vector. Although we generally consider X to be a univariate random variable, for several of our examples it is a random vector.

As in Chapter 4, our point estimator of θ is $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$, where $\widehat{\theta}$ maximizes the function $L(\theta)$. We call $\widehat{\theta}$ the maximum likelihood estimator (mle) of θ. In Section 4.1, several motivating examples were given, including the binomial and normal probability models. Later we give several more examples, but first we offer a theoretical justification for considering the mle. Let θ_0 denote the *true value* of θ. Theorem 6.1.1 shows that the maximum of $L(\theta)$ asymptotically separates the true model at θ_0 from models at $\theta \neq \theta_0$. To prove this theorem, we assume certain assumptions, usually called *regularity conditions.*

Assumptions 6.1.1 (Regularity Conditions). *Regularity conditions (R0)–(R1) are given by*

(R0) *The pdfs are distinct; i.e., $\theta \neq \theta' \Rightarrow f(x_i;\theta) \neq f(x_i;\theta')$.*

(R1) *The pdfs have common support for all θ.*

(R2) *The point θ_0 is an interior point in Ω.*

The first assumption states that the parameter identifies the pdf. The second assumption implies that the support of X_i does not depend on θ. This is restrictive, and some examples and exercises cover models in which (R1) is not true.

Theorem 6.1.1. *Let θ_0 be the true parameter. Under assumptions (R0) and (R1),*

$$\lim_{n\to\infty} P_{\theta_0}[L(\theta_0,\mathbf{X}) > L(\theta,\mathbf{X})] = 1, \quad \textit{for all } \theta \neq \theta_0. \tag{6.1.3}$$

Proof: By taking logs, the inequality $L(\theta_0,\mathbf{X}) > L(\theta,\mathbf{X})$ is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right] < 0.$$

Since the summands are iid with finite expectation and the function $\phi(x) = -\log(x)$ is strictly convex, it follows from the Law of Large Numbers (Theorem 5.1.1) and Jensen's inequality (Theorem 1.10.5) that, when θ_0 is the true parameter,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right] \xrightarrow{P} E_{\theta_0}\left[\log \frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right] < \log E_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right].$$

But

$$E_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right] = \int \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)\,dx = 1.$$

Because $\log 1 = 0$, the theorem follows. Note that common support is needed to obtain the last equalities. ■

Theorem 6.1.1 says that asymptotically the likelihood function is maximized at the true value θ_0. So in considering estimates of θ_0, it seems natural to consider the value of θ which maximizes the likelihood.

Definition 6.1.1 (Maximum Likelihood Estimator). *We say that $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ is a* **maximum likelihood estimator** *(mle) of θ if*

$$\widehat{\theta} = Argmax\, L(\theta;\mathbf{X}). \tag{6.1.4}$$

The notation Argmax means that $L(\theta;\mathbf{X})$ achieves its maximum value at $\widehat{\theta}$.

As in Chapter 4, to determine the mle, we often take the log of the likelihood and determine its critical value; that is, letting $l(\theta) = \log L(\theta)$, the mle solves the equation

$$\frac{\partial l(\theta)}{\partial\theta} = 0. \tag{6.1.5}$$

This is an example of an **estimating equation**, which we often label as an EE. This is the first of several EEs in the text.

Example 6.1.1 (Laplace Distribution). Let $X_1, \ldots, X_n$ be iid with density

$$f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}, \quad -\infty < x < \infty, -\infty < \theta < \infty. \tag{6.1.6}$$

This pdf is referred to as either the *Laplace* or the *double exponential distribution.* The log of the likelihood simplifies to

$$l(\theta) = -n\log 2 - \sum_{i=1}^{n}|x_i - \theta|.$$

The first partial derivative is

$$l'(\theta) = \sum_{i=1}^{n}\operatorname{sgn}(x_i - \theta), \tag{6.1.7}$$

where $\operatorname{sgn}(t) = 1, 0,$ or -1 depending on whether $t > 0, t = 0,$ or $t < 0$. Note that we have used $\frac{d}{dt}|t| = \operatorname{sgn}(t)$, which is true unless $t = 0$. Setting equation (6.1.7) to 0, the solution for θ is $\operatorname{med}\{x_1, x_2, \ldots, x_n\}$, because the median makes half the terms of the sum in expression (6.1.7) nonpositive and half nonnegative. Recall that we denote the median of a sample by Q_2 (the second quartile of the sample). Hence, $\widehat{\theta} = Q_2$ is the mle of θ for the Laplace pdf (6.1.6). ■

There is no guarantee that the mle exists or, if it does, whether it is unique. This is often clear from the application as in the next two examples. Other examples are given in the exercises.

Example 6.1.2 (Logistic Distribution). Let $X_1, \ldots, X_n$ be iid with density

$$f(x;\theta) = \frac{\exp\{-(x-\theta)\}}{(1+\exp\{-(x-\theta)\})^2}, \quad -\infty < x < \infty, -\infty < \theta < \infty. \tag{6.1.8}$$

The log of the likelihood simplifies to

$$l(\theta) = \sum_{i=1}^{n}\log f(x_i;\theta) = n\theta - n\overline{x} - 2\sum_{i=1}^{n}\log(1+\exp\{-(x_i-\theta)\}).$$