Bayesian Statistical Methods

CHAPMAN & HALL/CRC

Texts in Statistical Science Series

Recently Published Titles

Theory of Stochastic Objects
Probability, Stochastic Processes and Inference
Athanasios Christou Micheas

Linear Models and the Relevant Distributions and Matrix Algebra
David A. Harville

An Introduction to Generalized Linear Models, Fourth Edition
Annette J. Dobson and Adrian G. Barnett

Graphics for Statistics and Data Analysis with R
Kevin J. Keen

Statistics in Engineering, Second Edition
With Examples in MATLAB and R
Andrew Metcalfe, David A. Green, Tony Greenfield, Mahayaudin Mansor, Andrew Smith, and Jonathan Tuke

A Computational Approach to Statistical Learning
Taylor Arnold, Michael Kane, and Bryan W. Lewis

Introduction to Probability, Second Edition
Joseph K. Blitzstein and Jessica Hwang

A Computational Approach to Statistical Learning
Taylor Arnold, Michael Kane, and Bryan W. Lewis

Theory of Spatial Statistics
A Concise Introduction
M.N.M van Lieshout

Bayesian Statistical Methods
Brian J. Reich, Sujit K. Ghosh

Time Series
A Data Analysis Approach Using R
Robert H. Shumway, David S. Stoffer

The Analysis of Time Series
An Introduction, Seventh Edition
Chris Chatfield, Haipeng Xing

Probability and Statistics for Data Science
Math + R + Data
Norman Matloff

Sampling
Design and Analysis, Second Edition
Sharon L. Lohr

Practical Multivariate Analysis, Sixth Edition
Abdelmonem Afifi, Susanne May, Robin A. Donatello, Virginia A. Clark

For more information about this series, please visit: https://www.crcpress.com/go/textsseries

Bayesian Statistical Methods

Brian J. Reich
Sujit K. Ghosh

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20190313

International Standard Book Number-13: 978-0-815-37864-8 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Michelle, Sophie, Swagata, and Sabita

Contents

Preface

Bayesian methods are standard in various fields of science including biology, engineering, finance and genetics, and thus they are an essential addition to an analyst's toolkit. In this book, we cover the material we deem indispensable for a practicing Bayesian data analyst. The book covers the most common statistical methods including the t-test, multiple linear regression, mixed models and generalized linear models from a Bayesian perspective and includes many examples and code to implement the analyses. To illustrate the flexibility of the Bayesian approach, the later chapters explore advanced topics such as nonparametric regression, missing data and hierarchical models. In addition to these important practical matters, we provide sufficient depth so that the reader can defend his/her analysis and argue the relative merits of Bayesian and classical methods.

The book is intended to be used as a one-semester course for advanced undergraduate and graduate students. At North Carolina State University (NCSU) this book is used for a course comprised of undergraduate statistics majors, non-statistics graduate students from all over campus (e.g., engineering, ecology, psychology, etc.) and students from the Masters of Science in Statistics Program. Statistics PhD students take a separate course that covers much of the same material but at a more theoretical and technical level. We hope this book and associated computing examples also serve as a useful resource to practicing data analysts. Throughout the book we have included case studies from several fields to illustrate the flexibility and utility of Bayesian methods in practice.

It is assumed that the reader is familiar with undergraduate-level calculus including limits, integrals and partial derivatives and some basic linear algebra. Derivation of some key results are given in the main text when this helps to communicate ideas, but the vast majority of derivations are relegated to the Appendix for interested readers. Knowledge of introductory statistical concepts through multiple regression would also be useful to contextualize the material, but this background is not assumed and thus not required. Fundamental concepts are covered in detail but with references held to a minimum in favor of clarity; advanced concepts are described concisely at a high level with references provided for further reading.

The book begins with a review of probability in the first section of Chapter 1. A solid understanding of this material is essential to proceed through the book, but this section may be skipped for readers with the appropriate background. The remainder of Chapter 1 through Chapter 5 form the core

of the book. Chapter 1 introduces the central concepts of and motivation for Bayesian inference. Chapter 2 provides ways to select the prior distribution which is the genesis of a Bayesian analysis. For all but the most basic methods, advanced computing tools are required, and Chapter 3 covers these methods with the most weight given to Markov chain Monte Carlo which is used for the remainder of the book. Chapter 4 applies these tools to common statistical models including multiple linear regression, generalized linear models and mixed effects models (and more complex regression models in Section 4.5 which may be skipped if needed). After cataloging numerous statistical methods in Chapter 4, Chapter 5 treats the problem of selecting an appropriate model for a given dataset and verifying and validating that the model fits the data well. Chapter 6 introduces hierarchical modeling as a general framework to extend Bayesian methods to complex problems, and illustrates this approach using detailed case studies. The final chapter investigates the theoretical properties of Bayesian methods, which is important to justify their use but can be omitted if the course is intended for non-PhD students.

The choice of software is crucial for any modern textbook or statistics course. We elected to use `R` as the software platform due to its immense popularity in the statistics community, and access to online tutorials and assistance. Fortunately, there are now many options within `R` to conduct a Bayesian analysis and we compare several including `JAGS`, `BUGS`, `STAN` and `NIMBLE`. We selected the package `JAGS` as the primary package for no particularly strong reason other than we have found it works well for the courses taught at our university. In our assessment, `JAGS` provides the nice combination of ease of use, speed and flexibility for the size and complexity of problems we consider. A repository of code and datasets used in the book is available at

https://bayessm.org/.

Throughout the book we use `R/JAGS`, but favor conceptual discussions over computational details and these concepts transcend software. The course webpage also includes latex/beamer slides.

We wish to thank our NCSU colleagues Kirkwood Cloud, Qian Guan, Margaret Johnson, Ryan Martin, Krishna Pacifici and Ana-Maria Staicu for providing valuable feedback. We also thank the students in Bayesian courses taught at NCSU for their probing questions that helped shape the material in the book. Finally, we thank our families and friends for their enduring support, as exemplified by the original watercolor painting by Swagata Sarkar that graces the front cover of the book.

1

Basics of Bayesian inference

CONTENTS

1.1 Probability background

Understanding basic probability theory is essential to any study of statistics. Generally speaking, the field of probability assumes a mathematical model for the process of interest and uses the model to compute the probability of events (e.g., what is the probability of flipping five straight heads using a fair coin?). In contrast, the field of statistics uses data to refine the probability model and test hypotheses related to the underlying process that generated the data (e.g., given we observe five straight heads, can we conclude the coin is biased?). Therefore, probability theory is a key ingredient to a statistical analysis, and in this section we review the most relevant concepts of probability for a Bayesian analysis.

Before developing probability mathematically, we briefly discuss probability from a conceptual perspective. The objective is to compute the probability of an event, $\mathcal{A}$, denoted Prob($\mathcal{A}$). For example, we may be interested in the

probability that the random variable X (random variables are generally represented with capital letters) takes the specific value x (lower-case letter), denoted $\text{Prob}(X = x)$, or the probability that X will fall in the interval $[a, b]$, denoted $\text{Prob}(X \in [a, b])$. There are two leading interpretations of this statement: objective and subjective. An objective interpretation views $\text{Prob}(\mathcal{A})$ as a purely mathematical statement. A frequentist interpretation is that if we repeated the experiment many times and recorded the sample proportion of the times $\mathcal{A}$ occurred, this proportion would eventually converge to the number $\text{Prob}(\mathcal{A}) \in [0, 1]$ as the number of samples increases. A subjective interpretation is that $\text{Prob}(\mathcal{A})$ represents an individual's degree of belief, which is often quantified in terms of the amount the individual would be willing to wager that $\mathcal{A}$ will occur. As we will see, these two conceptual interpretations of probability parallel the two primary statistical frameworks: frequentist and Bayesian. However, a Bayesian analysis makes use of both of these concepts.

1.1.1 Univariate distributions

The random variable X's support $\mathcal{S}$ is the smallest set so that $X \in \mathcal{S}$ with probability one. For example, if X the number of successes in n trials then $\mathcal{S} = \{0, 1, ..., n\}$. Probability equations differ based on whether $\mathcal{S}$ is a countable set: X is a discrete random variable if $\mathcal{S}$ is countable, and X is continuous otherwise. Discrete random variables can have a finite (rainy days in the year) or an infinite (number lightning strikes in a year) number of possible outcomes as long as the number is countable, e.g., a random count $X \in \mathcal{S} = \{0, 1, 2, ...\}$ has an infinite but countable number of possible outcomes and is thus discrete. An example of a continuous random variable is the amount of rain on a given day which can be any real non-negative number and so $\mathcal{S} = [0, \infty)$.

1.1.1.1 Discrete distributions

For a discrete random variable, the probability mass function (PMF) $f(x)$ assigns a probability to each element of X's support, that is,

$$\text{Prob}(X = x) = f(x). \tag{1.1}$$

A PMF is valid if all probabilities are non-negative, $f(x) \geq 0$, and sum to one, $\sum_{x \in \mathcal{S}} f(x) = 1$. The PMF can also be used to compute probabilities of more complex events by summing over the PMF. For example, the probability that X is either x_1 or x_2, i.e., $X \in \{x_1, x_2\}$, is

$$\text{Prob}(X = x_1 \text{ or } X = x_2) = f(x_1) + f(x_2). \tag{1.2}$$

Generally, the probability of the event that X falls in a set $\mathcal{S}' \subset \mathcal{S}$ is the sum over elements in $\mathcal{S}'$,

$$\text{Prob}(X \in \mathcal{S}') = \sum_{x \in \mathcal{S}'} f(x). \tag{1.3}$$

Using this fact defines the cumulative distribution function (CDF)

$$F(x) = \text{Prob}(X \leq x) = \sum_{c \leq x} f(c). \tag{1.4}$$

A PMF is a function from the support of X to the probability of events. It is often useful to summarize the function using a few interpretable quantities such as the mean and variance. The expected value or mean value is

$$\text{E}(X) = \sum_{x \in \mathcal{S}} x f(x) \tag{1.5}$$

and measures the center of the distribution. The variance measures the spread around the mean via the expected squared deviation from the center of the distribution,

$$\text{Var}(X) = \text{E}\{[X - \text{E}(X)]^2\} = \sum_{x \in \mathcal{S}} [x - E(X)]^2 f(x). \tag{1.6}$$

The variance is often converted to the standard deviation $\text{SD}(X) = \sqrt{\text{Var}(X)}$ to express the variability on the same scale as the random variable.

The central concept of statistics is that the PMF and its summaries (such as the mean) describe the population of interest, and a statistical analysis uses a sample from the population to estimate these functions of the population. For example, we might take a sample of size n from the population. Denote the i^{th} sample value as $X_i \sim f$ ("$\sim$" means "is distributed as"), and $X_1, ..., X_n$ as the complete sample. We might then approximate the population mean $\text{E}(X)$ with the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, the probability of an outcome $f(x)$ with the sample proportion of the n observations that equal x, and the entire PMF $f(x)$ with a sample histogram. However, even for a large sample, $\bar{X}$ will likely not equal $\text{E}(X)$, and if we repeat the sampling procedure again we might get a different $\bar{X}$ while $\text{E}(X)$ does not change. The distribution of a statistic, i.e., a summary of the sample, such as $\bar{X}$ across random samples from the population is called the statistic's sampling distribution.

A statistical analysis to infer about the population from a sample often proceeds under the assumption that the population belongs to a parametric family of distributions. This is called a parametric statistical analysis. In this type of analysis, the entire PMF is assumed to be known up to a few unknown parameters denoted $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ (or simply θ if there is only $p = 1$ parameter). We then denote the PMF as $f(x|\boldsymbol{\theta})$. The vertical bar "|" is read as "given," and so $f(x|\boldsymbol{\theta})$ gives the probability that $X = x$ given the parameters $\boldsymbol{\theta}$. For example, a common parametric model for count data with $\mathcal{S} = \{0, 1, 2, ...\}$ is the Poisson family. The Poisson PMF with unknown parameter $\theta \geq 0$ is

$$\text{Prob}(X = x|\theta) = f(x|\theta) = \frac{\exp(-\theta)\theta^x}{x!}. \tag{1.7}$$

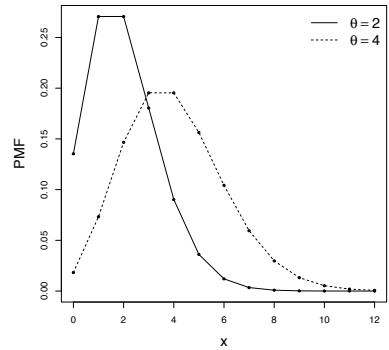

FIGURE 1.1
Poisson probability mass function. Plot of the PMF $f(x|\theta) = \frac{\exp(-\theta)\theta^x}{x!}$ for $\theta = 2$ and $\theta = 4$. The PMF is connected by lines for visualization, but the probabilities are only defined for $x = \{0, 1, 2, ...\}$.

Clearly $f(x|\theta) > 0$ for all $x \in \mathcal{S}$ and it can be shown that $\sum_{x=0}^{\infty} f(x|\theta) = 1$ so this is a valid PMF. As shown in Figure 1.1, changing the parameter θ changes the PMF and so the Poisson is not a single distribution but rather a family of related distributions indexed by θ.

