Moments

Let $X_1, X_2, \ldots, X_n$ be a random sample from the probability distribution $f(x)$ where $f(x)$ can be a discrete probability mass function or a continuous probability density function. The kth **population moment** (or **distribution moment**) is $E(X^k)$, $k = 1, 2, \ldots$. The corresponding kth **sample moment** is $(1/n)\sum_{i=1}^{n} X_i^k$, $k = 1, 2, \ldots$.

To illustrate, the first population moment is $E(X) = \mu$, and the first sample moment is $(1/n)\sum_{i=1}^{n} X_i = \bar{X}$. Thus, by equating the population and sample moments, we find that $\hat{\mu} = \bar{X}$. That is, the sample mean is the **moment estimator** of the population mean. In the general case, the population moments will be functions of the unknown parameters of the distribution, say, $\theta_1, \theta_2, \ldots, \theta_m$.

Moment Estimators

Let $X_1, X_2, \ldots, X_n$ be a random sample from either a probability mass function or a probability density function with m unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$. The **moment estimators** $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_m$ are found by equating the first m population moments to the first m sample moments and solving the resulting equations for the unknown parameters.

Example 7-7 **Exponential Distribution Moment Estimator** Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Now there is only one parameter to estimate, so we must equate $E(X)$ to $\bar{X}$. For the exponential, $E(X) = 1/\lambda$. Therefore, $E(X) = \bar{X}$ results in $1/\lambda = \bar{X}$, so $\lambda = 1/\bar{X}$, is the moment estimator of λ.

As an example, suppose that the time to failure of an electronic module used in an automobile engine controller is tested at an elevated temperature to accelerate the failure mechanism. The time to failure is exponentially distributed. Eight units are randomly selected and tested, resulting in the following failure time (in hours): $x_1 = 11.96$, $x_2 = 5.03$, $x_3 = 67.40$, $x_4 = 16.07$, $x_5 = 31.50$, $x_6 = 7.73$, $x_7 = 11.10$, and $x_8 = 22.38$. Because $\bar{x} = 21.65$, the moment estimate of λ is $\hat{\lambda} = 1/\bar{x} = 1/21.65 = 0.0462$.

Example 7-8 **Normal Distribution Moment Estimators** Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with parameters μ and σ^2. For the normal distribution, $E(X) = \mu$ and $E(X^2) = \mu^2 + \sigma^2$. Equating $E(X)$ to X and $E(X^2)$ to $\frac{1}{n}\sum_i^n X_i^2$ gives

$$\mu = \bar{X}, \qquad \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Solving these equations gives the moment estimators

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n}$$

Practical Conclusion: Notice that the moment estimator of σ^2 is not an unbiased estimator.

Example 7-9 **Gamma Distribution Moment Estimators** Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a gamma distribution with parameters r and λ. For the gamma distribution, $E(X) = r/\lambda$ and $E(X^2) = r(r+1)/\lambda^2$. The moment estimators are found by solving

$$r/\lambda = \bar{X}, \quad r(r+1)/\lambda^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

The resulting estimators are

$$\hat{r} = \frac{\bar{X}^2}{(1/n)\sum_{i=1}^{n} X_i^2 - \bar{X}^2} \qquad \hat{\lambda} = \frac{\bar{X}}{(1/n)\sum_{i=1}^{n} X_i^2 - \bar{X}^2}$$

To illustrate, consider the time to failure data introduced following Example 7-7. For these data, $\bar{x} = 21.65$ and $\sum_{i=1}^{8} x_i^2 = 6639.40$, so the moment estimates are

$$\hat{r} = \frac{(21.65)^2}{(1/8)6645.43 - (21.65)^2} = 1.29, \qquad \hat{\lambda} = \frac{21.65}{(1/8)6645.43 - (21.65)^2} = 0.0598$$

Interpretation: When $r = 1$, the gamma reduces to the exponential distribution. Because $\hat{r}$ slightly exceeds unity, it is quite possible that either the gamma or the exponential distribution would provide a reasonable model for the data.

