CHAPTER 8

Estimation of Parameters and Fitting of Probability Distributions

8.1 Introduction

In this chapter, we discuss fitting probability laws to data. Many families of probability laws depend on a small number of parameters; for example, the Poisson family depends on the parameter λ (the mean number of counts), and the Gaussian family depends on two parameters, μ and σ. Unless the values of parameters are known in advance, they must be estimated from data in order to fit the probability law.

After parameter values have been chosen, the model should be compared to the actual data to see if the fit is reasonable; Chapter 9 is concerned with measures and tests of goodness of fit.

In order to introduce and illustrate some of the ideas and to provide a concrete basis for later theoretical discussions, we will first consider a classical example—the fitting of a Poisson distribution to radioactive decay. The concepts introduced in this example will be elaborated in this and the next chapter.

8.2 Fitting the Poisson Distribution to Emissions of Alpha Particles

Records of emissions of alpha particles from radioactive sources show that the number of emissions per unit of time is not constant but fluctuates in a seemingly random fashion. If the underlying rate of emission is constant over the period of observation (which will be the case if the half-life is much longer than the time period of observation) and if the particles come from a very large number of independent sources (atoms), the Poisson model seems appropriate. For this reason, the Poisson distribution is frequently used as a model for radioactive decay. You should recall that the

Poisson distribution as a model for random counts in space or time rests on three assumptions: (1) the underlying rate at which the events occur is constant in space or time, (2) events in disjoint intervals of space or time occur independently, and (3) there are no multiple events.

Berkson (1966) conducted a careful analysis of data obtained from the National Bureau of Standards. The source of the alpha particles was americium 241. The experimenters recorded 10,220 times between successive emissions. The observed mean emission rate (total number of emissions divided by total time) was .8392 emissions per sec. The clock used to record the times was accurate to .0002 sec.

The first two columns of the following table display the counts, n, that were observed in 1207 intervals, each of length 10 sec. In 18 of the 1207 intervals, there were 0, 1, or 2 counts; in 28 of the intervals there were 3 counts, etc.

n	Observed	Expected
0–2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9
17+	5	7.1
	1207	1207

In fitting a Poisson distribution to the counts shown in the table, we view the 1207 counts as 1207 independent realizations of Poisson random variables, each of which has the probability mass function

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

In order to fit the Poisson distribution, we must estimate a value for λ from the observed data. Since the average count in a 10-second interval was 8.392, we take this as an estimate of λ (recall that the $E(X) = \lambda$) and denote it by $\hat{\lambda}$.

Before continuing, we want to mention some issues that will be explored in depth in subsequent sections of this chapter. First, observe that if the experiment

were to be repeated, the counts would be different and the estimate of λ would be different; it is thus appropriate to regard the estimate of λ as a random variable which has a probability distribution referred to as its **sampling distribution.** The situation is entirely analogous to tossing a coin 10 times and regarding the number of heads as a binomially distributed random variable. Doing so and observing 6 heads generates one realization of this random variable; in the same sense 8.392 is a realization of a random variable. The question thus arises: what is the sampling distribution? This is of some practical interest, since the spread of the sampling distribution reflects the variability of the estimate. We could ask crudely, to what decimal place is the estimate 8.392 accurate? Second, later in this chapter we will discuss the rationale for choosing to estimate λ as we have done. Although estimating λ as the observed mean count is quite reasonable on its face, in principle there might be better procedures.

We now turn to assessing goodness of fit, a subject that will be taken up in depth in the next chapter. Consider the 16 cells into which the counts are grouped. Under the hypothesized model, the probability that a random count falls in any one of the cells may be calculated from the Poisson probability law. The probability that an observation falls in the first cell (0, 1, or 2 counts) is

$$p_1 = \pi_0 + \pi_1 + \pi_2$$

The probability that an observation falls in the second cell is $p_2 = \pi_3$. The probability that an observation falls in the 16th cell is

$$p_{16} = \sum_{k=17}^{\infty} \pi_k$$

Under the assumption that $X_1, \ldots, X_{1207}$ are independent Poisson random variables, the number of observations out of 1207 falling in a given cell follows a binomial distribution with a mean, or expected value, of $1207\,p_k$, and the joint distribution of the counts in all the cells is multinomial with $n = 1207$ and probabilities $p_1, p_2, \ldots, p_{16}$. The third column of the preceding table gives the expected number of counts in each cell; for example, because $p_4 = .0786$, the expected count in the corresponding cell is $1207 \times .0786 = 94.9$. Qualitatively, there is good agreement between the expected and observed counts. Quantitative measures will be presented in Chapter 9.

8.3 Parameter Estimation

As was illustrated in the example of alpha particle emissions, in order to fit a probability law to data, one typically has to estimate parameters associated with the probability law from the data. The following examples further illustrate this point.

EXAMPLE A *Normal Distribution*

The normal, or Gaussian, distribution involves two parameters, μ and σ, where μ is the mean of the distribution and σ^2 is the variance:

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \qquad -\infty < x < \infty$$

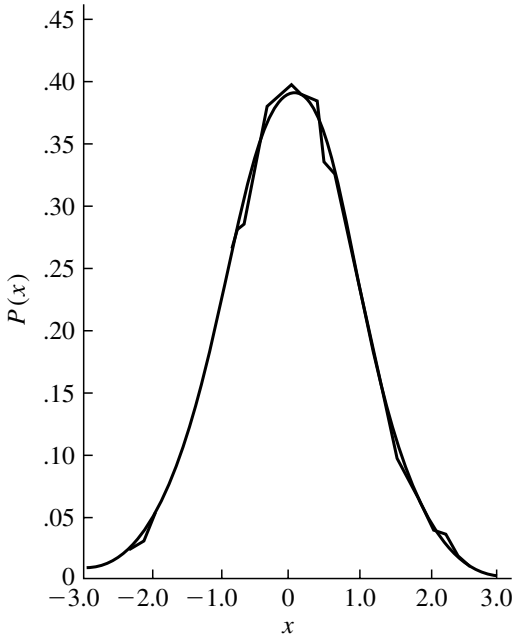

FIGURE **8.1** Gaussian fit of current flow across a cell membrane to a frequency polygon.

