CHAPTER 1

REVIEW

This chapter reviews notation and background material in mathematics, probability, and statistics. Readers may wish to skip this chapter and turn directly to Chapter 2, returning here only as needed.

1.1 MATHEMATICAL NOTATION

We use boldface to distinguish a vector $\mathbf{x} = (x_1, \ldots, x_p)$ or a matrix $\mathbf{M}$ from a scalar variable x or a constant M. A vector-valued function $\mathbf{f}$ evaluated at $\mathbf{x}$ is also boldfaced, as in $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x}))$. The transpose of $\mathbf{M}$ is denoted $\mathbf{M}^{\mathrm{T}}$.

Unless otherwise specified, all vectors are considered to be column vectors, so, for example, an $n \times p$ matrix can be written as $\mathbf{M} = (\mathbf{x}_1 \ \ldots \ \mathbf{x}_n)^{\mathrm{T}}$. Let $\mathbf{I}$ denote an identity matrix, and $\mathbf{1}$ and $\mathbf{0}$ denote vectors of ones and zeros, respectively.

A symmetric square matrix $\mathbf{M}$ is *positive definite* if $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} > 0$ for all nonzero vectors $\mathbf{x}$. Positive definiteness is equivalent to the condition that all eigenvalues of $\mathbf{M}$ are positive. $\mathbf{M}$ is *nonnegative definite* or *positive semidefinite* if $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} \geq 0$ for all nonzero vectors $\mathbf{x}$.

The derivative of a function f, evaluated at x, is denoted $f'(x)$. When $\mathbf{x} = (x_1, \ldots, x_p)$, the *gradient* of f at $\mathbf{x}$ is

$$\mathbf{f}'(\mathbf{x}) = \left(\frac{df(\mathbf{x})}{dx_1}, \ldots, \frac{df(\mathbf{x})}{dx_p} \right).$$

The *Hessian matrix* for f at $\mathbf{x}$ is $\mathbf{f}''(\mathbf{x})$ having (i, j)th element equal to $d^2 f(\mathbf{x})/(dx_i\, dx_j)$. The negative Hessian has important uses in statistical inference.

Let $\mathbf{J}(\mathbf{x})$ denote the *Jacobian matrix* evaluated at $\mathbf{x}$ for the one-to-one mapping $\mathbf{y} = \mathbf{f}(\mathbf{x})$. The (i, j)th element of $\mathbf{J}(\mathbf{x})$ is equal to $df_i(\mathbf{x})/dx_j$.

A *functional* is a real-valued function on a space of functions. For example, if $T(f) = \int_0^1 f(x)\, dx$, then the functional T maps suitably integrable functions onto the real line.

The indicator function $1_{\{A\}}$ equals 1 if A is true and 0 otherwise. The real line is denoted $\Re$, and p-dimensional real space is $\Re^p$.

Computational Statistics, Second Edition. Geof H. Givens and Jennifer A. Hoeting.

1.2 TAYLOR'S THEOREM AND MATHEMATICAL LIMIT THEORY

First, we define standard "big oh" and "little oh" notation for describing the relative orders of convergence of functions. Let the functions f and g be defined on a common, possibly infinite interval. Let z_0 be a point in this interval or a boundary point of it (i.e., $-\infty$ or ∞). We require $g(z) \neq 0$ for all $z \neq z_0$ in a neighborhood of z_0. Then we say

$$f(z) = \mathcal{O}(g(z)) \tag{1.1}$$

if there exists a constant M such that $|f(z)| \leq M|g(z)|$ as $z \to z_0$. For example, $(n+1)/(3n^2) = \mathcal{O}(n^{-1})$, and it is understood that we are considering $n \to \infty$. If $\lim_{z \to z_0} f(z)/g(z) = 0$, then we say

$$f(z) = \mathcal{o}(g(z)). \tag{1.2}$$

For example, $f(x_0 + h) - f(x_0) = hf'(x_0) + \mathcal{o}(h)$ as $h \to 0$ if f is differentiable at x_0. The same notation can be used for describing the convergence of a sequence $\{x_n\}$ as $n \to \infty$, by letting $f(n) = x_n$.

Taylor's theorem provides a polynomial approximation to a function f. Suppose f has finite $(n+1)$th derivative on (a, b) and continuous nth derivative on $[a, b]$. Then for any $x_0 \in [a, b]$ distinct from x, the Taylor series expansion of f about x_0 is

$$f(x) = \sum_{i=0}^{n} \frac{1}{i!} f^{(i)}(x_0)(x - x_0)^i + R_n, \tag{1.3}$$

where $f^{(i)}(x_0)$ is the ith derivative of f evaluated at x_0, and

$$R_n = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x - x_0)^{n+1} \tag{1.4}$$

for some point ξ in the interval between x and x_0. As $|x - x_0| \to 0$, note that $R_n = \mathcal{O}(|x - x_0|^{n+1})$.

The multivariate version of Taylor's theorem is analogous. Suppose f is a real-valued function of a p-dimensional variable $\mathbf{x}$, possessing continuous partial derivatives of all orders up to and including $n+1$ with respect to all coordinates, in an open convex set containing $\mathbf{x}$ and $\mathbf{x}_0 \neq \mathbf{x}$. Then

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \sum_{i=1}^{n} \frac{1}{i!} D^{(i)}(f; \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0) + R_n, \tag{1.5}$$

where

$$D^{(i)}(f; \mathbf{x}, \mathbf{y}) = \sum_{j_1=1}^{p} \cdots \sum_{j_i=1}^{p} \left\{ \left(\frac{d^i}{dt_{j_1} \cdots dt_{j_i}} f(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{x}} \right) \prod_{k=1}^{i} y_{j_k} \right\} \tag{1.6}$$

and

$$R_n = \frac{1}{(n+1)!} D^{(n+1)}(f; \boldsymbol{\xi}, \mathbf{x} - \mathbf{x}_0) \tag{1.7}$$

for some $\boldsymbol{\xi}$ on the line segment joining $\mathbf{x}$ and $\mathbf{x}_0$. As $|\mathbf{x} - \mathbf{x}_0| \to 0$, note that $R_n = \mathcal{O}(|\mathbf{x} - \mathbf{x}_0|^{n+1})$.

