estimated standard error

$$s_{\hat{\theta}} = \frac{1}{\sqrt{n}}\sigma(\hat{\theta})$$

We now claim that the consistency of $\hat{\theta}$ implies that $s_{\hat{\theta}} \approx \sigma_{\hat{\theta}}$. More precisely,

$$\lim_{n\to\infty} \frac{s_{\hat{\theta}}}{\sigma_{\hat{\theta}}} = 1$$

provided that the function $\sigma(\theta)$ is continuous in θ. The result follows since if $\hat{\theta} \to \theta_0$, then $\sigma(\hat{\theta}) \to \sigma(\theta_0)$. Of course, this is just a limiting result and we always have a finite value of n in practice, but it does provide some hope that the ratio will be close to 1 and that the estimated standard error will be a reasonable indication of variability.

Let us summarize the results of this section. We have shown how the method of moments can provide estimates of the parameters of a probability distribution based on a "sample" (an i.i.d. collection) of random variables from that distribution. We addressed the question of variability or reliability of the estimates by observing that if the sample is random, the parameter estimates are random variables having distributions that are referred to as their sampling distributions. The standard deviation of the sampling distribution is called the *standard error of the estimate.* We then faced the problem of how to ascertain the variability of an estimate from the sample itself. In some cases the sampling distribution was of an explicit form depending upon the unknown parameters (Examples A and B); in these cases we could substitute our estimates for the unknown parameters in order to approximate the sampling distribution. In other cases the form of the sampling distribution was not so obvious, but we realized that even if we didn't know it explicitly, we could simulate it. By using the bootstrap we avoided doing perhaps difficult analytic calculations by sitting back and instructing a computer to generate random numbers.

8.5 The Method of Maximum Likelihood

As well as being a useful tool for parameter estimation in our current context, the method of maximum likelihood can be applied to a great variety of other statistical problems, such as curve fitting, for example. This general utility is one of the major reasons for the importance of likelihood methods in statistics. We will later see that maximum likelihood estimates have nice theoretical properties as well.

Suppose that random variables $X_1, \ldots, X_n$ have a joint density or frequency function $f(x_1, x_2, \ldots, x_n|\theta)$. Given observed values $X_i = x_i$, where $i = 1, \ldots, n$, the likelihood of θ as a function of $x_1, x_2, \ldots, x_n$ is defined as

$$\text{lik}(\theta) = f(x_1, x_2, \ldots, x_n|\theta)$$

Note that we consider the joint density as a function of θ rather than as a function of the x_i. If the distribution is discrete, so that f is a frequency function, the likelihood function gives the probability of observing the given data as a function of the parameter θ. The **maximum likelihood estimate (mle)** of θ is that value of θ that maximizes the likelihood—that is, makes the observed data "most probable" or "most likely."

If the X_i are assumed to be i.i.d., their joint density is the product of the marginal densities, and the likelihood is

$$\text{lik}(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

Rather than maximizing the likelihood itself, it is usually easier to maximize its natural logarithm (which is equivalent since the logarithm is a monotonic function). For an i.i.d. sample, the **log likelihood** is

$$l(\theta) = \sum_{i=1}^{n} \log[f(X_i|\theta)]$$

(In this text, "log" will always mean the natural logarithm.)

Let us find the maximum likelihood estimates for the examples first considered in Section 8.4.

EXAMPLE **A** *Poisson Distribution*

If X follows a Poisson distribution with parameter λ, then

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

If $X_1, \ldots, X_n$ are i.i.d. and Poisson, their joint frequency function is the product of the marginal frequency functions. The log likelihood is thus

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n} (X_i \log \lambda - \lambda - \log X_i!) \\ &= \log \lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log X_i! \end{aligned}$$