The use of the normal distribution as a model is usually justified using some version of the central limit theorem, which says that the sum of a large number of independent random variables is approximately normally distributed. For example, Bevan, Kullberg, and Rice (1979) studied random fluctuations of current across a muscle cell membrane. The cell membrane contained a large number of channels, which opened and closed at random and were assumed to operate independently. The net current resulted from ions flowing through open channels and was therefore the sum of a large number of roughly independent currents. As the channels opened and closed, the net current fluctuated randomly. Figure 8.1 shows a smoothed histogram of values obtained from 49,152 observations of the net current and an approximating Gaussian curve. The fit of the Gaussian distribution is quite good, although the smoothed histogram seems to show a slight skewness. In this application, information about the characteristics of the individual channels, such as conductance, was extracted from the estimated parameters μ and σ^2. ■

EXAMPLE **B** *Gamma Distribution*

The gamma distribution depends on two parameters, α and λ:

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

The family of gamma distributions provides a flexible set of densities for nonnegative random variables.

Figure 8.2 shows how the gamma distribution fits to the amounts of rainfall from different storms (Le Cam and Neyman 1967). Gamma distributions were fit

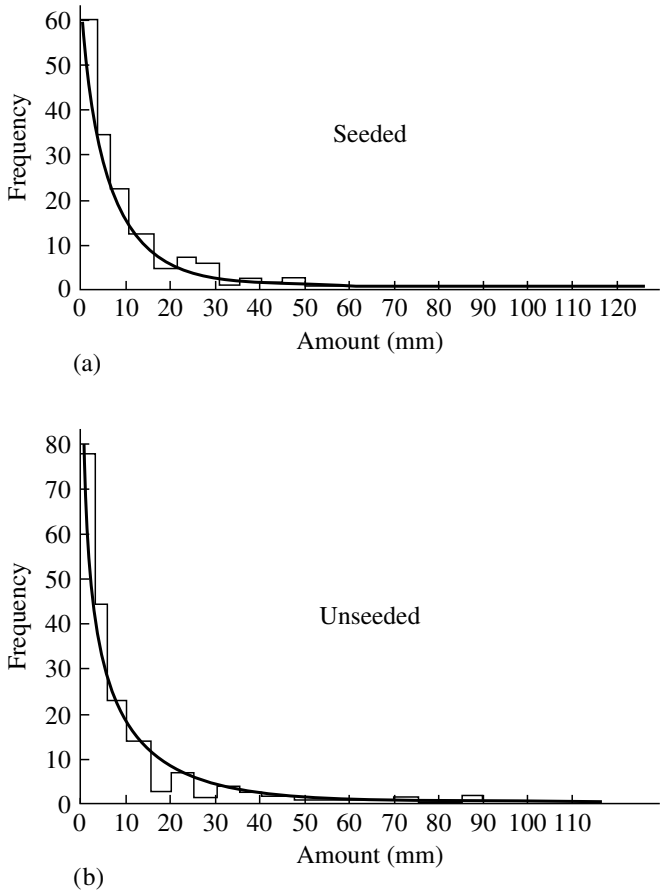

FIGURE **8.2** Fit of gamma densities to amounts of rainfall for (a) seeded and (b) unseeded storms.

to rainfall amounts from storms that were seeded and unseeded in an experiment to determine the effects, if any, of seeding. Differences in the distributions between the seeded and unseeded conditions should be reflected in differences in the parameters α and λ. ■

As these examples illustrate, there are a variety of reasons for fitting probability laws to data. A scientific theory may suggest the form of a probability distribution and the parameters of that distribution may be of direct interest to the scientific investigation; the examples of alpha particle emission and Example A are of this character. Example B is typical of situations in which a model is fit for essentially descriptive purposes as a method of data summary or compression. A probability model may play a role in a complex modeling situation; for example, utility companies interested in projecting patterns of consumer demand find it useful to model daily temperatures as random variables from a distribution of a particular form. This distribution may then be used in simulations of the effects of various pricing and generation schemes. In a similar way, hydrologists planning uses of water resources use stochastic models of rainfall in simulations.

We will take the following basic approach to the study of parameter estimation. The observed data will be regarded as realizations of random variables $X_1, X_2, \ldots, X_n$, whose joint distribution depends on an unknown parameter θ. Note that θ may be a vector, such as (α, λ) in Example B. Usually the X_i will be modeled as independent random variables all having the same distribution $f(x|\theta)$, in which case their joint distribution is $f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$. We will refer to such X_i as independent and identically distributed, or i.i.d. An estimate of θ will be a function of $X_1, X_2, \ldots, X_n$ and will hence be a random variable with a probability distribution called its **sampling distribution.** We will use approximations to the sampling distribution to assess the variability of the estimate, most frequently through its standard deviation, which is commonly called its **standard error.**

It is desirable to have general procedures for forming estimates so that each new problem does not have to be approached *ab initio*. We will develop two such procedures, the method of moments and the method of maximum likelihood, concentrating primarily on the latter, because it is the more generally useful.

The advanced theory of statistics is heavily concerned with "optimal estimation," and we will touch lightly on this topic. The essential idea is that given a choice of many different estimation procedures, we would like to use that estimate whose sampling distribution is most concentrated around the true parameter value.

Before going on to the method of moments, let us note that there are strong similarities of the subject matter of this and the previous chapter. In Chapter 7 we were concerned with estimating population parameters, such as the mean and total, and the process of random sampling created random variables whose probability distributions depended on those parameters. We were concerned with the sampling distributions of the estimates and with assessing variability via standard errors and confidence intervals. In this chapter we consider models in which the data are generated from a probability distribution. This distribution usually has a more hypothetical status than that of Chapter 7, where the distribution was induced by deliberate randomization. In this chapter we will also be concerned with sampling distributions and with assessing variability through standard errors and confidence intervals.

