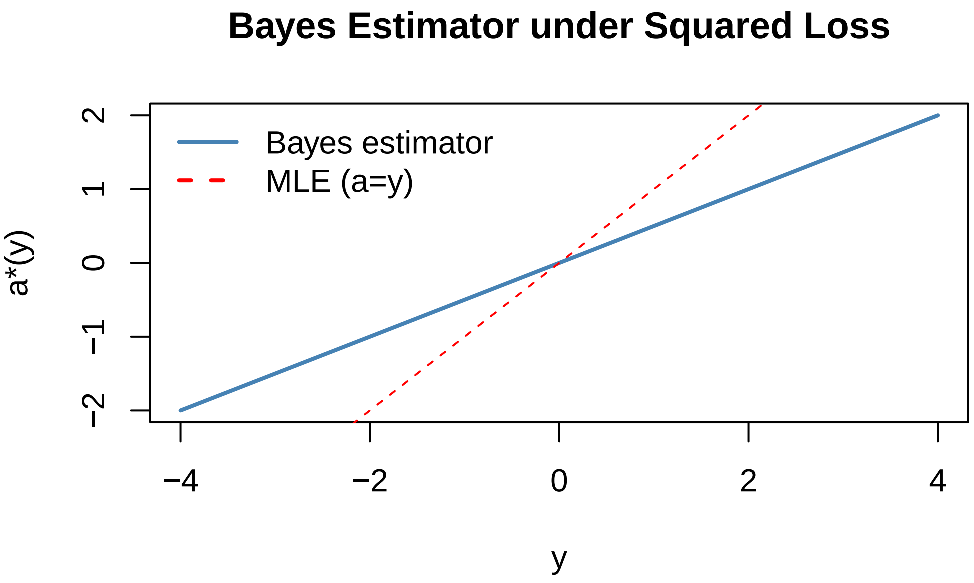
Suppose $y \mid \theta \sim N(\theta, 1)$with prior $\theta \sim N(0, 1)$.

Posterior: $\theta \mid y \sim N(\frac{y}{2}, \frac{1}{2})$.

Bayes estimator under squared loss:

$$a^*(y) = E[\theta \mid y] = \frac{y}{2}.$$

```r
set.seed(7)
y <- seq(-4, 4, length=100)
bayes_est <- y/2
plot(y, bayes_est, type="l", lwd=2, col="steelblue",
     main="Bayes Estimator under Squared Loss", xlab="y", ylab="a*(y)")
abline(a=0, b=1, col="red", lty=2)
legend("topleft", legend=c("Bayes estimator","MLE (a=y)"),
       col=c("steelblue","red"), lwd=2, lty=c(1,2), bty="n")
```



Interpretation: The Bayes rule shrinks the estimate toward zero (the prior mean), especially for small |y|.

8.2.5 1.5 Decision Rules and Risk

The **Bayes risk** is the expected loss averaged over data and parameters:

$$r(a) = E[L(a(Y), \Theta)] = \int\int L(a(y), \theta)\, p(y, \theta)\, dy\, d\theta.$$

A decision rule minimizing Bayes risk across all priors is **admissible** (cannot be uniformly improved).

8.2.6 1.6 Example — Hypothesis Testing with 0–1 Loss

We test $H_0 : \theta \leq 0$vs $H_1 : \theta\, 0$.

Loss: $L(a, \theta) = \begin{cases} 0 & \text{if correct,} \\ 1 & \text{if wrong.} \end{cases}$

Posterior decision rule:

$$\text{Accept } H_1 \text{ if } P(\theta > 0 \mid y) > 0.5.$$

```r
set.seed(8)
theta_draws <- rnorm(5000, mean=1, sd=1)
mean(theta_draws > 0)   # posterior probability of H1
```

```
[1] 0.8406
```

8.3 Lecture 2 — Applications and Extensions

8.3.1 2.1 Bayesian Credible Intervals as Decision Regions

For a given loss that penalizes excluding the true parameter,
a **credible interval** minimizing posterior expected loss corresponds to the shortest interval containing a fixed posterior probability (e.g. 95%).

```r
theta_post <- rnorm(5000, mean=2, sd=1)
quantile(theta_post, c(0.025, 0.975))
```

```
        2.5%        97.5%
-0.004383415  4.024907476
```

8.3.2 2.2 Decision Theory for Classification

For a two-class problem with class probabilities $p_1 = P(Y = 1 \mid x)$and $p_0 = 1 - p_1$: Minimize expected loss $L(a, y)$using a **loss matrix**.