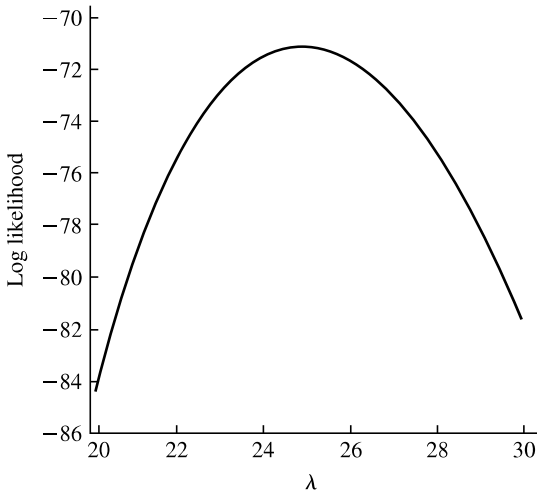

FIGURE **8.5** Plot of the log likelihood function of λ for asbestos data.

Figure 8.5 is a graph of $l(\lambda)$ for the asbestos counts of Example A in Section 8.4. Setting the first derivative of the log likelihood equal to zero, we find

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = 0$$

The mle is then

$$\hat{\lambda} = \overline{X}$$

We can check that this is indeed a maximum (in fact, $l(\lambda)$ is a concave function of λ; see Figure 8.5). The maximum likelihood estimate agrees with the method of moments for this case and thus has the same sampling distribution. ■

EXAMPLE **B** *Normal Distribution*

If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, their joint density is the product of their marginal densities:

$$f(x_1, x_2, \ldots, x_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma}\right]^2\right)$$

Regarded as a function of μ and σ, this is the likelihood function. The log likelihood is thus

$$l(\mu, \sigma) = -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

The partials with respect to μ and σ are

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^{n}(X_i - \mu)^2$$

Setting the first partial equal to zero and solving for the mle, we obtain

$$\hat{\mu} = \overline{X}$$

Setting the second partial equal to zero and substituting the mle for μ, we find that the mle for σ is

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Again, these estimates and their sampling distributions are the same as those obtained by the method of moments. ■

EXAMPLE C *Gamma Distribution*

Since the density function of a gamma distribution is

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}, \qquad 0 \le x < \infty$$

the log likelihood of an i.i.d. sample, $X_1, \ldots, X_n$, is

$$\begin{aligned} l(\alpha, \lambda) &= \sum_{i=1}^{n}[\alpha \log \lambda + (\alpha - 1)\log X_i - \lambda X_i - \log \Gamma(\alpha)] \\ &= n\alpha \log \lambda + (\alpha - 1)\sum_{i=1}^{n} \log X_i - \lambda \sum_{i=1}^{n} X_i - n \log \Gamma(\alpha) \end{aligned}$$

The partial derivatives are

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= n \log \lambda + \sum_{i=1}^{n} \log X_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ \frac{\partial l}{\partial \lambda} &= \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i \end{aligned}$$

Setting the second partial equal to zero, we find

$$\hat{\lambda} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n} X_i} = \frac{\hat{\alpha}}{\overline{X}}$$

But when this solution is substituted into the equation for the first partial, we obtain a nonlinear equation for the mle of α:

$$n \log \hat{\alpha} - n \log \overline{X} + \sum_{i=1}^{n} \log X_i - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0$$

This equation cannot be solved in closed form; an iterative method for finding the roots has to be employed. To start the iterative procedure, we could use the initial value obtained by the method of moments.

For this example, the two methods do not give the same estimates. The mle's are computed from the precipitation data of Example C in Section 8.4 by an iterative procedure (a combination of the secant method and the method of bisection) using the method of moments estimates as starting values. The resulting estimates are $\hat{\alpha} = .441$ and $\hat{\lambda} = 1.96$. In Example C in Section 8.4, the method of moments estimates were found to be $\hat{\alpha} = .375$ and $\hat{\lambda} = 1.674$. Figure 8.3 shows fitted densities from both types of estimates of α and λ. There is clearly little practical difference, especially if we keep in mind that the gamma distribution is only a possible model and should not be taken as being literally true.

