STAT8310 - Bayesian Data Analysis

Chi-Kuang Yeh

2026-01-22

Table of contents

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

10:00–13:00 on Monday, or by appointment.

Grade Distribution

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

Assignment

☐ A1, due on Jan 29, 2026

Midterm

☐ March 3, 2026

Topics and Corresponding Lectures

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

Status	Topic	Lecture Covered
	Welcome and Overview	1
	Intro to R Programming	2
	Probability and Exchangeability	3–

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 Belief function and Probability Review

Leading objectives:

be familiar with the following concepts

- Belief Functions
- Probability
- Bayes' Rule
- Random Variables
- Exchangeability

2.1 Belief functions

Probability is a way to express rational beliefs.

A **belief function** Be$(\cdot)$ is a function that assigns number to statements such that the large the number, the higher the degree of belief.

Let F, G, and H be three possibly overlapping statements about the world.

For example:

- F = { a person owns a smartphone }
- G = { a person uses social media daily }
- H = { a person works remotely at least part of the time }

or

- F = { a person has a graduate degree }
- G = { a person works in a STEM field }
- H = { a person is employed in the private sector }

The perference over bets involving these statements can be used to define a belief function

- Be$(F) >$ Be(G) means you prefer a bet F is true over that G is true.

Also, we want Be$(\cdot)$ to describe our beliefs under certain conditions

- $\text{Be}(F \mid H) > \text{Be}(G \mid H)$ means that if we knew that H were true, then we would perfer to bet that F is also true over G is also true.
- $\text{Be}(F \mid G) > \text{Be}(F \mid H)$ means that if we bet on F, we would perfer to do it under the condition that G is true rather than H is true.

Some more notations:

- Let $\neg$ denote negation. That is, $\neg F$ is the statement that F is not true.
- Let $F \vee G$ denote the disjunction (or) of statements F and G, meaning that at least one of F or G is true.
- Let $F \wedge G$ denote the conjunction (and) of statements F and G, meaning that both F and G are true.

It has been argued by many that any function that is to numerically represent our beliefs should have the following properties:

- **B1**: $\text{Be}(\neg H \mid H) \leq \text{Be}(F \mid H) \leq \text{Be}(H \mid H)$
- **B2**: $\text{Be}(F \vee G \mid H) \geq \max\{\text{Be}(F \mid H), \text{Be}(G \mid H)\}$
- **B3**: $\text{Be}(F \wedge G \mid H)$ can be derived from $\text{Be}(G \mid H)$ and $\text{Be}(F \mid G \wedge H)$

How should we interpret these properties, and do they make sense?

- B1 means that the number we assign to $\text{Be}(F \mid H)$, our conditional belief in F given H, is bounded below and above by the numbers we assign to complete disbelief $\text{Be}(\neg H \mid H)$ and complete belief $\text{Be}(H \mid H)$.
- B2 says that our belief that the truth lies in a given set of possibilities should not decrease as we add to the set of possibilities.
- B3 is a bit trickier. To see why it makes sense, imagine you have to decide whether or not F and G are true, knowing that H is true. You could do this by first deciding whether or not G is true given H, and if so, then deciding whether or not F is true given G and H.

Recall the notation from (elementary) probability that, $F \cup G$ means F or G, and $F \cap G$ means F and G, and $\emptyset$ is the empty set.

- **P1**:
$$0 = \Pr(\neg H \mid H) \leq \Pr(F \mid H) \leq \Pr(H \mid H) = 1$$
- **P2**:
$$\Pr(F \cup G \mid H) = \Pr(F \mid H) + \Pr(G \mid H), \quad \text{if} \quad F \cap G = \emptyset$$
- **P3**:
$$\Pr(F \cap G \mid H) = \Pr(G \mid H)\Pr(F \mid G \cap H)$$

2.1.1 Conclusion

You can see that, a probability function satisfy P1–P3 also satisfies B1–B3. Therefore, probability functions are a special case of belief functions, and we can use it to describe our belief.

2.2 Events, Partitions and Bayes' Rule

A collectiion of sets $\{H_1, \dots, H_K\}$ is a partition of another set $\mathcal{H}$ if

1. $H_i \cap H_j = \emptyset$ for all $i \neq j$ (mutually exclusive);
2. $\bigcup_{i=1}^{K} H_i = \mathcal{H}$ (collectively exhaustive).

In the context of indetifying which of several statements is true, if $\mathcal{H}$ is the set of all possible truths and $\{H_1, \dots, H_K\}$ is a partition of $\mathcal{H}$, then exactly one set H_j contains the truth.

Let $\mathcal{H}$ be the status of a statistical model.

Valid partitions include:

- {correctly specified, misspecified}
- {underfitting, well-specified, overfitting}

2.2.1 Partition and Probability

Suppose $\{H_1, \dots, H_K\}$ is a partition of $\mathcal{H}$,$\Pr(\mathcal{H}) = 1$ and E is some specific event. Then, by the axioms of probability, we have

- Law of total probability

$$\sum_{k=1}^{K} \Pr(H_k) = \Pr\left(\bigcup_{k=1}^{K} H_k\right) = \Pr(\mathcal{H}) = 1$$

- Law of marginal probability

$$\Pr(E) = \sum_{k=1}^{K} \Pr(E \cap H_k) = \sum_{k=1}^{K} \Pr(E \mid H_k)\Pr(H_k)$$

- Bayes' rule

$$\Pr(H_j \mid E) = \frac{\Pr(E \mid H_j)\Pr(H_j)}{\Pr(E)} = \frac{\Pr(E \mid H_j)\Pr(H_j)}{\sum_{k=1}^{K} \Pr(E \mid H_k)\Pr(H_k)}$$

A subset of the 1996 General Social Survey includes data on the education level and income for a sample of males over 30 years of age. Let {H1,H2,H3,H4} be the events that a randomly selected person in this sample is in, respectively, the lower 25th percentile, the second 25th percentile, the third 25th percentile and the upper 25th percentile in terms of income. By definition,

$$\{\Pr(H_1), \Pr(H_2), \Pr(H_3), \Pr(H_4)\} = \{.25, .25, .25, .25\}.$$

Note that $\{H1, H2, H3, H4\}$ is a partition and so these probabilities sum to 1. Let E be the event that a randomly sampled person from the survey has a college education. From the survey data, we have

$$\{\Pr(E \mid H_1), \Pr(E \mid H_2), \Pr(E \mid H_3), \Pr(E \mid H_4)\} = \{.11, .19, .31, .53\}.$$

These probabilities do not sum to 1 - they represent the proportions of people with college degrees in the four different income subpopulations H_1, H_2, H_3 and H_4. Now let's consider the income distribution of the college-educated population. Using Bayes' rule we can obtain

$\{\Pr(H_1 \mid E), \Pr(H_2 \mid E), \Pr(H_3 \mid E), \Pr(H_4 \mid E)\} = \{0.09, 0.17, 0.27, 0.47\}$, and we see that the income distribution for people in the college-educated population differs markedly from $\{0.25, 0.25, 0.25, 0.25\}$, the distribution for the general population. Note that these probabilities do sum to 1 - they are the conditional probabilities of the events in the partition, given E.

