STAT8310 - Bayesian Data Analysis

Chi-Kuang Yeh

2026-02-14

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
☐ A2, due on Feb 13, 2026

Midterm

☐ March 3, 2026

Topics and Corresponding Lectures

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

Status	Chapter	Topic	Lecture
	Ch. 1	Welcome and Overview	1
	—	Intro to R Programming	2
	Ch. 2	Probability and Exchangeability	3–5
	Ch. 3	One Parameter Models	6–10
	Ch. 4	Monte Carlo	11–

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 Ramdon 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 for $\Pr(Y = y)$ 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 **cumulative 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:
$$E(Y) = \int_{y} yf(y)dy.$$

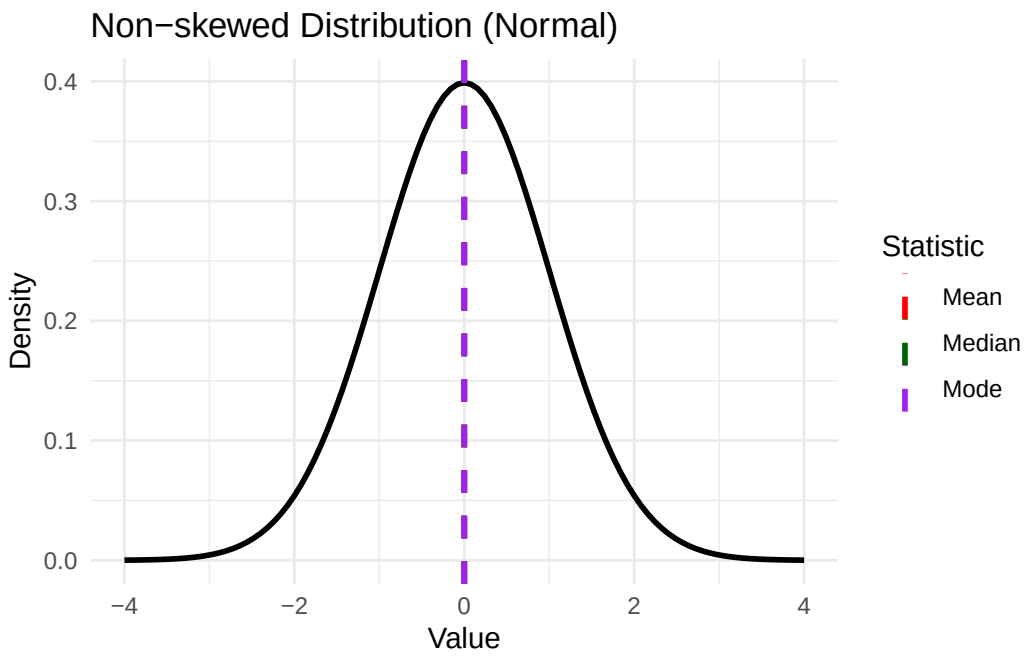
> **Difference between mean, mode and median**
>
> - Mean: the centre of mass of the distribution
> - Mode: The most probable value of Y
> - Median: The value of Y in the middle of the distribution.