8.4 The Method of Moments

The kth moment of a probability law is defined as

$$\mu_k = E(X^k)$$

where X is a random variable following that probability law (of course, this is defined only if the expectation exists). If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from that distribution, the kth **sample moment** is defined as

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

We can view $\hat{\mu}_k$ as an estimate of μ_k. The method of moments estimates parameters by finding expressions for them in terms of the lowest possible order moments and then substituting sample moments into the expressions.

Suppose, for example, that we wish to estimate two parameters, θ_1 and θ_2. If θ_1 and θ_2 can be expressed in terms of the first two moments as

$$\theta_1 = f_1(\mu_1, \mu_2)$$
$$\theta_2 = f_2(\mu_1, \mu_2)$$

then the method of moments estimates are

$$\hat{\theta}_1 = f_1(\hat{\mu}_1, \hat{\mu}_2)$$
$$\hat{\theta}_2 = f_2(\hat{\mu}_1, \hat{\mu}_2)$$

The construction of a method of moments estimate involves three basic steps:

1. Calculate low order moments, finding expressions for the moments in terms of the parameters. Typically, the number of low order moments needed will be the same as the number of parameters.
2. Invert the expressions found in the preceding step, finding new expressions for the parameters in terms of the moments.
3. Insert the sample moments into the expressions obtained in the second step, thus obtaining estimates of the parameters in terms of the sample moments.

To illustrate this procedure, we consider some examples.

EXAMPLE **A** *Poisson Distribution*

The first moment for the Poisson distribution is the parameter $\lambda = E(X)$. The first sample moment is

$$\hat{\mu}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is, therefore, the method of moments estimate of λ: $\hat{\lambda} = \overline{X}$.

As a concrete example, let us consider a study done at the National Institute of Science and Technology (Steel et al. 1980). Asbestos fibers on filters were counted as part of a project to develop measurement standards for asbestos concentration. Asbestos dissolved in water was spread on a filter, and 3-mm diameter punches were taken from the filter and mounted on a transmission electron microscope. An operator counted the number of fibers in each of 23 grid squares, yielding the following counts:

31	29	19	18	31	28
34	27	34	30	16	18
26	27	27	18	24	22
28	24	21	17	24	

The Poisson distribution would be a plausible model for describing the variability from grid square to grid square in this situation and could be used to characterize the inherent variability in future measurements. The method of moments estimate of λ is simply the arithmetic mean of the counts listed above, these or $\hat{\lambda} = 24.9$.

If the experiment were to be repeated, the counts—and therefore the estimate—would not be exactly the same. It is thus natural to ask how stable this estimate is.

A standard statistical technique for addressing this question is to derive the sampling distribution of the estimate or an approximation to that distribution. The statistical model stipulates that the individual counts X_i are independent Poisson random variables with parameter λ_0. Letting $S = \sum X_i$, the parameter estimate $\hat{\lambda} = S/n$ is a random variable, the distribution of which is called its sampling distribution. Now from Example E in Section 4.5, the distribution of the sum of independent Poisson random variables is Poisson distributed, so the distribution of S is Poisson $(n\lambda_0)$. Thus the probability mass function of $\hat{\lambda}$ is

$$\begin{aligned} P(\hat{\lambda} = v) &= P(S = nv) \\ &= \frac{(n\lambda_0)^{nv} e^{-n\lambda_0}}{(nv)!} \end{aligned}$$

for v such that nv is a nonnegative integer.

Since S is Poisson, its mean and variance are both $n\lambda_0$, so

$$\begin{aligned} E(\hat{\lambda}) &= \frac{1}{n} E(S) = \lambda_0 \\ \text{Var}(\hat{\lambda}) &= \frac{1}{n^2} \text{Var}(S) = \frac{\lambda_0}{n} \end{aligned}$$

From Example A in Section 5.3, if $n\lambda_0$ is large, the distribution of S is approximately normal; hence, that of $\hat{\lambda}$ is approximately normal as well, with mean and variance given above. Because $E(\hat{\lambda}) = \lambda_0$, we say that the estimate is **unbiased:** the sampling distribution is centered at λ_0. The second equation shows that the sampling distribution becomes more concentrated about λ_0 as n increases. The standard deviation of this distribution is called the **standard error** of $\hat{\lambda}$ and is

$$\sigma_{\hat{\lambda}} = \sqrt{\frac{\lambda_0}{n}}$$

Of course, we can't know the sampling distribution or the standard error of $\hat{\lambda}$ because they depend on λ_0, which is unknown. However, we can derive an approximation by substituting $\hat{\lambda}$ and λ_0 and use it to assess the variability of our estimate. In particular, we can calculate the **estimated standard error** of $\hat{\lambda}$ as

$$s_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{n}}$$

For this example, we find

$$s_{\hat{\lambda}} = \sqrt{\frac{24.9}{23}} = 1.04$$

At the end of this section, we will present a justification for using $\hat{\lambda}$ in place of λ_0.

In summary, we have found that the sampling distribution of $\hat{\lambda}$ is approximately normal, centered at the true value λ_0 with standard deviation 1.04. This gives us a reasonable assessment of the variability of our parameter estimate. For example, because a normally distributed random variable is unlikely to be more than two standard deviations away from its mean, the error in our estimate of λ is unlikely to be more than 2.08. We thus have not only an estimate of λ_0, but also an understanding of the inherent variability of that estimate.