In Bayesian inference, $H_1, ..., H_K$ often refer to disjoint hypotheses or states of nature and E refers to the outcome of a survey, study or experiment. To compare hypotheses *post-experimentally*, we often calculate the following ratio:

$$\begin{aligned}
\frac{\Pr(H_i \mid E)}{\Pr(H_j \mid E)} &= \frac{\Pr(E \mid H_i)\Pr(H_i) / \Pr(E)}{\Pr(E \mid H_j)\Pr(H_j) / \Pr(E)} \\
&= \frac{\Pr(E \mid H_i)\Pr(H_i)}{\Pr(E \mid H_j)\Pr(H_j)} \\
&= \frac{\Pr(E \mid H_i)}{\Pr(E \mid H_j)} \times \frac{\Pr(H_i)}{\Pr(H_j)} \\
&= \text{ "Bayes factor" } \times \text{ "prior beliefs".}
\end{aligned}$$

This calculation reminds us that Bayes' rule does not determine what our *beliefs should be* after seeing the data, it only tells us how they *should change after seeing the data.*

2.3 Independence

Two events F and G are conditionally independent, if given H, we have $\Pr(F \cap G \mid H) = \Pr(F \mid H)\Pr(G \mid H)$.

How do we interpret conditional independence? By Axiom P3, the following is always true:

$$\begin{aligned}
\Pr(G \mid H)\Pr(F \mid H \cap G) &\overset{\text{always}}{=} \Pr(F \cap G \mid H) &&\overset{\text{independence}}{=} \Pr(F \mid H)\Pr(G \mid H) \\
\Pr(G \mid H)\Pr(F \mid H \cap G) &= && \Pr(F \mid H)\Pr(G \mid H) \\
\Pr(F \mid H \cap G) &= && \Pr(F \mid H).
\end{aligned}$$

Thus, conditional independence implies that $\Pr(F \mid H \cap G) = \Pr(F \mid H)$. In other words, if we know H is true, and F and G are conditionally independent given H, then knowing G does not change our belief about F.

Let's consider the conditional dependence of F and G when H is assumed to be true in the following two situations:

Siutation 1:

- F = { a hospital patient is a smoker }
- G = { a hospital patient has lung cancer }
- H = { smoking causes lung cancer}

Situation 2:

- F = { a student studies regularly for an exam }
- G = { a student receives a high exam score }
- H = { studying improves exam performance }

Think: In both of these situations, H being true implies a relationship between F and G. What about when H is not true?

2.4 Random Variables

In Bayesian inference a random variable is defined as an unknown numerical quantity about which we make probability statements. For example, the quantitative outcome of a survey, experiment or study is a random variable before the study is performed. Additionally, a fixed but unknown population parameter is also a random variable

2.4.1 Discrete Ramdom variables

Let Y be a random variable and let $\mathcal{Y}$ be the set of all possible values that Y can take. If $\mathcal{Y}$ is countable, meaning that $\mathcal{Y} = \{y_1, y_2, ...\}$, then Y is a discrete random variable.

The event that the outcome Y of our survey has the value Y is expressed as $\{Y = y\}$. For each $y \in \mathcal{Y}$, the shorthand notation fofr $\Pr(Y = y)$ is $p(y)$, and $p(\cdot)$ is called the **probability mass function** of Y, and with two properties

1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$,
2. $\sum_{y \in \mathcal{Y}} p(y) = 1$.

General probability statements about Y can be derived from the pdf/pmf, for example, for any subset $A \subseteq \mathcal{Y}$, we have $\Pr(Y \in A) = \sum_{y \in A} p(y)$. When we have two disjoint subsets A and B of $\mathcal{Y}$, we have

$$\Pr(Y \in A \cup B) = \Pr(Y \in A) + \Pr(Y \in B) = \sum_{y \in A} p(y) + \sum_{y \in B} p(y).$$

Let Y be the number of successes in n independent Bernoulli trials, each with probability of success θ. Then, Y follows a Binomial distribution with parameters n and θ, denoted as $Y \sim \text{Binomial}(n, p)$. The probability mass function of Y is given by

$$p(y) = \Pr(Y = y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad y = 0, 1, 2, \ldots, n.$$

If $\theta = 0.3$ and $n = 3$, then the probability of observing exactly 2 successes is

$$p(2) = \Pr(Y = 2 \mid \theta = 0.3) = \binom{3}{2} (0.3)^2 (0.7)^1 = 3 \cdot 0.09 \cdot 0.7 = 0.189.$$

2.4.2 Continuous random variables

If $\mathcal{Y}$ is uncountable, for example, $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = (0, 1)$, then Y is a continuous random variable. In this case, we cannot list all possible values of Y and assign probabilities to each value. Instead, we use a probability distribution to describe the distribution of Y. That is, the cummulative distribution function (cdf) defined as follows.

The **cummulative distribution function** (cdf) of a continuous random variable Y is defined as

$$F(y) = \Pr(Y \leq y), \quad y \in \mathcal{Y}.$$

Note that, for the cdf $F(y)$, we have the following properties:

- $0 \leq F(y) \leq 1$ for all $y \in \mathcal{Y}$,
- $F(y)$ is non-decreasing, meaning that if $y_1 < y_2$, then $F(y_1) \leq F(y_2)$,
- $\lim_{y \to -\infty} F(y) = 0$
- $\lim_{y \to \infty} F(y) = 1$.