The *Euler–Maclaurin formula* is useful in many asymptotic analyses. If f has $2n$ continuous derivatives in [0, 1], then

$$\begin{aligned}\int_0^1 f(x)\,dx = &\frac{f(0) + f(1)}{2}\\ &- \sum_{i=0}^{n-1} \frac{b_{2i}(f^{(2i-1)}(1) - f^{(2i-1)}(0))}{(2i)!} - \frac{b_{2n} f^{(2n)}(\xi)}{(2n)!},\end{aligned} \tag{1.8}$$

where $0 \le \xi \le 1$, $f^{(j)}$ is the jth derivative of f, and $b_j = B_j(0)$ can be determined using the recursion relation

$$\sum_{j=0}^{m} \binom{m+1}{j} B_j(z) = (m+1) z^m \tag{1.9}$$

initialized with $B_0(z) = 1$. The proof of this result is based on repeated integrations by parts [376].

Finally, we note that it is sometimes desirable to approximate the derivative of a function numerically, using finite differences. For example, the ith component of the gradient of f at $\mathbf{x}$ can be approximated by

$$\frac{df(\mathbf{x})}{dx_i} \approx \frac{f(\mathbf{x} + \epsilon_i \mathbf{e}_i) - f(\mathbf{x} - \epsilon_i \mathbf{e}_i)}{2\epsilon_i}, \tag{1.10}$$

where ϵ_i is a small number and $\mathbf{e}_i$ is the unit vector in the ith coordinate direction. Typically, one might start with, say, $\epsilon_i = 0.01$ or 0.001 and approximate the desired derivative for a sequence of progressively smaller ϵ_i. The approximation will generally improve until ϵ_i becomes small enough that the calculation is degraded and eventually dominated by computer roundoff error introduced by subtractive cancellation. Introductory discussion of this approach and a more sophisticated Richardson extrapolation strategy for obtaining greater precision are provided in [376]. Finite differences can also be used to approximate the second derivative of f at $\mathbf{x}$ via

$$\begin{aligned}\frac{df(\mathbf{x})}{dx_i\,dx_j} \approx \frac{1}{4\epsilon_i\epsilon_j}\Big(& f(\mathbf{x} + \epsilon_i \mathbf{e}_i + \epsilon_j \mathbf{e}_j) - f(\mathbf{x} + \epsilon_i \mathbf{e}_i - \epsilon_j \mathbf{e}_j)\\ &- f(\mathbf{x} - \epsilon_i \mathbf{e}_i + \epsilon_j \mathbf{e}_j) + f(\mathbf{x} - \epsilon_i \mathbf{e}_i - \epsilon_j \mathbf{e}_j)\Big)\end{aligned} \tag{1.11}$$

with similar sequential precision improvements.

1.3 STATISTICAL NOTATION AND PROBABILITY DISTRIBUTIONS

We use capital letters to denote random variables, such as Y or $\mathbf{X}$, and lowercase letters to represent specific realized values of random variables such as y or $\mathbf{x}$. The probability density function of X is denoted f; the cumulative distribution function is F. We use the notation $X \sim f(x)$ to mean that X is distributed with density $f(x)$. Frequently, the dependence of $f(x)$ on one or more parameters also will be denoted with a conditioning bar, as in $f(x|\alpha, \boldsymbol{\beta})$. Because of the diversity of topics covered in this book, we want to be careful to distinguish when $f(x|\alpha)$ refers to a density function as opposed to the evaluation of that density at a point x. When the meaning is unclear from the context, we will be explicit, for example, by using $f(\cdot|\alpha)$ to denote the function. When it is important to distinguish among several densities, we may adopt subscripts referring to specific random variables, so that the density functions for X and Y are f_X and f_Y, respectively. We use the same notation for distributions of discrete random variables and in the Bayesian context.

The conditional distribution of X given that Y equals y (i.e., $X|Y = y$) is described by the density denoted $f(x|y)$, or $f_{X|Y}(x|y)$. In this case, we write that $X|Y$ has density $f(x|Y)$. For notational simplicity we allow density functions to be implicitly specified by their arguments, so we may use the same symbol, say f, to refer to many distinct functions, as in the equation $f(x, y|\mu) = f(x|y, \mu)f(y|\mu)$. Finally, $f(X)$ and $F(X)$ are random variables: the evaluations of the density and cumulative distribution functions, respectively, at the random argument X.

The expectation of a random variable is denoted $E\{X\}$. Unless specifically mentioned, the distribution with respect to which an expectation is taken is the distribution of X or should be implicit from the context. To denote the probability of an event A, we use $P[A] = E\{1_{\{A\}}\}$. The conditional expectation of $X|Y = y$ is $E\{X|y\}$. When Y is unknown, $E\{X|Y\}$ is a random variable that depends on Y. Other attributes of the distribution of X and Y include $\text{var}\{X\}$, $\text{cov}\{X, Y\}$, $\text{cor}\{X, Y\}$, and $\text{cv}\{X\} = \text{var}\{X\}^{1/2}/E\{X\}$. These quantities are the variance of X, the covariance and correlation of X and Y, and the coefficient of variation of X, respectively.

A useful result regarding expectations is *Jensen's inequality*. Let g be a *convex function* on a possibly infinite open interval I, so

$$g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y) \tag{1.12}$$

for all $x, y \in I$ and all $0 < \lambda < 1$. Then Jensen's inequality states that $E\{g(X)\} \geq g(E\{X\})$ for any random variable X having $P[X \in I] = 1$.

Tables 1.1, 1.2, and 1.3 provide information about many discrete and continuous distributions used throughout this book. We refer to the following well-known combinatorial constants:

$$n! = n(n - 1)(n - 2) \cdots (3)(2)(1) \quad \text{with } 0! = 1, \tag{1.13}$$

$$\binom{n}{k} = \frac{n!}{k!(n - k)!}, \tag{1.14}$$

TABLE 1.1 Notation and description for common probability distributions of discrete random variables.