Because the maximum likelihood estimates are not given in closed form, obtaining their exact sampling distribution would appear to be intractable. We thus use the bootstrap to approximate these distributions, just as we did to approximate the sampling distributions of the method of moments estimates. The underlying rationale is the same: If we knew the "true" values, α_0 and λ_0, say, we could approximate

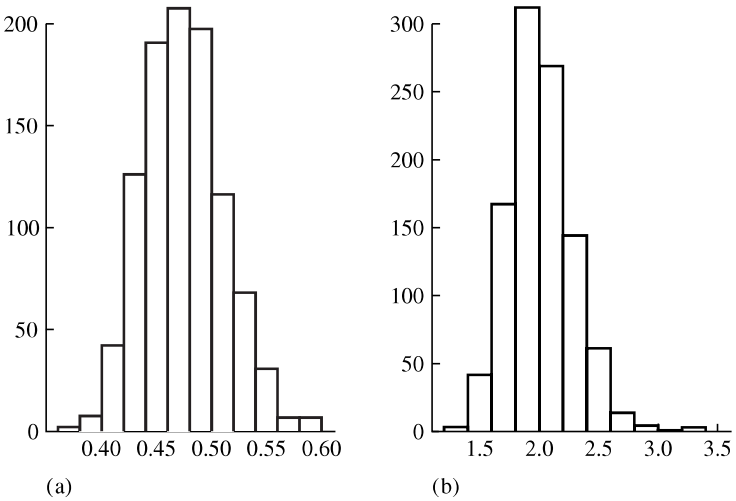

FIGURE **8.6** Histograms of 1000 simulated maximum likelihood estimates of (a) α and (b) λ.

the sampling distribution of their maximum likelihood estimates by generating many, many samples of size $n = 227$ from a gamma distribution with parameters α_0 and λ_0, forming the maximum likelihood estimates from each sample, and displaying the results in histograms. Since, of course, we don't know the true values, we let our maximum likelihood estimates play their role: We generated 1000 samples each of size $n = 227$ of gamma distributed random variables with $\alpha = .471$ and $\lambda = 1.97$. For each of these samples, the maximum likelihood estimates of α and λ were calculated. Histograms of these 1000 estimates are shown in Figure 8.6; we regard these histograms as approximations to the sampling distribution of the maximum likelihood estimates $\hat{\alpha}$ and $\hat{\lambda}$.

Comparison of Figures 8.6 and 8.4 is interesting. We see that the sampling distributions of the maximum likelihood estimates are substantially less dispersed than those of the method of moments estimates, which indicates that in this situation, the method of maximum likelihood is more precise than the method of moments. The standard deviations of the values displayed in the histograms are the estimated standard errors of the maximum likelihood estimates; we find $s_{\hat{\alpha}} = .03$ and $s_{\hat{\lambda}} = .26$. Recall that in Example C of Section 8.4 the corresponding estimated standard errors for the method of moments estimates were found to be .06 and .34. ■

EXAMPLE **D** *Muon Decay*

From the form of the density given in Example D in Section 8.4, the log likelihood is

$$l(\alpha) = \sum_{i=1}^{n} \log(1 + \alpha X_i) - n \log 2$$

Setting the derivative equal to zero, we see that the mle of α satisfies the following

nonlinear equation:

$$\sum_{i=1}^{n} \frac{X_i}{1 + \hat{\alpha} X_i} = 0$$

Again, we would have to use an iterative technique to solve for $\hat{\alpha}$. The method of moments estimate could be used as a starting value. ■

In Examples C and D, in order to find the maximum likelihood estimate, we would have to solve a nonlinear equation. In general, in some problems involving several parameters, systems of nonlinear equations must be solved to find the mle's. We will not discuss numerical methods here; a good discussion is found in Chapter 6 of Dahlquist and Bjorck (1974).