A parametric assumption greatly simplifies the analysis because we only have to estimate a few parameters and we can compute the probability of any x in $\mathcal{S}$. Of course, this assumption is only useful if the assumed distribution provides a reasonable fit to the observed data, and thus a statistician needs a large catalog of distributions to be able to find an appropriate family for a given analysis. Appendix A.1 provides a list of parametric distributions, and we discuss a few discrete distributions below.

Bernoulli: If X is binary, i.e., $\mathcal{S} = \{0, 1\}$, then X follows a Bernoulli(θ) distribution. A binary random variable is often used to model the result of a trial where a success is recorded as a one and a failure as zero. The parameter $\theta \in [0, 1]$ is the success probability $\text{Prob}(X = 1|\theta) = \theta$, and to be a valid PMF we must have the failure probability $\text{Prob}(X = 0|\theta) = 1 - \theta$. These two cases can be written concisely as

$$f(x|\theta) = \theta^x(1 - \theta)^{1-x}. \tag{1.8}$$

This gives mean

$$\text{E}(X|\theta) = \sum_{x=0}^{1} xf(x|\theta) = f(1|\theta) = \theta \tag{1.9}$$

and variance $\text{Var}(X|\theta) = \theta(1 - \theta)$.

Binomial: The binomial distribution is a generalization of the Bernoulli to

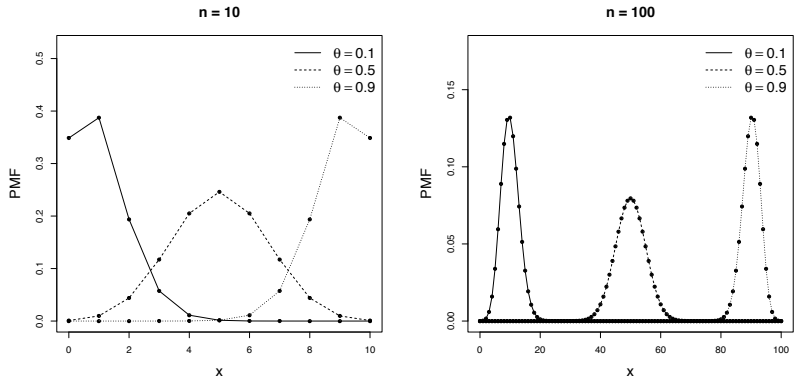

FIGURE 1.2
Binomial probability mass function. Plot of the PMF $f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ for combinations of the number of trials n and the success probability θ. The PMF is connected by lines for visualization, but the probabilities are only defined for $x = \{0, 1, ..., n\}$.

the case of $n \geq 1$ independent trials. Specifically, if $X_1, ..., X_n$ are the binary results of the n independent trials each with success probability θ (so that $X_i \sim \text{Bernoulli}(\theta)$ for all $i = 1, ..., n$) and $X = \sum_{i=1}^{n} X_i$ is the total number of successes, then X's support is $\mathcal{S} = \{0, 1, ..., n\}$ and $X \sim \text{Binomial}(n, \theta)$. The PMF is

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \tag{1.10}$$

where $\binom{n}{x}$ is the binomial coefficient. This gives mean and variance $\text{E}(X|\theta) = n\theta$ and $\text{Var}(X|\theta) = n\theta(1-\theta)$. It is certainly reasonable that the expected number of successes in n trials is n times the success probability for each trial, and Figure 1.2 shows that the variance is maximized with $\theta = 0.5$ when the outcome of each trial is the least predictable. Appendix A.1 provides other distributions with support $\mathcal{S} = \{0, 1, ..., n\}$ including the discrete uniform and beta-binomial distributions.

Poisson: When the support is the counting numbers $\mathcal{S} = \{0, 1, 2, ...\}$, a common model is the Poisson PMF defined above (and plotted in Figure 1.1). The Poisson PMF can be motivated as the distribution of the number of events that occur in a time interval of length T if the events are independent and equally likely to occur at any time with rate θ/T events per unit of time. The mean and variance are $\text{E}(X|\theta) = \text{Var}(X|\theta) = \theta$. Assuming that the mean and variance are equal is a strong assumption, and Appendix A.1 provides alternatives with support $\mathcal{S} = \{0, 1, 2, ...\}$ including the geometric and negative binomial distributions.

1.1.1.2 Continuous distributions

The PMF does not apply for continuous distributions with support $\mathcal{S}$ that is a subinterval of the real line. To see this, assume that X is the daily rainfall (inches) and can thus be any non-negative real number, $\mathcal{S} = [0, \infty)$. What is the probability that X is exactly $\pi/2$ inches? Well, within some small range around $\pi/2$, $\mathcal{T} = (\pi/2 - \epsilon, \pi/2 + \epsilon)$ with say $\epsilon = 0.001$, it seems reasonable to assume that all values in $\mathcal{T}$ are equally likely, say $\text{Prob}(X = x) = q$ for all $x \in \mathcal{T}$. But since there are an uncountable number of values in $\mathcal{T}$ when we sum the probability over the values in $\mathcal{T}$ we get infinity and thus the probabilities are invalid unless $q = 0$. Therefore, for continuous random variables $\text{Prob}(X = x) = 0$ for all x and we must use a more sophisticated method for assigning probabilities.

Instead of defining the probability of outcomes directly using a PMF, for continuous random variables we define probabilities indirectly through the cumulative distribution function (CDF)

$$F(x) = \text{Prob}(X \leq x). \tag{1.11}$$

The CDF can be used to compute probabilities for any interval, e.g., in the rain example $\text{Prob}(X \in \mathcal{S}) = \text{Prob}(X < \pi/2 + \epsilon) - \text{Prob}(X < \pi/2 - \epsilon) = F(\pi/2 + \epsilon) - F(\pi/2 - \epsilon)$, which converges to zero as ϵ shrinks if F is a continuous function. Defining the probability of X falling in an interval resolves the conceptual problems discussed above, because it is easy to imagine the proportion of days with rainfall in an interval converging to a non-zero value as the sample size increases.

The probability on a small interval is

$$\text{Prob}(x - \epsilon < X < x + \epsilon) = F(x + \epsilon) - F(x - \epsilon) \approx 2\epsilon f(x) \tag{1.12}$$

where $f(x)$ is the derivative of $F(x)$ and is called the probability density function (PDF). If $f(x)$ is the PDF of X, then the probability of $X \in [a, b]$ is the area under the density curve between a and b (Figure 1.3),

$$\text{Prob}(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)dx. \tag{1.13}$$

The distributions of random variables are usually defined via the PDF. To ensure that the PDF produces valid probability statements we must have $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)dx = 1$. Because $f(x)$ is not a probability, but rather a function used compute probabilities via integration, the PDF can be greater than one for some x so long as it integrates to one.

The formulas for the mean and variance of a continuous random variable resemble those for a discrete random variable except that the sum over the

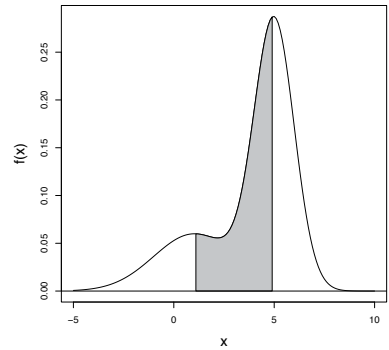

FIGURE 1.3
Computing probabilities using a PDF. The curve is a PDF $f(x)$ and the shaded area is $\text{Prob}(1 < X < 5) = \int_1^5 f(x)dx$.

PMF is replaced with an integral over the PDF,

$$\begin{aligned} \text{E}(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ \text{Var}(X) &= \text{E}\{[X - E(X)]^2\} = \int_{-\infty}^{\infty} [x - \text{E}(X)]^2 f(x)dx. \end{aligned}$$

Another summary that is defined for both discrete and continuous random variables (but is much easier to define in the continuous case) is the quantile function $Q(\tau)$. For $\tau \in [0, 1]$, $Q(\tau)$ is the solution to

$$\text{Prob}[X \leq Q(\tau)] = F[Q(\tau)] = \tau. \tag{1.14}$$

That is, $Q(\tau)$ is the value so that the probability of X being no larger than $Q(\tau)$ is τ. The quantile function is the inverse of the distribution function, $Q(\tau) = F^{-1}(\tau)$, and gives the median $Q(0.5)$ and a $(1 - \alpha)\%$ equal-tailed interval $[Q(\alpha/2), Q(1-\alpha/2)]$ so that $\text{Prob}[Q(\alpha/2) \leq X \leq Q(1-\alpha/2)] = 1-\alpha$.

Gaussian: As with discrete data, parametric models are typically assumed for continuous data and practitioners must be familiar with several parametric families. The most common parametric family with support $\mathcal{S} = (-\infty, \infty)$ is the normal (Gaussian) family. The normal distribution has two parameters, the mean $\text{E}(X|\boldsymbol{\theta}) = \mu$ and variance $\text{Var}(X|\boldsymbol{\theta}) = \sigma^2$, and the familiar bell-shaped PDF

$$f(x|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right], \tag{1.15}$$

where $\boldsymbol{\theta} = (\mu, \sigma^2)$. The Gaussian distribution is famous because of the central

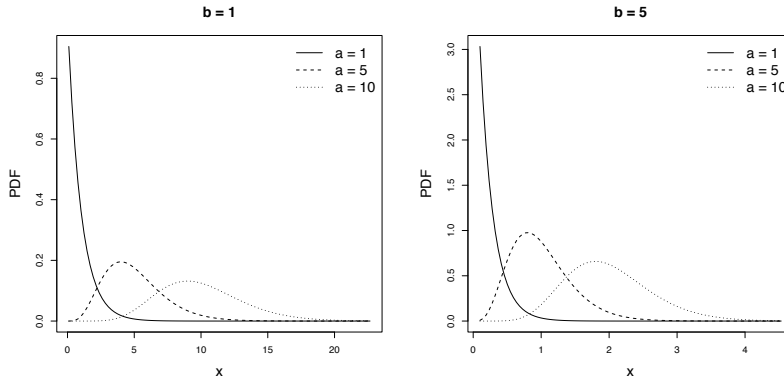

FIGURE 1.4
Plots of the gamma PDF. Plots of the gamma density function $f(x|\boldsymbol{\theta}) = \frac{b^a}{\Gamma(a)} x^a \exp(-bx)$ for several combinations of a and b.

limit theorem (CLT). The CLT applies to the distribution of the sample mean $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$, where $X_1, ..., X_n$ are independent samples from some distribution $f(x)$. The CLT says that under fairly general conditions, for large n the distribution of $\bar{X}_n$ is approximately normal even if $f(x)$ is not. Therefore, the Gaussian distribution is a natural model for data that are defined as averages, but can be used for other data as well. Appendix A.1 gives other continuous distributions with $\mathcal{S} = (-\infty, \infty)$ including the double exponential and student-t distributions.

Gamma: The gamma distribution has $\mathcal{S} = [0, \infty)$. The PDF is

$$f(x|\boldsymbol{\theta}) = \begin{cases} \frac{b^a}{\Gamma(a)} x^a \exp(-bx) & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{1.16}$$

where Γ is the gamma function and $a > 0$ and $b > 0$ are the two parameters in $\boldsymbol{\theta} = (a, b)$. Beware that the gamma PDF is also written with b in the denominator of the exponential function, but we use the parameterization above. Under the parameterization in (1.16) the mean is a/b and the variance is a/b^2. As shown in Figure 1.4, a is the shape parameter and b is the scale. Setting $a = 1$ gives the exponential distribution with PDF $f(x|\boldsymbol{\theta}) = b \exp(-bx)$ which decays exponentially from the origin and large a gives approximately a normal distribution. Varying b does not change the shape of the PDF but only affects its spread. Appendix A.1 gives other continuous distributions with $\mathcal{S} = [0, \infty)$ including the inverse-gamma distribution.

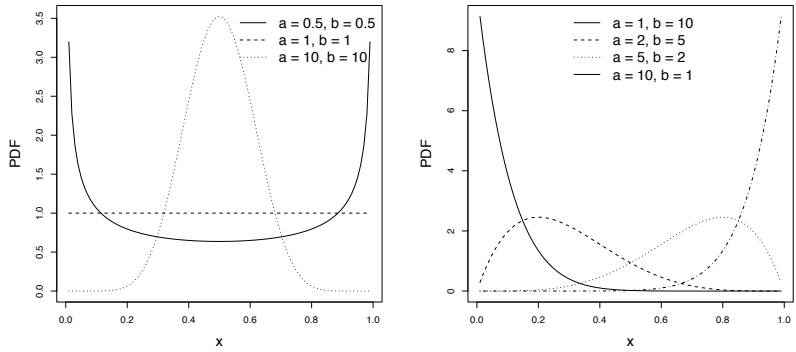

FIGURE 1.5
Plots of the beta PDF. Plots of the beta density function $f(x|\boldsymbol{\theta}) = \frac{\Gamma(a,b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ for several a and b.