Name	Notation and Parameter Space	Density and Sample Space	Mean and Variance
Bernoulli	$X \sim \text{Bernoulli}(p)$ $0 \leq p \leq 1$	$f(x) = p^x(1-p)^{1-x}$ $x = 0$ or 1	$E\{X\} = p$ $\text{var}\{X\} = p(1-p)$
Binomial	$X \sim \text{Bin}(n, p)$ $0 \leq p \leq 1$ $n \in \{1, 2, \ldots\}$	$f(x) = \binom{n}{x} p^x(1-p)^{n-x}$ $x = 0, 1, \ldots n$	$E\{X\} = np$ $\text{var}\{X\} = np(1-p)$
Multinomial	$\mathbf{X} \sim \text{Multinomial}(n, \mathbf{p})$ $\mathbf{p} = (p_1, \ldots, p_k)$ $0 \leq p_i \leq 1$ and $n \in \{1, 2, \ldots\}$ $\sum_{i=1}^k p_i = 1$	$f(\mathbf{x}) = \binom{n}{x_1 \ \ldots \ x_k} \prod_{i=1}^k p_i^{x_i}$ $\mathbf{x} = (x_1, \ldots, x_k)$ and $x_i \in \{0, 1, \ldots n\}$ $\sum_{i=1}^k x_i = n$	$E\{\mathbf{X}\} = n\mathbf{p}$ $\text{var}\{X_i\} = np_i(1-p_i)$ $\text{cov}\{X_i, X_j\} = -np_i p_j$
Negative Binomial	$X \sim \text{NegBin}(r, p)$ $0 \leq p \leq 1$ $r \in \{1, 2, \ldots\}$	$f(x) = \binom{x+r-1}{r-1} p^r(1-p)^x$ $x \in \{0, 1, \ldots\}$	$E\{X\} = r(1-p)/p$ $\text{var}\{X\} = r(1-p)/p^2$
Poisson	$X \sim \text{Poisson}(\lambda)$ $\lambda > 0$	$f(x) = \dfrac{\lambda^x}{x!} \exp\{-\lambda\}$ $x \in \{0, 1, \ldots\}$	$E\{X\} = \lambda$ $\text{var}\{X\} = \lambda$

TABLE 1.2 Notation and description for some common probability distributions of continuous random variables.

Name	Notation and Parameter Space	Density and Sample Space	Mean and Variance
Beta	$X \sim \text{Beta}(\alpha, \beta)$ $\alpha > 0$ and $\beta > 0$	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $0 \le x \le 1$	$E\{X\} = \frac{\alpha}{\alpha+\beta}$ $\text{var}\{X\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Cauchy	$X \sim \text{Cauchy}(\alpha, \beta)$ $\alpha \in \Re$ and $\beta > 0$	$f(x) = \frac{1}{\pi\beta\left[1+\left(\frac{x-\alpha}{\beta}\right)^2\right]}$ $x \in \Re$	$E\{X\}$ is nonexistent $\text{var}\{X\}$ is nonexistent
Chi-square	$X \sim \chi^2_\nu$ $\nu > 0$	$f(x) = \text{Gamma}(\nu/2, 1/2)$ $x > 0$	$E\{X\} = \nu$ $\text{var}\{X\} = 2\nu$
Dirichlet	$\mathbf{X} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ $\alpha_i > 0$ $\alpha_0 = \sum_{i=1}^k \alpha_i$	$f(\mathbf{x}) = \frac{\Gamma(\alpha_0)\prod_{i=1}^k x_i^{\alpha_i-1}}{\prod_{i=1}^k \Gamma(\alpha_i)}$ $\mathbf{x} = (x_1, \ldots, x_k)$ and $0 \le x_i \le 1$ $\sum_{i=1}^k x_i = 1$	$E\{\mathbf{X}\} = \boldsymbol{\alpha}/\alpha_0$ $\text{var}\{X_i\} = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$ $\text{cov}\{X_i, X_j\} = \frac{-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$
Exponential	$X \sim \text{Exp}(\lambda)$ $\lambda > 0$	$f(x) = \lambda \exp\{-\lambda x\}$ $x > 0$	$E\{X\} = 1/\lambda$ $\text{var}\{X\} = 1/\lambda^2$
Gamma	$X \sim \text{Gamma}(r, \lambda)$ $\lambda > 0$ and $r > 0$	$f(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)} \exp\{-\lambda x\}$ $x > 0$	$E\{X\} = r/\lambda$ $\text{var}\{X\} = r/\lambda^2$

TABLE 1.3 Notation and description for more common probability distributions of continuous random variables.

Name	Notation and Parameter Space	Density and Sample Space	Mean and Variance
Lognormal	$X \sim \text{Lognormal}(\mu, \sigma^2)$ $\mu \in \Re$ and $\sigma > 0$	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{\log\{x\} - \mu}{\sigma}\right)^2\right\}$ $x \in \Re$	$E\{X\} = \exp\{\mu + \sigma^2/2\}$ $\text{var}\{X\} = \exp\{2\mu + 2\sigma^2\} - \exp\{2\mu + \sigma^2\}$
Multivariate Normal	$\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \Re^k$ $\boldsymbol{\Sigma}$ positive definite	$f(\mathbf{x}) = \frac{\exp\{-(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})/2\}}{(2\pi)^{k/2}\lvert\boldsymbol{\Sigma}\rvert^{1/2}}$ $\mathbf{x} = (x_1, \ldots, x_k) \in \Re^k$	$E\{\mathbf{X}\} = \boldsymbol{\mu}$ $\text{var}\{\mathbf{X}\} = \boldsymbol{\Sigma}$
Normal	$X \sim \text{N}(\mu, \sigma^2)$ $\mu \in \Re$ and $\sigma > 0$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\}$ $x \in \Re$	$E\{X\} = \mu$ $\text{var}\{X\} = \sigma^2$
Student's *t*	$X \sim t_\nu$ $\nu > 0$	$f(x) = \frac{\Gamma((\nu + 1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$ $x \in \Re$	$E\{X\} = 0$ if $\nu > 1$ $\text{var}\{X\} = \frac{\nu}{\nu + 2}$ if $\nu > 2$
Uniform	$X \sim \text{Unif}(a, b)$ $a, b \in \Re$ and $a < b$	$f(x) = \frac{1}{b - a}$ $x \in [a, b]$	$E\{X\} = (a + b)/2$ $\text{var}\{X\} = (b - a)^2/12$
Weibull	$X \sim \text{Weibull}(a, b)$ $a > 0$ and $b > 0$	$f(x) = abx^{b-1} \exp\{-ax^b\}$ $x > 0$	$E\{X\} = \frac{\Gamma(1 + 1/b)}{a^{1/b}}$ $\text{var}\{X\} = \frac{\Gamma(1 + 2/b) - \Gamma(1 + 1/b)^2}{a^{2/b}}$

$$\binom{n}{k_1 \quad \ldots \quad k_m} = \frac{n!}{\prod_{i=1}^m k_i!} \quad \text{where } n = \sum_{i=1}^m k_i, \tag{1.15}$$

$$\Gamma(r) = \begin{cases} (r-1)! & \text{for } r = 1, 2, \ldots \\ \int_0^\infty t^{r-1} \exp\{-t\}\, dt & \text{for general } r > 0. \end{cases} \tag{1.16}$$

It is worth knowing that

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi} \quad \text{and} \quad \Gamma\left(n + \tfrac{1}{2}\right) = 1 \times 3 \times 5 \times \cdots \times (2n-1)\sqrt{\pi}/2^n$$

for positive integer n.