8.5.1 Maximum Likelihood Estimates of Multinomial Cell Probabilities

The method of maximum likelihood is often applied to problems involving multinomial cell probabilities. Suppose that $X_1, \ldots, X_m$, the counts in cells $1, \ldots, m$, follow a multinomial distribution with a total count of n and cell probabilities $p_1, \ldots, p_m$. We wish to estimate the p's from the x's. The joint frequency function of $X_1, \ldots, X_m$ is

$$f(x_1, \ldots, x_m | p_1, \ldots, p_m) = \frac{n!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_i^{x_i}$$

Note that the marginal distribution of each X_i is binomial (n, p_i), and that since the X_i are not independent (they are constrained to sum to n), their joint frequency function is not the product of the marginal frequency functions, as it was in the examples considered in the preceding section. We can, however, still use the method of maximum likelihood since we can write an expression for the joint distribution. We assume n is given, and we wish to estimate $p_1, \ldots, p_m$ with the constraint that the p_i sum to 1. From the joint frequency function just given, the log likelihood is

$$l(p_1, \ldots, p_m) = \log n! - \sum_{i=1}^{m} \log x_i! + \sum_{i=1}^{m} x_i \log p_i$$

To maximize this likelihood subject to the constraint, we introduce a Lagrange multiplier and maximize

$$L(p_1, \ldots, p_m, \lambda) = \log n! - \sum_{i=1}^{m} \log x_i! + \sum_{i=1}^{m} x_i \log p_i + \lambda \left(\sum_{i=1}^{m} p_i - 1 \right)$$

Setting the partial derivatives equal to zero, we have the following system of equations:

$$\hat{p}_j = -\frac{x_j}{\lambda}, \qquad j = 1, \ldots, m$$

Summing both sides of this equation, we have

$$1 = \frac{-n}{\lambda}$$

or

$$\lambda = -n$$

Therefore,

$$\hat{p}_j = \frac{x_j}{n}$$

which is an obvious set of estimates. The sampling distribution of $\hat{p}_j$ is determined by the distribution of x_j, which is binomial.

In some situations, such as frequently occur in the study of genetics, the multinomial cell probabilities are functions of other unknown parameters θ; that is, $p_i = p_i(\theta)$. In such cases, the log likelihood of θ is

$$l(\theta) = \log n! - \sum_{i=1}^{m} \log x_i! + \sum_{i=1}^{m} x_i \log p_i(\theta)$$

EXAMPLE A *Hardy-Weinberg Equilibrium*

If gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur in a population with frequencies $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2, according to the Hardy-Weinberg law. In a sample from the Chinese population of Hong Kong in 1937, blood types occurred with the following frequencies, where M and N are erythrocyte antigens:

	Blood Type			
	M	MN	N	Total
Frequency	342	500	187	1029

There are several possible ways to estimate θ from the observed frequencies. For example, if we equate θ^2 with 187/1029, we obtain .4263 as an estimate of θ. Intuitively, however, it seems that this procedure ignores some of the information in the other cells. If we let X_1, X_2, and X_3 denote the counts in the three cells and let $n = 1029$, the log likelihood of θ is (you should check this):

$$\begin{aligned} l(\theta) &= \log n! - \sum_{i=1}^{3} \log X_i! + X_1 \log(1-\theta)^2 + X_2 \log 2\theta(1-\theta) + X_3 \log \theta^2 \\ &= \log n! - \sum_{i=1}^{3} \log X_i! + (2X_1 + X_2)\log(1-\theta) \\ &\quad + (2X_3 + X_2)\log\theta + X_2 \log 2 \end{aligned}$$

In maximizing $l(\theta)$, we do not need to explicitly incorporate the constraint that the cell probabilities sum to 1 since the functional form of $p_i(\theta)$ is such that $\sum_{i=1}^{3} p_i(\theta) = 1$.