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

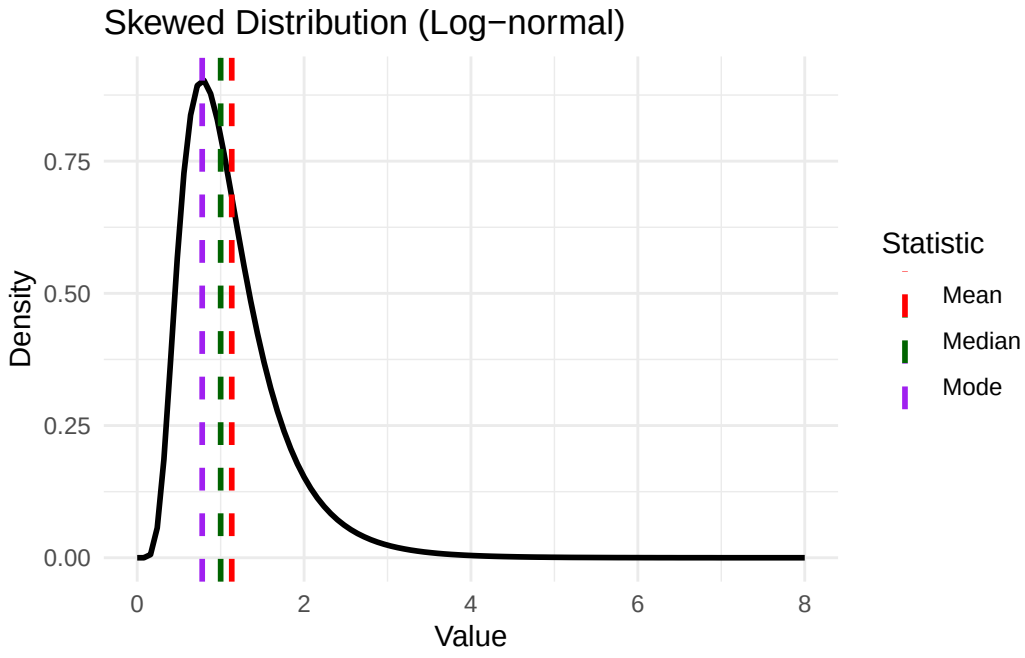
```
library(ggplot2)

# ----------------------------
# 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")
```

```r
# ----------------------------
# Normal distribution
# ----------------------------
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 definition 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 $Y_1, Y_2, \ldots, Y_n$ is **exchangeable** if for any permutation π of the indices $\{1, 2, \ldots, n\}$, we have

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

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 $Y_1, Y_2, \ldots, Y_{10}$.

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 $Y_1, \ldots, Y_9$.

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,

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

Then

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

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

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

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 $Y_1, \dots, Y_n$ are conditionally i.i.d. given θ, then $Y_1, \dots, Y_n$ (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,

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

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 of this and the previous section is as follows

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

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 replacement**
    3. $Y_1, \dots, Y_n$ are sampled from an infinite population without replacement.

> **i** 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

Recall the important ingredients of Bayesian inference:

1. **Prior distribution:** $\pi(\theta)$
2. **Likelihood function:** $p(y \mid \theta)$
3. **Posterior distribution:** $p(\theta \mid y) \propto p(y \mid \theta)\pi(\theta)$

3.1 Three basic ingredients of Bayesian inference

3.1.1 Prior

The prior distribution encodes our beliefs about the parameter θ *before* conduct any experiments.

> **Prior and Data are independent**
>
> Note that, the prior distribution is independent of the data. It represents our knowledge or beliefs about the parameter before seeing the data.

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)

3.1.2 Likelihood

The likelihood function represents the probability of observing the data given the parameter θ. It can be derived from the assumed statistical model for the data or experiment, i.e., $y \sim p(y \mid \theta)$, or we can estimate this non-parametrically (i.e., without assuming the underlying distribution is the one we know.).

> **Likelihood is NOT a probability distribution for θ**
>
> Note that, the likelihood function is not a probability distribution for θ itself. It is a function of θ for fixed data y.

3.1.3 Posterior

The posterior distribution combines the prior and likelihood to update our beliefs about θ after observing the data. It is given by Bayes' theorem:

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

where $p(y) = \int p(y \mid \theta)\pi(\theta)d\theta$ is the marginal likelihood or evidence.

3.1.4 An simple example

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

Let's look a simple example to illustrate the convenience of conjugate priors. Consider a Binomial model with unknown success probability θ and known number of trials n. We can use a Beta prior for θ.

Suppose we have a Binomial model with known number of trials n and unknown success probability θ. We can use a Beta prior for θ.

- **Prior:** $\theta \sim \text{Beta}(\alpha, \beta)$
- **Likelihood:** $y \mid \theta \sim \text{Binomial}(n, \theta)$

The derivation of the posterior is as follows:

$$p(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y},$$
$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(\alpha, \beta)$ is the Beta function. Then the posterior is proportional to:

$$p(\theta \mid y) \propto p(y \mid \theta)\pi(\theta) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}.$$

This is the kernel of a Beta distribution with parameters $(\alpha + y, \beta + n - y)$. Thus, the posterior distribution is:

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

Thus, the **Posterior** is $\theta \mid y \sim \text{Beta}(\alpha + y, \beta + n - y)$.

3.2 Happiness Data – the first example of Bayesian inference procedure

We study Bayesian inference for a binomial proportion θ when the sample size n is fixed. In this example, we want to see what is the procedure of doing Bayesian inference

In the 1998 General Social Survey, each female respondent aged 65 or over was asked whether she was generally happy.

Define the response variable

$$Y_i = \begin{cases} 1, & \text{if respondent } i \text{ reports being generally happy,} \\ 0, & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n,$$

where $n = 129$.

Because we lack information that distinguishes individuals, it is reasonable to treat the responses as **exchangeable**.
That is, before observing the data, the labels or ordering of respondents carry no information.

Since the sample size n is small relative to the population size N of senior women, results from the previous chapter justify the following modeling approximation.

Modeling Assumptions: Our beliefs about $(Y_1, \ldots, Y_{129})$ are described by:

- **An unknown population proportion**

$$\theta = \frac{1}{N}\sum_{i=1}^{N} Y_i,$$

where θ represents the proportion of generally happy individuals in the population.

- **A sampling model given θ**

Conditional on θ, the responses $Y_1, \ldots, Y_{129}$ are independent and identically distributed Bernoulli random variables with

$$\Pr(Y_i = 1 \mid \theta) = \theta.$$

Given the population proportion θ, each respondent independently reports being happy with probability θ.

Likelihood: Under this model, the probability of observing data $\{y_1, \ldots, y_{129}\}$ given θ is

$$p(y_1, \ldots, y_{129} \mid \theta) = \theta^{\sum_{i=1}^{129} y_i}(1-\theta)^{129-\sum_{i=1}^{129} y_i}.$$

This expression depends on the data only through the sufficient statistic

$$S = \sum_{i=1}^{129} Y_i,$$

the total number of respondents who report being generally happy.

For the happiness data,

$$S = 118,$$

so the likelihood simplifies to

$$p(y_1, \ldots, y_{129} \mid \theta) = \theta^{118}(1-\theta)^{11}.$$

Q: Which prior to be used?

A prior distribution is **conjugate** to a likelihood if the posterior distribution belongs to the same family as the prior. For the binomial likelihood, the **Beta distribution** is conjugate. But we have another choice of prior, to use *non-informative prior*.

A Uniform Prior Distribution: Suppose our prior information about θ is very weak, in the sense that all subintervals of $[0, 1]$ with equal length are equally plausible.
Symbolically, for any $0 \le a < b < b + c \le 1$,

$$\Pr(a \le \theta \le b) = \Pr(a + c \le \theta \le b + c).$$

This implies a **uniform prior**:

$$\pi(\theta) = 1, \qquad 0 \leq \theta \leq 1.$$

Posterior Distribution: Bayes' rule gives

$$p(\theta \mid y_1, \dots, y_{129}) = \frac{p(y_1, \dots, y_{129} \mid \theta)\, \pi(\theta)}{p(y_1, \dots, y_{129})}.$$

With a uniform prior, this reduces to

$$p(\theta \mid y_1, \dots, y_{129}) \propto \theta^{118}(1-\theta)^{11}.$$

Key idea: with a uniform prior, the posterior has the **same shape** as the likelihood.

To obtain a proper probability distribution, we must normalize.

Normalizing Constant and the Beta Distribution: Using the identity

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

we find

$$p(y_1, \dots, y_{129}) = \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}.$$

Therefore, the posterior density is

$$p(\theta \mid y_1, \dots, y_{129}) = \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)}\theta^{119-1}(1-\theta)^{12-1}.$$

That is,

$$\theta \mid y \sim \text{Beta}(119,\, 12).$$

Recall that, a random variable $\theta \sim \text{Beta}(a, b)$ distribution if

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

For $\theta \sim \text{Beta}(a, b)$, the expectation (i.e., mean or the first moment) is $\mathbb{E}(\theta) = \frac{a}{a+b}$, and the variance is $\text{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$.

In our example, the happiness data, the posterior distribution is

$$\theta \mid y \sim \text{Beta}(119, 12).$$

Thus, the posterior mean is $\mathbb{E}(\theta \mid y) = 0.915$, and the posterior standard deviation is $\text{sd}(\theta \mid y) = 0.025$.

These summaries quantify both our **best estimate** of the population proportion and our **remaining uncertainty** after observing the data.

3.2.1 Inference about exchangeable binary data

Posterior Inference under a Uniform Prior

Suppose $Y_1, \ldots, Y_n \mid \theta \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, and we place a uniform prior on θ. The posterior distribution of θ given the observed data $y_1, \ldots, y_n$ is proportional to

$$\begin{aligned} p(\theta \mid y_1, \ldots, y_n) &= \frac{p(y_1, \ldots, y_n \mid \theta)\pi(\theta)}{p(y_1, \ldots, y_n)} \\ &= \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i} \times \frac{\pi(\theta)}{p(y_1, \ldots, y_n)} \\ &\propto \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}. \end{aligned}$$

Consider two parameter values θ_a and θ_b. The ratio of their posterior densities is

$$\begin{aligned} \frac{p(\theta_a \mid y_1, \ldots, y_n)}{p(\theta_b \mid y_1, \ldots, y_n)} &= \frac{\theta_a^{\sum y_i}\left(1-\theta_a\right)^{n-\sum y_i} \times p\left(\theta_a\right) / p\left(y_1, \ldots, y_n\right)}{\theta_b^{\sum y_i}\left(1-\theta_b\right)^{n-\sum y_i} \times p\left(\theta_b\right) / p\left(y_1, \ldots, y_n\right)} \\ &= \left(\frac{\theta_a}{\theta_b}\right)^{\sum_i y_i}\left(\frac{1-\theta_a}{1-\theta_b}\right)^{n-\sum_i y_i} \frac{p(\theta_a)}{p(\theta_b)}. \end{aligned}$$

This expression shows that the data affect the posterior distribution **only through the sum of the data** $\sum_{i=1}^n y_i$ based on the relative probability density at θ_a to θ_b.

As a result, for any set A, one can show that

$$\Pr(\theta \in A \mid Y_1 = y_1, \ldots, Y_n = y_n) = \Pr\left(\theta \in A \mid \sum_{i=1}^n Y_i = \sum_{i=1}^n y_i\right).$$

This means that $\sum_{i=1}^n Y_i$ contains **all the information** in the data relevant for inference about θ. We therefore say that $Y = \sum_{i=1}^n Y_i$ is a **sufficient statistic** for θ. The term *sufficient* is used because knowing $\sum_{i=1}^n Y_i$ is sufficient to carry out inference about θ; no additional information from the individual observations $Y_1, \ldots, Y_n$ is required.

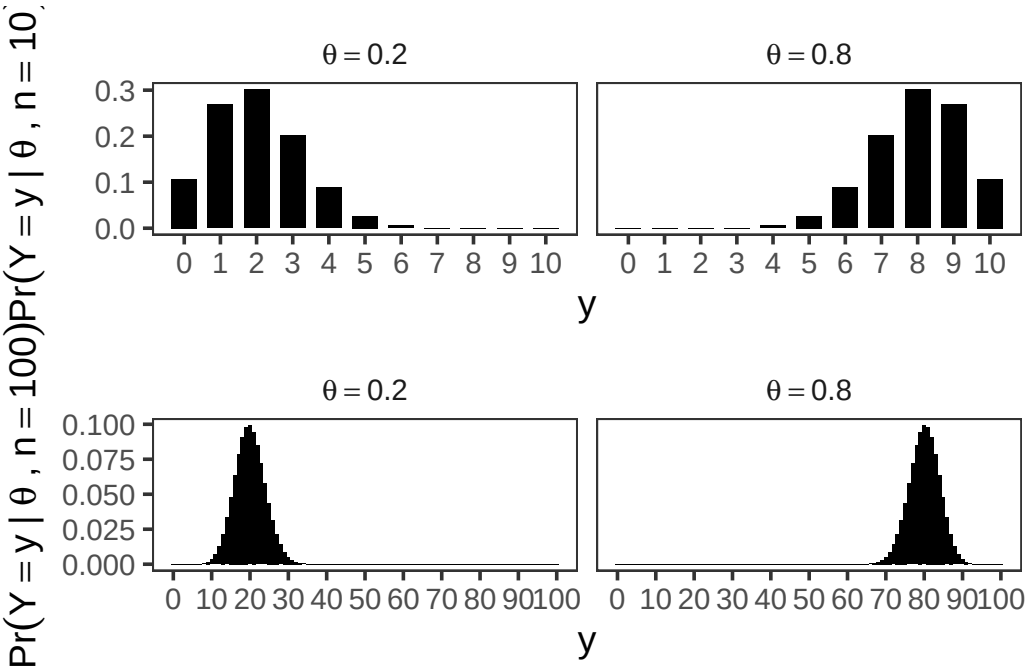
In the case where $Y_1, \ldots, Y_n \mid \theta$ are i.i.d. Bernoulli(θ) random variables, the sufficient statistic $Y = \sum_{i=1}^n Y_i$ follows a **binomial distribution** with parameters (n, θ).

The Binomial Model

Because each Y_i is Bernoulli(θ) and the observations are independent, the sufficient statistic $Y = \sum_{i=1}^n Y_i$ follows a **binomial distribution** with parameters (n, θ).

That is, $\Pr(Y = y \mid \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$, $y = 0, 1, \ldots, n$. For a binomial(n, θ) random variable Y,

- $\mathbb{E}[Y \mid \theta] = n\theta$,
- $\mathrm{Var}(Y \mid \theta) = n\theta(1-\theta)$.



Posterior inference under a uniform prior distribution

Having observed $Y = y$ our task is to obtain the posterior distribution of θ. By Bayes' theorem,

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

For a binomial model with $Y \sim \text{Binomial}(n, \theta)$, the likelihood is

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

Therefore,

$$p(\theta \mid y) = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}\pi(\theta)}{p(y)} = c(y)\theta^y(1-\theta)^{n-y}\pi(\theta),$$

where $c(y)$ is a normalizing constant that depends only on y, not on θ. When using the uniform

distribution, $\pi(\theta)$, we can calculate $c(y)$ easily as

$$\begin{aligned} 1 &= \int_0^1 c(y)\theta^y(1-\theta)^{n-y}d\theta \\ &= c(y)\int_0^1 \theta^y(1-\theta)^{n-y}d\theta \quad . \\ &= c(y)\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} \end{aligned}$$

Hence, $c(y) = \Gamma(n+2)/\{\Gamma(y+1)\Gamma(n-y+1)\}$, and the posterior distribution is

$$\begin{aligned} p(\theta \mid y) &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y} \\ &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{(y+1)-1}(1-\theta)^{(n-y+1)-1}, \end{aligned}$$

Which is exactly the beta$(y+1, n-y+1)$. In the happiness example, we have $n = 129$ and $Y = \sum Y_i = 118$, so the posterior distribution is beta$(119, 12)$, written as

$$n = 129, Y \equiv \sum Y_i = 118 \quad \Rightarrow \quad \theta \mid \{Y = 118\} \sim \text{beta}(119, 12).$$

This confirms the sufficiency result for this model and prior distribution, by showing that if $\sum y_i = y = 118$, $p(\theta \mid y_1, \ldots y_n) = p(\theta \mid y) = \text{beta}(119, 12)$. That is, the information contained in $\{Y_1 = y_1, \ldots, Y_n = y_n\}$ is the same as the information contained in $\{Y = y\}$, where $Y = \sum Y_i$ and $y = \sum y_i$. This show the posterior when we use **uniform prior**. One may ask, what if we use a different prior?

Posterior distributions under beta prior distributions