Probability of various events can be derived from the cdf. For example, for any interval $A = (a, b] \subseteq \mathcal{Y}$, we have

$$\Pr(Y \in A) = \Pr(a < Y \leq b) = F(b) - F(a).$$

Also, $\Pr(Y \leq a) = F(a)$ and $\Pr(Y > a) = 1 - F(a)$.

2.4.3 Description of distributions through quantiles and moments

In this subsection, we discuss a few ways to describe probability distributions: quantiles and moments. They are used to describe the behaviour of the distribution compressing them into summary statistics.

The **expectation** or **mean** of a random variable Y can be thought as the centre of mass or the location of the distribution, which is defined as

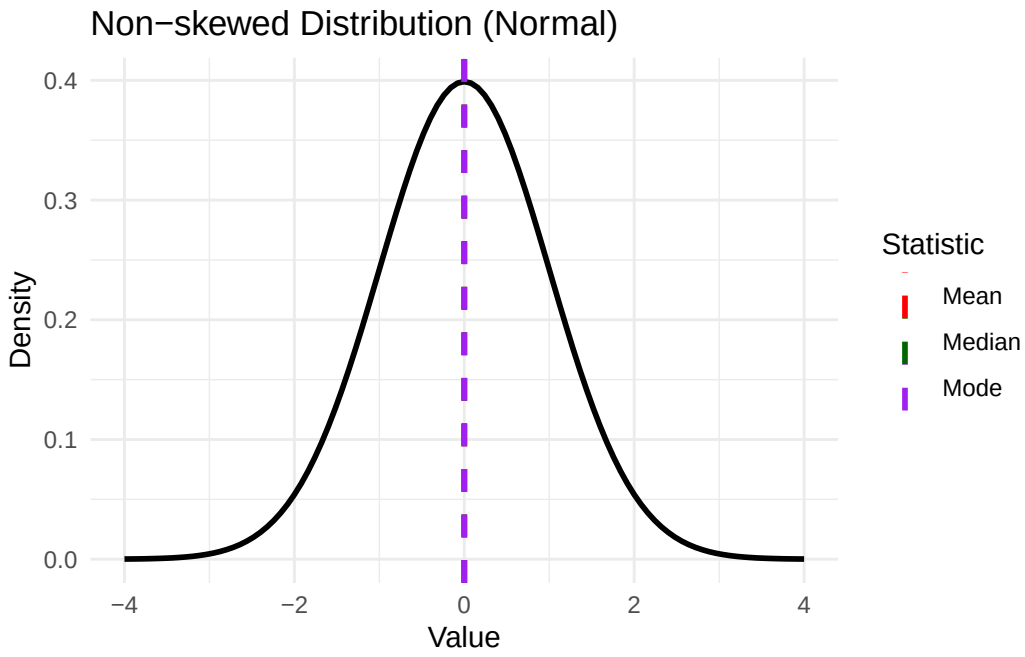
- For discrete random variable:
$$E(Y) = \sum_{y \in \mathcal{Y}} yp(y).$$
- For continuous random variable:
$$E(Y) = \int_{y} yf(y)dy.$$

> **i** Difference bewtwen mean, mode and median
>
> - Mean: the centre of mass of the distribution
> - Mode: The most probable value of Y
> - Median: The value of Y in the middle of the distribution.

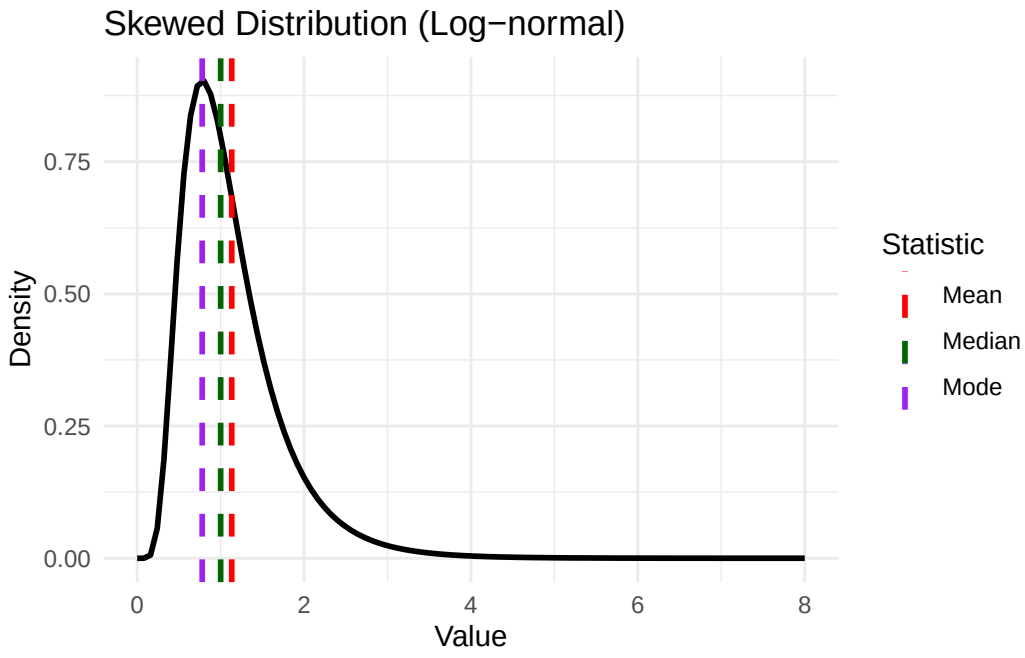
In skewed distribution, the three will not equal to each other.

```
library(ggplot2)

# ----------------------------
# Theoretical reference lines
# ----------------------------
lines_normal <- data.frame(
  value = c(0, 0, 0),
  Statistic = c("Mean", "Median", "Mode")
)

lines_lognormal <- data.frame(
  value = c(exp(1/8), 1, exp(-1/4)),
  Statistic = c("Mean", "Median", "Mode")
)

cols <- c("Mean" = "red", "Median" = "darkgreen", "Mode" = "purple")

# ----------------------------
# Normal distribution
```

```r
# ------------------------------
p1 <- ggplot() +
  stat_function(fun = dnorm, size = 1, color = "black") +
  geom_vline(
    data = lines_normal,
    aes(xintercept = value, color = Statistic),
    linetype = "dashed",
    size = 1
  ) +
  scale_color_manual(values = cols) +
  scale_x_continuous(limits = c(-4, 4)) +
  labs(
    title = "Non-skewed Distribution (Normal)",
    x = "Value", y = "Density", color = "Statistic"
  ) +
  theme_minimal()

# ------------------------------
# Log-normal distribution: LN(0, 0.5)
# ------------------------------
p2 <- ggplot() +
  stat_function(
    fun = function(x) dlnorm(x, meanlog = 0, sdlog = 0.5),
    size = 1,
    color = "black"
  ) +
  geom_vline(
    data = lines_lognormal,
    aes(xintercept = value, color = Statistic),
    linetype = "dashed",
    size = 1
  ) +
  scale_color_manual(values = cols) +
  scale_x_continuous(limits = c(0, 8)) +
  labs(
    title = "Skewed Distribution (Log-normal)",
    x = "Value", y = "Density", color = "Statistic"
  ) +
  theme_minimal()

p1
```