In Chapter 9, we will address the question of whether the Poisson distribution really fits these data. Clearly, we could calculate the average of any batch of numbers, whether or not they were well fit by the Poisson distribution. ■

EXAMPLE **B** *Normal Distribution*

The first and second moments for the normal distribution are

$$\mu_1 = E(X) = \mu$$
$$\mu_2 = E(X^2) = \mu^2 + \sigma^2$$

Therefore,

$$\mu = \mu_1$$
$$\sigma^2 = \mu_2 - \mu_1^2$$

The corresponding estimates of μ and σ^2 from the sample moments are

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

From Section 6.3, the sampling distribution of $\overline{X}$ is $N(\mu, \sigma^2/n)$ and $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$. Furthermore, $\overline{X}$ and $\hat{\sigma}^2$ are independently distributed. We will return to these sampling distributions later in the chapter. ■

EXAMPLE **C** *Gamma Distribution*

The first two moments of the gamma distribution are

$$\mu_1 = \frac{\alpha}{\lambda}$$
$$\mu_2 = \frac{\alpha(\alpha + 1)}{\lambda^2}$$

(see Example B in Section 4.5). To apply the method of moments, we must express α and λ in terms of μ_1 and μ_2. From the second equation,

$$\mu_2 = \mu_1^2 + \frac{\mu_1}{\lambda}$$

or

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$$

Also, from the equation for the first moment given here,

$$\alpha = \lambda\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$$

The method of moments estimates are, since $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$,

$$\hat{\lambda} = \frac{\overline{X}}{\hat{\sigma}^2}$$

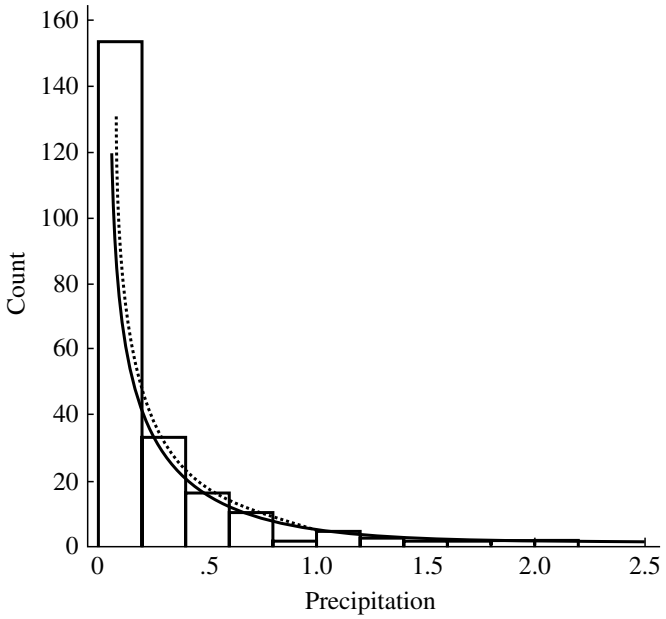

FIGURE **8.3** Gamma densities fit by the methods of moments and by the method of maximum likelihood to amounts of precipitation; the solid line shows the method of moments estimate and the dotted line the maximum likelihood estimate.

and

$$\hat{\alpha} = \frac{\overline{X}^2}{\hat{\sigma}^2}$$

As a concrete example, let us consider the fit of the amounts of precipitation during 227 storms in Illinois from 1960 to 1964 to a gamma distribution (Le Cam and Neyman 1967). The data, listed in Problem 42 at the end of Chapter 10, were gathered and analyzed in an attempt to characterize the natural variability in precipitation from storm to storm. A histogram shows that the distribution is quite skewed, so a gamma distribution is a natural candidate for a model. For these data, $\overline{X} = .224$ and $\hat{\sigma}^2 = .1338$, and therefore $\hat{\alpha} = .375$ and $\hat{\lambda} = 1.674$.

The histogram with the fitted density is shown in Figure 8.3. Note that, in order to make visual comparison easy, the density was normalized to have a total area equal to the total area under the histogram, which is the number of observations times the bin width of the histogram, or $227 \times .2 = 45.4$. Alternatively, the histogram could have been normalized to have a total area of 1. Qualitatively, the fit in Figure 8.3 looks reasonable; we will examine it in more detail in Example C in Section 9.9. ■

We now turn to a discussion of the sampling distributions of $\hat{\alpha}$ and $\hat{\lambda}$. In the previous two examples, we were able to use known theoretical results in deriving sampling distributions, but it appears that it would be difficult to derive the exact forms of the sampling distributions of $\hat{\lambda}$ and $\hat{\alpha}$, because they are each rather complicated functions of the sample values $X_1, X_2, \ldots, X_n$. However, the problem can be approached by simulation. Imagine for the moment that we knew the true values λ_0 and α_0. We could generate many, many samples of size $n = 227$ from the gamma distribution with

these parameter values, and from each of these samples we could calculate estimates of λ and α. A histogram of the values of the estimates of λ, for example, should then give us a good idea of the sampling distribution of $\hat{\lambda}$.

The only problem with this idea is that it requires knowing the true parameter values. (Notice that we faced a problem very much like this in Example A.) So we substitute our estimates of λ and α for the true values; that is we draw many, many samples of size $n = 227$ from a gamma distribution with parameters $\alpha = .375$ and $\lambda = 1.674$. The results of drawing 1000 such samples of size $n = 227$ are displayed in Figure 8.4. Figure 8.4(a) is a histogram of the 1000 estimates of α so obtained and Figure 8.4(b) shows the corresponding histogram for λ. These histograms indicate the variability that is inherent in estimating the parameters from a sample of this size. For example, we see that if the true value of α is .375, then it would not be very unusual for the estimate to be in error by .1 or more. Notice that the shapes of the histograms suggest that they might be approximated by normal densities.

The variability shown by the histograms can be summarized by calculating the standard deviations of the 1000 estimates, thus providing estimated standard errors of $\hat{\alpha}$ and $\hat{\lambda}$. To be precise, if the 1000 estimates of α are denoted by $\alpha_i^*, i = 1, 2, \ldots, 1000$, the standard error of $\hat{\alpha}$ is estimated as

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\alpha_i^* - \overline{\alpha})^2}$$

where $\overline{\alpha}$ is the mean of the 1000 values. The results of this calculation and the corresponding one for $\hat{\lambda}$ are $s_{\hat{\alpha}} = .06$ and $s_{\hat{\lambda}} = .34$. These standard errors are concise quantifications of the amount of variability of the estimates $\hat{\alpha} = .375$ and $\hat{\lambda} = 1.674$ displayed in Figure 8.4.