The uniform prior distribution has $\pi(\theta) = 1$ for all $\theta \in [0, 1]$. This distribution can be thought of as a beta prior distribution with parameters $a = 1, b = 1$

$$\pi(\theta) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{1\times 1}1 \times 1 = 1$$

for all $\theta \in [0, 1]$.

The gamma function is defined as

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, \quad x > 0.$$

It satisfies the following properties:

- $\Gamma(n) = (n-1)!$ for any positive integer n.
- $\Gamma(x+1) = x\Gamma(x)$ for any $x > 0$.

- $\Gamma(1/2) = \sqrt{\pi}$.
- $\Gamma(1) = 1$ by convention.

Now, from the previous part, recall that we have,

$$\text{if } \left\{ \begin{array}{c} \theta \sim \text{beta}(1,1) \text{ (uniform)} \\ Y \sim \text{binomial}(n, \theta) \end{array} \right\}, \text{ then } \{\theta \mid Y = y\} \sim \text{beta}(1 + y, 1 + n - y).$$

To get the posterior distribution under a general beta prior distribution, we just need to add the number of 1's to the α parameter and the number of 0's to the β parameter. To see this, assume $\theta \sim \text{beta}(\alpha, \beta)$, and $Y \mid \theta \sim \text{binomial}(n, \theta)$. Then, once we observed $\{Y = y\}$, by Bayes' theorem, the posterior distribution is

$$\begin{aligned} p(\theta \mid y) &= \frac{\pi(\theta) p(y \mid \theta)}{p(y)} \\ &= \frac{1}{p(y)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \times \binom{n}{y} \theta^{y}(1-\theta)^{n-y} \\ &= c(n, y, a, b) \times \theta^{a+y-1}(1-\theta)^{b+n-y-1} \\ &\propto \beta(a + y, b + n - y). \end{aligned}$$

> **One-to-one correspondence between the distribution**
>
> Note that, there is a one-to-one correspondence between the prior distribution parameters and the posterior distribution parameters. Two distributions are said to be the same if
>
> - Their CDFs are the same.
> - Their PDFs are the same.
> - All of their moments are the same. This implies that they are equal if and only if the moment generating function or the probability generating functions are the same.

We have seen the beta-binomial example twice, which is an example of **conjugate prior**, let's definite this formally,

A class $\mathcal{P}$ of prior distribution for θ is said **conjugate** for the likelihood function $p(y \mid \theta)$ if for every prior distribution $\pi(\theta) \in \mathcal{P}$, the corresponding posterior distribution $p(\theta \mid y)$ is also in $\mathcal{P}$, that is

$$\pi(\theta) \in \mathcal{P} \Rightarrow p(\theta \mid y) \in \mathcal{P}.$$

> **Note**
>
> Conjugate priors simplify posterior calculations, but they may not accurately reflect genuine prior beliefs. Still, mixtures of conjugate priors offer substantially greater flexibility while remaining computationally tractable.

If the likelihood $\theta \mid \{Y = y\} \sim beta(a + y, b + n - y)$, recall that

- $\mathrm{E}[\theta \mid y] = \frac{a+y}{a+b+n}$
- $\mathrm{mode}[\theta \mid y] = \frac{a+y-1}{a+b+n-2}$
- $\mathrm{Var}[\theta \mid y] = \frac{\mathrm{E}[\theta|y]\mathrm{E}[1-\theta|y]}{a+b+n+1}$

The posterior mean can be expressed as a weighted average of the prior mean and the maximum likelihood estimate (MLE) of θ:

$$\begin{aligned}
\mathrm{E}[\theta \mid y] &= \frac{a+y}{a+b+n} \\
&= \frac{a+b}{a+b+n} \times \frac{a}{a+b} + \frac{n}{a+b+n} \times \frac{y}{n} \\
&= \frac{a+b}{a+b+n} \times \text{ prior expectation } + \frac{n}{a+b+n} \times \text{ data mean}
\end{aligned}$$

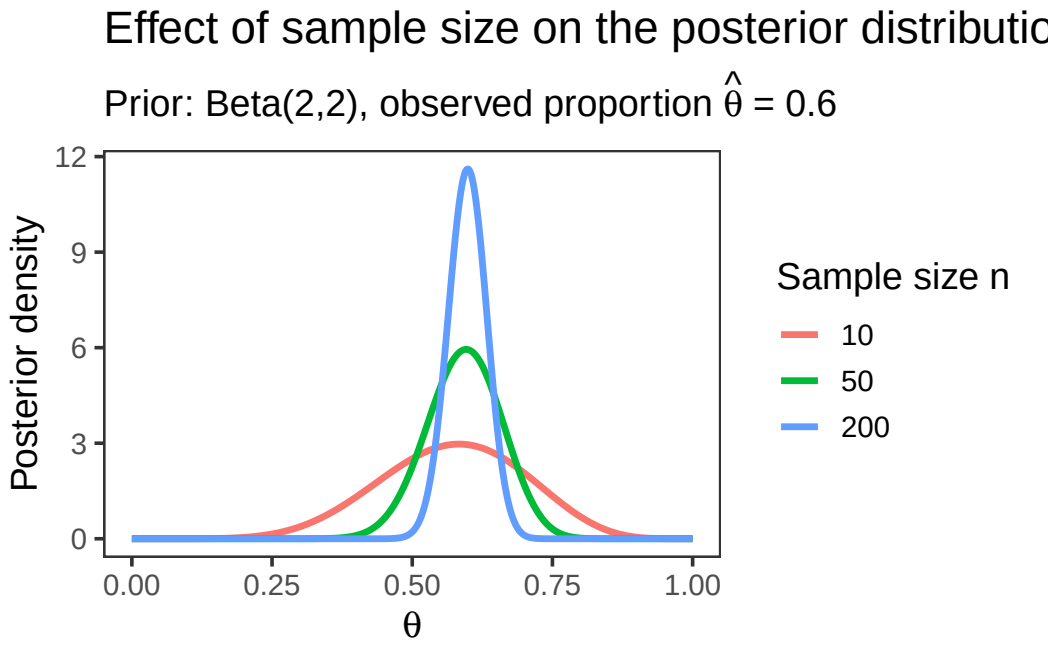
For this model and prior distribution, the posterior expectation (also known as the posterior mean) can be expressed as a weighted average of the prior expectation and the sample mean. The weights are proportional to the prior sample size a + b and the observed sample size n, respectively. This representation leads to a natural interpretation of the Beta prior parameters as prior data:

- $a \approx$ "prior # of 1's,"
- $b \approx$ "prior # of 0's,"
- $a + b \approx$ "prior sample size."

When $n \gg a + b$, it is reasonable to expect that most of the information about θ should come from the data rather than from the prior distribution. This intuition is confirmed mathematically. In particular, when $n \gg a + b$,

- $\frac{a+b}{a+b+n} \approx 0$,
- $\mathbb{E}[\theta \mid y] \approx \frac{y}{n}$,
- $\mathrm{Var}(\theta \mid y) \approx \frac{1}{n} \frac{y}{n} \left(1 - \frac{y}{n}\right)$.

Thus, in large samples, the posterior mean approaches the sample proportion and the posterior variance shrinks at rate $1/n$, reflecting increasing information from the data.



Prediction

An important feature of Bayesian inference is the existence of a **predictive distribution** for new observations.

The posterior predictive distribution for a new observation Y_{new} given the observed data y is obtained by integrating over the posterior distribution of θ.

Returning to our notation for binary data, let $y_1, \ldots, y_n$ be the observed outcomes from a sample of n binary rvs, and let $\tilde{Y} \in \{0, 1\}$ denote a future observation from the same population that has not yet been observed. The **predictive distribution** of $\tilde{Y}$ is defined as the conditional distribution of $\tilde{Y}$ given the observed data $\{Y_1 = y_1, \ldots, Y_n = y_n\}$. For conditionally i.i.d. binary observations, the predictive distribution can be derived by integrating out the unknown parameter θ:

$$\begin{aligned}
\Pr\left(\tilde{Y} = 1 \mid y_1, \ldots, y_n\right) &= \int \Pr\left(\tilde{Y} = 1, \theta \mid y_1, \ldots, y_n\right) d\theta \\
&= \int \Pr\left(\tilde{Y} = 1 \mid \theta, y_1, \ldots, y_n\right) p\left(\theta \mid y_1, \ldots, y_n\right) d\theta \\
&= \int p\left(\theta \mid y_1, \ldots, y_n\right) \theta d\theta \\
&= \mathrm{E}\left[\theta \mid y_1, \ldots, y_n\right] \\
&= \frac{a + \sum_{i=1}^{n} y_i}{a + b + n}.
\end{aligned}$$

Hence, we also have,

$$\Pr\left(\tilde{Y}=0 \mid y_1,\dots,y_n\right) = 1 - \mathrm{E}\left[\theta \mid y_1,\dots,y_n\right] = \frac{b+\sum_{i=1}^{n}(1-y_i)}{a+b+n}.$$

> **Properties of the predictive distribution**
>
> 1. It **does not depend on any unknown quantities.** If it did, it could not be used to make predictions.
> 2. It **depends on the observed data.** In particular, $\tilde{Y}$ is not independent of $Y_1,\dots,Y_n$, because the observed data provide information about θ, which in turn influences $\tilde{Y}$. If $\tilde{Y}$ were independent of the observed data, learning from data would be impossible.

The uniform prior distribution on [0,1], also known as the Beta$(1,1)$ prior, can be interpreted as containing the same information as a hypothetical prior dataset consisting of one success ("1") and one failure ("0").

Under this prior, the posterior predictive probability of a future success is

$$\Pr(\tilde{Y}=1 \mid Y=y) = \mathbb{E}[\theta \mid Y=y] = \frac{2}{2+n}\cdot\frac{1}{2} + \frac{n}{2+n}\cdot\frac{y}{n}.$$

This expression highlights that the predictive probability is a weighted average of:

- the prior mean 1/2, and
- the sample proportion y/n,

with weights proportional to the prior sample size 2 and the observed sample size n, respectively.

The posterior mode under this prior is

$$\mathrm{mode}(\theta \mid Y=y) = \frac{y}{n},$$

where

$$Y = \sum_{i=1}^{n} Y_i.$$

At first glance, the discrepancy between these two posterior summaries may seem surprising. However, it reflects the fact that different summaries capture different features of the posterior distribution.

To see this clearly, consider the case $Y=0$. In this case,

$$\mathrm{mode}(\theta \mid Y=0) = 0,$$

but the predictive probability remains

$$\Pr(\tilde{Y} = 1 \mid Y = 0) = \frac{1}{2+n}.$$

Thus, even when no successes have been observed, the Bayesian predictive distribution assigns a positive probability to a future success due to the prior information. This illustrates how Bayesian prediction naturally balances prior beliefs with observed data.

3.2.2 Confidence Regions: Bayesian v.s. Frequentist

If it often desirable to identify the regions of the parameter space that are likely to contain the true value of the parameter. To do this, after observing the data $Y = y$, we can construct an interval $[\ell(y), u(y)]$ that is likely to contain the true value of θ, i.e., the probability that $\ell(y) < \theta < u(y)$ is large. There are two different ways to interpret this probability, leading to the concepts of **Bayesian coverage** and **frequentist coverage**.

An interval $[\ell(y), u(y)]$, based on the observed data $Y = y$, has 100(1-α)% Bayesian coverage for θ if

$$\Pr(\ell(y) < \theta < u(y) \mid Y = y) = 1 - \alpha.$$

A random interval $[\ell(Y), u(Y)]$ has 100(1-α)% frequentist coverage for θ if, before the data are gathered,

$$\Pr(\ell(Y) < \theta < u(Y) \mid \theta) = 1 - \alpha.$$

> **Note**
>
> In a sense, the frequentist and Bayesian notions of coverage describe **pre** experimental and **post** experimental perspectives, respectively.

3.3 Frequentist vs Bayesian Coverage

You may recall an important point often emphasized in introductory statistics courses. Suppose we observe data $Y = y$ and compute a frequentist confidence interval

$$[l(y),\ u(y)].$$

Once the data are observed, the parameter θ is **treated as fixed, not random**. Therefore,

$$\Pr(l(y) < \theta < u(y) \mid \theta) = \begin{cases} 1, & \text{if } \theta \in [l(y), u(y)], \\ 0, & \text{if } \theta \notin [l(y), u(y)]. \end{cases}$$

This highlights a key limitation of frequentist confidence intervals:

They do not admit a post-experimental probability interpretation.

After observing the data, it is *not meaningful*, from a frequentist perspective, to say that there is a 95% probability that θ lies in the computed interval.

What Frequentist Coverage Means

Although this interpretation may feel unsatisfying, frequentist coverage is still useful in many situations. Imagine repeatedly running many independent experiments and constructing a confidence interval for each one.
If each interval procedure has 95% frequentist coverage, then:

About 95% of the intervals will contain the true parameter value.

This is a **long-run, repeated-sampling interpretation**, not a statement about any single observed interval.

Can Bayesian and Frequentist Coverage Agree?

A natural question is whether a confidence interval can simultaneously have:

- a Bayesian interpretation, i.e., a 100(1-α)% posterior probability that θ lies in the interval, and
- approximately 100(1-α)% frequentist coverage.

Hartigan (1966) showed that, for the types of intervals considered in Hopf (2009), an interval that has 95% Bayesian coverage additionally has the property that

$$\Pr(l(Y) < \theta < u(Y) \mid \theta) = 0.95 + \varepsilon_n,$$

where the error term satisfies $|\varepsilon_n| < a/n$ for some constant a. This result implies that, an interval with 95% Bayesian coverage, will also have approximately 95% frequentist coverage, at least asymptotically, as the sample size n grows.

In other words, under suitable conditions, **Bayesian credible intervals and frequentist confidence intervals can agree in large samples**, even though their interpretations are fundamentally different. Keep in mind that most non-Bayesian methods of constructing 100(1-α)% confidence intervals also only achieve their nominal coverage probability asymptotically.

> **i** Reminder
>
> This reconciliation is important, but it should not obscure the conceptual distinction:
>
> - frequentist coverage is a *pre-experimental* property of a procedure,
> - Bayesian coverage is a *post-experimental* probability statement about θ given the data.

For further discussion of the similarities between Bayesian and frequentist intervals, see Severini (1991) and Sweeting (2001).

3.4 Posterior Quantile Intervals

One of the simplest ways to construct a Bayesian credible interval is to use **posterior quantiles**. To form a $100(1-\alpha)\%$ credible interval for θ, find numbers $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ such that

1. $\Pr(\theta < \theta_{\alpha/2} \mid Y = y) = \alpha/2$,
2. $\Pr(\theta > \theta_{1-\alpha/2} \mid Y = y) = \alpha/2$,

where $\theta_{\alpha/2}$ and $\theta_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ posterior quantiles of θ. By construction,

$$\begin{aligned}\Pr\left(\theta \in \left[\theta_{\alpha/2}, \theta_{1-\alpha/2}\right] \mid Y = y\right) &= 1 - \Pr\left(\theta \notin \left[\theta_{\alpha/2}, \theta_{1-\alpha/2}\right] \mid Y = y\right) \\ &= 1 - \left[\Pr\left(\theta < \theta_{\alpha/2} \mid Y = y\right) + \Pr\left(\theta > \theta_{1-\alpha/2} \mid Y = y\right)\right] \\ &= 1 - \alpha.\end{aligned}$$

Suppose we observe $n = 10$ conditionally independent Bernoulli trials and obtain $Y = 2$ successes. Using a uniform prior for θ, $\theta \sim \text{Beta}(1, 1)$, the posterior distribution is

$$\theta \mid \{Y = 2\} \sim \text{Beta}(1 + 2,\ 1 + 8) = \text{Beta}(3, 9).$$

A 95% posterior confidence interval can be obtained from by 2.5% and 97.5% quantiles of this Beta distribution $[\theta_{0.025}, \theta_{0.975}]$. In this case,

$$\theta_{0.025} \approx 0.06, \qquad \theta_{0.975} \approx 0.52,$$

so

$$\Pr(0.06 \leq \theta \leq 0.52 \mid Y = 2) = 0.95.$$

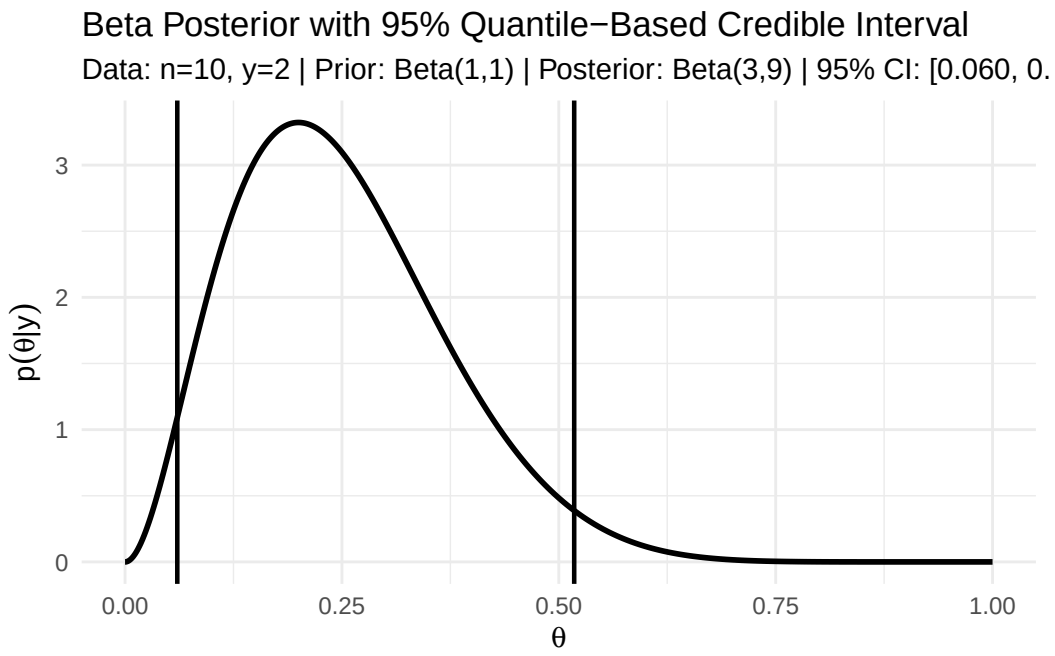
This interval has a direct probabilistic interpretation: **given the observed data**, there is a 95% posterior probability that θ lies in this range.