Beta: The beta distribution has $\mathcal{S} = [0,1]$ and PDF

$$f(x|\boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(a,b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} & x \in [0,1] \\ 0 & x < 0 \text{ or } \ x > 1, \end{cases} \tag{1.17}$$

where $a > 0$ and $b > 0$ are the two parameters in $\boldsymbol{\theta} = (a,b)$. As shown in Figure 1.5, the beta distribution is quite flexible and can be left-skewed, right-skewed, or symmetric. The beta distribution also includes the uniform distribution $f(x) = 1$ for $x \in [0,1]$ by setting $a = b = 1$.

1.1.2 Multivariate distributions

Most statistical analyses involve multiple variables with the objective of studying relationships between variables. To model relationships between variables we need multivariate extensions of mass and density functions. Let $X_1, ..., X_p$ be p random variables, $\mathcal{S}_j$ be the support of X_j so that $X_j \in \mathcal{S}_j$, and $\mathbf{X} = (X_1, ..., X_p)$ be the random vector. Table 1.1 describes the joint distribution between the $p = 2$ variables: X_1 indicates that the patient has primary health outcome and X_2 indicates the patient has a side effect. If all variables are discrete, then the joint PMF is

$$\text{Prob}(X_1 = x_1, ..., X_p = x_p) = f(x_1, ..., x_p). \tag{1.18}$$

To be a valid PMF, f must be non-negative $f(x_1, ..., x_p) \geq 0$ and sum to one $\sum_{x_1 \in \mathcal{S}_1}, ..., \sum_{x_p \in \mathcal{S}_1} f(x_1, ..., x_p) = 1$.

As in the univariate case, probabilities for continuous random variables are

TABLE 1.1
Hypothetical joint PMF. This PMF $f(x_1, x_2)$ gives the probabilities that a patient has a positive ($X_1 = 1$) or negative ($X_1 = 0$) primary health outcome and the patient having ($X_2 = 1$) or not having ($X_2 = 0$) a negative side effect.

		X_2		
		0	1	$f_1(x_1)$
X_1	0	0.06	0.14	0.20
	1	0.24	0.56	0.80
	$f_2(x_2)$	0.30	0.70	

computed indirectly via the PDF $f(x_1, ..., x_p)$. In the univariate case, probabilities are computed as the area under the density curve. For $p > 1$, probabilities are computed as the volume under the density surface. For example, for $p = 2$ random variables, the probability of $a_1 < X_1 < b_1$ and $a_2 < X_2 < b_2$ is

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) dx_1 dx_2. \tag{1.19}$$

This gives the probability of the random vector $\mathbf{X} = (X_1, X_2)$ lying in the rectangle defined by the endpoints a_1, b_1, a_2, and b_2. In general, the probability of the random vector $\mathbf{X}$ falling in region $\mathcal{A}$ is the p-dimensional integral $\int_{\mathcal{A}} f(x_1, ..., x_p) dx_1, ..., dx_p$. As an example, consider the bivariate PDF on the unit square with $f(x_1, x_2) = 1$ for $\mathbf{X}$ with $x_1 \in [0, 1]$ and $x_2 \in [0, 1]$ and $f(x_1, x_2) = 0$ otherwise. Then $\text{Prob}(X_1 < .5, X_2 < .1) = \int_0^{0.5} \int_0^{0.1} f(x_1, x_2) dx_1 dx_2 = 0.05$.

1.1.3 Marginal and conditional distributions

The marginal and conditional distributions that follow from a multivariate distribution are key to a Bayesian analysis. To introduce these concepts we assume discrete random variables, but extensions to the continuous case are straightforward by replacing sums with integrals. Further, we assume a bivariate PMF with $p = 2$. Again, extensions to high dimensions are conceptually straightforward by replacing sums over one or two dimensions with sums over $p - 1$ or p dimensions.

The marginal distribution of X_j is simply the distribution of X_j if we consider only a univariate analysis of X_j and disregard the other variable. Denote $f_j(x_j) = \text{Prob}(X_j = x_j)$ as the marginal PMF of X_j. The marginal distribution is computed by summing over the other variable in the joint PMF

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2) \quad \text{and} \quad f_2(x_2) = \sum_{x_1} f(x_1, x_2). \tag{1.20}$$

These are referred to as the marginal distributions because in a two-way table

such as Table 1.1, the marginal distributions are the row and column totals of the joint PMF that appear along the table's margins.

As with any univariate distribution, the marginal distribution can be summarized with its mean and variance,

$$\begin{aligned}\mu_j &= \mathrm{E}(X_j) = \sum_{x_j \in \mathcal{S}_j} x_j f_j(x_j) = \sum_{x_1}\sum_{x_2} x_j f(x_1, x_2) \\ \sigma_j^2 &= \mathrm{Var}(X_j) = \sum_{x_j} (x_j - \mu_j)^2 f_j(x_j) = \sum_{x_1}\sum_{x_2} (x_j - \mu_j)^2 f(x_1, x_2).\end{aligned}$$

The marginal mean and variance measure the center and spread of the marginal distributions, respectively, but do not capture the relationship between the two variables. The most common one-number summary of the joint relationship is covariance, defined as

$$\begin{aligned}\sigma_{12} &= \mathrm{Cov}(X_1, X_2) = \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \sum_{x_1}\sum_{x_2} (x_1 - \mu_1)(x_2 - \mu_2) f(x_1, x_2).\end{aligned} \tag{1.21}$$

The covariance is often hard to interpret because it depends on the scale of both X_1 and X_2, i.e., if we double X_1 we double the covariance. Correlation is a scale-free summary of the joint relationship, $\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$. In vector notation, the mean of the random vector $\mathbf{X}$ is $\mathrm{E}(\mathbf{X}) = (\mu_1, \mu_2)^T$, the covariance matrix is $\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$, and the correlation matrix is $\mathrm{Cor}(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix}$. Generalizing for $p > 2$, the mean vector becomes $\mathrm{E}(\mathbf{X}) = (\mu_1, ..., \mu_p)^T$ and the covariance matrix becomes the symmetric $p \times p$ matrix with diagonal elements $\sigma_1^2, ..., \sigma_p^2$ and (i, j) off-diagonal element σ_{ij}.

While the marginal distributions sum over columns or rows of a two-way table, the conditional distributions focus only on a single column or row. The conditional distribution of X_1 given that the random variables X_2 is fixed at x_2 is denoted $f_{1|2}(x_1|X_2 = x_2)$ or simply $f(x_1|x_2)$. Referring to Table 1.1, the knowledge that $X_2 = x_2$ restricts the domain to a single column of the two-way table. However, the probabilities in a single column do not define a valid PMF because their sum is less than one. We must rescale these probabilities to sum to one by dividing the column total, which we have previously defined as $f_2(x_2)$. Therefore, the general expression for the conditional distributions of $X_1|X_2 = x_2$, and similarly $X_2|X_1 = x_1$, is

$$f_{1|2}(x_1|X_2 = x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \quad \text{and} \quad f_{2|1}(x_2|X_1 = x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}. \tag{1.22}$$

Atlantic hurricanes example: Table 1.2 provides the counts of Atlantic hurricanes that made landfall between 1990 and 2016 tabulated by their intensity category (1–5) and whether they hit the US or elsewhere. Of course, these are only sample proportions and not true probabilities,

TABLE 1.2
Table of the Atlantic hurricanes that made landfall between 1990 and 2016. The counts are tabulated by their maximum Saffir–Simpson intensity category and whether they made landfall in the US or elsewhere. The counts are downloaded from `http://www.aoml.noaa.gov/hrd/hurdat/`.

(a) Counts

	Category					
	1	2	3	4	5	Total
US	14	13	10	1	1	39
Not US	46	19	20	17	3	105
Total	60	32	30	18	4	144

(b) Sample proportions

	Category					
	1	2	3	4	5	Total
US	0.0972	0.0903	0.0694	0.0069	0.0069	0.2708
Not US	0.3194	0.1319	0.1389	0.1181	0.0208	0.7292
Total	0.4167	0.2222	0.2083	0.1250	0.0278	1.0000

but for this example we treat Table 1.2b as the joint PMF of location, $X_1 \in \{\text{US}, \text{Not US}\}$, and intensity category, $X_2 \in \{1, 2, 3, 4, 5\}$. The marginal distribution of X_1 is given in the final column and is simply the row sums of the joint PMF. The marginal probability of a hurricane making landfall in the US is $\text{Prob}(X_1 = \text{US}) = 0.2708$, which is the proportion calculation as if we had never considered the storms' category. Similarly, the column sums are the marginal probability of intensity averaging over location, e.g., $\text{Prob}(X_2 = 5) = 0.0278$.

The conditional distributions tell us about the relationship between the two variables. For example, the marginal probability of a hurricane reaching category 5 is 0.0278, but given that the storm hits the US, the conditional distribution is slightly lower, $f_{2|1}(5|\text{US}) = \text{Prob}(X_1 = \text{US}, X_2 = 5)/\text{Prob}(X_1 = \text{US}) = 0.0069/0.2708 = 0.0255$. By definition, the conditional probabilities sum to one,

$$\begin{aligned} f_{2|1}(1|\text{US}) &= \frac{0.0972}{0.2708} = 0.3589,\ f_{2|1}(2|\text{US}) = \frac{0.0903}{0.2708} = 0.3334 \\ f_{2|1}(3|\text{US}) &= \frac{0.0694}{0.2708} = 0.2562,\ f_{2|1}(4|\text{US}) = \frac{0.0069}{0.2708} = 0.0255 \\ f_{2|1}(5|\text{US}) &= \frac{0.0069}{0.2708} = 0.0255. \end{aligned}$$

Given that a storm hits the US, the probability of a category 2 or 3 storm

increases, while the probability of a category 4 or 5 storm decreases, and so there is a relationship between landfall location and intensity.

Independence example: Consider the joint PMF of the primary health outcome (X_1) and side effect (X_2) in Table 1.1. In this example, the marginal probability of a positive primary health outcome is $\text{Prob}(X_1 = 1) = 0.80$, as is the conditional probability $f_{1|2}(1|X_2 = 1) = 0.56/0.70 = 0.80$ given the patient has the side effect and the conditional probability $f_{1|2}(1|0) = 0.24/0.30 = 0.80$ given the patient does not have the side effect. In other words, both with and without knowledge of side effect status, the probability of a positive health outcome is 0.80, and thus side effect is not informative about the primary health outcome. This is an example of two random variables that are independent.

Generally, X_1 and X_2 are independent if and only if the joint PMF (or PDF for continuous random variables) factors into the product of the marginal distributions,

$$f(x_1, x_2) = f_1(x_1) f_2(x_2). \tag{1.23}$$

From this expression it is clear that if X_1 and X_2 are independent then $f_{1|2}(x_1|X_2 = x_2) = f(x_1, x_2)/f_2(x_2) = f_1(x_1)$, and thus X_2 is not informative about X_1. A special case of joint independence is if all marginal distributions are same, $f_j(x) = f(x)$ for all j and x. In this case, we say that X_1 and X_2 are independent and identically distributed ("iid"), which is denoted $X_j \overset{iid}{\sim} f$. If variables are not independent then they are said to be dependent.

Multinomial: Parametric families are also useful for multivariate distributions. A common parametric family for discrete data is the multinomial, which, as the name implies, is a generalization of the binomial. Consider the case of n independent trials where each trial results in one of p possible outcomes (e.g., $p = 3$ and each result is either win, lose or draw). Let $X_j \in \{0, 1, ..., n\}$ be the number of trails that resulted in outcome j, and $\mathbf{X} = (X_1, ..., X_p)$ be the vector of counts. If we assume that θ_j is the probability of outcome j for each trial, with $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ and $\sum_{j=1}^{p} \theta_j = 1$, then $\mathbf{X}|\boldsymbol{\theta} \sim \text{Multinomial}(n, \boldsymbol{\theta})$ with

$$f(x_1, ..., x_p) = \frac{n!}{x_1! \cdot ... \cdot x_p!} \theta_1^{x_1} \cdot ... \cdot \theta_p^{x_p} \tag{1.24}$$

where $n = \sum_{j=1}^{p} x_j$. If there are only $p = 2$ categories then this would be a binomial experiment $X_2 \sim \text{Binomial}(n, \theta_2)$.

Multivariate normal: The multivariate normal distribution is a generalization of the normal distribution to $p > 1$ random variables. For $p = 2$, the bivariate normal has five parameters, $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$: a mean parameter for each variable $\text{E}(X_j) = \mu_j$, a variance for each variable $\text{Var}(X_j) = \sigma_j^2 > 0$, and the correlation $\text{Cor}(X_1, X_2) = \rho \in [-1, 1]$. The density function is

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2(1-\rho^2)}\right], \tag{1.25}$$

where $z_j = (x_j - \mu_j)/\sigma_j$. As shown in Figure 1.6 the density surface is elliptical with center determined by $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and shape determined by the covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}$.

A convenient feature of the bivariate normal distribution is that the marginal and conditional distributions are also normal. The marginal distribution of X_j is Gaussian with mean μ_j and variance σ_j^2. The conditional distribution, shown in Figure 1.7, is

$$X_2|X_1 = x_1 \sim \text{Normal}\left[\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), (1-\rho^2)\sigma_2^2\right]. \tag{1.26}$$

If $\rho = 0$ then the conditional distribution is the marginal distribution, as expected. If $\rho > 0$ ($\rho < 0$) then the conditional mean increases (decreases) with x_1. Also, the conditional variance $(1-\rho^2)\sigma_2^2$ is less than the marginal variance σ_2^2, especially for ρ near -1 or 1, and so conditioning on X_1 reduces uncertainty in X_2 when there is strong correlation.