Many of the distributions commonly used in statistics are members of an *exponential family*. A k-parameter exponential family density can be expressed as

$$f(x|\boldsymbol{\gamma}) = c_1(x)c_2(\boldsymbol{\gamma}) \exp\left\{\sum_{i=1}^k y_i(x)\theta_i(\boldsymbol{\gamma})\right\} \tag{1.17}$$

for nonnegative functions c_1 and c_2. The vector $\boldsymbol{\gamma}$ denotes the familiar parameters, such as λ for the Poisson density and p for the binomial density. The real-valued $\theta_i(\boldsymbol{\gamma})$ are the *natural*, or *canonical, parameters*, which are usually transformations of $\boldsymbol{\gamma}$. The $y_i(x)$ are the sufficient statistics for the canonical parameters. It is straightforward to show

$$E\{\mathbf{y}(X)\} = \boldsymbol{\kappa}'(\boldsymbol{\theta}) \tag{1.18}$$

and

$$\text{var}\{\mathbf{y}(X)\} = \boldsymbol{\kappa}''(\boldsymbol{\theta}), \tag{1.19}$$

where $\boldsymbol{\kappa}(\boldsymbol{\theta}) = -\log c_3(\boldsymbol{\theta})$, letting $c_3(\boldsymbol{\theta})$ denote the reexpression of $c_2(\boldsymbol{\gamma})$ in terms of the canonical parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, and $\mathbf{y}(X) = (y_1(X), \ldots, y_k(X))$. These results can be rewritten in terms of the original parameters $\boldsymbol{\gamma}$ as

$$E\left\{\sum_{i=1}^k \frac{d\theta_i(\boldsymbol{\gamma})}{d\gamma_j} y_i(X)\right\} = -\frac{d}{d\gamma_j} \log c_2(\boldsymbol{\gamma}) \tag{1.20}$$

and

$$\text{var}\left\{\sum_{i=1}^k \frac{d\theta_i(\boldsymbol{\gamma})}{d\gamma_j} y_i(X)\right\} = -\frac{d^2}{d\gamma_j^2} \log c_2(\boldsymbol{\gamma}) - E\left\{\sum_{i=1}^k \frac{d^2\theta_i(\boldsymbol{\gamma})}{d\gamma_j^2} y_i(X)\right\}. \tag{1.21}$$

Example 1.1 (Poisson) The Poisson distribution belongs to the exponential family with $c_1(x) = 1/x!$, $c_2(\lambda) = \exp\{-\lambda\}$, $y(x) = x$, and $\theta(\lambda) = \log \lambda$. Deriving moments in terms of θ, we have $\kappa(\theta) = \exp\{\theta\}$, so $E\{X\} = \kappa'(\theta) = \exp\{\theta\} = \lambda$ and $\text{var}\{X\} = \kappa''(\theta) = \exp\{\theta\} = \lambda$. The same results may be obtained with (1.20) and (1.21), noting that $d\theta/d\lambda = 1/\lambda$. For example, (1.20) gives $E\{X/\lambda\} = 1$. □

It is also important to know how the distribution of a random variable changes when it is transformed. Let $\mathbf{X} = (X_1, \ldots, X_p)$ denote a p-dimensional random variable with continuous density function f. Suppose that

$$\mathbf{U} = \mathbf{g}(\mathbf{X}) = (g_1(\mathbf{X}), \ldots, g_p(\mathbf{X})) = (U_1, \ldots, U_p), \tag{1.22}$$

where $\mathbf{g}$ is a one-to-one function mapping the support region of f onto the space of all $\mathbf{u} = g(\mathbf{x})$ for which $\mathbf{x}$ satisfies $f(\mathbf{x}) > 0$. To derive the probability distribution of $\mathbf{U}$ from that of $\mathbf{X}$, we need to use the Jacobian matrix. The density of the transformed variables is

$$f(\mathbf{u}) = f(\mathbf{g}^{-1}(\mathbf{u}))\,|\mathbf{J}(\mathbf{u})|\,, \tag{1.23}$$

where $|\mathbf{J}(\mathbf{u})|$ is the absolute value of the determinant of the Jacobian matrix of $\mathbf{g}^{-1}$ evaluated at $\mathbf{u}$, having (i, j)th element dx_i/du_j, where these derivatives are assumed to be continuous over the support region of $\mathbf{U}$.

1.4 LIKELIHOOD INFERENCE

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and identically distributed (i.i.d.) each having density $f(\mathbf{x} \mid \boldsymbol{\theta})$ that depends on a vector of p unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, then the joint likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\theta}). \tag{1.24}$$

When the data are not i.i.d., the joint likelihood is still expressed as the joint density $f(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta})$ viewed as a function of $\boldsymbol{\theta}$.

The observed data, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, might have been realized under many different values for $\boldsymbol{\theta}$. The parameters for which observing $\mathbf{x}_1, \ldots, \mathbf{x}_n$ would be most likely constitute the *maximum likelihood estimate* of $\boldsymbol{\theta}$. In other words, if $\hat{\boldsymbol{\vartheta}}$ is the function of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that maximizes $L(\boldsymbol{\theta})$, then $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\vartheta}}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the *maximum likelihood estimator (MLE)* for $\boldsymbol{\theta}$. MLEs are invariant to transformation, so the MLE of a transformation of $\boldsymbol{\theta}$ equals the transformation of $\hat{\boldsymbol{\theta}}$.