```
b <- a <- 1 # prior parameter
n <- 10 ; y <- 2 # data
qbeta(c(0.025, 0.975), a + y, b + n - y)
```

```
[1] 0.06021773 0.51775585
```

```
a_post <- a + y
b_post <- b + (n - y)

# 95% quantile-based credible interval
ci <- qbeta(c(0.025, 0.975), a_post, b_post)
```

```r
ci_low <- ci[1]
ci_high <- ci[2]

# Grid for plotting posterior density
theta <- seq(0, 1, length.out = 2000)
df <- data.frame(theta = theta, density = dbeta(theta, a_post, b_post))

# Plot: posterior density curve + two vertical CI bars
ggplot(df, aes(x = theta, y = density)) +
  geom_line(linewidth = 1) +
  geom_vline(xintercept = ci_low, linewidth = 0.8) +
  geom_vline(xintercept = ci_high, linewidth = 0.8) +
  labs(
    title = "Beta Posterior with 95% Quantile-Based Credible Interval",
    subtitle = sprintf(
      "Data: n=%d, y=%d | Prior: Beta(%d,%d) | Posterior: Beta(%d,%d) | 95%% CI: [%.3f, %.3f]
      n, y, a, b, a_post, b_post, ci_low, ci_high
    ),
    x = expression(theta),
    y = expression(p(theta * "|" * y))
  ) +
  theme_minimal()
```



Highest posterior density (HPD) region

The Figure above illustrates the posterior distribution of θ for the binomial example with a uniform prior, together with a 95% quantile-based credible interval. Notice an important feature of the plot:

> There exist values of θ *outside* the quantile-based interval that have *higher posterior density* than some values *inside* the interval.

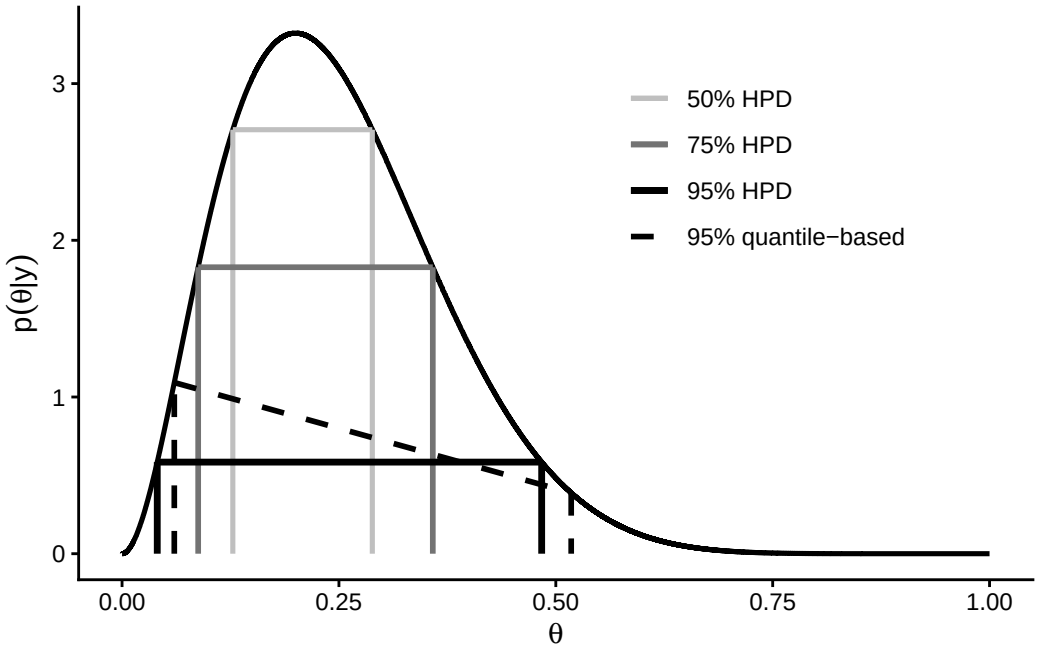
This observation suggests that the quantile-based interval may not be the most efficient way to summarize posterior uncertainty. In particular, it motivates a more restrictive type of credible region that concentrates on the most plausible parameter values.

A $100(1-\alpha)\%$ *HPD region* is a subset of the sample space, $s(y) \subset \Theta$ such that:

1. $\Pr(\theta \in s(y) \mid Y = y) = 1 - \alpha$, and
2. If $\theta_a \in s(y)$ and $\theta_b \notin s(y)$, then $p(\theta_a \mid Y = y) \geq p(\theta_b \mid Y = y)$.

In words, an HPD region contains the parameter values with the *largest posterior density*, subject to containing probability mass $1 - \alpha$.

Observed that, all points inside an HPD region are at least as plausible as any point outside the region, according to the posterior distribution. This property distinguishes HPD regions from quantile-based intervals, which are defined purely by cumulative probability and may include low-density values while excluding higher-density ones.



An HPD region can be constructed conceptually as follows:

> **Algorithm to construct an HPD region**
>
> 1. Begin with a horizontal line above the posterior density curve.
> 2. Gradually lower the line.
> 3. At each height, include all values of θ whose posterior density exceeds the line.
> 4. Stop lowering the line once the total posterior probability of the included region reaches $1 - \alpha$.

This procedure guarantees that the retained region consists of the most probable values of θ.

HPD Regions and Multimodality

If the posterior density is **unimodal**, the HPD region will typically be a single interval. However, if the posterior density is **multimodal** (having multiple peaks), the HPD region need not be an interval; it may consist of several disjoint subsets of the parameter space.

In the binomial example with $n = 10$, $Y = 2$, and a uniform prior, the posterior distribution is $\text{Beta}(3, 9)$.

For this posterior:

- The 95% quantile-based credible interval is approximately [0.06, 0.52].
- The 95% HPD region is approximately [0.04, 0.48].

The HPD region is *narrower*, and therefore more precise, than the quantile-based interval, while still containing 95% of the posterior probability.

Both intervals are valid Bayesian credible intervals, but they summarize posterior uncertainty in different ways.

3.5 The Poisson Model

Another commonly used distribution is the *Poisson*, in this case, the measurement are the integer numbers. Some examples include number of coin tosses, the number of friends they have, or the number of birthday celebrations have a person have. In these situations, the sample space is $\mathcal{Y} = \{0, 1, 2, ...\}$. There are other possible models for those situation, but perhaps the simplest probability model on $\mathcal{Y}$ is the Poisson model.

Poisson distribution

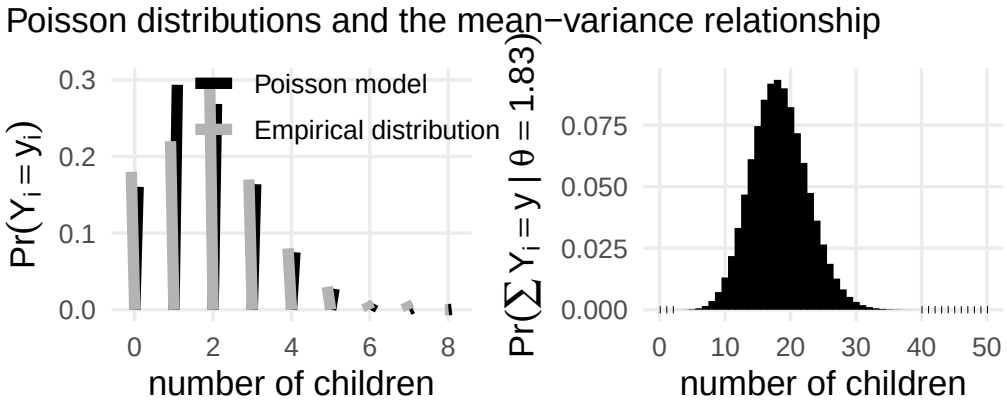
Recall that a random variable Y has a Poisson distribution with mean θ if

$$\Pr(Y = y \mid \theta) = \text{dpois}(y, \theta) = \frac{\theta^y e^{-\theta}}{y!}, \qquad y \in \{0, 1, 2, ...\}.$$

For such a random variable, we have,

- $\mathbb{E}(Y) = \theta$
- $\text{Var}(Y) = \theta$.

People sometimes use the Poisson distribution to model count data because of its simplicity and its ability to model events that occur independently over a fixed interval of time or space. The Poisson distribution is particularly useful when the events being counted are rare or infrequent, and when the average rate of occurrence is known. Note that, in this model, the mean and the variance are the same, which is a property that can be useful in certain applications; One may call this property as "mean-variance relationship".



Left: Poisson pmf with mean . = 1.83 (black) overlaid with an empirical distribution (grey

Right: Distribution of the sum of 10 i.i.d. Poisson(1.83) variables; by additivity this is Pois

The increased spread illustrates the Poisson mean−variance relationship: larger means

3.5.1 Inference on the Posterior for Poisson Model

Suppose we observe data $Y_1, \ldots, Y_n$ and model them as conditionally independent Poisson random variables with common mean θ, i.e.,

$$Y_i \mid \theta \sim \text{Poisson}(\theta), \qquad i = 1, \ldots, n.$$

The joint probability mass function of the data, given θ, is

$$\Pr(Y_1 = y_1, \ldots, Y_n = y_n \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta).$$

Using the Poisson pmf,

$$p(y_i \mid \theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!},$$

we obtain

$$\begin{aligned}\Pr(Y_1 = y_1, \dots, Y_n = y_n \mid \theta) &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= c(y_1, \dots, y_n)\, \theta^{\sum_{i=1}^n y_i} e^{-n\theta},\end{aligned}$$

where

$$c(y_1, \dots, y_n) = \prod_{i=1}^{n} \frac{1}{y_i!},$$

which does not depend on θ. This expression shows that the likelihood depends on the data only through the statistic $S = \sum_{i=1}^n Y_i$.

As in the binomial model, the statistic $S = \sum_{i=1}^n Y_i$ contains all information in the data about θ.
Indeed,

$$\sum_{i=1}^{n} Y_i \mid \theta \sim \text{Poisson}(n\theta),$$

and we therefore say that S is a sufficient statistic for θ.

3.5.2 Comparing posterior beliefs

To compare two values θ_a and θ_b *a posteriori*, consider the posterior odds:

$$\frac{p(\theta_a \mid y_1, \dots, y_n)}{p(\theta_b \mid y_1, \dots, y_n)}.$$

By Bayes' rule,

$$p(\theta \mid y_1, \dots, y_n) \propto \pi(\theta)\, p(y_1, \dots, y_n \mid \theta) \propto \pi(\theta)\, \theta^{\sum_{i=1}^n y_i} e^{-n\theta}.$$

Therefore,

$$\frac{p(\theta_a \mid y)}{p(\theta_b \mid y)} = \frac{\theta_a^{\sum y_i} e^{-n\theta_a} p(\theta_a)}{\theta_b^{\sum y_i} e^{-n\theta_b} p(\theta_b)}.$$

This expression highlights how posterior beliefs balance prior information with evidence from the data.

Conjugate prior for the Poisson model

We now seek a prior distribution for θ that leads to a posterior distribution of the same functional form. From the likelihood,

$$p(\theta \mid y) \propto p(\theta)\, \theta^{\sum y_i} e^{-n\theta},$$

we see that a conjugate prior must involve terms of the form

$$\theta^{c_1} e^{-c_2 \theta}$$

for some constants c_1 and c_2. The simplest family of distributions with this structure is the family of Gamma distributions, also called **Gamma family**.

3.5.3 Gamma distribution

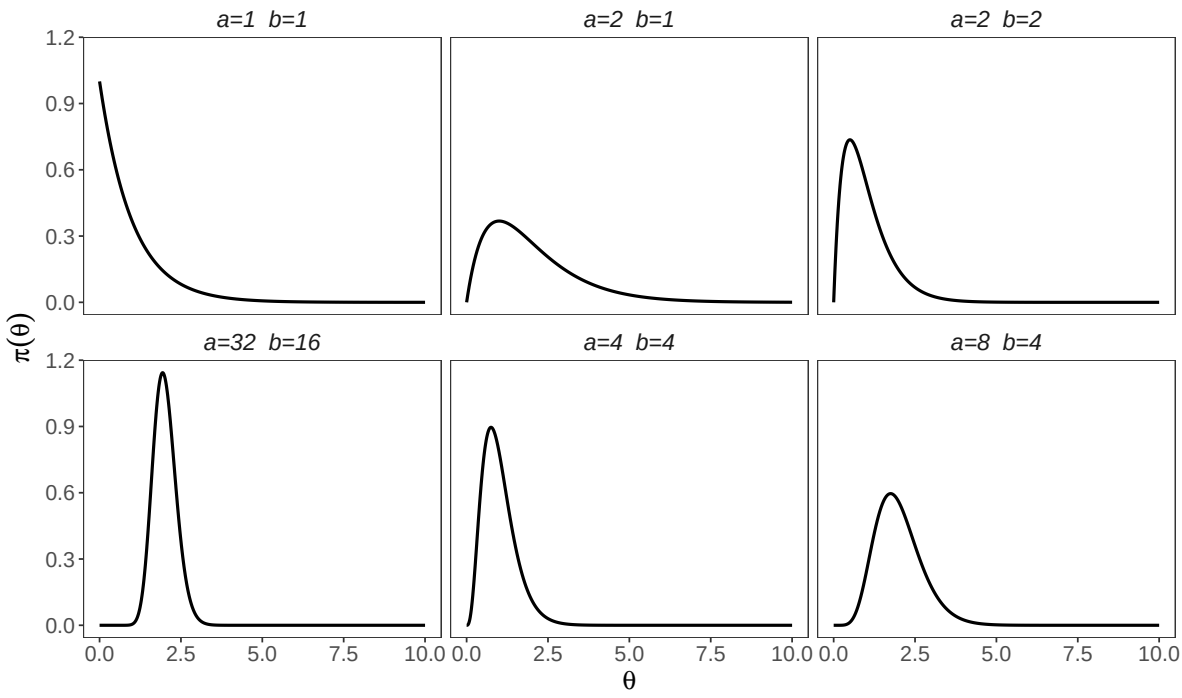
A positive random variable θ has a Gamma(a, b) distribution if

$$\pi(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}, \qquad \theta > 0,$$

where $a > 0$ is the shape parameter and $b > 0$ is the rate parameter.

For a Gamma(a, b) random variable,

- $\mathbb{E}(\theta) = a/b$,
- $\mathrm{Var}(\theta) = a/b^2$.
- $\mathrm{mode}[\theta] = \begin{cases} (a-1)/b & \text{if } a > 1 \\ 0 & \text{if } a \leq 1 \end{cases}$.