Setting the derivative equal to zero, we have

$$-\frac{2X_1 + X_2}{1-\theta} + \frac{2X_3 + X_2}{\theta} = 0$$

Solving this, we obtain the mle:

$$\begin{aligned}\hat{\theta} &= \frac{2X_3 + X_2}{2X_1 + 2X_2 + 2X_3} \\ &= \frac{2X_3 + X_2}{2n} \\ &= \frac{2 \times 187 + 500}{2 \times 1029} = .4247\end{aligned}$$

How precise is this estimate? Do we have faith in the accuracy of the first, second, third, or fourth decimal place? We will address these questions by using the bootstrap to estimate the sampling distribution and the standard error of $\hat{\theta}$. The bootstrap logic is as follows: If θ were known, then the three multinomial cell probabilities, $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2, would be known. To find the sampling distribution of $\hat{\theta}$, we could simulate many multinomial random variables with these probabilities and $n = 1029$, and for each we could form an estimate of θ. A histogram of these estimates would be an approximation to the sampling distribution. Since, of course, we don't know the actual value of θ to use in such a simulation, the bootstrap principle tells us to use $\hat{\theta} = .4247$ in its place. With this estimated value of θ the three cell probabilities (M, MN, N) are .331, .489, and .180. One thousand multinomial random counts, each with total count 1029, were simulated with these probabilities (see problem 35 at the end of the chapter for the method of generating these random counts). From each of these 1000 computer "experiments," a value θ^* was determined. A histogram of the estimates (Figure 8.7) can be regarded as an estimate of the sampling distribution of $\hat{\theta}$. The estimated standard error of $\hat{\theta}$ is the standard deviation of these 1000 values: $s_{\hat{\theta}} = .011$. ■

8.5.2 Large Sample Theory for Maximum Likelihood Estimates

In this section we develop approximations to the sampling distribution of maximum likelihood estimates by using limiting arguments as the sample size increases. The theory we shall sketch shows that under reasonable conditions, maximum likelihood estimates are consistent. We also develop a useful and important approximation for the variance of a maximum likelihood estimate and argue that for large sample sizes, the sampling distribution is approximately normal.

The rigorous development of this large sample theory is quite technical; we will simply state some results and give very rough, heuristic arguments for the case of an i.i.d. sample and a one-dimensional parameter. (The arguments for Theorems A and B may be skipped without loss of continuity. Rigorous proofs may be found in Cramér (1946).)

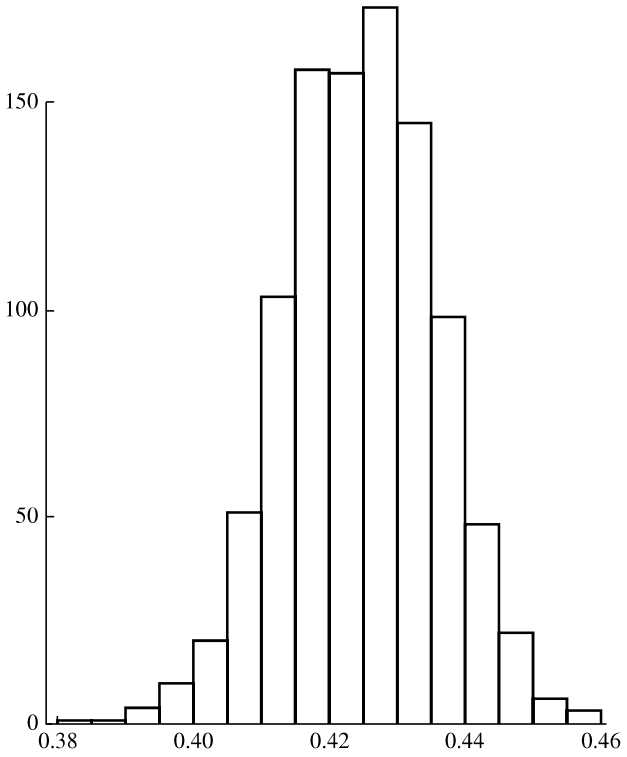

FIGURE **8.7** Histogram of 1000 simulated maximum likelihood estimates of θ described in Example A.