Non−skewed Distribution (Normal)
Density
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
Value
Statistic
Mean
Median
Mode

```
p2
```


Skewed Distribution (Log−normal)
Density
0.75
0.50
0.25
0.00
0
2
4
6
8
Value
Statistic
Mean
Median
Mode

Why use mean?

The mean is widely used in statistics and data analysis for several reasons:

1. Mathematical properties: The mean has desirable mathematical properties, such as linearity, which makes it easier to work with in various statistical analyses and models.
2. Sensitivity to all values: The mean takes into account all values in the dataset, providing a comprehensive measure of central tendency. It is also a scaled version of the total, which is often an interest
3. Foundation for other statistical measures: The mean serves as the basis for many other statistical measures, such as variance and standard deviation, which are essential for understanding the spread and variability of data.
4. Mean minimizes the sum of squared deviations: The mean is the value that minimizes the sum of squared deviations (i.e., the expected penalty by choosing one value) from itself, making it a natural choice for summarizing data.
5. May contains full information: In some distributions (e.g., bernoulli distribution), the mean contains all the information about the distribution, making it a sufficient statistic for inference.

The **variance** of a random variable Y measures the spread or dispersion of the distribution, and is defined as

$$\text{Var}(Y) = E\left[(Y - E(Y))^2\right] = E[Y^2] - E^2[Y].$$

The standard deviation is the square root of the variance, denoted as $\text{SD}(Y) = \sqrt{\text{Var}(Y)}$.

The **quantile** of order α of a random variable Y is defined as the value y_α such that

$$\Pr(Y \leq y_\alpha) = F(y_\alpha) = \alpha$$

for $0 < \alpha < 1$.

For example, the median is the quantile of order 0.5, denoted as $y_{0.5}$, which satisfies $\Pr(Y \leq y_{0.5}) = 0.5$. Also, $(y_{0.025}, y_{0.975})$ and $(y_{0.25}, y_{0.75})$ contains 95% and 50% of the mass of the distribution, respectively.

2.5 Joint Disitrubiton

2.5.1 Discrete random variables

Let Y_1 and Y_2 be two random variables with possible values in $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively. The **joint distribution** of Y_1 and Y_2 describes the probability of various combinations of values

that (Y_1, Y_2) can take.

Joint beliefs about Y_1 and Y_2 can be represented with probabilities. For example, for subsets $A \subset \mathcal{Y}_1$ and $B \subset \mathcal{Y}_2$, $\Pr(\{Y_1 \in A\} \cap \{Y_2 \in B\})$ represents our belief that Y_1 takes a value in A and Y_2 takes a value in B. The *joint pdf* or *joint density* of Y_1 and Y_2 is defined as

$$p_{Y_1Y_2}(y_1, y_2) = \Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}), \text{ for } y_1 \in \mathcal{Y}_1, y_2 \in \mathcal{Y}_2.$$

The *marginal density* of Y_1 can be computed from the joint density:

$$\begin{aligned} p_{Y_1}(y_1) &\equiv \Pr(Y_1 = y_1) \\ &= \sum_{y_2 \in \mathcal{Y}_2} \Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}) \\ &\equiv \sum_{y_2 \in \mathcal{Y}_2} p_{Y_1Y_2}(y_1, y_2) \end{aligned}$$

The *conditional density* of Y_2 given $\{Y_1 = y_1\}$ can be computed from the joint density and the marginal density:

$$\begin{aligned} p_{Y_2|Y_1}(y_2 \mid y_1) &= \frac{\Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})}{\Pr(Y_1 = y_1)} \\ &= \frac{p_{Y_1Y_2}(y_1, y_2)}{p_{Y_1}(y_1)}. \end{aligned}$$

You should be able to see that

- $\left\{p_{Y_1}, p_{Y_2|Y_1}\right\}$ can be derived from $p_{Y_1Y_2}$,
- $\left\{p_{Y_2}, p_{Y_1|Y_2}\right\}$ can be derived from $p_{Y_1Y_2}$
- $p_{Y_1Y_2}$ can be derived from $\left\{p_{Y_1}, p_{Y_2|Y_1}\right\}$
- $p_{Y_1Y_2}$ can be derived from $\left\{p_{Y_2}, p_{Y_1|Y_2}\right\}$

BUT

- $p_{Y_1Y_2}$ cannot be derived from $\left\{p_{Y_1}, p_{Y_2}\right\}$.

The subscripts of density functions are often dropped, in which case the type of density function is determined by the arguments. For example,

- $p(y_1, y_2) = p_{Y_1Y_2}(y_1, y_2)$ is the joint density of Y_1 and Y_2,
- $p(y_1) = p_{Y_1}(y_1)$ is the marginal density of Y_1
- $p(y_2 \mid y_1) = p_{Y_2|Y_1}(y_2 \mid y_1)$ is the conditional density of Y_2 given $\{Y_1 = y_1\}$, and so on.

Suppose a sociological study reports the following joint distribution of parents' education level and children's income level in a population.

Joint distribution of education and income Suppose a sociological study reports the following **joint distribution** of **parents' education level** and **children's income level** in a population as shown in the Table below

Parent \ Child	Low Income	Middle Income	High Income
High School or Less	0.18	0.22	0.10
College	0.08	0.20	0.12
Graduate School	0.04	0.06	0.10

Suppose we randomly sample a **parent–child pair** from this population.

Let
- Y_1 be the parent's education level
- Y_2 be the child's income level

We are interested in the conditional probability that the child has **high income**, given that the parent has a **college education**.