3.5.4 Posterior distribution of θ

If the prior is $\theta \sim \text{Gamma}(a, b)$, then combining the prior with the Poisson likelihood yields

$$\begin{aligned} p(\theta \mid y_1, \ldots, y_n) &= \pi(\theta) \times p(y_1, \ldots, y_n \mid \theta) / p(y_1, \ldots, y_n) \\ &= \left\{\theta^{a-1} e^{-b\theta}\right\} \times \left\{\theta^{\sum y_i} e^{-n\theta}\right\} \times c(y_1, \ldots, y_n, a, b) \\ &= \left\{\theta^{a+\sum y_i - 1} e^{-(b+n)\theta}\right\} \times c(y_1, \ldots, y_n, a, b) \\ &\propto \theta^{a+\sum y_i - 1} e^{-(b+n)\theta}. \end{aligned}$$

Thus, by the uniqueness theorem (of the density), the posterior distribution is

$$\theta \mid y_1, \ldots, y_n \sim \text{Gamma}(a + \sum_{i=1}^{n} y_i,\ b + n).$$

This shows that the Gamma distribution is **conjugate** to the Poisson likelihood.

Interpretation

Posterior inference for the Poisson model is therefore straightforward:

- The data enter only through the sufficient statistic $\sum Y_i$;

- The posterior mean is

$$\begin{aligned}\mathrm{E}\left[\theta \mid y_1, \dots, y_n\right] &= \frac{a + \sum y_i}{b + n} \\ &= \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum y_i}{n}\end{aligned}$$

- This decomposition shows that it is a **convex combination** of the prior mean a/b and the sample mean $\bar{y}$, and gives a useful information
- b acts like the **number of prior observations**;
- a acts like the **total count** from those b observations;
- a/b is the prior mean.
- Increasing the sample size n reduces posterior uncertainty, because the ifnromation from the data *dominates* the prior belief. To see this, we have, for $n \gg b$, we have
- $\mathrm{E}[\theta \mid y] \approx \bar{y}$,
- $\mathrm{Var}(\theta \mid y) \approx \bar{y}/n$.

This conjugate structure makes the Poisson–Gamma model a convenient and interpretable starting point for Bayesian analysis of *count data.*

3.5.5 Posterior predictive distribution for Poisson Model

We have seen that, the Bayesian prediction for a future observation $\tilde{y}$ is based on the **posterior predictive distribution**. In Gamma-Poisson model, we have

$$p(\tilde{y} \mid y_1, \dots, y_n) = \int_0^\infty p(\tilde{y} \mid \theta, y)\, p(\theta \mid y_1, \dots, y_n)\, d\theta.$$

For the Poisson model,

$$p(\tilde{y} \mid \theta) = \mathrm{Poisson}(\theta), \qquad p(\theta \mid y) = \mathrm{Gamma}\Big(a + \sum y_i,\ b + n\Big)\,.$$

Substituting,

$$p(\tilde{y} \mid y) = \int_0^\infty \mathrm{dpois}(\tilde{y}, \theta)\, \mathrm{dgamma}\Big(\theta,\ a + \sum y_i,\ b + n\Big)\ d\theta.$$

Writing this integral explicitly,

$$p(\tilde{y} \mid y) = \int_0^\infty \frac{\theta^{\tilde{y}} e^{-\theta}}{\tilde{y}!} \cdot \frac{(b+n)^{a+\sum y_i}}{\Gamma(a + \sum y_i)} \theta^{a+\sum y_i - 1} e^{-(b+n)\theta}\, d\theta.$$

Combining terms, we have

$$p(\tilde{y} \mid y) = \frac{(b+n)^{a+\sum y_i}}{\tilde{y}!\,\Gamma(a+\sum y_i)} \int_0^\infty \theta^{a+\sum y_i+\tilde{y}-1} e^{-(b+n+1)\theta}\, d\theta.$$

The integral term seems difficult to evaluate in a glance, but there is actually a clear "trick" to help use. Recall the density of a Gamma distribution:

$$1 = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}\, d\theta, \quad , \alpha, \beta > 0,$$

which implies

$$\int_0^\infty \theta^{\alpha-1} e^{-\beta\theta}\, d\theta = \frac{\Gamma(\alpha)}{\beta^\alpha}, \qquad \alpha, \beta > 0.$$

Applying this with

$$\alpha = a + \sum y_i + \tilde{y}, \qquad \beta = b + n + 1,$$

we obtain

$$\int_0^\infty \theta^{a+\sum y_i+\tilde{y}-1} e^{-(b+n+1)\theta}\, d\theta = \frac{\Gamma(a+\sum y_i+\tilde{y})}{(b+n+1)^{a+\sum y_i+\tilde{y}}}.$$

Substituting back and simplifying,

$$p(\tilde{y} \mid y_1, \dots, y_n) = \frac{\Gamma(a+\sum y_i+\tilde{y})}{\Gamma(a+\sum y_i)\,\tilde{y}!} \left(\frac{b+n}{b+n+1}\right)^{a+\sum y_i} \left(\frac{1}{b+n+1}\right)^{\tilde{y}},$$

for $\tilde{y} \in \{0, 1, 2, ...\}$. We relieazed that it is a **negative binomial distribution** with parameters $(a + \sum y_i, b + n)$. That is, $\tilde{Y} \mid y_1, \dots, y_n \sim \text{NB}(a + \sum y_i, b + n)$. As a result, we have the mean and variance of the posterior predictive distribution

$$\mathbb{E}(\tilde{Y} \mid y) = \frac{a+\sum y_i}{b+n} = \mathbb{E}[\theta \mid y],$$

and

$$\begin{aligned} \text{Var}\left[\tilde{Y} \mid y_1, \dots, y_n\right] &= \frac{a+\sum y_i}{b+n} \frac{b+n+1}{b+n} \\ &= \text{Var}\left[\theta \mid y_1, \dots y_n\right] \times (b+n+1) \\ &= \mathbb{E}\left[\theta \mid y_1, \dots, y_n\right] \times \frac{b+n+1}{b+n} \quad , \\ &= \frac{a+\sum y_i}{b+n} \times \frac{b+n+1}{b+n} \end{aligned}$$

respectively

Interpretation and Take away

Recall that the predictive variance is to some extent a measure of our posterior uncertainty about a new observation $\tilde{Y}$. It reflects **two sources of uncertainty**:

1. **Sampling variability** (Sampling) For a Poisson model, the variance of Y given θ is equal to θ.
2. **Parameter uncertainty** (Population) When θ is unknown, uncertainty about θ inflates the variance of future observations.

For large n, the data dominate the prior:

$$\frac{b+n+1}{b+n} \approx 1,$$

so predictive uncertainty is driven primarily by sampling variability. In this case, the uncertainy abbout θ is small which for the Possion model is equal to *theta*

For small n, posterior uncertainty about θ is substantial, and

$$\frac{b+n+1}{b+n} > 1,$$

leading to larger predictive variance than under a fixed-θ Poisson model (i.e., larger than just the sampling variability).

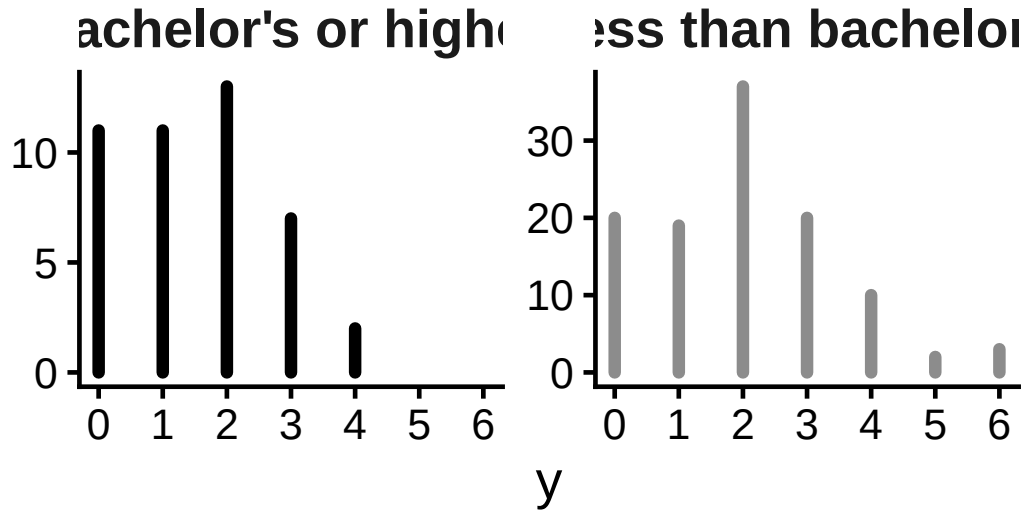
- Bayesian prediction naturally incorporates both **sampling variability** and **parameter uncertainty**.

This completes posterior inference and prediction for the Gamma-Poisson model.

3.6 Example: Birth rates

Over the course of the 1990s, the General Social Survey collected data on the educational attainment and number of children of women who were 40 years old at the time of participation in the survey. These women were in their 20s during the 1970s, a period of historically low fertility rates in the United States.

In this example, we compare women **with** a bachelor's degree to those **without** a bachelor's degree in terms of their numbers of children.



Numbers of children for the two groups.

Sampling models

Let

- $Y_{i,1}$ denote the number of children for woman i **without** a bachelor's degree, $i = 1, \dots, n_1$,
- $Y_{i,2}$ denote the number of children for woman i **with** a bachelor's degree, $i = 1, \dots, n_2$.

Then, since it is a coint data, we assume the following Poisson sampling models:

$$Y_{1,1}, \dots, Y_{n_1,1} \mid \theta_1 \sim \text{i.i.d. Poisson}(\theta_1),$$

$$Y_{1,2}, \dots, Y_{n_2,2} \mid \theta_2 \sim \text{i.i.d. Poisson}(\theta_2).$$

Here, θ_1 and θ_2 represent the population mean birth rates for the two groups.

Data summaries

The empirical summaries of the data are:

- No college degree:

$$n_1 = 111, \qquad \sum_{i=1}^{n_1} Y_{i,1} = 217, \qquad \bar{Y}_1 = 1.95$$

- Bachelor's degree or higher:

$$n_2 = 44, \qquad \sum_{i=1}^{n_2} Y_{i,2} = 66, \qquad \bar{Y}_2 = 1.50$$

Prior distributions

We place independent Gamma priors on the two population means:

$$\theta_1, \theta_2 \sim \text{i.i.d. Gamma}(a = 2, b = 1),$$

where the Gamma distribution is parameterized by **shape** a and **rate** b.

The prior mean is $a/b = 2$, representing weak prior information corresponding to roughly one prior observation with an average count of two. (Why?)

Posterior distributions

Under the Poisson–Gamma conjugate model, the posterior distributions are

$$\theta_1 \mid y \sim \text{Gamma}(a + \sum Y_{i,1},\, b + n_1) = \text{Gamma}(219, 112),$$

$$\theta_2 \mid y \sim \text{Gamma}(a + \sum Y_{i,2},\, b + n_2) = \text{Gamma}(68, 45).$$

These posteriors summarize our updated beliefs about the two population mean birth rates after observing the data. From the Gamma posterior distributions, we can compute posterior means, modes, and 95% quantile-based credible intervals.

- Posterior means:

$$\mathbb{E}(\theta_1 \mid y) = \frac{219}{112} \approx 1.96, \qquad \mathbb{E}(\theta_2 \mid y) = \frac{68}{45} \approx 1.51$$

- Posterior modes:

$$\text{mode}(\theta_1 \mid y) = \frac{218}{112} \approx 1.95, \qquad \text{mode}(\theta_2 \mid y) = \frac{67}{45} \approx 1.49$$