For an i.i.d. sample of size n, the log likelihood is

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

We denote the true value of θ by θ_0. It can be shown that under reasonable conditions $\hat{\theta}$ is a consistent estimate of θ_0; that is, $\hat{\theta}$ converges to θ_0 in probability as n approaches infinity.

THEOREM **A**

Under appropriate smoothness conditions on f, the mle from an i.i.d. sample is consistent.

Proof

The following is merely a sketch of the proof. Consider maximizing

$$\frac{1}{n} l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\theta)$$

As n tends to infinity, the law of large numbers implies that

$$\begin{aligned}\frac{1}{n}l(\theta) &\to E\log f(X|\theta)\\ &= \int \log f(x|\theta) f(x|\theta_0)\, dx\end{aligned}$$

It is thus plausible that for large n, the θ that maximizes $l(\theta)$ should be close to the θ that maximizes $E\log f(X|\theta)$. (An involved argument is necessary to establish this.) To maximize $E\log f(X|\theta)$, we consider its derivative:

$$\frac{\partial}{\partial\theta}\int \log f(x|\theta) f(x|\theta_0)\, dx = \int \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} f(x|\theta_0)\, dx$$

If $\theta = \theta_0$, this equation becomes

$$\int \frac{\partial}{\partial\theta} f(x|\theta_0)\, dx = \frac{\partial}{\partial\theta}\int f(x|\theta_0)\, dx = \frac{\partial}{\partial\theta}(1) = 0$$

which shows that θ_0 is a stationary point and hopefully a maximum. Note that we have interchanged differentiation and integration and that the assumption of smoothness on f must be strong enough to justify this. ■

We will now derive a useful intermediate result.

LEMMA **A**

Define $I(\theta)$ by

$$I(\theta) = E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2$$

Under appropriate smoothness conditions on f, $I(\theta)$ may also be expressed as

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$$

Proof

First, we observe that since $\int f(x|\theta)\, dx = 1$,

$$\frac{\partial}{\partial\theta}\int f(x|\theta)\, dx = 0$$

Combining this with the identity

$$\frac{\partial}{\partial\theta} f(x|\theta) = \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)$$

we have

$$0 = \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx$$

where we have interchanged differentiation and integration (some assumptions must be made in order to do this). Taking second derivatives of the preceding expressions, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx \\ &= \int \left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] f(x|\theta)\,dx + \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^2 f(x|\theta)\,dx \end{aligned}$$

From this, the desired result follows. ■

The large sample distribution of a maximum likelihood estimate is approximately normal with mean θ_0 and variance $1/[nI(\theta_0)]$. Since this is merely a limiting result, which holds as the sample size tends to infinity, we say that the mle is **asymptotically unbiased** and refer to the variance of the limiting normal distribution as the **asymptotic variance of the mle.**

THEOREM **B**

Under smoothness conditions on f, the probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0)$ tends to a standard normal distribution.

Proof

The following is merely a sketch of the proof; the details of the argument are beyond the scope of this book. From a Taylor series expansion,

$$\begin{aligned} 0 = l'(\hat{\theta}) &\approx l'(\theta_0) + (\hat{\theta}-\theta_0)l''(\theta_0) \\ (\hat{\theta}-\theta_0) &\approx \frac{-l'(\theta_0)}{l''(\theta_0)} \\ n^{1/2}(\hat{\theta}-\theta_0) &\approx \frac{-n^{-1/2}l'(\theta_0)}{n^{-1}l''(\theta_0)} \end{aligned}$$

First, we consider the numerator of this last expression. Its expectation is

$$\begin{aligned} E[n^{-1/2}l'(\theta_0)] &= n^{-1/2}\sum_{i=1}^{n} E\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right] \\ &= 0 \end{aligned}$$

as in Theorem A. Its variance is

$$\begin{aligned}\text{Var}[n^{-1/2}l'(\theta_0)] &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right]^2 \\ &= I(\theta_0)\end{aligned}$$

Next, we consider the denominator:

$$\frac{1}{n}l''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta_0)$$

By the law of large numbers, the latter expression converges to

$$E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0)\right] = -I(\theta_0)$$

from Lemma A.