We may answer this question using the conditional probability formula:

$$\Pr(Y_2 = \text{High Income} \mid Y_1 = \text{College}) = \frac{\Pr(Y_2 = \text{High Income} \cap Y_1 = \text{College})}{\Pr(Y_1 = \text{College})}$$

From the table,

$$\Pr(Y_2 = \text{High Income} \cap Y_1 = \text{College}) = 0.12$$

$$\Pr(Y_1 = \text{College}) = 0.08 + 0.20 + 0.12 = 0.40$$

Therefore,

$$\Pr(Y_2 = \text{High Income} \mid Y_1 = \text{College}) = \frac{0.12}{0.40} = 0.30$$

Thus, our conclusion from the table is, among children whose parents have a college education, **30%** attain high income.

2.5.2 Continuous random variables

Let Y_1 and Y_2 be two continuous random variables with possible values in $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively. The **joint distribution** of Y_1 and Y_2 describes the probability of various combinations of values that (Y_1, Y_2) can take. We again work with the cumulative distribution function (cdf). The defintion is given as follows.

Given a continuous joint cdf $F_{Y_1Y_2}(y_1, y_2)$, there is a function p_{Y_1,Y_2} such that

$$F_{Y_1,Y_2}(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} p_{Y_1,Y_2}(y_1, y_2) dy_2 dy_1,$$

and $p_{Y_1,Y_2}(y_1, y_2)$ is called the *joint density function* of Y_1 and Y_2.

Similar to the discrete case, we can derive marginal and conditional densities from the joint density as

- Marginal density of Y_1:
$$p_{Y_1}(y_1) = \int_{y_2} p_{Y_1,Y_2}(y_1, y_2) dy_2,$$
- Conditional density of Y_2 given $\{Y_1 = y_1\}$:
$$p_{Y_2|Y_1}(y_2 \mid y_1) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_1}(y_1)}.$$

Think about why $p_{Y_2|Y_1}(y_2 \mid y_1)$ is an actual pdf.

2.5.3 Mixed continuous and discrete variables

It is possible to have joint distributions involving both discrete and continuous random variables. For example, let Y_1 be a discrete random variable taking values in $\mathcal{Y}_1$ and Y_2 be a continuous random variable taking values in $\mathcal{Y}_2$. The joint distribution of Y_1 and Y_2 can be described by the joint density function $p_{Y_1,Y_2}(y_1, y_2)$, which gives the probability that Y_1 takes the value y_1 and Y_2 takes a value in an infinitesimal interval around y_2. One such as example is that Y_1 is a binary variable indicating the presence or absence of a disease, and Y_2 is a continuous variable representing the severity of symptoms. Suppose we define

- Marginal density p_{Y_1} from our belief $\Pr(Y_1 = y_1)$
- a conditional density $p_{Y_2|Y_1}$ from $\Pr(Y_2 \leq y_2 \mid Y_1 = y_1) \doteq F_{Y_2|Y_1}(y_2 \mid y_1)$.

Then, the joint density can be derived as

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1)p_{Y_2|Y_1}(y_2 \mid y_1),$$

and the probability can be calculated as

$$\Pr(Y_1 \in A, Y_2 \in B) = \int_{y_2 \in B} \left\{ \sum_{y_1 \in A} p_{Y_1,Y_2}(y_1, y_2) \right\} dy_2.$$

2.5.4 Bayes' rule and parameter estimation

Let

- θ: proportion of people in a large population who have a certain charactersitic.
- Y: number of people in a small random sample from the population who have the charactersitic

Then, in this case, we may threat θ as continuous random variable taking values in $\Theta = (0, 1)$, and Y as a discrete random variable taking values in $\mathcal{Y} = \{0, 1, 2, ..., n\}$, where n is the sample size. *Bayesian estimation of the parameter θ derives from the calculate of $p(\theta \mid y)$ where y is the observed value of Y.* In Bayesian, this calculation first requires that we have a joint density $p(y, \theta)$ representing our belief about θ and the survey outcome Y. Often, it is natural to construct this joint density from

- $p(\theta)$: our prior belief about θ before seeing the data, and
- $p(y \mid \theta)$: belief about Y given θ, often called the likelihood function.

Once we observed $\{Y = y\}$, we need to compute our updated belief about θ, represented by the **posterior density** $p(\theta \mid y)$ as

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

If we have two values θ_1 and θ_2 in Θ that may be true, then the ratio of their posterior densities is given by

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \frac{p(y \mid \theta_1)p(\theta_1)/p(y)}{p(y \mid \theta_2)p(\theta_2)/p(y)} = \frac{p(y \mid \theta_1)p(\theta_1)}{p(y \mid \theta_2)p(\theta_2)}.$$

> **Note**
>
> From this calculation, we notice when we are calculating the relative posterior probability between two parameter values **we do not need** calculate $p(y)$ out.

Another way to think about this is, for a function of θ,

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

> **Note**
>
> We will see that the numerator is the important part, while the denominator is just a normalizing constant to make sure the posterior density integrates to 1.

2.6 Independence Random Variables

Let $Y_1, \dots, Y_n$ be random variables with joint density $p(y_1, \dots, y_n)$, and θ is the parameter describe the conditions under which the random variables are generated. We say that $Y_1, \dots, Y_n$ are conditionally independent given θ if every collection of n sets $\{A_1, \dots, A_n\}$ satisfies

$$\Pr(Y_1 \in A_1, \dots, Y_n \in A_n \mid \theta) = \prod_{i=1}^{n} \Pr(Y_i \in A_i \mid \theta).$$

If we have independence property, then

$$\Pr(Y_i \in A_i \mid \theta, Y_j \in A_j) = \Pr(Y_i \in A_i \mid \theta),$$

so the conditional indpenddence can be interpreted as meaning that Y_j gives no additional information about Y_i once we know θ. Also, under independence, the joint density can be factorized as

$$p(y_1, \dots, y_n \mid \theta) = \prod_{i=1}^{n} p_{Y_i}(y_i \mid \theta).$$

If the samples are also identically distributed, meaning that each Y_i has the same marginal density $p_Y(y \mid \theta)$, then the joint density can be further simplified as

$$p(y_1, \dots, y_n \mid \theta) = \prod_{i=1}^{n} p_Y(y_i \mid \theta).$$

In this case , we say that $Y_1, \dots, Y_n$ are **independent and identically distributed** (i.i.d.) given θ, with notation

$$Y_1, \dots, Y_n \mid \theta \overset{i.i.d.}{\sim} p_Y(y \mid \theta).$$

2.7 Exchangeability

A sequence of random variables $Y_1, Y_2, \dots, Y_n$ is **exchangeable** if for any permutation π of the indices $\{1, 2, \dots, n\}$, we have

$$p(y_1, y_2, \dots, y_n) = p(y_{\pi(1)}, y_{\pi(2)}, \dots, y_{\pi(n)}).$$

In other words, the joint density of an exchangeable sequence is invariant to the order of the random variables. That is, the labels contains no information about the outcome.