The multivariate normal PDF for $p > 2$ is most concisely written using matrix notation. The multivariate normal PDF for the random vector $\mathbf{X}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is

$$f(\mathbf{X}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right] \tag{1.27}$$

where $|\mathbf{A}|$, $\mathbf{A}^T$ and $\mathbf{A}^{-1}$ are the determinant, transpose and inverse, respectively, of the matrix $\mathbf{A}$. From this expression it is clear that the contours of the log PDF are elliptical. All conditional and marginal distributions are normal, as are all linear combinations $\sum_{j=1}^p w_j X_j$ for any $w_1, ..., w_p$.

1.2 Bayes' rule

As the name implies, Bayes' rule (or Bayes' theorem) is fundamental to Bayesian statistics. However, this rule is a general result from probability and follows naturally from the definition of a conditional distribution. Consider two random variables X_1 and X_2 with joint PMF (or PDF as the result holds for both discrete and continuous data) density function $f(x_1, x_2)$. The definition of a conditional distribution gives $f(x_1|x_2) = f(x_2, x_1)/f(x_2)$ and $f(x_2|x_1) = f(x_1, x_2)f(x_1)$. Combining these two expressions gives Bayes' rule

$$f(x_2|x_1) = \frac{f(x_1|x_2)f(x_2)}{f(x_1)}. \tag{1.28}$$

This result is useful as a means to reverse conditioning from $X_1|X_2$ to $X_2|X_1$, and also indicates the need to define a joint distribution for this inversion to be valid.

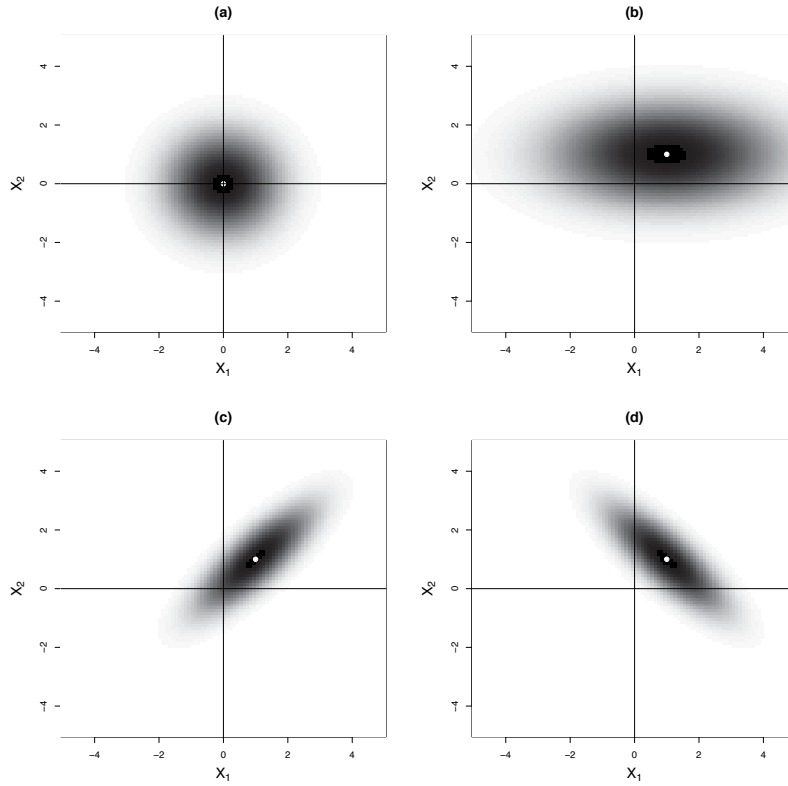

FIGURE 1.6
Plots of the bivariate normal PDF. Panel (a) plots the bivariate normal PDF for $\boldsymbol{\mu} = (0, 0)$, $\sigma_1 = 1$, $\sigma_2 = 1$ and $\rho = 0$. The other panels modify Panel (a) as follows: (b) has $\mu = (1, 1)$ and $\sigma_1 = 2$, (c) has $\mu = (1, 1)$ and $\rho = 0.8$, and (d) has $\mu = (1, 1)$ and $\rho = -0.8$. The plots are shaded according to the PDF with white indicating the PDF near zero and black indicating the areas with highest PDF; the white dot is the mean vector $\boldsymbol{\mu}$.

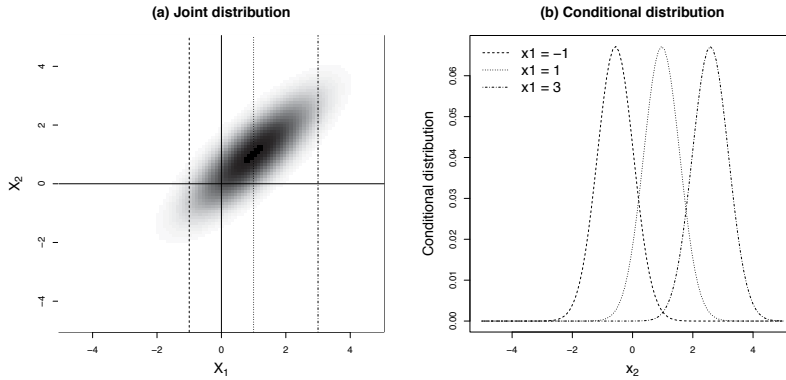

FIGURE 1.7
Plots of the joint and conditional bivariate normal PDF. Panel (a) plots the bivariate normal PDF for $\boldsymbol{\mu} = (1, 1)$, $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.8$. The plot is shaded according to the PDF with white indicating the PDF near zero and black indicating the areas with highest PDF; the vertical lines represent $x_1 = -1$, $x_1 = 1$ and $x_1 = 3$. The conditional distribution of $X_2|X_1 = x_1$ for these three values of x_1 are plotted in Panel (b).

1.2.1 Discrete example of Bayes' rule

You have a scratchy throat and so you go to the doctor who administers a rapid strep throat test. Let $Y \in \{0, 1\}$ be the binary indicator of a positive test, i.e., $Y = 1$ if the test is positive for strep and $Y = 0$ if the test is negative. The test is not perfect. The false positive rate $p \in [0, 1]$ is the probability of testing positive if you do not have strep, and the false negative rate $q \in [0, 1]$ is the probability of testing negative given that you actually have strep. To express these probabilities mathematically we must define the true disease status $\theta \in \{0, 1\}$, where $\theta = 1$ if you are truly infected and $\theta = 0$ otherwise. This unknown variable we hope to estimate is called a parameter. Given these error rates and the definition of the model parameter, the data distribution can be written

$$\text{Prob}(Y = 1|\theta = 0) = p \qquad \text{and} \qquad \text{Prob}(Y = 0|\theta = 1) = q. \tag{1.29}$$

Generally, the PMF (or PDF) of the observed data given the model parameters is called the likelihood function.

To formally analyze this problem we must determine which components should be treated as random variables. Is the test result Y a random variable? Before the exam, Y is clearly random and (1.29) defines its distribution. This is aleatoric uncertainty because the results may differ if we repeat the test. However, after the learning of the test results, Y is determined and you must

decide how to proceed given the value of Y at hand. In this sense, Y is known and no longer random at the analysis stage.

Is the true disease status θ a random variable? Certainly θ is not a random variable in the sense that it changes from second-to-second or minute-to-minute, and so it is reasonable to assume that the true disease status is a fixed quantity for the purpose of this analysis. However, because our test is imperfect we do not know θ. This is epistemic uncertainty because θ is a quantity that we could theoretically know, but at the analysis stage we do not and cannot know θ using only noisy data. Despite our uncertainty about θ, we have to decide what to do next and so it is useful to quantify our uncertainty using the language of probability. If the test is reliable and p and q are both small, then in light of a positive test we might conclude that θ is more likely to be one than zero. But how much more likely? Twice as likely? Three times? In Bayesian statistics we quantify uncertainty about fixed but unknown parameters using probability theory by treating them as random variables. As (1.28) suggests, for formal inversion of conditional probabilities we would need to treat both variables as random.

The probabilities in (1.29) supply the distribution of the test result given disease status, $Y|\theta$. However, we would like to quantify uncertainty in the disease status given the test results, that is, we require the distribution of $\theta|Y$. Since this is the uncertainty distribution after collecting the data this is referred to as the posterior distribution. As discussed above, Bayes' rule can be applied to reverse the order of conditioning,

$$\text{Prob}(\theta = 1|Y = 1) = \frac{\text{Prob}(Y = 1|\theta = 1)\text{Prob}(\theta = 1)}{\text{Prob}(Y = 1)}, \tag{1.30}$$

where the marginal probability $\text{Prob}(Y = 1)$ is

$$\sum_{\theta=0}^{1} f(1, \theta) = \text{Prob}(Y = 1|\theta = 1)\text{Prob}(\theta = 1) + \text{Prob}(Y = 1|\theta = 0)\text{Prob}(\theta = 0). \tag{1.31}$$

To apply Bayes' rule requires specifying the unconditional probability of having strep throat, $\text{Prob}(\theta = 1) = \pi \in [0, 1]$. Since this is the probability of infection before we conduct the test, we refer to this as the prior probability. We can then compute the posterior using Bayes' rule,

$$\text{Prob}(\theta = 1|Y = 1) = \frac{(1 - q)\pi}{(1 - q)\pi + p(1 - \pi)}. \tag{1.32}$$

To understand this equation consider a few extreme scenarios. Assuming the error rates p and q are not zero or one, if $\pi = 1$ ($\pi = 0$) then the posterior probability of $\theta = 1$ ($\theta = 0$) is one for any value of Y. That is, if we have no prior uncertainty then the imperfect data does not update the prior. Conversely, if the test is perfect and $q = p = 0$ then for any prior $\pi \in (0, 1)$ the posterior probability that θ is Y is one. That is, with perfect data the prior

TABLE 1.3
Strep throat data. Number of patients that are truly positive and tested positive in the rapid strep throat test data taken from Table 1 of [26].

	Truly positive, test positive	Truly positive, test negative	Truly negative, test positive	Truly negative, test negative
Children	80	38	23	349
Adults	43	10	14	261
Total	123	48	37	610

is irrelevant. Finally, if $p = q = 1/2$, then the test is a random coin flip and the posterior is the prior $\text{Prob}(\theta = 1|Y) = \pi$.

For a more realistic scenario we use the data in Table 1.3 taken from [26]. We plug in the sample error rates from these data for $p = 37/(37 + 610) = 0.057$ and $q = 48/(48 + 123) = 0.281$. Of course these data represent only a sample and the sample proportions are not exactly the true error rates, but for illustration we assume these error rates are correct. Then if we assume prior probability of disease is $\pi = 0.5$, the posterior probabilities are $\text{Prob}(\theta = 1|Y = 0) = 0.230$ and $\text{Prob}(\theta = 1|Y = 1) = 0.927$. Therefore, beginning with a prior probability of 0.5, a negative test moves the probability down to 0.230 and a positive test increases the probability to 0.927.

Of course, in reality the way individuals process test results is complicated and subjective. If you have had strep many times before and you went to the doctor because your current symptoms resemble previous bouts with the disease, then perhaps your prior is $\pi = 0.8$ and the posterior is $\text{Prob}(\theta = 1|Y = 1) = 0.981$. On the other hand, if you went to the doctor only at the urging of your friend and your prior probability is $\pi = 0.2$, then $\text{Prob}(\theta = 1|Y = 1) = 0.759$.

This simple example illustrates a basic Bayesian analysis. The objective is to compute the posterior distribution of the unknown parameters θ. The posterior has two ingredients: the likelihood of the data given the parameters and the prior distribution. Selection of these two distributions is thus largely the focus of the remainder of this book.