```r
# ============================================================
# Poisson-Gamma model: two-group posterior summaries
# Prior:  theta ~ Gamma(a, b)  with shape = a, rate = b
# Data:   Yi | theta ~ i.i.d. Poisson(theta)
# Posterior: theta | y ~ Gamma(a + sum(y), b + n)
# ============================================================

# ----------------------------
# 1) Inputs: prior + data
# ----------------------------
a <- 2
b <- 1

group <- data.frame(
  group = c("Less than bachelor's", "Bachelor's or higher"),
  n     = c(111, 44),
  sum_y = c(217, 66)
)

# ----------------------------
# 2) Posterior functions
# ----------------------------
post_shape <- function(a, sum_y) a + sum_y
post_rate  <- function(b, n)     b + n

post_mean <- function(shape, rate) shape / rate
post_mode <- function(shape, rate) ifelse(shape > 1, (shape - 1) / rate, 0)

post_ci <- function(shape, rate, level = 0.95) {
  alpha <- (1 - level) / 2
  qgamma(c(alpha, 1 - alpha), shape = shape, rate = rate)
}

# ----------------------------
# 3) Compute summaries
# ----------------------------
out <- within(group, {
  shape_post <- post_shape(a, sum_y)
  rate_post  <- post_rate(b, n)

  mean_post  <- post_mean(shape_post, rate_post)
  mode_post  <- post_mode(shape_post, rate_post)
```

```r
  ci_post    <- t(mapply(post_ci, shape_post, rate_post))
  ci_lower   <- ci_post[, 1]
  ci_upper   <- ci_post[, 2]
})

# ----------------------------
# 4) Print results clearly
# ----------------------------
cat("Prior: theta ~ Gamma(shape = ", a, ", rate = ", b, ")\n\n", sep = "")
```

```
Prior: theta ~ Gamma(shape = 2, rate = 1)
```

```r
for (i in seq_len(nrow(out))) {
  cat("------------------------------------------------------------\n")
  cat("Group: ", out$group[i], "\n", sep = "")
  cat("n = ", out$n[i], ",  sum(y) = ", out$sum_y[i], "\n", sep = "")
  cat("Posterior: theta | y ~ Gamma(shape = ", out$shape_post[i],
      ", rate = ", out$rate_post[i], ")\n", sep = "")
  cat(sprintf("Posterior mean = %.6f\n", out$mean_post[i]))
  cat(sprintf("Posterior mode = %.6f\n", out$mode_post[i]))
  cat(sprintf("Posterior 95%% CI = [%.6f, %.6f]\n", out$ci_lower[i], out$ci_upper[i]))
}
```

```
------------------------------------------------------------
Group: Less than bachelor's
n = 111,  sum(y) = 217
Posterior: theta | y ~ Gamma(shape = 219, rate = 112)
Posterior mean = 1.955357
Posterior mode = 1.946429
Posterior 95% CI = [1.704943, 2.222679]
------------------------------------------------------------
Group: Bachelor's or higher
n = 44,  sum(y) = 66
Posterior: theta | y ~ Gamma(shape = 68, rate = 45)
Posterior mean = 1.511111
Posterior mode = 1.488889
Posterior 95% CI = [1.173437, 1.890836]
```

```r
summary_tbl <- out[, c("group", "n", "sum_y", "mean_post", "mode_post", "ci_lower", "ci_upper
print(summary_tbl, row.names = FALSE)
```

```
                group   n sum_y mean_post mode_post ci_lower ci_upper
 Less than bachelor's 111   217  1.955357  1.946429 1.704943 2.222679
 Bachelor's or higher  44    66  1.511111  1.488889 1.173437 1.890836
```

The posterior distributions indicate **strong evidence** that the mean birth rate is higher for women without a bachelor's degree (i.e., $\theta_1 > \theta_2$). For example,

$$\Pr(\theta_1 > \theta_2 \mid y) \approx 0.97.$$

Posterior predictive distributions

Now consider two randomly sampled future individuals:

- $\tilde{Y}_1$: a woman without a bachelor's degree,
- $\tilde{Y}_2$: a woman with a bachelor's degree.

  Question: To what extent do we expect the one without the college degree to have more children than the other?

The posterior predictive distributions integrate over uncertainty in θ:

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

Under the Poisson–Gamma model, the posterior predictive distributions are **Negative Binomial**:

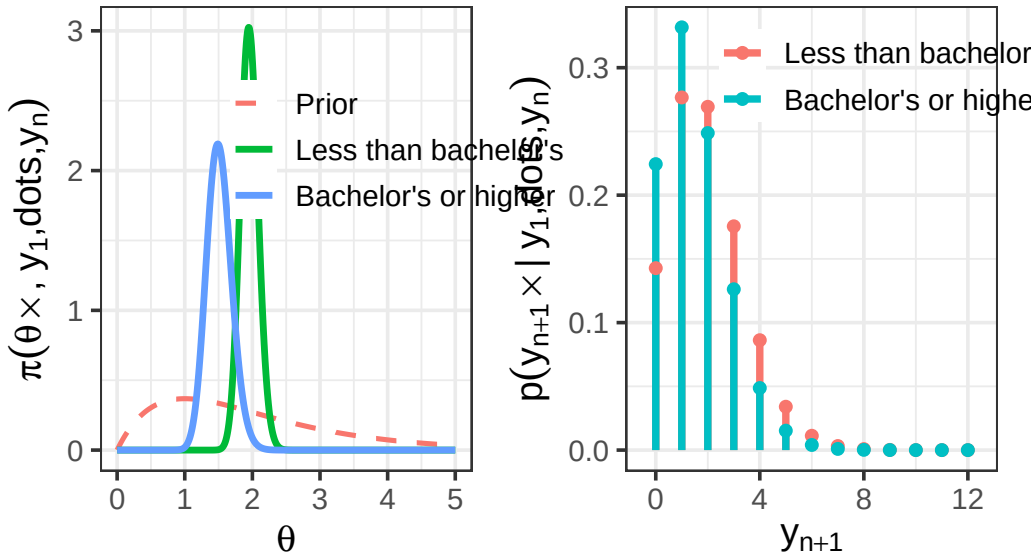
$$\tilde{Y}_1 \mid y \sim \text{NegBin}(a + \sum Y_{i,1},\, b + n_1),$$

$$\tilde{Y}_2 \mid y \sim \text{NegBin}(a + \sum Y_{i,2},\, b + n_2).$$

```
# A tibble: 2 x 9
  group                   n sum_y shape  rate post_mean post_mode  q025  q975
  <chr>               <dbl> <dbl> <dbl> <dbl>     <dbl>     <dbl> <dbl> <dbl>
1 Less than bachelor's   111   217   219   112      1.96      1.95  1.70  2.22
2 Bachelor's or higher    44    66    68    45      1.51      1.49  1.17  1.89
```

```
y <- 0:10
```

```
Group 1: Less than bachelor's
```



posterior predictive distributions for the number of children in each group.

Figure 3.1

```
size = 219   mu = 1.955357

 [1] 1.427473e-01 2.766518e-01 2.693071e-01 1.755660e-01 8.622930e-02
 [6] 3.403387e-02 1.124423e-02 3.198421e-03 7.996053e-04 1.784763e-04
[11] 3.601115e-05


Group 2: Bachelor's or higher

size = 68   mu = 1.511111

 [1] 2.243460e-01 3.316420e-01 2.487315e-01 1.261681e-01 4.868444e-02
 [6] 1.524035e-02 4.030961e-03 9.263700e-04 1.887982e-04 3.465861e-05
[11] 5.801551e-06

    y Less.than.bachelor.s Bachelor.s.or.higher
1   0         1.427473e-01         2.243460e-01
2   1         2.766518e-01         3.316420e-01
3   2         2.693071e-01         2.487315e-01
```