Suppose a factory produces a large batch of items. Each item may be either **defective** or **non-defective**.

Let

$$Y_i = \begin{cases} 1, & \text{if the } i\text{th inspected item is defective,} \\ 0, & \text{otherwise.} \end{cases}$$

We inspect $n = 10$ items chosen at random from the batch and record $Y_1, Y_2, \dots, Y_{10}$.

Consider the following three observed sequences:

1. $p(1, 0, 1, 0, 1, 0, 0, 1, 0, 1)$
2. $p(0, 1, 0, 1, 0, 1, 1, 0, 0, 1)$
3. $p(1, 1, 0, 0, 1, 0, 1, 0, 0, 1)$

Each sequence contains **5 defective items** and **5 non-defective items**.

Question: Is there a reason to assign these three sequences *different probabilities*?

If the inspection order conveys no additional information about quality, then *only the number of defective items matters*, not their positions in the sequence. This motivates the concept of exchangeability.

2.7.1 Independence versus dependence

Consider the probability assignments

$$\begin{cases} \Pr(Y_{10} = 1) = a, \\ \Pr(Y_{10} = 1 \mid Y_1 = \cdots = Y_9 = 1) = b. \end{cases}$$

If $a \neq b$, then Y_{10} is **not independent** of $Y_1, \dots, Y_9$.

However, lack of independence does **not** imply lack of exchangeability.

Question: should we have $a = b$, $a > b$ or $a < b$?

2.7.2 A latent-parameter model

Suppose the defect rate θ of the factory is unknown.

Conditional on θ,

$$Y_1, \ldots, Y_{10} \mid \theta \sim \text{i.i.d. Bernoulli}(\theta).$$

Then

$$\Pr(Y_1 = y_1, \ldots, Y_{10} = y_{10} \mid \theta) = \theta^{\sum y_i}(1-\theta)^{10-\sum y_i}.$$

If our uncertainty about θ is described by a prior distribution $p(\theta)$, the marginal joint distribution is

$$p(y_1, \ldots, y_{10}) = \int \theta^{\sum y_i}(1-\theta)^{10-\sum y_i} p(\theta)\, d\theta.$$

This probability depends **only on the number of defective items**, not their order.

Thus, we have exchangeability, even though the Y_i are not independent under this model of belief.

Conditional i.i.d. given a latent parameter implies marginal exchangeability. That is, if $\theta \sim p(\theta)$ and $Y_1, \ldots, Y_n$ are conditionally i.i.d. given θ, then $Y_1, \ldots, Y_n$ (i.e., unconditional on θ) are exchangeable.

For the Proof, see page 28 in Hopf (2009).

2.8 de Finetti's Theorem

As of now, we have seen that conditional i.i.d. given a latent parameter implies marginal exchangeability. For example,

$$\begin{cases} Y_1, \ldots, Y_n \mid \theta \overset{i.i.d.}{\sim} \\ \theta \sim p(\theta) \end{cases} \implies Y_1, \ldots, Y_n \text{ are exchangeable.}$$

The converse is also true, as stated in de Finetti's theorem.

Let $Y_i \in \mathcal{Y}$ for all $i \in \{1, 2, \dots, n\}$ be an exchangeable sequence of random variables. Then, there exists a parameter space Θ and a prior distribution $p(\theta)$ on Θ such that the joint distribution of $Y_1, \dots, Y_n$ can be represented as

$$p(y_1, \dots, y_n) = \int_\Theta \left\{ \prod_{i=1}^{n} p_Y(y_i \mid \theta) \right\} p(\theta) d\theta,$$

where $p_Y(y \mid \theta)$ is a probability density function on $\mathcal{Y}$ for each $\theta \in \Theta$. The prior and sampling model depend on the form of the belief model $p(y_1, \dots, y_n)$.

The probability distribution $p(\theta)$ represents our belief about the outcomes $\{Y_1, Y_2, \dots, Y_n\}$, induced by our belief model $p(y_1, \dots, y_n)$. That is,

- $p(\theta)$ represents our belief about $\lim_{n\to\infty} \sum Y_i/n$ in the binary sense
- $p(\theta)$ represents our belief about $\lim_{n\to\infty} \sum (Y_i \le c)/n$ for each c in the general case.

The main idea ofthis and the previous section is as follows

$$\begin{aligned} Y_1, \dots, Y_n \mid \theta &\overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta), \\ \theta &\sim p(\theta) \end{aligned} \quad \Longleftrightarrow \quad Y_1, \dots, Y_n \text{ are exchangeable for all } n.$$

Question: When is the condition of "exchangeability for all n" reasonable?

- Have exchaneability and repeatability
  - Exchangeability holds if the labels convey no information
  - repeatability hold includes the follows
    1. $Y_1, \dots, Y_n$ are outcomes of a repeartable experiment
    2. $Y_1, \dots, Y_n$ are sampled from a finite population **with repleacement**
    3. $Y_1, \dots, Y_n$ are sampled from an infinite population without replacement.

> **In large finite population**
>
> Note, if $Y_1, \dots, Y_n$ are exchangeable and sampled from a finite population of size N that is way bigger than n without replacement, then they can be modelled as *approximate* being conditional i.i.d.

This Chapter follows closely with Chapter 2 in Hoff (2009).

3 Bayesian Inference for single parameter models

Leading objectives:

Understand how to perform Bayesian inference on a single parameter model.

- Binomial model with given n
- Poission model
- Exponential family

3.1 Binomial model with given n

3.2 Prior

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)

3.2.1 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

This Chapter follows closely with Chapter 3 in Hoff (2009).

4 Week 1 — Introduction and Bayesian Thinking

4.1 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.

4.2 Foundational Concepts

4.2.1 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.

4.3 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.