1.2.2 Continuous example of Bayes' rule

Let $\theta \in [0, 1]$ be the proportion of the population in a county that has health insurance. It is known that the proportion varies across counties following a $\text{Beta}(a, b)$ distribution and so the prior is $\theta \sim \text{Beta}(a, b)$. We take a sample of size $n = 20$ from your county and assume that the number of respondents with insurance, $Y \in \{0, 1, ..., n\}$, is distributed as $Y|\theta \sim \text{Binomial}(n, \theta)$. Joint

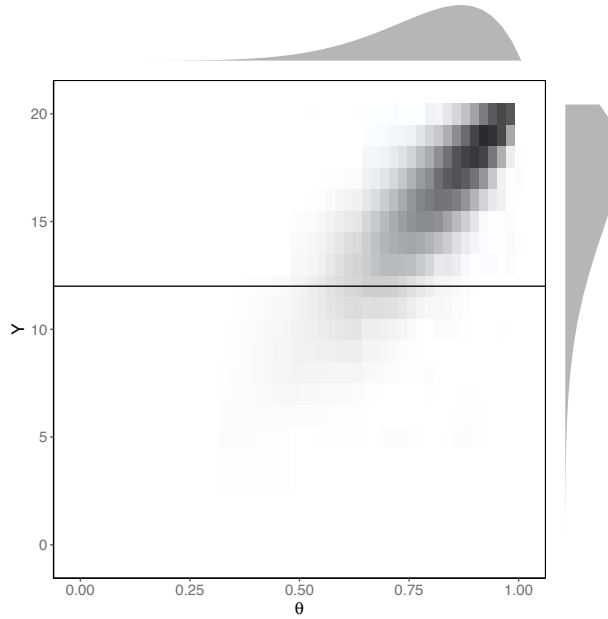

FIGURE 1.8
Joint distribution for the beta-binomial example. Plot of $f(\theta, y)$ for the example with $\theta \sim \text{Beta}(8, 2)$ and $Y|\theta \sim \text{Binomial}(20, \theta)$. The marginal distributions $f(\theta)$ (top) and $f(y)$ (right) are plotted in the margins. The horizontal line is the $Y = 12$ line.

probabilities for θ and Y can be computed from

$$\begin{aligned} f(\theta, y) &= f(y|\theta) f(\theta) \\ &= \left\{ \binom{n}{y} \theta^y (1-\theta)^{n-y} \right\} \left\{ \frac{\Gamma(a,b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \right\} \\ &= c\theta^{y+a-1} (1-\theta)^{n-y+b-1} \end{aligned}$$

where $c = \binom{n}{y} \Gamma(a, b)/[\Gamma(a)\Gamma(b)]$ is a constant that does not depend on θ. Figure 1.8 plots $f(\theta, y)$ and the marginal distributions for θ and Y. By the way we have defined the problem, the marginal distribution of θ, $f(\theta)$, is a Beta(a, b) PDF, which could also be derived by summing $f(\theta, y)$ over y. The marginal distribution of Y plotted on the right of Figure 1.8 is $f(y) = \int_0^1 f(\theta, y) d\theta$. In this case the marginal distribution of Y follows a beta-binomial distribution, but as we will see this is not needed in the Bayesian analysis.

In this problem we are given the unconditional distribution of the disease rate (prior) and the distribution of the sample given the true proportion (likelihood), and Bayes' rule gives the (posterior) distribution of the true proportion

given the sample. Say we observe $Y = 12$. The horizontal line in Figure 1.8 traces over the conditional distribution $f(\theta|Y = 12)$. The conditional distribution is centered around the sample proportion $Y/n = 0.60$ but has non-trivial mass from 0.4 to 0.8. More formally, the posterior is

$$\begin{aligned} f(\theta|y) &= \frac{f(y|\theta)f(\theta)}{f(y)} \\ &= \left[\frac{c}{f(y)}\right]\theta^{y+a-1}(1-\theta)^{n-y+b-1} \\ &= C\theta^{A-1}(1-\theta)^{B-1} \end{aligned} \tag{1.33}$$

where $C = c/f(y)$, $A = y + a$, and $B = n - y + b$.

We note the resemblance between $f(\theta|y)$ and the PDF of a Beta(A, B) density. Both include $\theta^{A-1}(1-\theta)^{B-1}$ but differ in the normalizing constant, C for $f(\theta|y)$ compared to $\Gamma(A, B)/[\Gamma(A)\Gamma(B)]$ for the Beta(A, B) PDF. Since both $f(\theta|y)$ and the Beta(A, B) PDF are proper, they both integrate to one, and thus

$$\int_0^1 C\theta^{A-1}(1-\theta)^{B-1}d\theta = \int_0^1 \frac{\Gamma(A,B)}{\Gamma(A)\Gamma(B)}\theta^{A-1}(1-\theta)^{B-1}d\theta = 1 \tag{1.34}$$

and so

$$C\int_0^1 \theta^{A-1}(1-\theta)^{B-1}d\theta = \frac{\Gamma(A,B)}{\Gamma(A)\Gamma(B)}\int_0^1 \theta^{A-1}(1-\theta)^{B-1}d\theta \tag{1.35}$$

and thus $C = \Gamma(A, B)/[\Gamma(A)\Gamma(B)]$. Therefore, $f(\theta|y)$ is in fact the Beta(A, B) PDF and $\theta|Y = y \sim$ Beta$(y + a, n - y + b)$.

Dealing with the normalizing constant makes posterior calculations quite tedious. Fortunately this can often be avoided by discarding terms that do not involve the parameter of interest and comparing the remaining terms with known distributions. The derivation above can be simplified to (using "$\propto$" to mean "proportional to")

$$f(\theta|y) \propto f(y|\theta)f(\theta) \propto \theta^{(y+a)-1}(1-\theta)^{(n-y+b)-1}$$

and immediately concluding that $\theta|Y = y \sim$ Beta$(y + a, n - y + b)$.

Figure 1.9 plots the posterior distribution for two priors and $Y \in \{0, 5, 10, 15, 20\}$. The plots illustrate how the posterior combines information from the prior and the likelihood. In both plots, the peak of the posterior distribution increases with the observation Y. Comparing the plots shows that the prior also contributes to the posterior. When we observe $Y = 0$ successes, the posterior under the Beta(8,2) prior (left) is pulled from zero to the right by the prior (thick line). Under the Beta(1,1), i.e., the uniform prior, when $Y = 0$ the posterior is concentrated around $\theta = 0$.

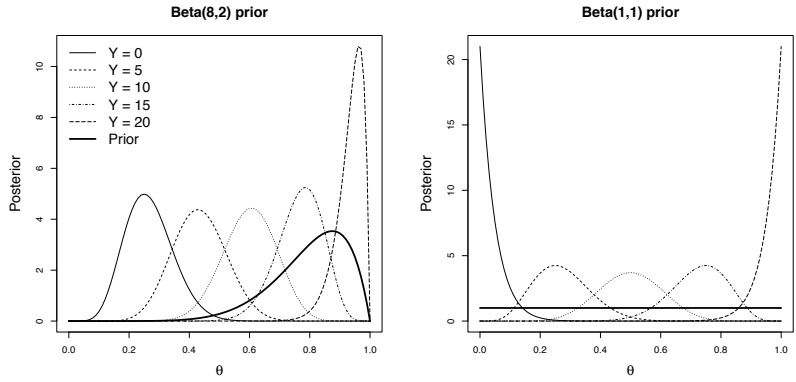

FIGURE 1.9
Posterior distribution for the beta-binomial example. The thick lines are the beta prior for success probability θ and the thin lines are the posterior assuming $Y|\theta \sim \text{Binomial}(20, \theta)$ for various values of Y.

1.3 Introduction to Bayesian inference

A parametric statistical analysis models the random process that produced the data, $\mathbf{Y} = (Y_1, ..., Y_n)$, in terms of fixed but unknown parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$. The PDF (or PMF) of the data given the parameters, $f(\mathbf{Y}|\boldsymbol{\theta})$, is called the likelihood function and links the observed data with the unknown parameters. Statistical inference is concerned with the inverse problem of using the likelihood function to estimate $\boldsymbol{\theta}$. Of course, if the data are noisy then we cannot perfectly estimate $\boldsymbol{\theta}$, and a Bayesian quantifies uncertainty about the unknown parameters by treating them as random variables. Treating $\boldsymbol{\theta}$ as a random variable requires specifying the prior distribution, $\pi(\boldsymbol{\theta})$, which represents our uncertainty about the parameters before we observe the data.

If we view $\boldsymbol{\theta}$ as a random variable, we can apply Bayes' tule to obtain the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\theta} \propto f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.36}$$

The posterior is proportional to the likelihood times the prior, and quantifies the uncertainty about the parameters that remain after accounting for prior knowledge and the new information in the observed data.

Table 1.4 establishes the notation we use throughout for the prior, likelihood and posterior. We will not adhere to the custom (e.g., Section 1.1) that random variables are capitalized because in a Bayesian analysis more often

TABLE 1.4
Notation used throughout the book for distributions involving the parameter vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ and data vector $\mathbf{Y} = (Y_1, ..., Y_n)$.

Prior density of $\boldsymbol{\theta}$:	$\pi(\boldsymbol{\theta})$
Likelihood function of $\mathbf{Y}$ given $\boldsymbol{\theta}$:	$f(\mathbf{Y}\|\boldsymbol{\theta})$
Marginal density of $\mathbf{Y}$:	$m(\mathbf{Y}) = \int f(\mathbf{Y}\|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$
Posterior density of $\boldsymbol{\theta}$ given $\mathbf{Y}$:	$p(\boldsymbol{\theta}\|\mathbf{Y}) = f(\mathbf{Y}\|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/m(\mathbf{Y})$

than not it is the parameters that are the random variables, and capital Greek letters, e.g., $\text{Prob}(\Theta = \theta)$, are unfamiliar to most readers. We will however follow the custom to use bold to represent vectors and matrices. Also, assume independence unless otherwise noted. For example, if we say "the priors are $\theta_1 \sim \text{Uniform}(0, 1)$ and $\theta_2 \sim \text{Gamma}(1, 1)$," you should assume that θ_1 and θ_2 have independent priors.

The Bayesian framework provides a logically consistent framework to use all available information to quantify uncertainty about model parameters. However, to apply Bayes' rule requires specifying the prior distribution and the likelihood function. *How do we pick the prior?* In many cases prior knowledge from experience, expert opinion or similar studies is available and can be used to specify an informative prior. It would be a waste to discard this information. In other cases where prior information is unavailable, then the prior should be uninformative to reflect this uncertainty. For instance, in the beta-binomial example in Section 1.2 we might use a uniform prior that puts equal mass on all possible parameter values. The choice of prior distribution is subjective, i.e., driven by the analyst's past experience and personal preferences. If a reader does not agree with your prior then they are unlikely to be persuaded by your analysis. Therefore, the prior, especially an informative prior, should be carefully justified, and a sensitivity analysis comparing the posteriors under different priors should be presented.

How do we pick the likelihood? The likelihood function is the same as in a classical analysis, e.g., a maximum likelihood analysis. The likelihood function for multiple linear regression is the product of Gaussian PDFs defined by the model

$$Y_i|\boldsymbol{\theta} \overset{indep}{\sim} \text{Normal}\left(\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j, \sigma^2\right) \tag{1.37}$$

where X_{ij} is the value of the j^{th} covariate for the i^{th} observation and $\boldsymbol{\theta} = (\beta_0, ..., \beta_p, \sigma^2)$ are the unknown parameters. A thoughtful application of multiple linear regression must consider many questions, including

- Which covariates to include?
- Are the errors Gaussian? Independent? Do they have equal variance?

- Should we include quadratic or interaction effects?
- Should we consider a transformation of the response (e.g., model $\log(Y_i)$)?
- Which observations are outliers? Should we remove them?
- How should we handle the missing observations?
- What p-value threshold should be used to define statistical significance?

As with specifying the prior, these concerns are arguably best resolved using subjective subject-matter knowledge. For example, while there are statistical methods to select covariates (Chapter 5), a more reliable strategy is to ask a subject-matter expert which covariates are the most important to include, at least as an initial list to be refined in the statistical analysis. As another example, it is hard to determine (without a natural ordering as in times series data) whether the observations are independent without consulting someone familiar with the data collection and the study population. Other decisions are made based on visual inspections of the data (such as scatter plots and histograms of the residuals) or ad hoc rules of thumb (threshold on outliers' z-scores or p-values for statistical significance). Therefore, in a typical statistical analysis there many subjective choices to be made, and the choice of prior is far from the most important.

Bayesian statistical methods are often criticized as being subjective. Perhaps an objective analysis that is free from personal preferences or beliefs is an ideal we should strive for (and this is the aim of objective Bayesian methods, see Section 2.3), but it is hard to make the case that non-Bayesian methods are objective, and it can be argued that almost any scientific knowledge and theories are subjective in nature. In an interesting article by Press and Tanur (2001), the authors cite many scientific theories (mainly from physics) where subjectivity played a major role and they concluded "*Subjectivity occurs, and should occur, in the work of scientists; it is not just a factor that plays a minor role that we need to ignore as a flaw...*" and they further added that "*Total objectivity in science is a myth. Good science inevitably involves a mixture of subjective and objective parts.*" The Bayesian inferential framework provides a logical foundation to accommodate objective and subjective parts involved in data analysis. Hence, a good scientific practice would be to state upfront all assumptions and then make an effort to validate such assumptions using the current data or preferable future test cases. There is nothing wrong to have a subjective but reasonably flexible model as long as we can exhibit some form of sensitivity analysis when the assumptions of the model are mildly violated.

In addition to explicitly acknowledging subjectivity, another important difference between Bayesian and frequentist (classical) methods is their notion of uncertainty. While a Bayesian considers only the data at hand, a frequentist views uncertainty as arising from repeating the process that generated the data many times. That is, a Bayesian might give a posterior probability that the population mean μ (a parameter) is positive given the data we have observed,

whereas a frequentist would give a probability that the sample mean $\bar{Y}$ (a statistic) exceeds a threshold given a specific value of the parameters if we repeated the experiment many times (as is done when computing a p-value).

The frequentist view of uncertainty is well-suited for developing procedures that have desirable error rates when applied broadly. This is reasonable in many settings. For instance, a regulatory agency might want to advocate statistical procedures that ensure only a small proportion of the medications made available to the public have adverse side effects. In some cases however it is hard to see why repeating the sampling is a useful thought experiment. For example, [14] study the relationship between a region's climate and the type of religion that emerged from that region. Assuming the data set consists of the complete list of known cultures, it is hard to imagine repeating the process that led to these data as it would require replaying thousands of years of human history.