It is typically easier to work with the *log likelihood function*,

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}), \tag{1.25}$$

which has the same maximum as the original likelihood, since log is a strictly monotonic function. Furthermore, any additive constants (involving possibly $\mathbf{x}_1, \ldots, \mathbf{x}_n$ but not $\boldsymbol{\theta}$) may be omitted from the log likelihood without changing the location of its maximum or differences between log likelihoods at different $\boldsymbol{\theta}$. Note that maximizing $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is equivalent to solving the system of equations

$$\boldsymbol{l}'(\boldsymbol{\theta}) = \mathbf{0}, \tag{1.26}$$

where

$$\mathbf{l}'(\boldsymbol{\theta}) = \left(\frac{dl(\boldsymbol{\theta})}{d\theta_1}, \ldots, \frac{dl(\boldsymbol{\theta})}{d\theta_n} \right)$$

is called the *score function*. The score function satisfies

$$E\{\boldsymbol{l}'(\boldsymbol{\theta})\} = \mathbf{0}, \tag{1.27}$$

where the expectation is taken with respect to the distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Sometimes an analytical solution to (1.26) provides the MLE; this book describes a variety of methods that can be used when the MLE cannot be solved for in closed form. It is worth noting that there are pathological circumstances where the MLE is not a solution of the score equation, or the MLE is not unique; see [127] for examples.

The MLE has a sampling distribution because it depends on the realization of the random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$. The MLE may be biased or unbiased for $\boldsymbol{\theta}$, yet under quite general conditions it is asymptotically unbiased as $n \to \infty$. The sampling variance of the MLE depends on the average curvature of the log likelihood: When the log likelihood is very pointy, the location of the maximum is more precisely known.

To make this precise, let $\mathbf{l}''(\boldsymbol{\theta})$ denote the $p \times p$ matrix having (i, j)th element given by $d^2 l(\boldsymbol{\theta})/(d\theta_i d\theta_j)$. The *Fisher information matrix* is defined as

$$\mathbf{I}(\boldsymbol{\theta}) = E\{\mathbf{l}'(\boldsymbol{\theta})\mathbf{l}'(\boldsymbol{\theta})^{\mathrm{T}}\} = -E\{\mathbf{l}''(\boldsymbol{\theta})\}, \tag{1.28}$$

where the expectations are taken with respect to the distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. The final equality in (1.28) requires mild assumptions, which are satisfied, for example, in exponential families. $\mathbf{I}(\boldsymbol{\theta})$ may sometimes be called the *expected Fisher information* to distinguish it from $-\mathbf{l}''(\boldsymbol{\theta})$, which is the *observed Fisher information*. There are two reasons why the observed Fisher information is quite useful. First, it can be calculated even if the expectations in (1.28) cannot easily be computed. Second, it is a good approximation to $\mathbf{I}(\boldsymbol{\theta})$ that improves as n increases.

Under regularity conditions, the asymptotic variance–covariance matrix of the MLE $\widehat{\boldsymbol{\theta}}$ is $\mathbf{I}(\boldsymbol{\theta}^*)^{-1}$, where $\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$. Indeed, as $n \to \infty$, the limiting distribution of $\hat{\boldsymbol{\theta}}$ is $N_p(\boldsymbol{\theta}^*, \mathbf{I}(\boldsymbol{\theta}^*)^{-1})$. Since the true parameter values are unknown, $\mathbf{I}(\boldsymbol{\theta}^*)^{-1}$ must be estimated in order to estimate the variance–covariance matrix of the MLE. An obvious approach is to use $\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$. Alternatively, it is also reasonable to use $-\mathbf{l}''(\widehat{\theta})^{-1}$. Standard errors for individual parameter MLEs can be estimated by taking the square root of the diagonal elements of the chosen estimate of $\mathbf{I}(\boldsymbol{\theta}^*)^{-1}$. A thorough introduction to maximum likelihood theory and the relative merits of these estimates of $\mathbf{I}(\boldsymbol{\theta}^*)^{-1}$ can be found in [127, 182, 371, 470].

Profile likelihoods provide an informative way to graph a higher-dimensional likelihood surface, to make inference about some parameters while treating others as nuisance parameters, and to facilitate various optimization strategies. The profile likelihood is obtained by constrained maximization of the full likelihood with respect to parameters to be ignored. If $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\phi})$, then the profile likelihood for $\boldsymbol{\phi}$ is

$$L(\boldsymbol{\phi}|\hat{\boldsymbol{\mu}}(\boldsymbol{\phi})) = \max_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{\phi}). \tag{1.29}$$

Thus, for each possible $\boldsymbol{\phi}$, a value of $\boldsymbol{\mu}$ is chosen to maximize $L(\boldsymbol{\mu}, \boldsymbol{\phi})$. This optimal $\boldsymbol{\mu}$ is a function of $\boldsymbol{\phi}$. The profile likelihood is the function that maps $\boldsymbol{\phi}$ to the value of the full likelihood evaluated at $\boldsymbol{\phi}$ and its corresponding optimal $\boldsymbol{\mu}$. Note that the $\hat{\boldsymbol{\phi}}$

that maximizes the profile likelihood $L(\boldsymbol{\phi}|\hat{\boldsymbol{\mu}}(\boldsymbol{\phi}))$ is also the MLE for $\boldsymbol{\phi}$ obtained from the full likelihood $L(\boldsymbol{\mu}, \boldsymbol{\phi})$. Profile likelihood methods are examined in [23].

1.5 BAYESIAN INFERENCE

In the Bayesian inferential paradigm, probability distributions are associated with the parameters of the likelihood, as if the parameters were random variables. The probability distributions are used to assign subjective relative probabilities to regions of parameter space to reflect knowledge (and uncertainty) about the parameters.

Suppose that $\mathbf{X}$ has a distribution parameterized by $\boldsymbol{\theta}$. Let $f(\boldsymbol{\theta})$ represent the density assigned to $\boldsymbol{\theta}$ before observing the data. This is called a *prior distribution*. It may be based on previous data and analyses (e.g., pilot studies), it may represent a purely subjective personal belief, or it may be chosen in a way intended to have limited influence on final inference.