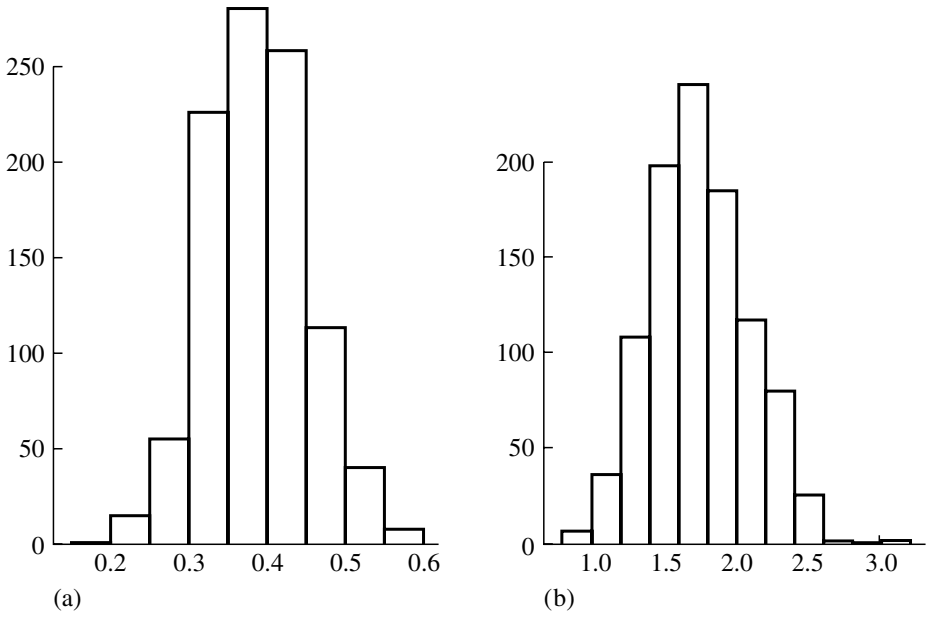

FIGURE **8.4** Histogram of 1000 simulated method of moment estimates of (a) α and (b) λ.

Our use of simulation (or Monte Carlo) here is an example of what in statistics is called the **bootstrap.** We will see more examples of this versatile method later.

EXAMPLE **D** *An Angular Distribution*

The angle θ at which electrons are emitted in muon decay has a distribution with the density

$$f(x|\alpha) = \frac{1+\alpha x}{2}, \qquad -1 \le x \le 1 \qquad \text{and} \qquad -1 \le \alpha \le 1$$

where $x = \cos\theta$. The parameter α is related to polarization. Physical considerations dictate that $|\alpha| \le \frac{1}{3}$, but we note that $f(x|\alpha)$ is a probability density for $|\alpha| \le 1$. The method of moments may be applied to estimate α from a sample of experimental measurements, $X_1, \ldots, X_n$. The mean of the density is

$$\mu = \int_{-1}^{1} x\frac{1+\alpha x}{2}\,dx = \frac{\alpha}{3}$$

Thus, the method of moments estimate of α is $\hat{\alpha} = 3\overline{X}$. Consideration of the sampling distribution of $\hat{\alpha}$ is left as an exercise (Problem 13). ■

Under reasonable conditions, method of moments estimates have the desirable property of consistency. An estimate, $\hat{\theta}$, is said to be a **consistent** estimate of a parameter, θ, if $\hat{\theta}$ approaches θ as the sample size approaches infinity. The following states this more precisely.

DEFINITION

Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges in probability to θ as n approaches infinity; that is, for any $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \qquad \text{as } n \to \infty$$ ■

The weak law of large numbers implies that the sample moments converge in probability to the population moments. If the functions relating the estimates to the sample moments are continuous, the estimates will converge to the parameters as the sample moments converge to the population moments.

The consistency of method of moments estimates can be used to provide a justification for a procedure that we used in estimating standard errors in the previous examples. We were interested in the variance (or its square root—the standard error) of a parameter estimate $\hat{\theta}$. Denoting the true parameter by θ_0, we had a relationship of the form

$$\sigma_{\hat{\theta}} = \frac{1}{\sqrt{n}}\sigma(\theta_0)$$

(In Example A, $\sigma_{\hat{\lambda}} = \sqrt{\lambda_0/n}$, so that $\sigma(\lambda) = \sqrt{\lambda}$.) We approximated this by the

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

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

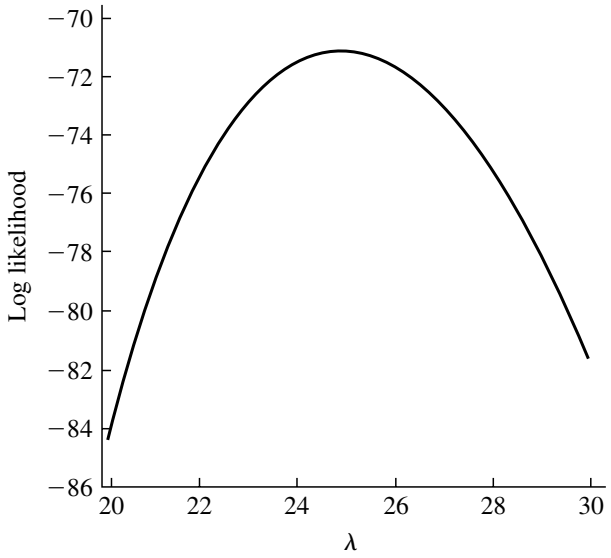

FIGURE **8.5** Plot of the log likelihood function of λ for asbestos data.