```
4   3          1.755660e-01          1.261681e-01
5   4          8.622930e-02          4.868444e-02
6   5          3.403387e-02          1.524035e-02
7   6          1.124423e-02          4.030961e-03
8   7          3.198421e-03          9.263700e-04
9   8          7.996053e-04          1.887982e-04
10  9          1.784763e-04          3.465861e-05
11 10          3.601115e-05          5.801551e-06
```

Interpretation and Conclusion

Although the posterior distributions of θ_1 and θ_2 are clearly separated, the posterior predictive distributions for $\tilde{Y}_1$ and $\tilde{Y}_2$ exhibit substantial overlap.

For example,

$$\Pr(\tilde{Y}_1 > \tilde{Y}_2 \mid y) \approx 0.48, \qquad \Pr(\tilde{Y}_1 = \tilde{Y}_2 \mid y) \approx 0.22.$$

This illustrates an important distinction:

> Strong evidence that population means differ does **not** imply that the difference is large or easily detectable at the individual level.

> **Note**
>
> Population-level effects and individual-level variability are fundamentally different sources of uncertainty.

3.7 Exponential Family

Many common sampling models belong to the **exponential family** of distributions, including the binomial and Poisson distribution we saw in this chapter. A one-parameter exponential family has probability density (or mass) function of the form

$$\begin{aligned} p(y \mid \phi) &= h(y) \exp\{\phi t(y) - A(\phi)\} \\ &= h(y) c(\phi) \exp\{\phi t(y)\}, \end{aligned}$$

where ϕ is the unknown parameter, $t(y)$ is a sufficient statistic, and $h(y)$, $A(\phi)$, and $c(\phi)$ are known functions. Diaconis and Ylvisaker (1979) studied conjugate priors for exponential family models and showed that they have the general form

$$p(\phi \mid n_0, t_0) = \kappa(n_0, t_0) c(\phi)^{n_0} \exp\{n_0 t_0 \phi\},$$

where $n_0 > 0$ and t_0 are hyperparameters.

With this result, and suppose the data consist of n independent observations $y_1, \dots, y_n$ are sampled from $Y_1, \dots, Y_n \overset{iid}{\sim} p(y \mid \theta)$. Combining the prior with the likelihood gives the posterior distribution

$$p(\phi \mid y_1, \dots, y_n) \;\propto\; p(\phi)\, p(y_1, \dots, y_n \mid \phi).$$

Substituting the exponential-family form,

$$\begin{aligned} p(\phi \mid y_1, \dots, y_n) &\propto c(\phi)^{\,n_0+n} \exp\left\{ \phi \left[n_0 t_0 + \sum_{i=1}^{n} t(y_i) \right] \right\} \\ &\propto p(\phi \mid n_0 + n,\; n_0 t_0 + n\,\bar{t}(y))\,, \end{aligned}$$

where

$$\bar{t}(y) = \frac{1}{n} \sum_{i=1}^{n} t(y_i).$$

Thus, the posterior distribution has the **same functional form** as the prior. This is why such priors are called *conjugate*.

3.7.1 Interpretation of n_0 and t_0

The similarity between the prior and posterior distributions suggests an interpretation of the hyperparameters:

- n_0 can be interpreted as a **prior sample size**,
- t_0 can be interpreted as a **prior guess** of $t(Y)$.

This interpretation can be made more precise. Diaconis and Ylvisaker (1979) show that

$$\mathbb{E}[t(Y)] = \mathbb{E}[\,\mathbb{E}[t(Y) \mid \phi]\,] = \mathbb{E}\left[-\frac{c'(\phi)}{c(\phi)} \right] = t_0.$$

(See also Exercise 3.6 in Hopf, 2009)

Thus, t_0 represents the **prior expected value** of the sufficient statistic $t(Y)$.

The parameter n_0 measures how informative the prior is. One way to see this is to note that, as a function of ϕ, $p(\phi \mid n_0, t_0)$ has the same shape as a likelihood $p(\tilde{y}_1, \dots, \tilde{y}_{n_0} \mid \phi)$ based on n_0 hypothetical "prior observations" $\tilde{y}_1, \dots, \tilde{y}_{n_0}$ satisfying

$$\frac{1}{n_0} \sum_{i=1}^{n_0} t(\tilde{y}_i) = t_0.$$

In this sense, the prior distribution $p(\phi \mid n_0, t_0)$ contains the same amount of information as would be obtained from n_0 independent observations from the population.

The exponential family representation of the binomial(θ) model can be obtained from the density of a single Bernoulli random variable:

$$p(y \mid \theta) = \theta^y (1-\theta)^{1-y}.$$

Rewrite this as

$$\begin{aligned} p(y \mid \theta) &= \left(\frac{\theta}{1-\theta}\right)^y (1-\theta) \\ &= \exp\left\{y \log\left(\frac{\theta}{1-\theta}\right)\right\}(1-\theta). \end{aligned}$$

Let

$$\phi = \log\left(\frac{\theta}{1-\theta}\right),$$

the **log-odds**. Then

$$\theta = \frac{e^\phi}{1+e^\phi}, \qquad 1-\theta = \frac{1}{1+e^\phi}.$$

Substituting gives

$$p(y \mid \phi) = e^{\phi y}(1+e^\phi)^{-1}.$$

Thus the Bernoulli model is an exponential family model with

- $t(y) = y$,
- $c(\phi) = (1+e^\phi)^{-1}$,
- $h(y) = 1$.

Conjugate prior

The conjugate prior for ϕ has the form

$$p(\phi \mid n_0, t_0) \propto (1+e^\phi)^{-n_0} \exp\{n_0 t_0 \phi\}.$$

Since $t(y) = y$, the parameter t_0 represents the prior expectation of Y, or equivalently,

$$t_0 = \mathbb{E}(\theta).$$

Using a change of variables back to θ, the prior becomes

$$p(\theta \mid n_0, t_0) \propto \theta^{n_0 t_0 - 1}(1-\theta)^{n_0(1-t_0)-1},$$

which is a Beta distribution:

$$\theta \sim \text{Beta}(n_0 t_0,\ n_0(1-t_0)).$$

A weakly informative prior can be obtained by setting

- t_0 equal to our prior expectation,
- $n_0 = 1$.

For example, if our prior expectation is $1/2$, this gives

$$\theta \sim \text{Beta}(1/2, 1/2),$$

which is Jeffreys' prior for the binomial model.

Under this prior, the posterior distribution is

$$\theta \mid y_1, \ldots, y_n \sim \text{Beta}\left(t_0 + \sum y_i,\ (1-t_0) + \sum (1-y_i)\right).$$

The Poisson(θ) model can also be written in exponential family form.

The Poisson pmf is

$$p(y \mid \theta) = \frac{\theta^y e^{-\theta}}{y!}.$$

Rewrite as

$$p(y \mid \theta) = \frac{1}{y!}\exp\{y \log\theta - \theta\}.$$

Thus it is an exponential family with

- $t(y) = y$,
- $\phi = \log\theta$,
- $c(\phi) = \exp(-e^{\phi})$,

- $h(y) = 1/y!$.

Conjugate prior

The conjugate prior for ϕ has the form

$$p(\phi \mid n_0, t_0) \propto \exp\{n_0 t_0 \phi - n_0 e^{\phi}\}.$$

Transforming back to $\theta = e^{\phi}$, this becomes

$$p(\theta \mid n_0, t_0) \propto \theta^{n_0 t_0 - 1} e^{-n_0 \theta},$$

which is a Gamma distribution:

$$\theta \sim \text{Gamma}(n_0 t_0,\ n_0).$$

A weakly informative prior can be obtained by setting

- t_0 equal to the prior expectation of θ,
- $n_0 = 1$,

giving

$$\theta \sim \text{Gamma}(t_0, 1).$$

Under this prior, the posterior distribution is

$$\theta \mid y_1, \ldots, y_n \sim \text{Gamma}\left(t_0 + \sum y_i,\ n_0 + n\right).$$

3.8 Discussion

The notion of conjugacy for classes of prior distributions was developed in Raiffa and Schlaifer (1961). Important results on conjugacy for exponential families appear in Diaconis and Ylvisaker (1979) and Diaconis and Ylvisaker (1985). They show that any prior distribution may be approximated by a mixture of conjugate priors.

Most authors refer to intervals of high posterior probability as **credible intervals**, as opposed to confidence intervals. However, credible intervals do not necessarily have frequentist coverage properties. In many common models they are numerically similar to classical intervals, but this is not guaranteed.

Some authors argue that accurate frequentist coverage can guide the construction of prior distributions. See Kass and Wasserman (1996) for a review of formal methods for selecting prior distributions.

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

4 Monte Carlo Method and its variations

Leading objectives:

- understand how Monte Carlo (MC) and Markov chain Monte Carlo (MCMC) methods differ
- implement a Metropolis–Hastings algorithm to draw samples from the posterior
- use output of MCMC to obtain estimates and standard errors
- use efficient proposals and tuning for MCMC

4.1 Background and Motivation

What we have seen in the last chapter up to now, is to use the `conjugate prior` to obtain closed form expressions for the posterior distribution. However, in many cases, conjugate priors are not available or not desirable. In such cases, we need to resort to numerical methods to approximate the posterior distribution.

> **Question.**
> How can we perform Bayesian inference when conjugate priors are not available and the posterior has no closed-form expression?

There are two broad classes of approaches:

- **Simulation-based methods**:
  accept–reject sampling, Markov chain Monte Carlo (MCMC), particle filters, and related algorithms.
- **Deterministic approximation methods**:
  Laplace approximations (including INLA), variational Bayes, expectation propagation, and related techniques.

We will be focusing on the Monte Carlo (MC) methods and its variation.

4.2 Overview

4.2.1 Monte Carlo (MC)

Monte Carlo methods approximate expectations or probabilities using **random sampling**.
If samples can be drawn **directly** from the target distribution, Monte Carlo methods provide simple and effective estimators.

Typical use: - Numerical integration - Bootstrap methods - Simulation-based probability estimation

4.2.2 Markov Chain Monte Carlo (MCMC)

MCMC methods are used when **direct sampling is infeasible**.
They construct a **Markov chain** whose stationary distribution is the target distribution, and use dependent samples from the chain after burn-in.

Typical use: - Bayesian posterior sampling - High-dimensional or unnormalized distributions

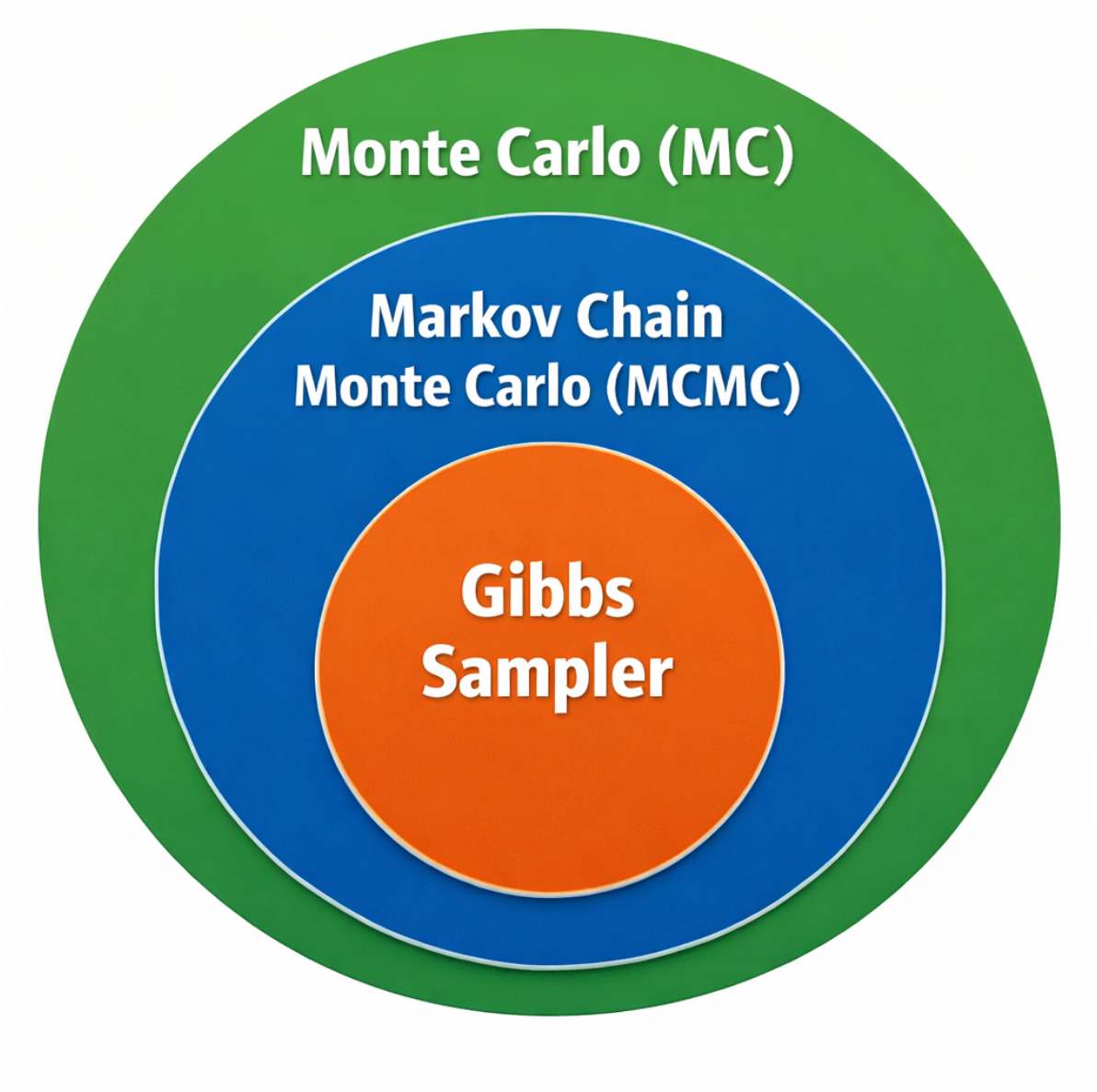
4.2.3 Gibbs Sampler

The Gibbs sampler is a **special case of MCMC** that samples sequentially from **full conditional distributions**.
Because proposals are drawn exactly from conditionals, all updates are automatically accepted.

Typical use: - Bayesian hierarchical models - Models with conjugate full conditionals

4.2.4 Relationship Between Monte Carlo, MCMC, and Gibbs Sampling

These three concepts are **not competing methods**, but rather form a **nested hierarchy** of ideas used for approximating expectations and probability distributions using randomness.



4.2.5 Summary Table

Method	Independent Samples	Uses Markov Chain	Accept–Reject Step	Typical Application
Monte Carlo (MC)	Yes	No	No	Direct simulation, integration

Method	Independent Samples	Uses Markov Chain	Accept–Reject Step	Typical Application
MCMC	No	Yes	Usually	Bayesian posterior sampling
Gibbs Sampler	No	Yes	No (always accept)	Bayesian models with tractable conditionals

i Key summary

- **Monte Carlo** is the general idea of using randomness for approximation.
- **MCMC** is Monte Carlo with dependent samples generated by a Markov chain.
- **Gibbs sampling** is a specific MCMC algorithm based on full conditional distributions.

4.3 Monte Carlo Methods

Motivation: why Monte Carlo?

In Bayesian inference we repeatedly encounter integrals such as

$$\mathbb{E}[g(\theta) \mid y] = \int g(\theta)\, p(\theta \mid y)\, d\theta, \qquad \Pr(\theta \in A \mid y) = \int_A p(\theta \mid y)\, d\theta,$$

that are not available in closed form. **Monte Carlo** replaces these integrals by averages of random draws.

i Note

Key idea: replace an intractable integral by an empirical mean.

4.4 The Monte Carlo Method

In the previous chapter, we obtained the following posterior distributions for the birth rates of women without and with bachelor's degrees:

$$\theta_1 \mid \sum_{i=1}^{111} Y_{i,1} = 217 \sim \text{Gamma}(219, 112),$$

$$\theta_2 \mid \sum_{i=1}^{44} Y_{i,2} = 66 \sim \text{Gamma}(68, 45).$$

It was claimed that

$$\Pr(\theta_1 > \theta_2 \mid \text{data}) = 0.97.$$

How do we compute such a probability? From Chapter 2, since θ_1 and θ_2 are conditionally independent given the data y, we have

$$\Pr(\theta_1 > \theta_2 \mid y) = \int_0^{\infty} \int_0^{\theta_1} p(\theta_1 \mid y) p(\theta_2 \mid y)\, d\theta_2\, d\theta_1.$$

Substituting the gamma densities gives

$$\int_0^{\infty} \int_0^{\theta_1} \text{dgamma}(\theta_1; 219, 112)\, \text{dgamma}(\theta_2; 68, 45)\, d\theta_2\, d\theta_1.$$

This integral can be evaluated numerically. However, in realistic Bayesian models, such integrals quickly become high-dimensional and analytically intractable. This motivates **Monte Carlo (MC) methods**.

4.4.1 MC Approximation

Suppose we wish to compute

$$\mathbb{E}[g(\theta) \mid y] = \int g(\theta)\, p(\theta \mid y)\, d\theta.$$

If we can generate independent samples

$$\theta^{(1)}, \ldots, \theta^{(S)} \sim p(\theta \mid y),$$

then we approximate the expectation by

$$\frac{1}{S} \sum_{s=1}^{S} g(\theta^{(s)}).$$

This is called a **Monte Carlo approximation**. By the Law of Large Numbers,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \;\longrightarrow\; \mathbb{E}[g(\theta) \mid y] = \int g(\theta)p(\theta \mid y)d\theta \quad \text{as } S \to \infty.$$

With the property above, we can calculate many quantities of interest about the posterior distribution. For example, suppose $\bar{\theta}$ is the average of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$, then as $S \to \infty$:

- $\bar{\theta} \to \mathbb{E}[\theta \mid y]$,
- $\frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \bar{\theta})^2 \to \text{Var}(\theta \mid y)$.
- $\frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{\theta^{(s)} \in A\} \to \Pr(\theta \in A \mid y)$.
- the empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ converges to $p(\theta \mid y)$.
- the sample median converges to the posterior median $\theta_{1/2}$.
- the sample α-quantile converges to θ_α.

4.4.2 Convergence Properties

Let $\theta^{(1)}, \ldots, \theta^{(S)} \sim p(\theta \mid y)$.

As $S \to \infty$:

- $\dfrac{\#\{\theta^{(s)} \leq c\}}{S} \;\longrightarrow\; \Pr(\theta \leq c \mid y)$
- The empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ converges to $p(\theta \mid y)$
- The sample median converges to the posterior median $\theta_{1/2}$
- The sample α-quantile converges to θ_α

Key message:
Almost any aspect of a posterior distribution can be approximated arbitrarily well using a sufficiently large Monte Carlo sample.

Thus Monte Carlo sampling allows us to approximate:

- posterior means,
- posterior variances,
- posterior probabilities,
- credible intervals,
- many more

To approximate

$$\Pr(\theta_1 > \theta_2 \mid y),$$

we can:

0. Choose a (large) number of samples S (e.g., $S = 10,000$).

1. Draw $\theta_1^{(s)} \sim \text{Gamma}(219, 112)$.
2. Draw $\theta_2^{(s)} \sim \text{Gamma}(68, 45)$.
3. Compute the indicator
$$I^{(s)} = \mathbf{1}\{\theta_1^{(s)} > \theta_2^{(s)}\}.$$

Then

$$\Pr(\theta_1 > \theta_2 \mid y) \approx \frac{1}{S}\sum_{s=1}^{S} I^{(s)}.$$

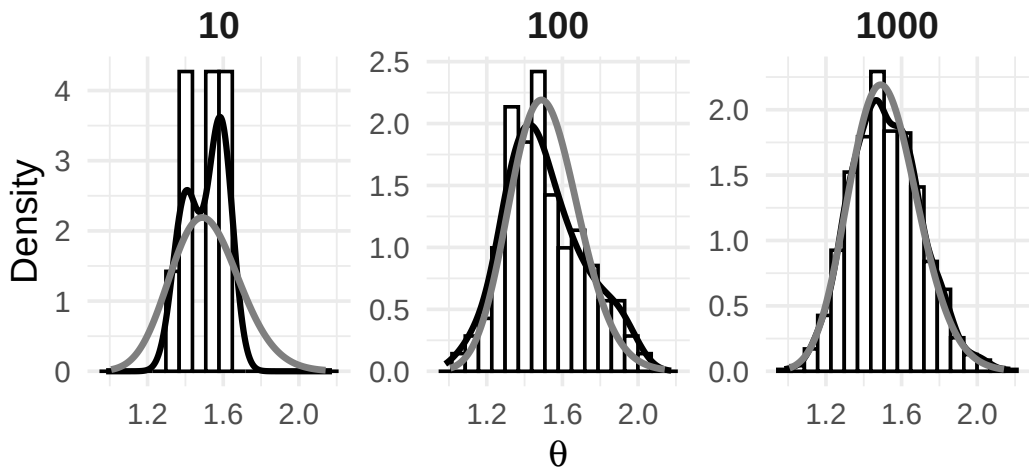
This avoids evaluating any double integrals.

> Why Monte Carlo?
>
> Works in high dimensions. Requires only the ability to simulate. Avoids symbolic integration. Scales to complex hierarchical models.

This is the foundation of modern Bayesian computation.