Bayesian inference is driven by the likelihood, often denoted $L(\boldsymbol{\theta}|\mathbf{x})$ in this context. Having established a prior distribution for $\boldsymbol{\theta}$ and subsequently observed data yielding a likelihood that is informative about $\boldsymbol{\theta}$, one's prior beliefs must be updated to reflect the information contained in the likelihood. The updating mechanism is Bayes' theorem:

$$f(\boldsymbol{\theta}|\mathbf{x}) = cf(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta}) = cf(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x}), \tag{1.30}$$

where $f(\boldsymbol{\theta}|\mathbf{x})$ is the *posterior density* of $\boldsymbol{\theta}$. The posterior distribution for $\boldsymbol{\theta}$ is used for statistical inference about $\boldsymbol{\theta}$. The constant c equals $1/\int f(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}$ and is often difficult to compute directly, although some inferences do not require c. This book describes a large variety of methods for enabling Bayesian inference, including the estimation of c.

Let $\tilde{\boldsymbol{\theta}}$ be the posterior mode, and let $\boldsymbol{\theta}^*$ be the true value of $\boldsymbol{\theta}$. The posterior distribution of $\tilde{\boldsymbol{\theta}}$ converges to $N(\boldsymbol{\theta}^*, \mathbf{I}(\boldsymbol{\theta}^*)^{-1})$ as $n \to \infty$, under regularity conditions. Note that this is the same limiting distribution as for the MLE. Thus, the posterior mode is of particular interest as a consistent estimator of $\boldsymbol{\theta}$. This convergence reflects the fundamental notion that the observed data should overwhelm any prior as $n \to \infty$.

Bayesian evaluation of hypotheses relies upon the *Bayes factor*. The ratio of posterior probabilities of two competing hypotheses or models, H_1 and H_2, is

$$\frac{P[H_2|\mathbf{x}]}{P[H_1|\mathbf{x}]} = \frac{P[H_2]}{P[H_1]} B_{2,1} \tag{1.31}$$

where $P[H_i|\mathbf{x}]$ denotes posterior probability, $P[H_i]$ denotes prior probability, and

$$B_{2,1} = \frac{f(\mathbf{x}|H_2)}{f(\mathbf{x}|H_1)} = \frac{\int f(\boldsymbol{\theta}_2|H_2)f(\mathbf{x}|\boldsymbol{\theta}_2, H_2)\, d\boldsymbol{\theta}_2}{\int f(\boldsymbol{\theta}_1|H_1)f(\mathbf{x}|\boldsymbol{\theta}_1, H_1)\, d\boldsymbol{\theta}_1} \tag{1.32}$$

with $\boldsymbol{\theta}_i$ denoting the parameters corresponding to the ith hypothesis. The quantity $B_{2,1}$ is the Bayes factor; it represents the factor by which the prior odds are multiplied to produce the posterior odds, given the data. The hypotheses H_1 and H_2 need not be nested as for likelihood ratio methods. The computation and interpretation of Bayes factors is reviewed in [365].

Bayesian interval estimation often relies on a 95% *highest posterior density* (HPD) region. The HPD region for a parameter is the region of shortest total length containing 95% of the posterior probability for that parameter for which the posterior density for every point contained in the interval is never lower than the density for every point outside the interval. For unimodal posteriors, the HPD is the narrowest possible interval containing 95% of the posterior probability. A more general interval for Bayesian inference is a *credible interval*. The $100(1-\alpha)\%$ credible interval is the region between the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution. When the posterior density is symmetric and unimodal, the HPD and the credible interval are identical.

A primary benefit of the Bayesian approach to inference is the natural manner in which resulting credibility intervals and other inferences are interpreted. One may speak of the posterior probability that the parameter is in some range. There is also a sound philosophical basis for the Bayesian paradigm; see [28] for an introduction. Gelman et al. provide a broad survey of Bayesian theory and methods [221].

The best prior distributions are those based on prior data. A strategy that is algebraically convenient is to seek conjugacy. A *conjugate prior* distribution is one that yields a posterior distribution in the same parametric family as the prior distribution. Exponential families are the only classes of distributions that have natural conjugate prior distributions.

When prior information is poor, it is important to ensure that the chosen prior distribution does not strongly influence posterior inferences. A posterior that is strongly influenced by the prior is said to be highly *sensitive* to the prior. Several strategies are available to reduce sensitivity. The simplest approach is to use a prior whose support is dispersed over a much broader region than the parameter region supported by the data, and fairly flat over it. A more formal approach is to use a Jeffreys prior [350]. In the univariate case, the Jeffreys prior is $f(\theta) \propto I(\theta)^{-1/2}$, where $I(\theta)$ is the Fisher information; multivariate extensions are possible. In some cases, the improper prior $f(\boldsymbol{\theta}) \propto 1$ may be considered, but this can lead to improper posteriors (i.e., not integrable), and it can be unintentionally informative depending on the parameterization of the problem.

Example 1.2 (Normal–Normal Conjugate Bayes Model) Consider Bayesian inference based on observations of i.i.d. random variables $X_1, \ldots, X_n$ with density $X_i|\theta \sim N(\theta, \sigma^2)$ where σ^2 is known. For such a likelihood, a normal prior for θ is conjugate. Suppose the prior is $\theta \sim N(\mu, \tau^2)$. The posterior density is

$$f(\theta|\mathbf{x}) \propto f(\theta) \prod_{i=1}^{n} f(x_i|\theta) \tag{1.33}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{(\theta-\mu)^2}{\tau^2} + \frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\sigma^2}\right)\right\} \tag{1.34}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\theta - \frac{\mu/\tau^2 + (n\bar{x})/\sigma^2}{1/\tau^2 + n/\sigma^2}\right)^2 \Bigg/ \left(\frac{1}{1/\tau^2 + n/\sigma^2}\right)\right\}, \tag{1.35}$$

where $\bar{x}$ is the sample mean. Recognizing (1.35) as being in the form of a normal distribution, we conclude that $f(\theta|\mathbf{x}) = N(\mu_n, \tau_n^2)$, where

$$\tau_n^2 = \frac{1}{1/\tau^2 + n/\sigma^2} \tag{1.36}$$

and

$$\mu_n = \left(\frac{\mu}{\tau^2} + \frac{n\bar{x}}{\sigma^2}\right)\tau_n^2. \tag{1.37}$$

Hence, a posterior 95% credibility interval for θ is $(\mu_n - 1.96\tau_n,\ \mu_n + 1.96\tau_n)$. Since the normal distribution is symmetric, this is also the posterior 95% HPD for θ.