We thus have

$$n^{1/2}(\hat{\theta}-\theta_0) \approx \frac{n^{-1/2}l'(\theta_0)}{I(\theta_0)}$$

Therefore,

$$E[n^{1/2}(\hat{\theta}-\theta_0)] \approx 0$$

Furthermore,

$$\begin{aligned}\text{Var}[n^{1/2}(\hat{\theta}-\theta_0)] &\approx \frac{I(\theta_0)}{I^2(\theta_0)} \\ &= \frac{1}{I(\theta_0)}\end{aligned}$$

and thus

$$\text{Var}(\hat{\theta}-\theta_0) \approx \frac{1}{nI(\theta_0)}$$

The central limit theorem may be applied to $l'(\theta_0)$, which is a sum of i.i.d. random variables:

$$l'(\theta_0) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta_0}\log f(X_i|\theta)$$ ■

Another interpretation of the result of Theorem B is as follows. For an i.i.d. sample, the maximum likelihood estimate is the maximizer of the log likelihood function,

$$l(\theta) = \sum_{i=1}^{n}\log f(X_i|\theta)$$

The asymptotic variance is

$$\frac{1}{nI(\theta_0)} = -\frac{1}{El''(\theta_0)}$$

when $El''(\theta_0)$ is large, $l(\theta)$ is, on average, changing very rapidly in a vicinity of θ_0 and the variance of the maximizer is small.

A corresponding result can be proved from the multidimensional case. The vector of maximum likelihood estimates is asymptotically normally distributed. The mean of the asymptotic distribution is the vector of true parameters, θ_0. The covariance of the estimates $\hat{\theta}_i$ and $\hat{\theta}_j$ is given by the ij entry of the matrix $n^{-1}I^{-1}(\theta_0)$, where $I(\theta)$ is the matrix with ij component

$$E\left[\frac{\partial}{\partial\theta_i}\log f(X|\theta)\frac{\partial}{\partial\theta_j}\log f(X|\theta)\right] = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X|\theta)\right]$$

Since we do not wish to delve deeply into technical details, we do not specify the conditions under which the results obtained in this section hold. It is worth mentioning, however, that the true parameter value, θ_0, is required to be an interior point of the set of all parameter values. Thus the results would not be expected to apply in Example D of Section 8.5 if $\alpha_0 = 1$, for example. It is also required that the support of the density or frequency function $f(x|\theta)$ [the set of values for which $f(x|\theta) > 0$] does not depend on θ. Thus, for example, the results would not be expected to apply to estimating θ from a sample of random variables that were uniformly distributed on the interval $[0, \theta]$.

The following sections will apply these results in several examples.

8.5.3 Confidence Intervals from Maximum Likelihood Estimates

In Chapter 7, confidence intervals for the population mean μ were introduced. Recall that the confidence interval for μ was a random interval that contained μ with some specified probability. In the current context, we are interested in estimating the parameter θ of a probability distribution. We will develop confidence intervals for θ based on $\hat{\theta}$; these intervals serve essentially the same function as they did in Chapter 7 in that they express in a fairly direct way the degree of uncertainty in the estimate $\hat{\theta}$. A confidence interval for θ is an interval based on the sample values used to estimate θ. Since these sample values are random, the interval is random and the probability that it contains θ is called the coverage probability of the interval. Thus, for example, a 90% confidence interval for θ is a random interval that contains θ with probability .9. A confidence interval quantifies the uncertainty of a parameter estimate.

We will discuss three methods for forming confidence intervals for maximum likelihood estimates: exact methods, approximations based on the large sample properties of maximum likelihood estimates, and bootstrap confidence intervals. The construction of confidence intervals for parameters of a normal distribution illustrates the use of exact methods.

EXAMPLE A We found in Example B of Section 8.5 that the maximum likelihood estimates of μ and σ^2 from an i.i.d. normal sample are

$$\hat{\mu} = \overline{X}$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$