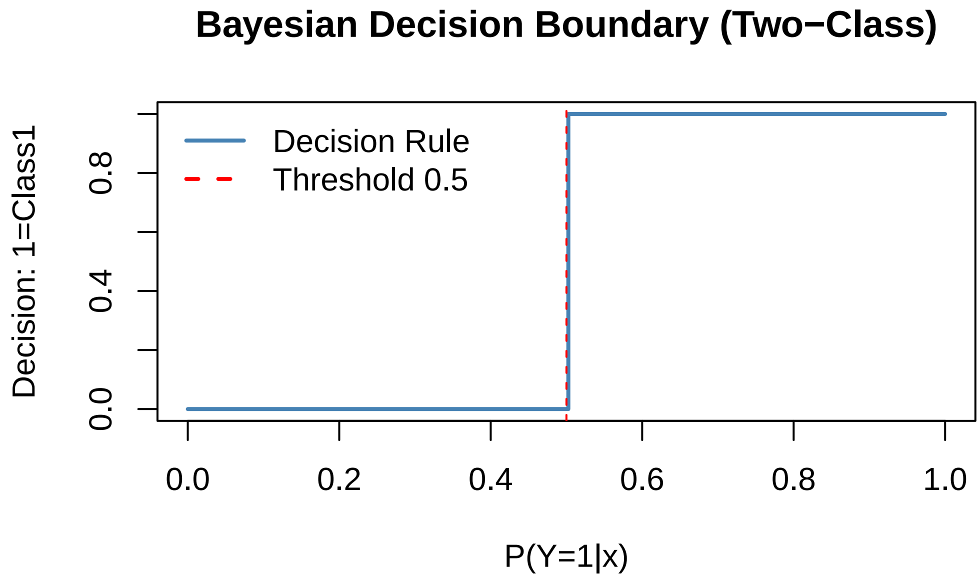
True Class	Predict 0	Predict 1
0	0	c_{10}
1	c_{01}	0

Bayes rule: choose class 1 if

$$\frac{p_1}{p_0} > \frac{c_{10}}{c_{01}}.$$

The usual 0–1 loss corresponds to $c\_10 = c\_01 = 1$, threshold $= 0.5$.

```r
p1 <- seq(0,1,length=200)
threshold <- 0.5
plot(p1, ifelse(p1>threshold,1,0), type="s", col="steelblue", lwd=2,
     main="Bayesian Decision Boundary (Two-Class)", xlab="P(Y=1|x)", ylab="Decision: 1=Class1
abline(v=threshold, col="red", lty=2)
legend("topleft", legend=c("Decision Rule","Threshold 0.5"),
       col=c("steelblue","red"), lwd=2, lty=c(1,2), bty="n")
```



8.3.3 2.3 Loss vs Utility

Utility $U(a, \theta)$is simply the negative of loss.
Maximizing expected utility is equivalent to minimizing expected loss:

$$a^*(y) = \arg\max_a E[U(a, \theta) \mid y].$$

This framing is often used in economics and decision analysis.

8.3.4 2.4 Connection to Frequentist Estimation

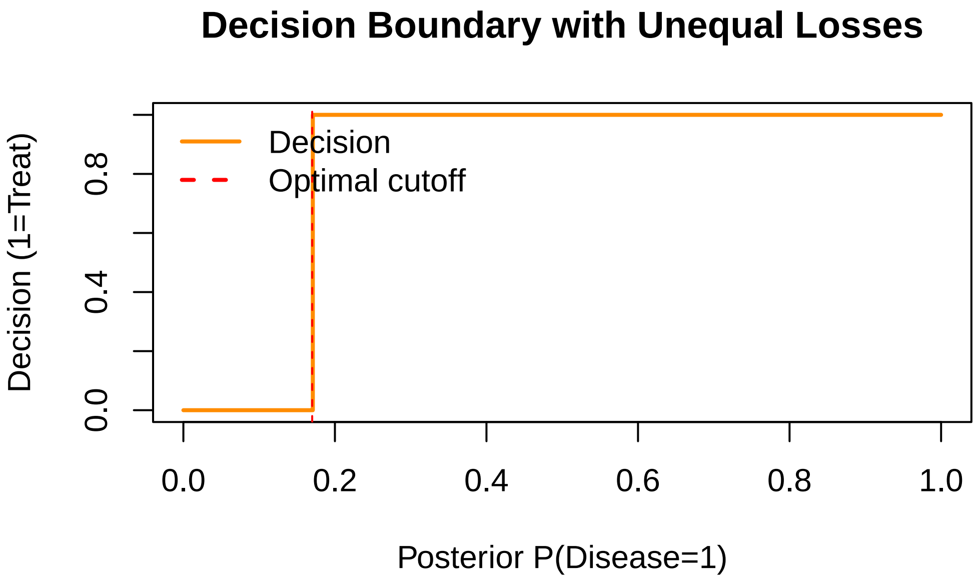
Under certain priors and symmetric losses, Bayes rules coincide with frequentist estimators (e.g. posterior mean = MLE for flat priors).
Bayesian decision theory thus **generalizes** classical estimation.