Bayesians can and do study the frequentist properties. This is critical to build trust in the methods. If a Bayesian weather forecaster gives the posterior predictive 95% interval every day for a year, but at the end of the year these intervals included the observed temperature only 25% of the time, then the forecaster would lose all credibility. It turns out that Bayesian methods often have desirable frequentist properties, and Chapter 7 examines these properties.

Developing Bayesian methods with good frequentist properties is often called calibrated Bayes (e.g., [52]). According to Rubin [52, 71]: "*The applied statistician should be Bayesian in principle and calibrated to the real world in practice - appropriate frequency calculations help to define such a tie... frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events.*"

1.4 Summarizing the posterior

The final output of a Bayesian analysis is the posterior distribution of the model parameters. The posterior contains all the relevant information from the data and the prior, and thus all statistical inference should be based on the posterior distribution. However, when there are many parameters, the posterior distribution is a high-dimensional function that is difficult to display graphically, and for complicated statistical models the mathematical form of the posterior may be challenging to work with. In this section, we discuss some methods to summarize a high-dimensional posterior with low-dimensional summaries.

1.4.1 Point estimation

One approach to summarizing the posterior is to use a point estimate, i.e., a single value that represents the best estimate of the parameters given the data and (for a Bayesian analysis) the prior. The posterior mean, median and mode are all sensible choices. Thinking of the Bayesian analysis as a procedure that can be applied to any dataset, the point estimator is an example of an *estimator*, i.e., a function that takes the data as input and returns an estimate of the parameter of interest. Bayesian estimators such as the posterior mean can then be seen as competitors to other estimators such as the sample mean estimator for a population mean or a sample variance for a population variance, or more generally as a competitor to the maximum likelihood estimator. We study the properties of these estimators in Chapter 7.

A common point estimator is the maximum a posteriori (MAP) estimator, defined as the value that maximizes the posterior (i.e., the posterior mode),

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \log[p(\boldsymbol{\theta}|\mathbf{Y})] = \arg\max_{\boldsymbol{\theta}} \log[f(\mathbf{Y}|\boldsymbol{\theta})] + \log[\pi(\boldsymbol{\theta})]. \quad (1.38)$$

The second equality holds because the normalizing constant $m(\mathbf{Y})$ does not depend on the parameters and thus does not affect the optimization.

If the prior is uninformative, i.e., mostly flat as a function of the parameters, then the MAP estimator should be similar to the maximum likelihood estimator (MLE)

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} \log[f(\mathbf{Y}|\boldsymbol{\theta})]. \quad (1.39)$$

In fact, this relationship is often used to intuitively justify maximum likelihood estimation. The addition of the log prior $\log[\pi(\boldsymbol{\theta})]$ in (1.38) can be viewed a regularization or penalty term to add stability or prior knowledge.

Point estimators are often useful as fast methods to estimate the parameters for purpose of making predictions. However, a point estimate alone does not quantify uncertainty about the parameters. Sections 1.4.2 and 1.4.3 provide more thorough summaries of the posterior for univariate and multivariate problems, respectively.

1.4.2 Univariate posteriors

A univariate posterior (i.e., from a model with $p = 1$ parameter) is best summarized with a plot because this retains all information about the parameter. Figure 1.10 shows a hypothetical univariate posterior with PDF centered at 0.8 and most of its mass on $\theta > 0.4$.

Point estimators such as the posterior mean or median summarize the center of the posterior, and should be accompanied by a posterior variance or standard deviation to convey uncertainty. The posterior standard deviation resembles a frequentist standard error in that if the posterior is approximately Gaussian then the posterior probability that the parameter is within two posterior standard deviation units of the posterior mean is roughly 0.95. However,

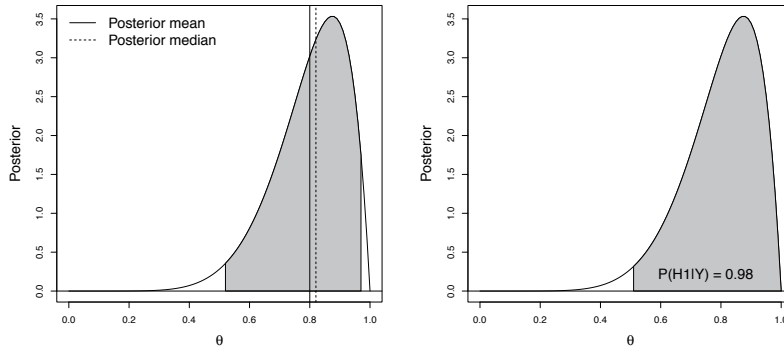

FIGURE 1.10
Summaries of a univariate posterior. The plot on the left gives the posterior mean (solid vertical line), median (dashed vertical line) and 95% equal-tailed interval (shaded area). The plot on the right shades the posterior probability of the hypothesis $H_1 : \theta > 0.5$.

the standard error is the standard deviation of the estimator (e.g., the sample mean) if we repeatedly sample different data sets and compute the estimator for each data set. In contrast, the posterior standard deviation quantifies uncertainty about the parameter given only the single data set under consideration.

The interval $\mathrm{E}(\theta|\mathbf{Y}) \pm 2\mathrm{SD}(\theta|Y)$ is an example of a credible interval. A $(1-\alpha)\%$ credible interval is any interval (l, u) so that $\mathrm{Prob}(l < \theta < u|\mathbf{Y}) = 1-\alpha$. There are infinitely many intervals with this coverage, but the easiest to compute is the equal-tailed interval with l and u set to the $\alpha/2$ and $1-\alpha/2$ posterior quantiles. An alternative is the highest posterior density interval which searches over all l and u to minimize the interval width $u-l$ while maintaining the appropriate posterior coverage. The HPD thus has the highest average posterior density of all intervals of the form (l, u) that have the nominal posterior probability. As opposed to equal-tailed intervals, the HPD requires an additional optimization step, but this can be computed using the `R` package `HDInterval`.

Interpreting a posterior credible interval is fairly straightforward. If (l, u) is a posterior 95% interval, this means "given my prior and the observed data, I am 95% sure that θ is between l and u." In a Bayesian analysis we express our subjective uncertainty about unknown parameters by treating them as random variables, and in this subjective sense it is reasonable to assign probabilities to θ. This is in contrast with frequentist confidence intervals which have a more nuanced interpretation. A confidence interval is a procedure that defines an

interval for a given data set in a way that ensures the procedure's intervals will include the true value 95% of the time when applied to random datasets.

The posterior distribution can also be used for hypothesis testing. Since the hypotheses are functions of the parameters, we can assign posterior probabilities to each hypothesis. Figure 1.10 (right) plots the posterior probability of the null hypothesis that $\theta < 0.5$ (white) and the posterior probability of the alternative hypothesis that $\theta > 0.5$ (shaded). These probabilities summarize the weight of evidence in support of each hypothesis, and can be used to guide future decisions. Hypothesis testing, and more generally model selection, is discussed in greater detail in Chapter 5.

Summarizing a univariate posterior using `R`: We have seen that if $Y|\theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$, then $\theta|Y \sim \text{Beta}(A, B)$, where $A = Y + a$ and $B = n - Y + b$. Listing 1.1 specifies a data set with $Y = 40$ and $n = 100$ and summarizes the posterior using `R`. Since the posterior is the beta density, the functions `dbeta`, `pbeta` and `qbeta` can be used to summarize the posterior. The posterior median and 95% credible set are 0.401 and (0.309, 0.498).

Monte Carlo sampling: Although univariate posterior distributions are best summarized by a plot, higher dimensional posterior distributions call for other methods such as Monte Carlo (MC) sampling and so we introduce this approach here. MC sampling draws S samples from the posterior, $\theta^{(1)}, ..., \theta^{(S)}$, and uses these samples to approximate the posterior. For example, the posterior mean is approximated using the mean of the S samples, $\text{E}(\theta|\mathbf{Y}) \approx \sum_{s=1}^{S} \theta^{(s)}/S$, the posterior 95% credible set is approximated using the 0.025 and 0.975 quantiles of the S samples, etc. Listing 1.1 provides an example of using MC sampling to approximate the posterior mean and 95% credible set.

1.4.3 Multivariate posteriors

Unlike the univariate case, a simple plot of the posterior will not suffice, especially for large p, because plotting in high dimensions is challenging. The typical remedy for this is to marginalize out other parameters and summarize univariate marginal distributions with plots, point estimates, and credible sets, and perhaps plots of a few bivariate marginal distributions (i.e., integrating over the other $p - 2$ parameters) of interest.

Consider the model $Y_i|\mu, \sigma \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ with independent priors $\mu \sim \text{Normal}(0, 100^2)$ and $\sigma \sim \text{Uniform}(0, 10)$ (other priors are discussed in Section 4.1.1). The likelihood

$$f(\mathbf{Y}|\mu, \sigma) \propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right] \propto \sigma^{-n} \exp\left[-\frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{2\sigma^2}\right] \quad (1.40)$$

factors as the product of n terms because the observations are assumed to be independent. The prior is $f(\mu, \sigma) = f(\mu)f(\sigma)$ because μ and σ have indepen-

Listing 1.1
Summarizing a univariate posterior in R.

```
# Load the data
 > n <- 100
 > Y <- 40
 > a <- 1
 > b <- 1
 > A <- Y + a
 > B <- n - Y + b

# Define a grid of points for plotting
 > theta <- seq(0,1,.001)

# Evaluate the density at these points
 > pdf <- dbeta(theta,A,B)

# Plot the posterior density
 > plot(theta,pdf,type="l",ylab="Posterior",xlab=expression(theta))

# Posterior mean
 > A/(A + B)
 [1] 0.4019608

# Posterior median (0.5 quantile)
 > qbeta(0.5,A,B)
 [1] 0.4013176

# Posterior probability P(theta<0.5|Y)
 > pbeta(0.5,A,B)
 [1] 0.976978

# Equal-tailed 95% credible interval
 > qbeta(c(0.025,0.975),A,B)
 [1] 0.3093085 0.4982559

# Monte Carlo approximation
 > S       <- 100000
 > samples <- rbeta(S,A,B)
 > mean(samples)
 [1] 0.402181
 >   quantile(samples,c(0.025,0.975))
      2.5%     97.5%
 0.3092051 0.4973871
```

TABLE 1.5
Bivariate posterior distribution. Summaries of the marginal posterior distributions for the model with $Y_i \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$, priors $\mu \sim \text{Normal}(0, 100^2)$ and $\sigma \sim \text{Uniform}(0, 10)$, and $n = 5$ observations $Y_1 = 2.68$, $Y_2 = 1.18$, $Y_3 = -0.97$, $Y_4 = -0.98$, $Y_5 = -1.03$.

Parameter	Posterior mean	Posterior SD	95% credible set
μ	0.17	1.31	(-2.49, 2.83)
σ	2.57	1.37	(1.10, 6.54)

dent priors and since $f(\sigma) = 1/10$ for all $\sigma \in [0, 10]$, the prior becomes

$$\pi(\mu, \sigma) \propto \exp\left(-\frac{\mu^2}{2 \cdot 100^2}\right) \tag{1.41}$$

for $\sigma \in [0, 10]$ and $f(\mu, \sigma) = 0$ otherwise. The posterior is proportional to the likelihood times the prior, and thus

$$p(\mu, \sigma|\mathbf{Y}) \propto \sigma^{-n} \exp\left[-\frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{2\sigma^2}\right] \exp\left(-\frac{\mu^2}{2 \cdot 100^2}\right) \tag{1.42}$$

for $\sigma \in [0, 10]$. Figure 1.11 plots this bivariate posterior assuming there are $n = 5$ observations: $Y_1 = 2.68$, $Y_2 = 1.18$, $Y_3 = -0.97$, $Y_4 = -0.98$, $Y_5 = -1.03$.

The two parameters in Figure 1.11 depend on each other. If $\sigma = 1.5$ (i.e., the conditional distribution traced by the horizontal line at $\sigma = 1.5$ in Figure 1.11) then the posterior of μ concentrates between -1 and 1, whereas if $\sigma = 3$ the posterior of μ spreads from -3 to 3. It is difficult to describe this complex bivariate relationship, so we often summarize the univariate marginal distributions instead. The marginal distributions

$$p(\mu|\mathbf{Y}) = \int_0^{10} p(\mu, \sigma|\mathbf{Y}) d\sigma \quad \text{and} \quad p(\sigma|\mathbf{Y}) = \int_{-\infty}^{\infty} p(\mu, \sigma|\mathbf{Y}) d\mu. \tag{1.43}$$

are plotted on the top (for μ) and right (for σ) of Figure 1.11; they are the row and columns sums of the joint posterior. By integrating over the other parameters, the marginal distribution of a parameter accounts for posterior uncertainty in the remaining parameters. The marginal distributions are usually summarized with point and interval estimates as in Table 1.5.