```
# Figure 4.1 - Monte Carlo Approximation (Gamma(68,45))
# Histograms + KDEs for S = 10, 100, 1000; true density in gray

set.seed(8670)
library(ggplot2)

# Posterior: Gamma(shape=68, rate=45)
shape_post <- 68
rate_post  <- 45

# MC samples
S_list <- c(10, 100, 1000)
mc_df <- do.call(rbind, lapply(S_list, function(S) {
  data.frame(theta = rgamma(S, shape = shape_post, rate = rate_post),
             S = factor(S, levels = S_list))
}))

# Grid for true density (choose a sensible range around the mass)
xgrid <- seq(
  qgamma(0.001, shape = shape_post, rate = rate_post),
  qgamma(0.999, shape = shape_post, rate = rate_post),
  length.out = 600
)
```

```r
true_df <- data.frame(
  theta = xgrid,
  dens  = dgamma(xgrid, shape = shape_post, rate = rate_post)
)

# Plot
ggplot(mc_df, aes(x = theta)) +
  # histogram (density scale so it overlays with densities)
  geom_histogram(aes(y = after_stat(density)),
                 bins = 18, color = "black", fill = "white") +
  # KDE from MC samples
  geom_density(linewidth = 1.1) +
  # True density (gray)
  geom_line(data = true_df, aes(x = theta, y = dens),
            linewidth = 1.2, color = "gray50") +
  facet_wrap(~ S, nrow = 1, scales = "free_y") +
  labs(
    title = "Monte Carlo Approximation",
  subtitle = paste(
    "Histograms and KDEs for Monte Carlo samples",
    "True Gamma(68, 45) density shown in gray",
    sep = "\n"
  ),
    x = expression(theta),
    y = "Density"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(size = 18, face = "bold"),
    plot.subtitle = element_text(size = 13),
    strip.text = element_text(size = 14, face = "bold")
  )
```



4.5 Numerical Evaluation

We now compare Monte Carlo approximations to quantities that can be computed analytically in this conjugate example.

Suppose

$$Y_1, \dots, Y_n \mid \theta \sim \text{Poisson}(\theta), \quad \theta \sim \text{Gamma}(a, b).$$

After observing $y_1, \dots, y_n$ with $\sum y_i = sy$ and sample size n, the posterior distribution is

$$\theta \mid y \sim \text{Gamma}(a + sy,\ b + n).$$

For the birth-rate example:

- $a = 2$
- $b = 1$
- $sy = 66$
- $n = 44$

Posterior:

$$\theta \mid y \sim \text{Gamma}(68, 45).$$

Posterior mean:

$$\mathbb{E}[\theta \mid y] = \frac{a + sy}{b + n} = \frac{68}{45} = 1.51.$$

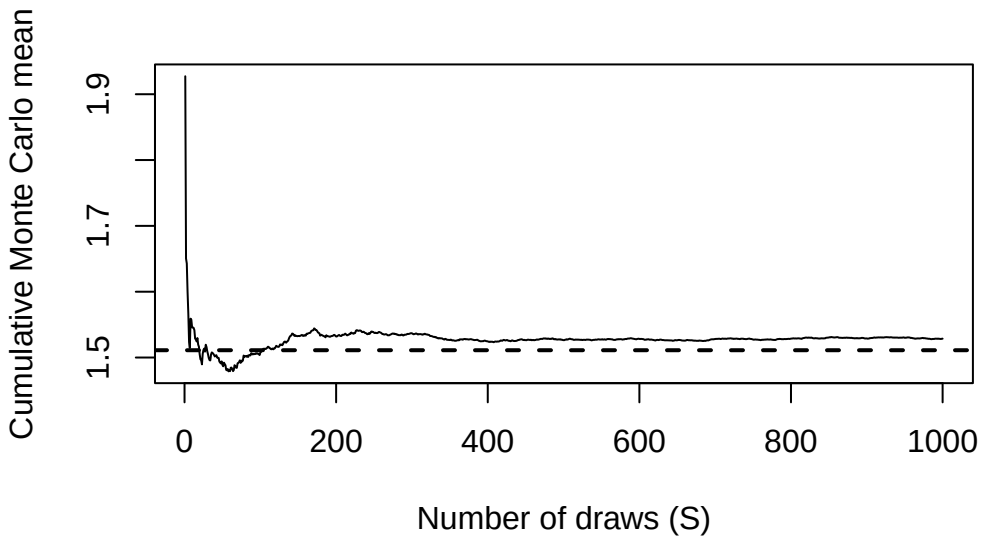
4.5.1 Monte Carlo in R

```r
set.seed(8670)

## Posterior parameters
a  <- 2
b  <- 1
sy <- 66
n  <- 44

shape_post <- a + sy
rate_post  <- b + n

## Exact quantities
mean_exact <- shape_post / rate_post
p_exact    <- pgamma(1.75, shape = shape_post, rate = rate_post)
ci_exact   <- qgamma(c(0.025, 0.975),
                     shape = shape_post,
                     rate  = rate_post)

## Monte Carlo samples
theta_mc10   <- rgamma(10,   shape_post, rate_post)
theta_mc100  <- rgamma(100,  shape_post, rate_post)
theta_mc1000 <- rgamma(1000, shape_post, rate_post)

## Function to summarize MC output
mc_summary <- function(theta_sample) {
  c(
    Mean        = mean(theta_sample),
    Prob_less   = mean(theta_sample < 1.75),
    CI_lower    = quantile(theta_sample, 0.025),
```

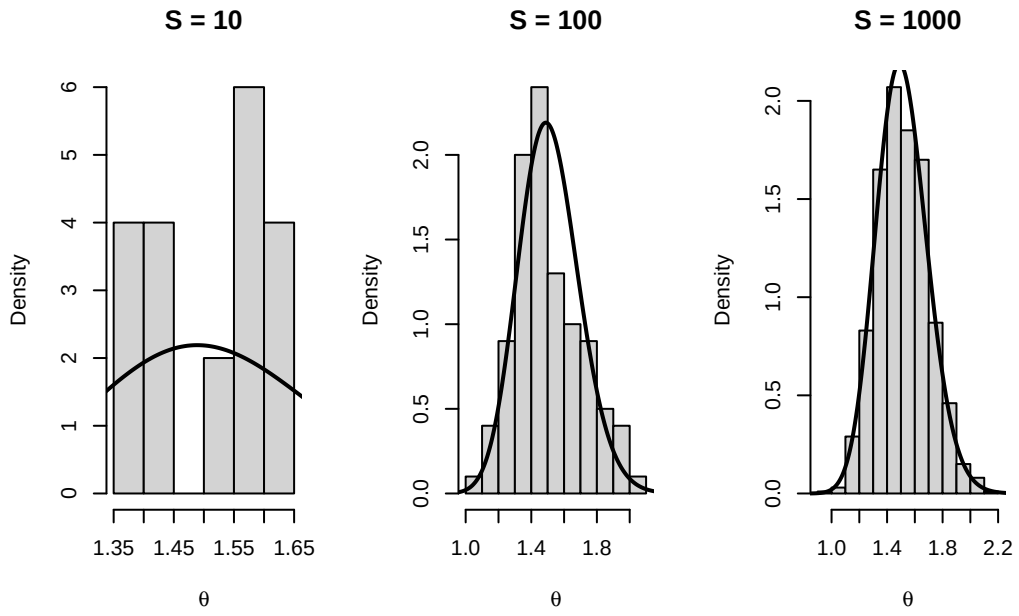
```
    CI_upper    = quantile(theta_sample, 0.975)
  )
}

## Build comparison table
results <- rbind(
  Exact   = c(Mean      = mean_exact,
              Prob_less = p_exact,
              CI_lower  = ci_exact[1],
              CI_upper  = ci_exact[2]),

  MC_10   = mc_summary(theta_mc10),
  MC_100  = mc_summary(theta_mc100),
  MC_1000 = mc_summary(theta_mc1000)
)

round(results, 4)
```

```
          Mean Prob_less CI_lower CI_upper
Exact   1.5111    0.8998   1.1734   1.8908
MC_10   1.5077    1.0000   1.3628   1.6228
MC_100  1.5033    0.8500   1.1536   1.9305
MC_1000 1.5180    0.8860   1.1810   1.8937
```

```
Smax <- 1000
theta_seq <- rgamma(Smax, shape = shape_post, rate = rate_post)
cum_mean  <- cumsum(theta_seq) / seq_along(theta_seq)

plot(cum_mean, type="l",
     xlab="Number of draws (S)",
     ylab="Cumulative Monte Carlo mean")
abline(h = mean_exact, lty = 2, lwd = 2)
```




```
xgrid <- seq(0.5, 2.5, length.out = 400)
true_pdf <- dgamma(xgrid, shape = shape_post, rate = rate_post)

par(mfrow=c(1,3))

for (S in c(10,100,1000)) {
  x <- get(paste0("theta_mc", S))
  hist(x, prob=TRUE,
       main=paste0("S = ", S),
       xlab=expression(theta),
       border="black")
  lines(xgrid, true_pdf, lwd=2)
}
```




```
par(mfrow=c(1,1))
```

4.5.2 There are much more about the MC method

- Variance reduction methods
- Antithetic variates
- Control variables
- Importance Sampling
- Stratified Sampling
- Stratified Importance Sampling
- etc...

You may refer to my notes in Chapter 4 in the Computational Methods in Statistics Course.

Here, we are gonig to shift our attention to a special kind of MC method, namely the MCMC method.

4.6 Markov Chain Monte Carlo (MCMC)

This Chapter borrows materials from Chapter 4 in Hoff (2009) and Chapter 4 in Computational Methods in Statistics Course

5 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

Part I

Appendix

6 Appendix: Introduction to R

6.1 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.

6.2 IDE

6.2.1 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.

6.2.2 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.

6.2.3 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.

6.3 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.

6.4 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`.

6.5 R Help

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

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

6.6 R Packages

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

6.6.1 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")
```

6.6.2 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")
```

6.6.3 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")
```

6.6.4 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.

6.7 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.

6.8 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
```

6.9 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