8.3.5 2.5 Example — Optimal Cutoff for a Diagnostic Test

Let θdenote disease presence (1 = disease).
If false negatives cost 5× more than false positives, the optimal threshold satisfies

$$\frac{p_1}{p_0} > \frac{1}{5} \Rightarrow p_1 > 0.17.$$

```r
p <- seq(0,1,length=200)
decision <- ifelse(p > 0.17, 1, 0)
plot(p, decision, type="s", col="darkorange", lwd=2,
     main="Decision Boundary with Unequal Losses", xlab="Posterior P(Disease=1)", ylab="Deci
abline(v=0.17, col="red", lty=2)
legend("topleft", legend=c("Decision","Optimal cutoff"), col=c("darkorange","red"),
       lwd=2, lty=c(1,2), bty="n")
```



8.3.6 2.6 Summary of Bayesian Decision Theory

Concept	Description
Loss function	Quantifies cost of wrong decisions
Posterior expected loss	Average loss given observed data
Bayes rule	Action minimizing posterior expected loss
Common losses	Squared, absolute, 0–1
Applications	Estimation, hypothesis testing, classification, decision support

---

8.4 Homework 7

1. **Conceptual**
   - Define loss and risk in the Bayesian framework.

- What is the relationship between posterior mean, median, and mode under different losses?

2. **Computational**
   - Simulate data from $N(\theta, 1)$with prior $N(0, 1)$.
   - Compute the Bayes estimator under squared loss and compare it with the MLE.
   - Repeat using absolute loss and report the posterior median.
3. **Reflection**
   - How does changing the loss function alter your decision?
   - Give a real-world example where asymmetric losses are important.

8.5 Key Takeaways

Concept	Summary
Decision Theory	Provides a unified framework linking inference to action.
Bayes Rule	Minimizes posterior expected loss.
Common Losses	Squared → mean; absolute → median; 0–1 → mode.
Applications	Estimation, testing, classification, optimal thresholds.
Perspective	Inference as a special case of decision-making under uncertainty.

Next Week: Advanced Bayesian Computation — Hamiltonian Monte Carlo (HMC) and Variational Inference.

9 Week 8 — Advanced Bayesian Computation

This week explores two major developments that enable scalable Bayesian inference for complex or high-dimensional models:
Hamiltonian Monte Carlo (HMC) and **Variational Inference (VI)**.
We study their principles, intuition, and practical use in software such as `Stan` and `brms`.

9.1 Learning Goals

By the end of this week, you should be able to:

- Explain the motivation for advanced sampling and approximation methods.
- Describe the mechanics and intuition of Hamiltonian Monte Carlo.
- Understand the trade-offs between exact (MCMC) and approximate (VI) inference.
- Run basic HMC and VI fits using modern R interfaces.
- Interpret diagnostics for both approaches.

9.2 Lecture 1 — Hamiltonian Monte Carlo (HMC)

9.2.1 1.1 Motivation

Traditional MCMC (e.g., Metropolis–Hastings, Gibbs) can mix slowly in high dimensions. **Hamiltonian Monte Carlo** accelerates exploration by using gradient information from the log posterior to simulate physical motion through parameter space.

9.2.2 1.2 Hamiltonian Dynamics

We introduce an auxiliary "momentum" variable p and define the **Hamiltonian**:

$$H(\theta, p) = U(\theta) + K(p),$$

where
- $U(\theta) = -\log p(\theta \mid y)$ (potential energy = negative log posterior),
- $K(p) = \frac{1}{2}p^\top M^{-1}p$ (kinetic energy, with mass matrix M).