4.3.1 Motivating Examples

4.3.1.1 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: $\widehat{\theta} = y/n = 8/20 = 0.4$ - 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

4.3.1.2 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.

4.3.2 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

4.3.3 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

4.3.4 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 $\int p(\theta \mid y)\, d\theta = 1$.

4.3.5 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 $\tilde{y}$:

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

4.4 One-Parameter Models

4.4.1 The Beta-Binomial Model

4.4.1.1 Setup

Consider binary outcome data: $y_1, \ldots, y_n$ 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:

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

4.4.1.2 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.

4.4.1.3 Posterior Distribution

By Bayes' theorem:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta) \times p(\theta) && (4.1)\\ &\propto \theta^y(1-\theta)^{n-y} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} && (4.2)\\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} && (4.3) \end{aligned}$$

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 $\frac{y}{n}$.

4.4.1.4 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.

4.4.2 The Normal Model with Known Variance

4.4.2.1 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:

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

4.4.2.2 Prior Distribution

We use a **Normal prior** for θ:

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

with density:

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

4.4.2.3 Posterior Distribution

By Bayes' theorem:

$$p(\theta \mid y) \propto p(y \mid \theta) \times p(\theta) \tag{4.4}$$

$$\propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\} \times \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\} \tag{4.5}$$

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.

4.4.2.4 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

---

4.4.3 Posterior Predictive Distribution

After observing data $y = (y_1, \ldots, y_n)$, 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.

4.4.3.1 Example: Beta-Binomial

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

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

4.4.3.2 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).

4.4.4 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

---

4.4.5 Practical Considerations

4.4.5.1 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)

4.4.5.2 Sensitivity Analysis

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

4.4.5.3 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

---

4.4.6 R Examples

4.4.6.1 Example 3.1: Beta-Binomial Model

```
# 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))
```

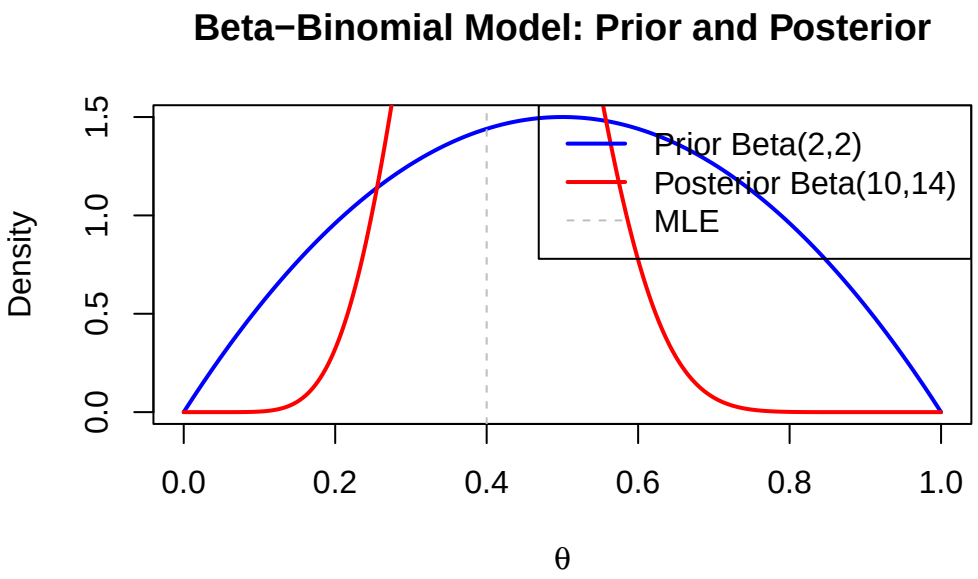


Figure 4.1: Prior, Likelihood, and Posterior for Beta-Binomial Model

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

```
Posterior mean: 0.4166667
```

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

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

4.4.6.2 Example 3.2: Normal Model

```r
# 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 )
```

```r
# 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))
```

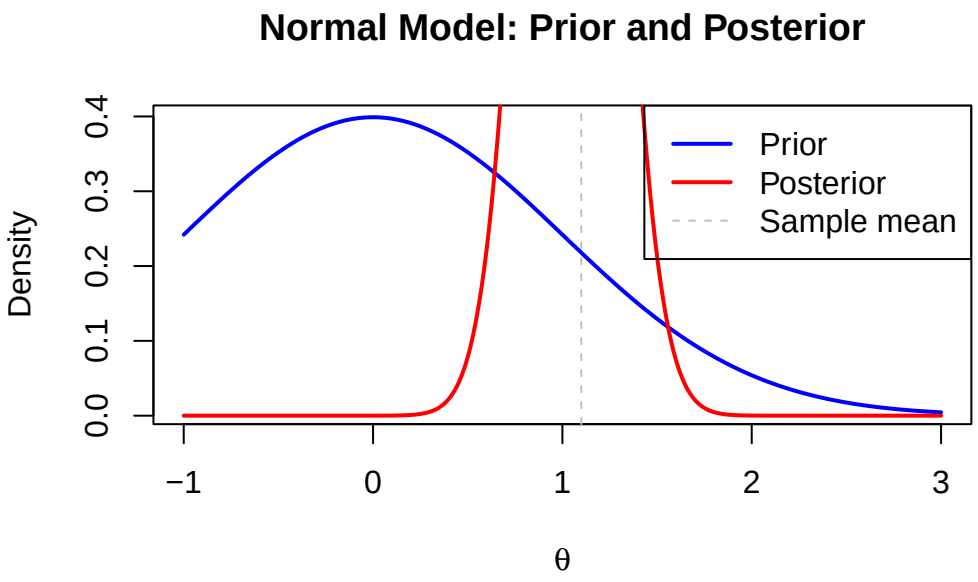


Figure 4.2: Prior, Likelihood, and Posterior for Normal Model

5 Week 2 — Conjugate Priors and Analytical Posteriors

5.1 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.

5.2 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.

5.3 Lecture 1: The Concept of Conjugacy

5.3.1 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.

5.3.2 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.

5.3.3 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

---

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

5.4.1 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.