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, $X_i, \ldots, X_n$, 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

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

$$\frac{\partial l}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i$$

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

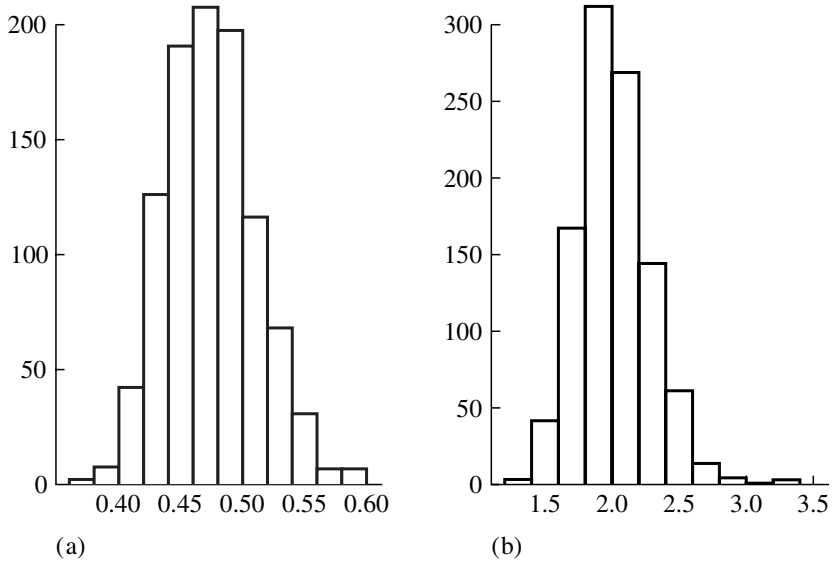

FIGURE **8.6** Histograms of 1000 simulated maximum likelihood estimates of (a) α and (b) λ.

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).)

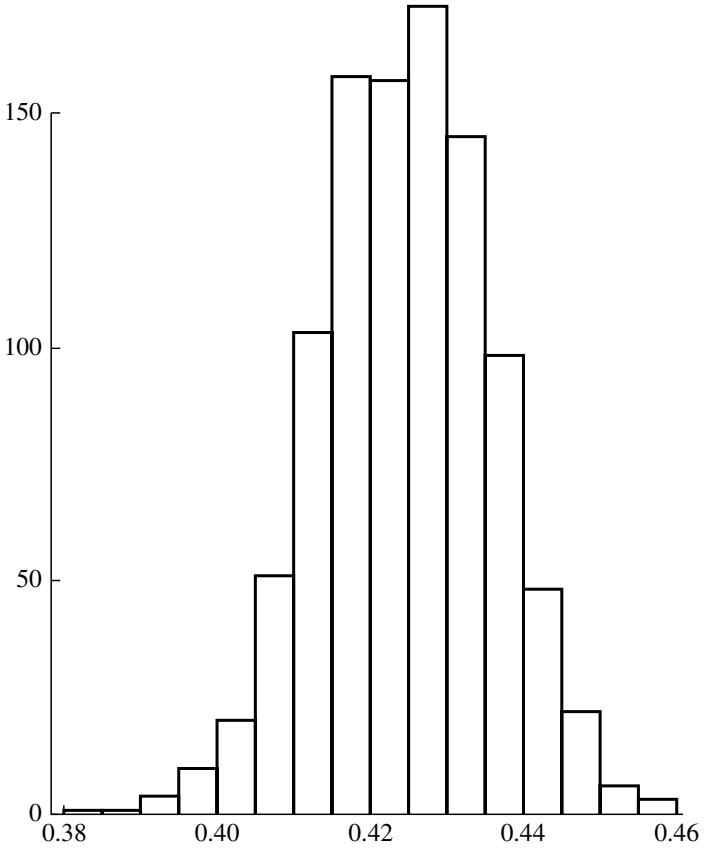

FIGURE **8.7** Histogram of 1000 simulated maximum likelihood estimates of θ described in Example A.

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

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

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:

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

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

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

A confidence interval for μ is based on the fact that

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t_{n-1}$$

where t_{n-1} denotes the t distribution with $n - 1$ degrees of freedom and

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

(see Section 6.3). Let $t_{n-1}(\alpha/2)$ denote that point beyond which the t distribution with $n - 1$ degrees of freedom has probability $\alpha/2$. Since the t distribution is symmetric about 0, the probability to the left of $-t_{n-1}(\alpha/2)$ is also $\alpha/2$. Then, by definition,

$$P\left(-t_{n-1}(\alpha/2) \le \frac{\sqrt{n}(\overline{X} - \mu)}{S} \le t_{n-1}(\alpha/2)\right) = 1 - \alpha$$

The inequality can be manipulated to yield

$$P\left(\overline{X} - \frac{S}{\sqrt{n}}t_{n-1}(\alpha/2) \le \mu \le \overline{X} + \frac{S}{\sqrt{n}}t_{n-1}(\alpha/2)\right) = 1 - \alpha$$

According to this equation, the probability that μ lies in the interval $\overline{X} \pm St_{n-1}(\alpha/2)/\sqrt{n}$ is $1 - \alpha$. Note that this interval is *random:* The center is at the random point $\overline{X}$ and the width is proportional to S, which is also random.

Now let us turn to a confidence interval for σ^2. From Section 6.3,

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

where χ^2_{n-1} denotes the chi-squared distribution with $n - 1$ degrees of freedom. Let $\chi^2_m(\alpha)$ denote the point beyond which the chi-square distribution with m degrees of freedom has probability α. It then follows by definition that

$$P\left(\chi^2_{n-1}(1 - \alpha/2) \le \frac{n\hat{\sigma}^2}{\sigma^2} \le \chi^2_{n-1}(\alpha/2)\right) = 1 - \alpha$$

Manipulation of the inequalities yields

$$P\left(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1 - \alpha/2)}\right) = 1 - \alpha$$

Therefore, a $100(1 - \alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)}, \quad \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1 - \alpha/2)}\right)$$

Note that this interval is not symmetric about $\hat{\sigma}^2$—it is not of the form $\hat{\sigma}^2 \pm c$, unlike the previous example.