The system evolves via Hamilton's equations:

$$\frac{d\theta}{dt} = \frac{\partial H}{\partial p}, \qquad \frac{dp}{dt} = -\frac{\partial H}{\partial \theta}.$$

9.2.3 1.3 Leapfrog Integration

To approximate continuous motion, HMC uses a **leapfrog integrator** with step size ϵ and L steps:

1. $p\_t + \epsilon/2 = p_t - \frac{\epsilon}{2}\nabla\_\theta U(\theta_t)$
2. $\theta\_t + \epsilon = \theta_t + \epsilon M^{-1}p\_t + \epsilon/2$
3. $p\_t + \epsilon = p\_t + \epsilon/2 - \frac{\epsilon}{2}\nabla\_\theta U(\theta\_t + \epsilon)$

After simulating this path, we apply a **Metropolis acceptance step** using the change in H.

9.2.4 1.4 Intuition

- The gradient $\nabla\_\theta U(\theta)$ guides proposals along high-density regions, avoiding random walk behavior.
- Proper tuning of step size ϵ and number of steps L yields efficient exploration.
- Modern implementations (e.g., *Stan*) adapt these automatically via the **No-U-Turn Sampler (NUTS)**.

9.2.5 1.5 Example — Logistic Regression with HMC (Stan)

```r
library(brms)
set.seed(8)

# Simulated logistic data
N <- 200
x <- rnorm(N)
y <- rbinom(N, 1, plogis(-1 + 2*x))
dat <- data.frame(x, y)

# Fit via Hamiltonian Monte Carlo (NUTS)
fit_hmc <- brm(y ~ x, data=dat, family=bernoulli(), chains=2, iter=2000, refresh=0)
summary(fit_hmc)
plot(fit_hmc)
```

`Stan`'s NUTS algorithm performs automatic adaptation of step size and trajectory length.

9.2.6 1.6 Diagnosing HMC Performance

Key diagnostics: - **Divergent transitions** → numerical instability (reduce step size or re-scale parameters).
- **Energy Bayesian Fraction of Missing Information (E-BFMI)** → low values (<0.3) indicate poor exploration.
- $\widehat{R}$ and effective sample size → check convergence and mixing.

```r
library(bayesplot)
mcmc_nuts_divergence(fit_hmc)
mcmc_trace(fit_hmc, pars=c("b_Intercept","b_x"))
```

9.2.7 1.7 Advantages of HMC

Feature	Benefit
Gradient-based proposals	Rapid movement through high-density regions
Higher acceptance rates	Fewer rejections than random-walk MH
Fewer tuning parameters	Automatic adaptation (NUTS)
Robust for high-dimensional models	Used in most modern Bayesian software

---

9.3 Lecture 2 — Variational Inference (VI)

9.3.1 2.1 Motivation

When exact sampling is too costly (e.g., massive datasets, deep models),
Variational Inference (VI) approximates the posterior by a simpler distribution $q\_\lambda(\theta)$ within a parameterized family.

9.3.2 2.2 Objective Function

We minimize the **Kullback–Leibler (KL) divergence**:

$$\mathrm{KL}(q_\lambda(\theta) \,\|\, p(\theta \mid y)) = \int q_\lambda(\theta) \log \frac{q_\lambda(\theta)}{p(\theta \mid y)} \, d\theta.$$

Equivalently, we **maximize the Evidence Lower Bound (ELBO):**

$$\mathrm{ELBO}(\lambda) = E_{q_\lambda}[\log p(y,\theta)] - E_{q_\lambda}[\log q_\lambda(\theta)].$$

The higher the ELBO, the closer $q\_\lambda(\theta)$ is to the true posterior.

9.3.3 2.3 Mean-Field Approximation

A common simplification assumes factorization:

$$q_\lambda(\theta) = \prod_j q_{\lambda_j}(\theta_j),$$

which allows coordinate-wise optimization of each factor.

9.3.4 2.4 Example — Variational Bayes for a Normal Mean

Assume $y_i \sim N(\theta, 1)$ with prior $\theta \sim N(0, 1)$.

Analytically, the posterior is $N\left(\frac{n\bar{y}}{n+1}, \frac{1}{n+1}\right)$.
We approximate it variationally by another normal $q(\theta) = N(m, s^2)$,
and find m, s^2 maximizing ELBO.

```
set.seed(9)
y <- rnorm(50, mean=1)
n <- length(y)
log_joint <- function(theta) sum(dnorm(y, theta, 1, log=TRUE)) + dnorm(theta, 0, 1, log=TRUE

# Closed-form optimal q is Normal(m,s^2) with same moments as true posterior:
m_vi <- n*mean(y)/(n+1)
s2_vi <- 1/(n+1)
c(mean=m_vi, sd=sqrt(s2_vi))
```

```
     mean        sd 
0.8884051 0.1400280 
```