The marginal distributions and their summaries above were computed by evaluating the joint posterior (1.42) for values of (μ, σ) that form a grid (i.e., pixels in Figure 1.11) and then simply summing over columns or rows of the grid. This is a reasonable approximation for $p = 2$ variables but quickly becomes unfeasible as p increases. Thus, it was only with the advent of more

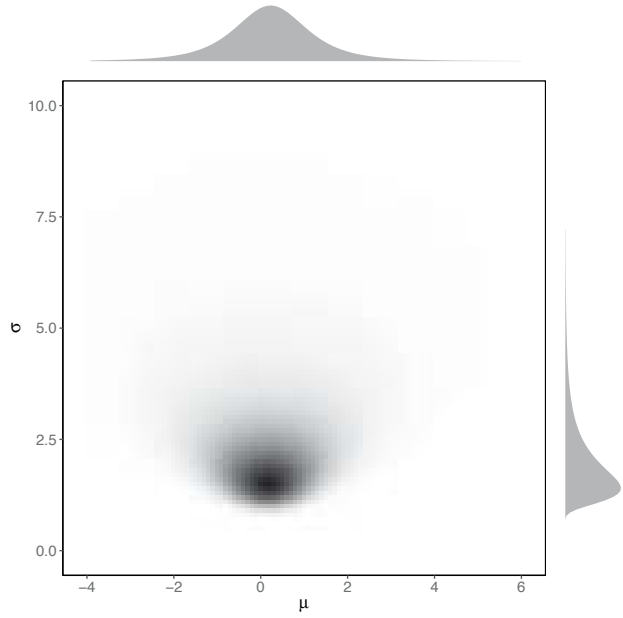

FIGURE 1.11
Bivariate posterior distribution. The bivariate posterior (center) and univariate marginal posteriors (top for μ and right for σ) for the model with $Y_i \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$, priors $\mu \sim \text{Normal}(0, 100^2)$ and $\sigma \sim \text{Uniform}(0, 10)$, and $n = 5$ observations $Y_1 = 2.68, Y_2 = 1.18, Y_3 = -0.97, Y_4 = -0.98, Y_5 = -1.03$.

efficient computing algorithms in the 1990s that Bayesian statistics became feasible for even medium-sized applications. These exciting computational developments are the subject of Chapter 3.

1.5 The posterior predictive distribution

Often the objective of a statistical analysis is to build a stochastic model that can be used to make predictions of future events or impute missing values. Let Y^* be the future observation we would like to predict. Assuming that the observations are independent given the parameters and that Y^* follows the same model as the observed data, then given $\boldsymbol{\theta}$ we have $Y^* \sim f(y|\boldsymbol{\theta})$ and prediction is straightforward. Unfortunately, we do not know $\boldsymbol{\theta}$ exactly, even after observing $\mathbf{Y}$. A remedy for this is to plug in a value of $\boldsymbol{\theta}$, say, the posterior mean $\hat{\boldsymbol{\theta}} = \mathrm{E}(\boldsymbol{\theta}|\mathbf{Y})$, and then sample $Y^* \sim f(Y|\hat{\boldsymbol{\theta}})$. However, this ignores uncertainty about the unknown parameters. If the posterior variance of $\boldsymbol{\theta}$ is small then its uncertainty is negligible, otherwise a better approach is needed.

For the sake of prediction, the parameters are not of interest themselves, but rather they serve as vehicles to transfer information from the data to the predictive model. We would rather bypass the parameters altogether and simply use the posterior predictive distribution (PPD)

$$Y^*|\mathbf{Y} \sim \mathrm{f}^*(Y^*|\mathbf{Y}). \tag{1.44}$$

The PPD is the distribution of a new outcome given the observed data.

In a parametric model, the PPD naturally accounts for uncertainty in the model parameters; this an advantage of the Bayesian framework. The PPD accounts for parametric uncertainty because it can be written

$$f^*(Y^*|\mathbf{Y}) = \int f^*(Y^*, \boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta} = \int f(Y^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta}, \tag{1.45}$$

where f is the likelihood density (here we assume that the observations are independent given the parameters and so $f(Y^*|\boldsymbol{\theta}) = f(Y^*|\boldsymbol{\theta}, \mathbf{Y})$, and p is the posterior density.

To further illustrate how the PPD accounts for parameteric uncertainty, we consider how to make a sample from the PPD. If we first draw posterior sample $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}|\mathbf{Y})$ and then a prediction from the likelihood, $Y^*|\boldsymbol{\theta}^* \sim f(Y|\boldsymbol{\theta}^*)$, then Y^* follows the PPD. A Monte Carlo approximation (Section 3.2) repeats these step many times to approximate the PPD. Unlike the plug-in predictor, each predictive uses a different value of the parameters and thus accurately reflects parametric uncertainty. It can be shown that $\mathrm{Var}(Y^*|\mathbf{Y}) \geq \mathrm{Var}(Y^*|\mathbf{Y}, \boldsymbol{\theta})$ with equality holding only if there is no posterior uncertainty in the mean of $Y^*|\mathbf{Y}, \boldsymbol{\theta}$.

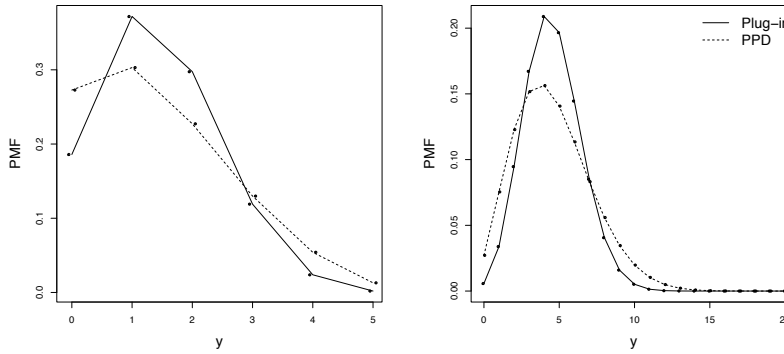

FIGURE 1.12
Posterior predictive distribution for a beta-binomial example. Plots of the posterior predictive distribution (PPD) from the model $Y|\theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(1, 1)$. The "plug-in" PMF is the binomial density evaluated at the posterior mean $\hat{\theta}$, $f(y|\hat{\theta})$. This is compared with the full PPD for $Y = 1$ success in $n = 5$ trials (left) and $Y = 4$ successes in $n = 20$ trials (right). The PMFs are connected by lines for visualization, but the probabilities are only defined for $y = \{0, 1, ..., n\}$.

As an example, consider the model $Y|\theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(1, 1)$. Given the data we have observed (Y and n) we would like to predict the outcome if we repeat the experiment, $Y^* \in \{0, 1, ..., n\}$. The posterior of θ is $\theta|Y \sim \text{Beta}(Y+1, n+1)$ and the posterior mean is $\hat{\theta} = (Y+1)/(n+2)$. The solid lines in Figure 1.12 show the plug-in prediction $Y^* \sim \text{Binomial}(n, \hat{\theta})$ versus the full PPD (Listing 1.2) that accounts for uncertainty in θ (which is a Beta-Binomial$(n, Y + 1, n + 1)$ distribution). For both $n = 5$ and $n = 20$, the PPD is considerably wider than the plug-in predictive distribution, as expected.

Listing 1.2
Summarizing a posterior predictive distribution (PPD) in R.

```
> # Load the data
> n <- 5
> Y <- 1
> a <- 1
> b <- 1
> A <- Y + a
> B <- n - Y + b

> # Plug-in estimator
> theta_hat  <- A/(A+B)
> y          <- 0:5
> PPD        <- dbinom(y,n,theta_hat)
> names(PPD) <- y
> round(PPD,2)
   0    1    2    3    4    5
0.19 0.37 0.30 0.12 0.02 0.00

> # Draws from the PPD, Y_star[i]~Binomial(n,theta_star[i])
> S           <- 100000
> theta_star  <- rbeta(S,A,B)
> Y_star      <- rbinom(S,n,theta_star)
> PPD         <- table(Y_star)/S
> round(PPD,2)
   0    1    2    3    4    5
0.27 0.30 0.23 0.13 0.05 0.01
```

1.6 Exercises

1. If X has support $X \in \mathcal{S} = [1, \infty]$, find the constant c (as a function of θ) that makes $f(x) = c\exp(-x/\theta)$ a valid PDF.
2. Assume that $X \sim \text{Uniform}(a, b)$ so the support is $\mathcal{S} = [a, b]$ and the PDF is $f(x) = 1/(b-a)$ for any $x \in \mathcal{S}$.
   (a) Prove that this is a valid PDF.
   (b) Derive the mean and variance of X.
3. Expert knowledge dictates that a parameter must be positive and that its prior distribution should have the mean 5 and variance 3. Find a prior distribution that satisfies these constraints.
4. X_1 and X_2 have joint PMF

x_1	x_2	$\text{Prob}(X_1 = x_1, X_2 = x_2)$
0	0	0.15
1	0	0.15
2	0	0.15
0	1	0.15
1	1	0.20
2	1	0.20

   (a) Compute the marginal distribution of X_1.
   (b) Compute the marginal distribution of X_2.
   (c) Compute the conditional distribution of $X_1|X_2$.
   (d) Compute the conditional distribution of $X_2|X_1$.
   (e) Are X_1 and X_2 independent? Justify your answer.
5. If (X_1, X_2) is bivariate normal with $\text{E}(X_1) = \text{E}(X_2) = 0$, $\text{Var}(X_1) = \text{Var}(X_2) = 1$, and $\text{Cor}(X_1, X_2) = \rho$:
   (a) Derive the marginal distribution of X_1.
   (b) Derive the conditional distribution of $X_1|X_2$.
6. Assume (X_1, X_2) have bivariate PDF

$$f(x_1, x_2) = \frac{1}{2\pi}\left(1 + x_1^2 + x_2^2\right)^{-3/2}.$$

   (a) Plot the conditional distribution of $X_1|X_2 = x_2$ for $x_2 \in \{-3, -2, -1, 0, 1, 2, 3\}$ (preferably on the same plot).
   (b) Do X_1 and X_2 appear to be correlated? Justify your answer.
   (c) Do X_1 and X_2 appear to be independent? Justify your answer.

7. According to `insurance.com`, the 2017 auto theft rate was 135 per 10,000 residents in Raleigh, NC compared to 214 per 10,000 residents in Durham/Chapel Hill. Assuming Raleigh's population is twice as large as Durham/Chapel Hill and a car has been stolen somewhere in the triangle (i.e., one of these two areas), what the probability it was stolen in Raleigh?
8. Your daily commute is distributed uniformly between 15 and 20 minutes if there no convention downtown. However, conventions are scheduled for roughly 1 in 4 days, and your commute time is distributed uniformly from 15 to 30 minutes if there is a convention. Let Y be your commute time this morning.
   (a) What is the probability that there was a convention downtown given $Y = 18$?
   (b) What is the probability that there was a convention downtown given $Y = 28$?
9. For this problem pretend we are dealing with a language with a six-word dictionary

   $$\{\text{fun, sun, sit, sat, fan, for}\}.$$

   An extensive study of literature written in this language reveals that all words are equally likely except that "for" is α times as likely as the other words. Further study reveals that:

   i. Each keystroke is an error with probability θ.
   ii. All letters are equally likely to produce errors.
   iii. Given that a letter is typed incorrectly it is equally likely to be any other letter.
   iv. Errors are independent across letters.

   For example, the probability of correctly typing "fun" (or any other word) is $(1-\theta)^3$, the probability of typing "pun" or "fon" when intending to type is "fun" is $\theta(1-\theta)^2$, and the probability of typing "foo" or "nnn" when intending to type "fun" is $\theta^2(1-\theta)$. Use Bayes' rule to develop a simple spell checker for this language. For each of the typed words "sun", "the", "foo", give the probability that each word in the dictionary was the intended word. Perform this for the parameters below:

   (a) $\alpha = 2$ and $\theta = 0.1$.
   (b) $\alpha = 50$ and $\theta = 0.1$.
   (c) $\alpha = 2$ and $\theta = 0.95$.

   Comment on the changes you observe in these three cases.