A simulation illustrates these ideas: The following experiment was done on a computer 20 times. A random sample of size $n = 11$ from normal distribution with mean $\mu = 10$ and variance $\sigma^2 = 9$ was generated. From the sample, $\overline{X}$ and $\hat{\sigma}^2$ were

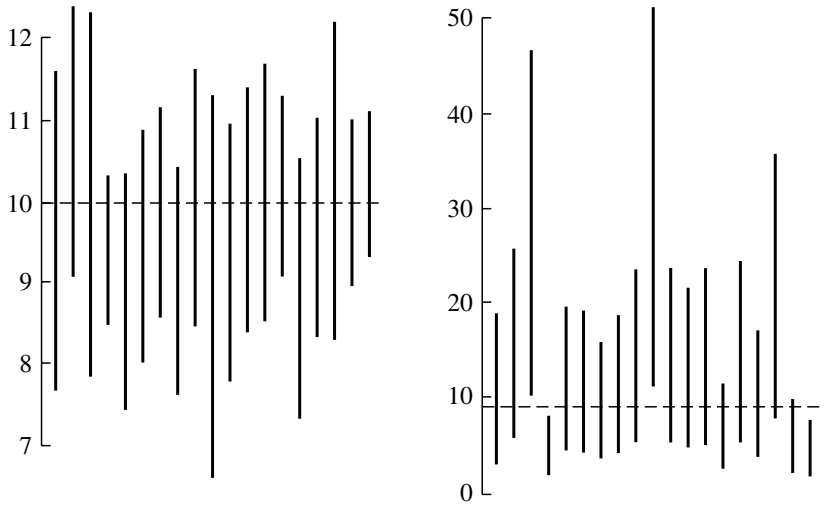

FIGURE **8.8** 20 confidence intervals for μ (left panel) and for σ^2 (right panel) as described in Example A. Horizontal lines indicate the true values.

calculated and 90% confidence intervals for μ and σ^2 were constructed, as described before. Thus at the end there were 20 intervals for μ and 20 intervals for σ^2. The 20 intervals for μ are shown as vertical lines in the left panel of Figure 8.8 and the 20 intervals for σ^2 are shown in the right panel. Horizontal lines are drawn at the true values $\mu = 10$ and $\sigma^2 = 9$. Since these are 90% confidence intervals, we expect the true parameter values to fall outside the intervals 10% of the time; thus on the average we would expect 2 of 20 intervals to fail to cover the true parameter value. From the figure, we see that all the intervals for μ actually cover μ, whereas four of the intervals of σ^2 failed to contain σ^2. ■

Exact methods such as that illustrated in the previous example are the exception rather than the rule in practice. To construct an exact interval requires detailed knowledge of the sampling distribution as well as some cleverness. A second method of constructing confidence intervals is based on the large sample theory of the previous section. According to the results of that section, the distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ is approximately the standard normal distribution. Since θ_0 is unknown, we will use $I(\hat{\theta})$ in place of $I(\theta_0)$; we have employed similar substitutions a number of times before—for example, in finding an approximate standard error in Example A of Section 8.4. It can be further argued that the distribution of $\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)$ is also approximately standard normal. Since the standard normal distribution is symmetric about 0,

$$P\left(-z(\alpha/2) \le \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \le z(\alpha/2)\right) \approx 1 - \alpha$$

Manipulation of the inequalities yields

$$\hat{\theta} \pm z(\alpha/2)\frac{1}{\sqrt{nI(\hat{\theta})}}$$

as an approximate $100(1-\alpha)\%$ confidence interval. We now illustrate this procedure with an example.

EXAMPLE **B** *Poisson Distribution*

The mle of λ from a sample of size n from a Poisson distribution is

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

Since the sum of independent Poisson random variables follows a Poisson distribution, the parameter of which is the sum of the parameters of the individual summands, $n\hat{\lambda} = \sum_{i=1}^{n} X_i$ follows a Poisson distribution with mean $n\lambda$. Also, the sampling distribution of $\hat{\lambda}$ is known, although it depends on the true value of λ, which is unknown. Exact confidence intervals for λ may be obtained by using this fact, and special tables are available (Pearson and Hartley 1966).

For large samples, confidence intervals may be derived as follows. First, we need to calculate $I(\lambda)$. Let $f(x|\lambda)$ denote the probability mass function of a Poisson random variable with parameter λ. There are two ways to do this. We may use the definition

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

We know that

$$\log f(x|\lambda) = x\log\lambda - \lambda - \log x!$$

and thus

$$I(\lambda) = E\left(\frac{X}{\lambda} - 1\right)^2$$

Rather than evaluate this quantity, we may use the alternative expression for $I(\lambda)$ given by Lemma A of Section 8.5.2:

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

Since

$$\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda) = -\frac{X}{\lambda^2}$$

$I(\lambda)$ is simply

$$\frac{E(X)}{\lambda^2} = \frac{1}{\lambda}$$

Thus, an approximate $100(1-\alpha)\%$ confidence interval for λ is

$$\overline{X} \pm z(\alpha/2)\sqrt{\frac{\overline{X}}{n}}$$

Note that the asymptotic variance is in fact the exact variance in this case. The confidence interval, however, is only approximate, since the sampling distribution of $\overline{X}$ is only approximately normal.

As a concrete example, let us return to the study that involved counting asbestos fibers on filters, discussed earlier. In Example A in Section 8.4, we found $\hat{\lambda} = 24.9$. The estimated standard error of $\hat{\lambda}$ is thus ($n = 23$)

$$s_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{n}} = 1.04$$

An approximate 90% confidence interval for λ is

$$\hat{\lambda} \pm 1.65 s_{\hat{\lambda}}$$

or (23.2, 26.6). This interval gives a good indication of the uncertainty inherent in the determination of the average asbestos level using the model that the counts in the grid squares are independent Poisson random variables. ■

In a similar way, approximate confidence intervals can be obtained for parameters estimated from random multinomial counts. The counts are not i.i.d., so the variance of the parameter estimate is not of the form $1/[nI(\theta)]$. However, it can be shown that

$$\text{Var}(\hat{\theta}) \approx \frac{1}{E[l'(\theta_0)^2]} = -\frac{1}{E[l''(\theta_0)]}$$

and the maximum likelihood estimate is approximately normally distributed. Example C illustrates this concept.