For fixed σ, consider increasingly large choices for the value of τ. The posterior variance for θ converges to σ^2/n as $\tau^2 \to \infty$. In other words, the influence of the prior on the posterior vanishes as the prior variance increases. Next, note that

$$\lim_{n\to\infty} \frac{\tau_n^2}{\sigma^2/n} = 1.$$

This shows that the posterior variance for θ and the sampling variance for the MLE, $\hat{\theta} = \bar{X}$, are asymptotically equal, and the effect of any choice for τ is washed out with increasing sample size.

As an alternative to the conjugate prior, consider using the improper prior $f(\theta) \propto 1$. In this case, $f(\theta|\mathbf{x}) = N(\bar{x}, \sigma^2/n)$, and a 95% posterior credibility interval corresponds to the standard 95% confidence interval found using frequentist methods. □

1.6 STATISTICAL LIMIT THEORY

Although this book is mostly concerned with a pragmatic examination of how and why various methods work, it is useful from time to time to speak more precisely about the limiting behavior of the estimators produced by some procedures. We review below some basic convergence concepts used in probability and statistics.

A sequence of random variables, $X_1, X_2, \ldots$, is said to *converge in probability* to the random variable X if $\lim_{n\to\infty} P[|X_n - X| < \epsilon] = 1$ for every $\epsilon > 0$. The sequence *converges almost surely* to X if $P[\lim_{n\to\infty} |X_n - X| < \epsilon] = 1$ for every $\epsilon > 0$. The variables converge in distribution to the distribution of X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all points x at which $F_X(x)$ is continuous. The variable X has property A *almost everywhere* if $P[A] = \int 1_{\{A\}} f_X(x)\, dx = 1$.

Some of the best-known convergence theorems in statistics are the laws of large numbers and the central limit theorem. For i.i.d. sequences of one-dimensional random variables $X_1, X_2, \ldots$, let $\bar{X}_n = \sum_{i=1}^n X_i/n$. The *weak law of large numbers* states that $\bar{X}_n$ converges in probability to $\mu = E\{X_i\}$ if $E\{|X_i|\} < \infty$. The *strong law of large numbers* states that $\bar{X}_n$ converges almost surely to μ if $E\{|X_i|\} < \infty$. Both results hold under the more stringent but easily checked condition that $\mathrm{var}\{X_i\} = \sigma^2 < \infty$.

If θ is a parameter and T_n is a statistic based on $X_1, \ldots, X_n$, then T_n is said to be *weakly* or *strongly consistent* for θ if T_n converges in probability or almost surely (respectively) to θ. T_n is *unbiased* for θ if $E\{T_n\} = \theta$; otherwise the *bias* is $E\{T_n\} - \theta$. If the bias vanishes as $n \to \infty$, then T_n is *asymptotically unbiased*.

A simple form of the *central limit theorem* is as follows. Suppose that i.i.d. random variables $X_1, \ldots, X_n$ have mean μ and finite variance σ^2, and that $E\{\exp\{tX_i\}\}$ exists in a neighborhood of $t = 0$. Then the random variable $T_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges in distribution to a normal random variable with mean zero and variance one, as $n \to \infty$. There are many versions of the central limit theorem for various situations. Generally speaking, the assumption of finite variance is critical, but the assumptions of independence and identical distributions can be relaxed in certain situations.

1.7 MARKOV CHAINS

We offer here a brief introduction to univariate, discrete-time, discrete-state-space Markov chains. We will use Markov chains in Chapters 7 and 8. A thorough introduction to Markov chains is given in [556], and higher-level study is provided in [462, 543].

Consider a sequence of random variables $\left\{X^{(t)}\right\}$, $t = 0, 1, \ldots$, where each $X^{(t)}$ may equal one of a finite or countably infinite number of possible values, called *states*. The notation $X^{(t)} = j$ indicates that the process is in state j at time t. The *state space*, $\mathcal{S}$, is the set of possible values of the random variable $X^{(t)}$.

A complete probabilistic specification of $X^{(0)}, \ldots, X^{(n)}$ would be to write their joint distribution as the product of conditional distributions of each random variable given its history, or

$$\begin{aligned} P\left[X^{(0)}, \ldots, X^{(n)}\right] = P&\left[X^{(n)} \middle| x^{(0)}, \ldots, x^{(n-1)}\right] \\ &\times P\left[X^{(n-1)} \middle| x^{(0)}, \ldots, x^{(n-2)}\right] \times \cdots \\ &\times P\left[X^{(1)} \middle| x^{(0)}\right] P\left[X^{(0)}\right]. \end{aligned} \tag{1.38}$$

A simplification of (1.38) is possible under the conditional independence assumption that

$$P\left[X^{(t)} \middle| x^{(0)}, \ldots, x^{(t-1)}\right] = P\left[X^{(t)} \middle| x^{(t-1)}\right]. \tag{1.39}$$

Here the next state observed is only dependent upon the present state. This is the *Markov property*, sometimes called "one-step memory." In this case,

$$\begin{aligned} P\left[X^{(0)}, \ldots, X^{(n)}\right] &= P\left[X^{(n)} \middle| x^{(n-1)}\right] \\ &\times P\left[X^{(n-1)} \middle| x^{(n-2)}\right] \cdots P\left[X^{(1)} \middle| x^{(0)}\right] P\left[X^{(0)}\right]. \end{aligned} \tag{1.40}$$

TABLE 1.4 San Francisco rain data considered in Example 1.3.

	Wet Today	Dry Today
Wet Yesterday	418	256
Dry Yesterday	256	884

Let $p_{ij}^{(t)}$ be the probability that the observed state changes from state i at time t to state j at time $t+1$. The sequence $\left\{X^{(t)}\right\}$, $t = 0, 1, \ldots$ is a *Markov chain* if

$$\begin{aligned} p_{ij}^{(t)} &= P\left[X^{(t+1)} = j \,\middle|\, X^{(0)} = x^{(0)}, X^{(1)} = x^{(1)}, \ldots, X^{(t)} = i\right] \\ &= P\left[X^{(t+1)} = j \,\middle|\, X^{(t)} = i\right] \end{aligned} \tag{1.41}$$

for all $t = 0, 1, \ldots$ and $x^{(0)}, x^{(1)}, \ldots, x^{(t-1)}, i, j \in \mathcal{S}$. The quantity $p_{ij}^{(t)}$ is called the *one-step transition probability*. If none of the one-step transition probabilities change with t, then the chain is called *time homogeneous*, and $p_{ij}^{(t)} = p_{ij}$. If any of the one-step transition probabilities change with t, then the chain is called *time-inhomogeneous*.