10. Let $X_1 \sim \text{Bernoulli}(\theta)$ be the indicator that a tree species occupies a forest and $\theta \in [0, 1]$ denote the prior occupancy probability. The researcher gathers a sample of n trees from the forest and X_2 belong to the species of interest. The model for the data is $X_2|X_1 \sim \text{Binomial}(n, \lambda X_1)$ where $\lambda \in [0, 1]$ the probability of detecting the species given it is present. Give expressions in terms of n, θ and λ for the following joint, marginal and conditional probabilities:
    (a) $\text{Prob}(X_1 = X_2 = 0)$.
    (b) $\text{Prob}(X_1 = 0)$.
    (c) $\text{Prob}(X_2 = 0)$.
    (d) $\text{Prob}(X_1 = 0|X_2 = 0)$.
    (e) $\text{Prob}(X_2 = 0|X_1 = 0)$.
    (f) $\text{Prob}(X_1 = 0|X_2 = 1)$.
    (g) $\text{Prob}(X_2 = 0|X_1 = 1)$.
    (h) Provide intuition for how (d)-(g) change with n, θ and λ.
    (i) Assuming $\theta = 0.5$, $\lambda = 0.1$, and $X_2 = 0$, how large must n be before we can conclude with 95% confidence that the species does not occupy the forest?
11. In a study that uses Bayesian methods to forecast the number of species that will be discovered in future years, [24] report that the number of marine bivalve species discovered each year from 2010-2015 was 64, 13, 33, 18, 30 and 20. Denoting Y_t as the number of species discovered in year t and assuming $Y_t|\lambda \overset{iid}{\sim} \text{Poisson}(\lambda)$ and $\lambda \sim \text{Uniform}(0, 100)$, plot the posterior distribution of λ.
12. Assume that (X, Y) follow the bivariate normal distribution and that both X and Y have marginal mean zero and marginal variance one. We observe six independent and identically distributed data points: (-3.3, -2.6), (0.1, -0.2), (-1.1, -1.5), (2.7, 1.5), (2.0, 1.9) and (-0.4, -0.3). Make a scatter plot of the data and, assuming the correlation parameter ρ has a $\text{Uniform}(-1, 1)$ prior, plot the posterior distribution of ρ.
13. The normalized difference vegetation index (NDVI) is commonly used to classify land cover using remote sensing data. Hypothetically, say that NDVI follows a $\text{Beta}(25, 10)$ distribution for pixels in a rain forest, and a $\text{Beta}(10, 15)$ distribution for pixels in a deforested area now used for agriculture. Assuming about 10% of the rain forest has been deforested, your objective is to build a rule to classify individual pixels as deforested based on their NDVI.
    (a) Plot the PDF of NDVI for forested and deforested pixels, and the marginal distribution of NDVI averaging over categories.

(b) Give an expression for the probability that a pixel is deforested given its NDVI value, and plot this probability by NDVI.

(c) You will classify a pixel as deforested if you are at least 90% sure it is deforested. Following this rule, give the range of NDVI that will lead to a pixel being classified as deforested.

14. Let n be the unknown number of customers that visit a store on the day of a sale. The number of customers that make a purchase is $Y|n \sim \text{Binomial}(n, \theta)$ where θ is the known probability of making a purchase given the customer visited the store. The prior is $n \sim \text{Poisson}(5)$. Assuming θ is known and n is the unknown parameter, plot the posterior distribution of n for all combinations of $Y \in \{0, 5, 10\}$ and $\theta \in \{0.2, 0.5\}$ and comment on the effect of Y and θ on the posterior.

15. Last spring your lab planted ten seedlings and two survived the winter. Let θ be the probability that a seedling survives the winter.

(a) Assuming a uniform prior distribution for θ, compute its posterior mean and standard deviation.

(b) Assuming the same prior as in (a), compute and compare the equal-tailed and highest density 95% posterior credible intervals.

(c) If you plant another 10 seedlings next year, what is the posterior predictive probability that at least one will survive the winter?

16. X_1 and X_2 are binary indicators of failure for two parts of a machine. Independent tests have shown that $X_1 \sim \text{Bernoulli}(1/2)$ and $X_2 \sim \text{Bernoulli}(1/3)$. Y_1 and Y_2 are binary indicators of two system failures. We know that $Y_1 = 1$ if both $X_1 = 1$ and $X_2 = 1$ and $Y_1 = 0$ otherwise, and $Y_2 = 0$ if both $X_1 = 0$ and $X_2 = 0$ and $Y_2 = 1$ otherwise. Compute the following probabilities:

(a) The probability that $X_1 = 1$ and $X_2 = 1$ given $Y_1 = 1$.

(b) The probability that $X_1 = 1$ and $X_2 = 1$ given $Y_2 = 1$.

(c) The probability that $X_1 = 1$ given $Y_1 = 1$.

(d) The probability that $X_1 = 1$ given $Y_2 = 1$.

17. The table below has the overall free throw proportion and results of free throws taken in pressure situations, defined as "clutch" (`https://stats.nba.com/`), for ten National Basketball Association players (those that received the most votes for the Most Valuable Player Award) for the 2016–2017 season. Since the overall proportion is computed using a large sample size, assume it is fixed and analyze the clutch data for each player separately using Bayesian methods. Assume a uniform prior throughout this problem.

Player	Overall proportion	Clutch makes	Clutch attempts
Russell Westbrook	0.845	64	75
James Harden	0.847	72	95
Kawhi Leonard	0.880	55	63
LeBron James	0.674	27	39
Isaiah Thomas	0.909	75	83
Stephen Curry	0.898	24	26
Giannis Antetokounmpo	0.770	28	41
John Wall	0.801	66	82
Anthony Davis	0.802	40	54
Kevin Durant	0.875	13	16

(a) Describe your model for studying the clutch success probability including the likelihood and prior.

(b) Plot the posteriors of the clutch success probabilities.

(c) Summarize the posteriors in a table.

(d) Do you find evidence that any of the players have a different clutch percentage than overall percentage?

(e) Are the results sensitive to your prior? That is, do small changes in the prior lead to substantial changes in the posterior?

18. In the early twentieth century, it was generally agreed that Hamilton and Madison (ignore Jay for now) wrote 51 and 14 Federalist Papers, respectively. There was dispute over how to attribute 12 other papers between these two authors. In the 51 papers attributed to Hamilton the word "upon" was used 3.24 times per 1,000 words, compared to 0.23 times per 1,000 words in the 14 papers attributed to Madison (for historical perspective on this problem, see [58]).

(a) If the word "upon" is used three times in a disputed text of length $1,000$ words and we assume the prior probability 0.5, what is the posterior probability the paper was written by Hamilton?

(b) Give one assumption you are making in (a) that is likely unreasonable. Justify your answer.

(c) In (a), if we changed the number of instances of "upon" to one, do you expect the posterior probability to increase, decrease or stay the same? Why?

(d) In (a), if we changed the text length to $10,000$ words and number of instances of "upon" to 30, do you expect the posterior probability to increase, decrease or stay the same? Why?

(e) Let Y be the number of observed number of instances of "upon" in 1,000 words. Compute the posterior probability the paper was written by Hamilton for each $Y \in \{0, 1, ..., 20\}$, plot these

posterior probabilities versus Y and give a rule for the number of instances of "upon" needed before the paper should be attributed to Hamilton.

Bibliography

[1] Helen Abbey. An examination of the Reed-Frost theory of epidemics. *Human Biology*, 24(3):201, 1952.

[2] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[3] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[4] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, 2014.

[5] Albert Barberán, Robert R Dunn, Brian J Reich, Krishna Pacifici, Eric B Laber, Holly L Menninger, James M Morton, Jessica B Henley, Jonathan W Leff, Shelly L Miller, and Noah Fierer. The ecology of microscopic life in household dust. In *Proceedings of the Royal Society B*, volume 282, page 20151139. The Royal Society, 2015.

[6] Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

[7] Daryl J Bem. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3):407, 2011.

[8] James Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

[9] James O Berger, Luis R Pericchi, JK Ghosh, Tapas Samanta, Fulvio De Santis, JO Berger, and LR Pericchi. Objective Bayesian methods for model selection: Introduction and comparison. *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pages 135–207, 2001.

[10] Jose M Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 113–147, 1979.

[11] Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

[12] Howard D Bondell and Brian J Reich. Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624, 2012.

[13] Dennis D Boos and Leonard A Stefanski. *Essential Statistical Inference: Theory and Methods*, volume 120. Springer Science & Business Media, 2013.

[14] Carlos A Botero, Beth Gardner, Kathryn R Kirby, Joseph Bulbulia, Michael C Gavin, and Russell D Gray. The ecology of religious beliefs. *Proceedings of the National Academy of Sciences*, 111(47):16784–16789, 2014.

[15] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. STAN: A probabilistic programming language. *Journal of Statistical Software*, 20(2):1–37, 2016.

[16] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[17] Fang Chen. Bayesian modeling using the MCMC procedure. In *Proceedings of the SAS Global Forum 2008 Conference, Cary NC: SAS Institute Inc*, 2009.

[18] Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[19] Ciprian M Crainiceanu, David Ruppert, and Matthew P Wand. Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14(1):1–24, 2005.

[20] A Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, pages 278–292, 1984.

[21] Gustavo de los Campos and Paulino Perez Rodriguez. *BLR: Bayesian Linear Regression*, 2014. R package version 1.4.

[22] Perry de Valpine, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017.

[23] Peter Diggle, Rana Moyeed, Barry Rowlingson, and Madeleine Thomson. Childhood malaria in the Gambia: A case-study in model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):493–506, 2002.

[24] Stewart M Edie, Peter D Smits, and David Jablonski. Probabilistic models of species discovery and biodiversity comparisons. *Proceedings of the National Academy of Sciences*, 114(14):3666–3671, 2017.

[25] Gregory M Erickson, Peter J Makovicky, Philip J Currie, Mark A Norell, Scott A Yerby, and Christopher A Brochu. Gigantism and comparative life-history parameters of tyrannosaurid dinosaurs. *Nature*, 430(7001):772–775, 2004.

[26] Kevin R. Forward, David Haldane, Duncan Webster, Carolyn Mills, Cherly Brine, and Diane Aylward. A comparison between the Strep A Rapid Test Device and conventional culture for the diagnosis of streptococcal pharyngitis. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 17:221–223, 2004.

[27] Seymour Geisser. Discussion on sampling and Bayes inference in scientific modeling and robustness (by GEP Box). *Journal of the Royal Statistical Society: Series A (General)*, 143:416–417, 1980.

[28] Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of Spatial Statistics*. CRC press, 2010.

[29] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[30] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[31] Andrew Gelman, Gareth O Roberts, and Walter R Gilks. Efficient Metropolis jumping rules. *Bayesian Statistics*, 5(599-608):42, 1996.

[32] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

[33] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

[34] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in Computer Vision*, pages 564–584. Elsevier, 1987.

[35] Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[36] John Geweke. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department, Minneapolis, MN, USA, 1991.

[37] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press, 2017.

[38] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[39] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006.

[40] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[41] Wilfred K Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[42] James S Hodges. *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models using Random Effects*. Chapman & Hall/CRC, 2016.

[43] James S Hodges and Brian J Reich. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334, 2010.

[44] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[45] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, pages 382–401, 1999.

[46] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[47] Bindu Kalesan, Matthew E Mobily, Olivia Keiser, Jeffrey A Fagan, and Sandro Galea. Firearm legislation and firearm mortality in the USA: A cross-sectional, state-level study. *The Lancet*, 387(10030):1847–1855, 2016.

[48] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[49] Robert E Kass and Larry Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934, 1995.

[50] Hong Lan, Meng Chen, Jessica B Flowers, Brian S Yandell, Donnie S Stapleton, Christine M Mata, Eric Ton-Keen Mui, Matthew T Flowers, Kathryn L Schueler, Kenneth F Manly, et al. Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2(1):e6, 2006.

[51] Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[52] Roderick Little. Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26(2):162–174, 2011.

[53] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[54] Peter McCullagh and John Nelder. *Generalized Linear Models, Second Edition*. Boca Raton: Chapman & Hall/CRC, 1989.

[55] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[56] Greg Miller. ESP paper rekindles discussion about statistics. *Science*, 331(6015):272–273, 2011.

[57] Antonietta Mira. On Metropolis-Hastings algorithms with delayed rejection. *Metron*, 59(3-4):231–241, 2001.

[58] Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[59] Radford M Neal. Slice sampling. *Annals of Statistics*, pages 705–741, 2003.

[60] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

[61] Radford M Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 2012.

[62] Jorge Nocedal and Stephen J Wright. *Sequential Quadratic Programming*. Springer, 2006.

[63] Krishna Pacifici, Brian J Reich, David AW Miller, Beth Gardner, Glenn Stauffer, Susheela Singh, Alexa McKerrow, and Jaime A Collazo. Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, 98(3):840–850, 2017.

[64] Anand Patil, David Huard, and Christopher J Fonnesbeck. PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, 35(4):1, 2010.

[65] LI Pettit. The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 175–184, 1990.

[66] Martyn Plummer. JAGS Version 4.0. 0 user manual. *See https://sourceforge. net/projects/mcmc-jags/files/Manuals/4. x*, 2015.

[67] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning*, pages 63–71. Springer, 2004.

[68] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[69] Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and Their Applications*, 49(2):207–216, 1994.

[70] Veronika Ročková and Edward I George. The spike-and-slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

[71] Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

[72] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

[73] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

[74] John R Sauer, James E Hines, and Jane E Fallon. *The North American Breeding Bird Survey, Results and Analysis 1966–2005*. Version 6.2.2006. USGS Patuxent Wildlife Research Center, Laurel, Maryland, USA, 2005.

[75] Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

[76] Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28, 2017.

[77] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[78] Mervyn Stone. Necessary and sufficient condition for convergence in probability to invariant posterior distributions. *The Annals of Mathematical Statistics*, pages 1349–1353, 1970.

[79] Sibylle Sturtz, Uwe Ligges, and Andrew Gelman. R2OpenBUGS: A package for running OpenBUGS from R. *URL http://cran. rproject. org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS. pdf*, 2010.

[80] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.

[81] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 267–288, 1996.

[82] Luke Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, pages 1701–1728, 1994.

[83] Arnold Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques*, 1986.

[84] Yan Zhang, Brian J Reich, and Howard D Bondell. High dimensional linear regression via the R2-D2 shrinkage prior. *arXiv preprint arXiv:1609.00046*, 2016.