EXAMPLE **C** *Hardy-Weinberg Equilibrium*

Let us return to the example of Hardy-Weinberg equilibrium discussed in Example A in Section 8.5.1. There we found $\hat{\theta} = .4247$. Now,

$$l'(\theta) = -\frac{2X_1 + X_2}{1 - \theta} + \frac{2X_3 + X_2}{\theta}$$

In order to calculate $E[l'(\theta)^2]$, we would have to deal with the variances and covariances of the X_i. This does not look too inviting; it turns out to be easier to calculate $E[l''(\theta)]$.

$$l''(\theta) = -\frac{2X_1 + X_2}{(1 - \theta)^2} - \frac{2X_3 + X_2}{\theta^2}$$

Since the X_i are binomially distributed, we have

$$\begin{aligned} E(X_1) &= n(1 - \theta)^2 \\ E(X_2) &= 2n\theta(1 - \theta) \\ E(X_3) &= n\theta^2 \end{aligned}$$

We find, after some algebra, that

$$E[l''(\theta)] = -\frac{2n}{\theta(1 - \theta)}$$

Since θ is unknown, we substitute $\hat{\theta}$ in its place and obtain the estimated standard

error of $\hat{\theta}$:

$$s_{\hat{\theta}} = \frac{1}{\sqrt{-I''(\hat{\theta})}}$$
$$= \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{2n}} = .011$$

An approximate 95% confidence interval for θ is $\hat{\theta} \pm 1.96 s_{\hat{\theta}}$, or (.403, .447). (Note that this estimated standard error of $\hat{\theta}$ agrees with that obtained by the bootstrap in Example 8.5.1A.) ■

Finally, we describe the use of the bootstrap for finding approximate confidence intervals. Suppose that $\hat{\theta}$ is an estimate of a parameter θ—the true, unknown value of which is θ_0—and suppose for the moment that the distribution of $\Delta = \hat{\theta} - \theta_0$ is known. Denote the $\alpha/2$ and $1 - \alpha/2$ quantiles of this distribution by $\underline{\delta}$ and $\overline{\delta}$; i.e.,

$$P(\hat{\theta} - \theta_0 \le \underline{\delta}) = \frac{\alpha}{2}$$
$$P(\hat{\theta} - \theta_0 \le \overline{\delta}) = 1 - \frac{\alpha}{2}$$

Then

$$P(\underline{\delta} \le \hat{\theta} - \theta_0 \le \overline{\delta}) = 1 - \alpha$$

and from manipulation of the inequalities,

$$P(\hat{\theta} - \overline{\delta} \le \theta_0 \le \hat{\theta} - \underline{\delta}) = 1 - \alpha$$

The preceding assumed that the distribution of $\hat{\theta} - \theta_0$ was known, which is typically not the case. If θ_0 were known, this distribution could be approximated arbitrarily well by simulation: Many, many samples of observations could be randomly generated on a computer with the true value θ_0; for each sample, the difference $\hat{\theta} - \theta_0$ could be recorded; and the two quantiles $\underline{\delta}$ and $\overline{\delta}$ could, consequently, be determined as accurately as desired. Since θ_0 is not known, the bootstrap principle suggests using $\hat{\theta}$ in its place: Generate many, many samples (say, B in all) from a distribution with value $\hat{\theta}$; and for each sample construct an estimate of θ, say θ_j^*, $j = 1, 2, \ldots, B$. The distribution of $\hat{\theta} - \theta_0$ is then approximated by that of $\theta^* - \hat{\theta}$, the quantiles of which are used to form an approximate confidence interval. Examples may make this clearer.

EXAMPLE **D** We first apply this technique to the Hardy-Weinberg equilibrium problem; we will find an approximate 95% confidence interval based on the bootstrap and compare the result to the interval obtained in Example C, where large-sample theory for maximum likelihood estimates was used. The 1000 bootstrap estimates of θ of Example A of Section 8.5.1 provide an estimate of the distribution of θ^*; in particular the 25th largest is .403 and the 975th largest is .446, which are our estimates of the .025 and

.975 quantiles of the distribution. The distribution of $\theta^* - \hat{\theta}$ is approximated by subtracting $\hat{\theta} = .425$ from each θ_i^*, so the .025 and .975 quantiles of this distribution are estimated as

$$\underline{\delta} = .403 - .425 = -.022$$
$$\overline{\delta} = .446 - .425 = .021$$

Thus our approximate 95% confidence interval is

$$(\hat{\theta} - \overline{\delta}, \hat{\theta} - \underline{\delta}) = (.404, .447)$$

Since the uncertainty in $\hat{\theta}$ is in the second decimal place, this interval and that found in Example C are identical for all practical purposes. ■

EXAMPLE **E** Finally, we apply the bootstrap to find approximate confidence intervals for the parameters of the gamma distribution fit in Example C of Section 8.5. Recall that the estimates were $\hat{\alpha} = .471$ and $\hat{\lambda} = 1.97$. Of the 1000 bootstrap values of $\alpha^*, \alpha_1^*, \alpha_2^*, \ldots, \alpha_{1000}^*$, the 50th largest was .419 and the 950th largest was .538; the .05 and .95 quantiles of the distribution of $\alpha^* - \hat{\alpha}$ are approximated by subtracting $\hat{\alpha}$ from these values, giving

$$\underline{\delta} = .419 - .471 = -.052$$
$$\overline{\delta} = .538 - .471 = .067$$

Our approximate 90% confidence interval for α_0 is thus

$$(\hat{\alpha} - \overline{\delta}, \hat{\alpha} - \underline{\delta}) = (.404, .523)$$

The 50th and 950th largest values of λ^* were 1.619 and 2.478, and the corresponding approximate 90% confidence interval for λ_0 is (1.462, 2.321). ■

We caution the reader that there are a number of different methods of using the bootstrap to find approximate confidence intervals. We have chosen to present the preceding method largely because the reasoning leading to its development is fairly direct. Another popular method, the *bootstrap percentile method,* uses the quantiles of the bootstrap distribution of $\hat{\theta}$ directly. Using this method in the previous example, the confidence interval for α would be (.419, .538). Although this direct equation of quantiles of the bootstrap sampling distribution with confidence limits may seem initially appealing, its rationale is somewhat obscure. If the bootstrap distribution is symmetric, the two methods are equivalent (see Problem 38).

8.6 The Bayesian Approach to Parameter Estimation

A preview of the Bayesian approach was given in Example E of Section 3.5.2, which should be reviewed before continuing.