A Markov chain is governed by a *transition probability matrix*. Suppose that the s states in $\mathcal{S}$ are, without loss of generality, all integer valued. Then $\mathbf{P}$ denotes $s \times s$ transition probability matrix of a time-homogeneous chain, and the (i, j)th element of $\mathbf{P}$ is p_{ij}. Each element in $\mathbf{P}$ must be between zero and one, and each row of the matrix must sum to one.

Example 1.3 (San Francisco Weather) Let us consider daily precipitation outcomes in San Francisco. Table 1.4 gives the rainfall status for 1814 pairs of consecutive days [488]. The data are taken from the months of November through March, starting in November of 1990 and ending in March of 2002. These months are when San Francisco receives over 80% of its precipitation each year, virtually all in the form of rain. We consider a binary classification of each day. A day is considered to be wet if more than 0.01 inch of precipitation is recorded and dry otherwise. Thus, $\mathcal{S}$ has two elements: "wet" and "dry." The random variable corresponding to the state for the tth day is $X^{(t)}$.

Assuming time homogeneity, an estimated transition probability matrix for $X^{(t)}$ would be

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.620 & 0.380 \\ 0.224 & 0.775 \end{bmatrix}. \tag{1.42}$$

Clearly, wet and dry weather states are not independent in San Francisco, as a wet day is more likely to be followed by a wet day and pairs of dry days are highly likely. □

The limiting theory of Markov chains is important for many of the methods discussed in this book. We now review some basic results.

A state to which the chain returns with probability 1 is called a *recurrent* state. A state for which the expected time until recurrence is finite is called *nonnull*. For finite state spaces, recurrent states are nonnull.

A Markov chain is *irreducible* if any state j can be reached from any state i in a finite number of steps for all i and j. In other words, for each i and j there must exist $m > 0$ such that $P\left[X^{(m+n)} = i \middle| X^{(n)} = j\right] > 0$. A Markov chain is *periodic* if it can visit certain portions of the state space only at certain regularly spaced intervals. State j has period d if the probability of going from state j to state j in n steps is 0 for all n not divisible by d. If every state in a Markov chain has period 1, then the chain is called *aperiodic*. A Markov chain is *ergodic* if it is irreducible, aperiodic, and all its states are nonnull and recurrent.

Let $\boldsymbol{\pi}$ denote a vector of probabilities that sum to one, with ith element π_i denoting the marginal probability that $X^{(t)} = i$. Then the marginal distribution of $X^{(t+1)}$ must be $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{P}$. Any discrete probability distribution $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{P} = \boldsymbol{\pi}^{\mathrm{T}}$ is called a *stationary* distribution for $\mathbf{P}$, or for the Markov chain having transition probability matrix $\mathbf{P}$. If $X^{(t)}$ follows a stationary distribution, then the marginal distributions of $X^{(t)}$ and $X^{(t+1)}$ are identical.

If a time-homogeneous Markov chain satisfies

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{1.43}$$

for all $i, j \in \mathcal{S}$, then $\boldsymbol{\pi}$ is a stationary distribution for the chain, and the chain is called *reversible* because the joint distribution of a sequence of observations is the same whether the chain is run forwards or backwards. Equation (1.43) is called the *detailed balance* condition.

If a Markov chain with transition probability matrix $\mathbf{P}$ and stationary distribution $\boldsymbol{\pi}$ is irreducible and aperiodic, then $\boldsymbol{\pi}$ is unique and

$$\lim_{n\to\infty} P\left[X^{(t+n)} = j \middle| X^{(t)} = i\right] = \pi_j, \tag{1.44}$$

where π_j is the jth element of $\boldsymbol{\pi}$. The π_j are the solutions of the following set of equations:

$$\pi_j \geq 0, \quad \sum_{i\in\mathcal{S}} \pi_i = 1, \quad \text{and} \quad \pi_j = \sum_{i\in\mathcal{S}} \pi_i p_{ij} \;\; \text{for each } j \in \mathcal{S}. \tag{1.45}$$

We can restate and extend (1.44) as follows. If $X^{(1)}, X^{(2)}, \ldots$ are realizations from an irreducible and aperiodic Markov chain with stationary distribution $\boldsymbol{\pi}$, then $X^{(n)}$ converges in distribution to the distribution given by $\boldsymbol{\pi}$, and for any function h,

$$\frac{1}{n}\sum_{t=1}^{n} h(X^{(t)}) \to E_{\boldsymbol{\pi}}\{h(X)\} \tag{1.46}$$

almost surely as $n \to \infty$, provided $E_{\boldsymbol{\pi}}\{|h(X)|\}$ exists [605]. This is one form of the *ergodic theorem*, which is a generalization of the strong law of large numbers.

We have considered here only Markov chains for discrete state spaces. In Chapters 7 and 8 we will apply these ideas to continuous state spaces. The principles and results for continuous state spaces and multivariate random variables are similar to the simple results given here.

1.8 COMPUTING

If you are new to computer programming, or wishing to learn a new language, there is no better time to start than now. Our preferred language for teaching and learning about statistical computing is R (freely available at www.r-project.org), but we avoid any language-specific limitations in this text. Most of the methods described in this book can also be easily implemented in other high-level computer languages for mathematics and statistics such as MATLAB. Programming in Java and low-level languages such as C++ and FORTRAN is also possible. The tradeoff between implementation ease for high-level languages and computation speed for low-level languages often guides this selection. Links to these and other useful software packages, including libraries of code for some of the methods described in this book, are available on the book website.

Ideally, your computer programming background includes a basic understanding of computer arithmetic: how real numbers and mathematical operations are implemented in the binary world of a computer. We focus on higher-level issues in this book, but the most meticulous implementation of the algorithms we describe can require consideration of the vagaries of computer arithmetic, or use of available routines that competently deal with such issues. Interested readers may refer to [383].