9.3.5 2.5 Automatic VI with `brms`

```
library(brms)
set.seed(10)
N <- 1000
x <- rnorm(N)
y <- 2 + 1.5*x + rnorm(N)
dat <- data.frame(x,y)

fit_vi <- brm(y ~ x, data=dat, family=gaussian(),
              algorithm="meanfield", iter=5000, refresh=0)
summary(fit_vi)
```

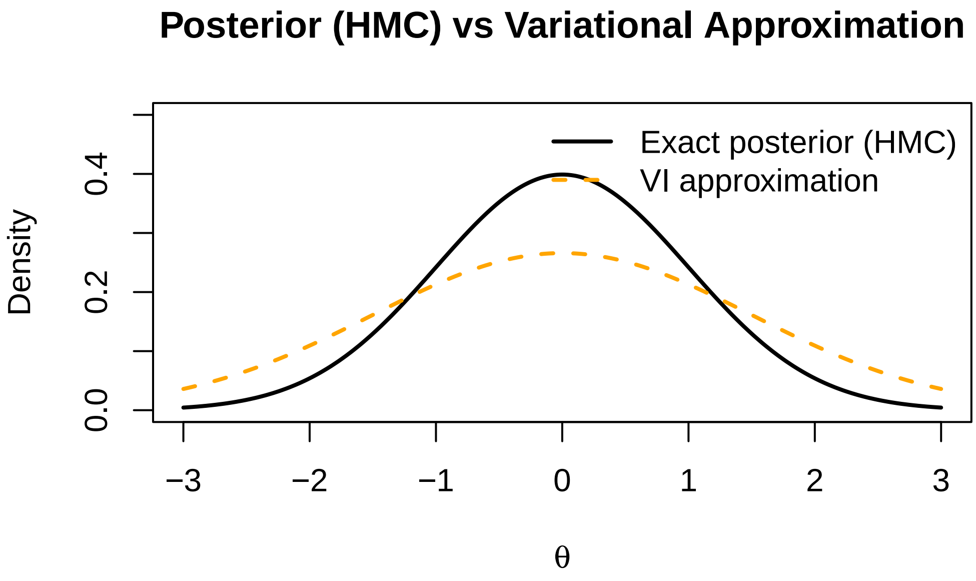
VI provides a fast deterministic approximation, trading off accuracy for scalability.

9.3.6 2.6 Comparison: HMC vs VI

Feature	HMC (NUTS)	Variational Inference
Type	Sampling (asymptotically exact)	Optimization (approximate)
Accuracy	Very high	Depends on variational family
Speed	Slower	Very fast
Diagnostics	Convergence via $\widehat{R}$, ESS	ELBO convergence
Use case	Complex or small data	Massive or real-time problems

9.3.7 2.7 Visual Comparison (Conceptual)

```r
theta <- seq(-3,3,length=400)
posterior <- dnorm(theta, 0, 1)              # true posterior
vi_approx <- dnorm(theta, 0, 1.5)            # wider variational approx
plot(theta, posterior, type="l", lwd=2, col="black", ylim=c(0,0.5),
     main="Posterior (HMC) vs Variational Approximation",
     ylab="Density", xlab=expression(theta))
lines(theta, vi_approx, col="orange", lwd=2, lty=2)
legend("topright", legend=c("Exact posterior (HMC)","VI approximation"),
       col=c("black","orange"), lwd=2, lty=c(1,2), bty="n")
```



9.3.8 2.8 Practical Advice

- Use **HMC (NUTS)** as the default for accuracy and diagnostics.
- Use **VI** for large-scale models, initialization, or quick exploration.
- Compare results: if VI and HMC differ substantially, favor HMC.

9.4 Homework 8

1. **Conceptual**
   - Explain the difference between sampling-based and optimization-based inference.
   - What role does the ELBO play in VI?
2. **Computational**

- Fit a simple linear regression using both HMC and VI in `brms`.
- Compare posterior means, standard deviations, and computation time.

3. **Reflection**
    - In what types of real-world problems might VI be preferred over HMC?
    - How would you check whether your VI approximation is adequate?

9.5 Key Takeaways

Concept	Summary
Hamiltonian Monte Carlo	Uses gradients to propose efficient moves through parameter space.
No-U-Turn Sampler (NUTS)	Adapts step size and trajectory automatically.
Variational Inference	Optimizes a tractable approximation to the posterior.
ELBO	Objective function for VI; measures closeness to the true posterior.
Trade-off	HMC = accuracy, VI = speed; choice depends on model and data size.

Next Week: Bayesian Model Averaging and Ensemble Learning — combining multiple Bayesian models for improved predictive performance.

References