5.4.2 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:

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

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:

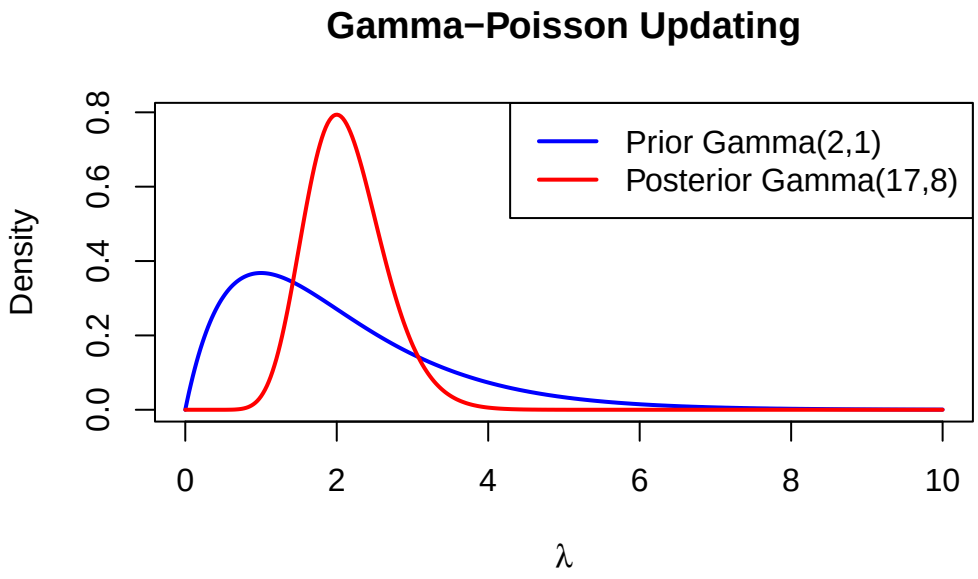
$$p(\tilde{y} \mid y) = \int \text{Poisson}(\tilde{y} \mid \lambda)\, p(\lambda \mid y)\, d\lambda,$$

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.

5.4.3 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)
```



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

---

6.1 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.

6.2 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.

6.3 Lecture 1: Motivation and Fundamentals of Monte Carlo

6.3.1 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.

6.3.2 1.2 Monte Carlo Idea

If we can sample $\theta^{(1)}, \ldots, \theta^{(M)}$ 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:

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

6.3.3 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.

6.3.4 1.4 Simple Example

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

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

```
[1] 0.2861808
```

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

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

7 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

Part I

Appendix

8 Appendix: Introduction to R

8.1 R

For conducting analyses with data sets of hundreds to thousands of observations, calculating by hand is not feasible and you will need a statistical software. **R** is one of those. **R** can also be thought of as a high-level programming language. In fact, **R** is one of the top languages to be used by data analysts and data scientists. There are a lot of analysis packages in **R** that are currently developed and maintained by researchers around the world to deal with different data problems. Most importantly, **R** is free! In this section, we will learn how to use **R** to conduct basic statistical analyses.

8.2 IDE

8.2.1 Rstudio

RStudio is an integrated development environment (IDE) designed specifically for working with the **R** programming language. It provides a user-friendly interface that includes a source editor, console, environment pane, and tools for plotting, debugging, version control, and package management. RStudio supports both **R** and Python and is widely used for data analysis, statistical modeling, and reproducible research. It also integrates seamlessly with tools like **R** Markdown, Shiny, and Quarto, making it popular among data scientists, statisticians, and educators.

8.2.2 Visual Studio Code (VS Code)

VS Code is a versatile code editor that supports multiple programming languages, including **R**. With the **R** extension for VS Code, users can write and execute **R** code, access **R**'s console, and utilize features like syntax highlighting, code completion, and debugging. While not as specialized as RStudio for **R** development, VS Code offers a lightweight alternative with extensive customization options and support for various programming tasks.

8.2.3 Positron

Positron IDE is the next-generation integrated development environment developed by Posit, the company behind RStudio. Designed to be a modern, extensible, and language-agnostic IDE, Positron builds on the strengths of RStudio while supporting a broader range of languages and workflows, including **R**, Python, and Quarto.

8.3 RStudio Layout

RStudio consists of several panes: - **Source**: Where you write scripts and markdown documents. - **Console**: Where you type and execute **R** commands. - **Environment/History**: Shows your variables and command history. - **Files/Plots/Packages/Help/Viewer**: For file management, viewing plots, managing packages, accessing help, and viewing web content.

8.4 R Scripts

R scripts are plain text files containing **R** code. You can create a new script in RStudio by clicking `File > New File > R Script`.

8.5 R Help

Use `?function_name` or `help(function_name)` to access help for any **R** function. For example:

```
?mean
help(mean)
```

8.6 R Packages

Packages extend **R**'s functionality. There are thousands of packages available in **R** ecosystem. You may install them from different sources.

8.6.1 With Comprehensive R Archive Network (CRAN)

CRAN is the primary repository for **R** packages. It contains thousands of packages that can be easily installed and updated.

Install a package with:

```
install.packages("package_name")
```

8.6.2 With Bioconductor

Bioconductor is a repository for bioinformatics packages in **R**. It provides tools for the analysis and comprehension of high-throughput genomic data.

Install Bioconductor packages using the `BiocManager` package:

```
BiocManager::install("package_name")
```

8.6.3 From GitHub

Many of the authors of **R** packages host their work on GitHub. You can install these packages using the `devtools` package:

```
devtools::install_github("username/package_name")
```

8.6.4 Load a package

Once a package is installed, you need to load it into your **R** session to use its functions:

```
library(package_name)
```

Alternatively, you may use a function in the package with `package_name::function_name()` without loading the entire package.

8.7 R Markdown

R Markdown allows you to combine text, code, and output in a single document. Create a new **R** Markdown file in RStudio via `File > New File > R Markdown...`.

Recently, the posit team has developed a new version of the **R** Markdown called quarto document, with the file extension `.qmd`. It is still under rapid development.

8.8 Vectors

Vectors are the most basic data structure in **R**.

```
x <- c(1, 2, 3, 4, 5)
x
```

```
[1] 1 2 3 4 5
```

You can perform operations on vectors:

```
x * 2
```

```
[1]  2  4  6  8 10
```

8.9 Data Sets

Data frames are used for storing data tables. Create a data frame:

```
df <- data.frame(Name = c("Alice", "Bob"), Score = c(90, 85))
df
```

```
   Name Score
1 Alice    90
2   Bob    85
```

You can import data from files using `read.csv()` or `read.table()`.

This appendix is adapted from Why R?.

References