7-4.2 Method of Maximum Likelihood

One of the best methods of obtaining a point estimator of a parameter is the method of maximum likelihood. This technique was developed in the 1920s by a famous British statistician, Sir R. A. Fisher. As the name implies, the estimator will be the value of the parameter that maximizes the **likelihood function**.

Maximum Likelihood Estimator

Suppose that X is a random variable with probability distribution $f(x;\theta)$ where θ is a single unknown parameter. Let $x_1, x_2, \ldots, x_n$ be the observed values in a random sample of size n. Then the **likelihood function** of the sample is

$$L(\theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdot \cdots \cdot f(x_n;\theta) \tag{7-10}$$

Note that the likelihood function is now a function of only the unknown parameter θ. The **maximum likelihood estimator** (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta)$.

In the case of a discrete random variable, the interpretation of the likelihood function is simple. The likelihood function of the sample $L(\theta)$ is just the probability

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

That is, $L(\theta)$ is just the probability of obtaining the sample values $x_1, x_2, \ldots, x_n$. Therefore, in the discrete case, the maximum likelihood estimator is an estimator that maximizes the probability of occurrence of the sample values. Maximum likelihood estimators are generally preferable to moment estimators because they possess good efficiency properties.

Example 7-10 **Bernoulli Distribution MLE** Let X be a Bernoulli random variable. The probability mass function is

$$f(x;p)=\begin{cases}p^x(1-p)^{1-x}, & x=0,1\\ 0, & \text{otherwise}\end{cases}$$

where p is the parameter to be estimated. The likelihood function of a random sample of size n is

$$L(p)=p^{x_1}(1-p)^{1-x_1}\,p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n}$$

$$=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}$$

We observe that if $\hat{p}$ maximizes $L(p)$, $\hat{p}$ also maximizes $\ln L(p)$. Therefore,

$$\ln L(p)=\left(\sum_{i=1}^{n}x_i\right)\ln p+\left(n-\sum_{i=1}^{n}x_i\right)\ln(1-p)$$

Now,

$$\frac{d\ln L(p)}{dp}=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{\left(n-\sum_{i=1}^{n}x_i\right)}{1-p}$$

Equating this to zero and solving for p yields $\hat{p}=(1/n)\sum_{i=1}^{n}x_i$. Therefore, the maximum likelihood estimator of p is

$$\hat{P}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Suppose that this estimator were applied to the following situation: n items are selected at random from a production line, and each item is judged as either defective (in which case we set $x_i=1$) or nondefective (in which case we set $x_i=0$). Then $\sum_{i=1}^{n}x_i$ is the number of defective units in the sample, and $\hat{p}$ is the sample proportion defective. The parameter p is the population proportion defective, and it seems intuitively quite reasonable to use $\hat{p}$ as an estimate of p.

Although the interpretation of the likelihood function just given is confined to the discrete random variable case, the method of maximum likelihood can easily be extended to a continuous distribution. We now give two examples of maximum likelihood estimation for continuous distributions.

Example 7-11 **Normal Distribution MLE** Let X be normally distributed with unknown μ and known variance σ^2. The likelihood function of a random sample of size n, say $X_1, X_2, \ldots, X_n$, is

$$L(\mu)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\,e^{-(x_i-\mu)^2/(2\sigma^2)}=\frac{1}{(2\pi\sigma^2)^{n/2}}\,e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Now

$$\ln L(\mu)=-(n/2)\ln(2\pi\sigma^2)-(2\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2$$

and

$$\frac{d \ln L(\mu)}{d\mu} = \left(\sigma^2\right)^{-1} \sum_{i=1}^{n} (x_i - \mu)$$

Equating this last result to zero and solving for μ yields

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

Conclusion: The sample mean is the maximum likelihood estimator of μ. Notice that this is identical to the moment estimator.

Example 7-12 **Exponential Distribution MLE** Let X be exponentially distributed with parameter λ. The likelihood function of a random sample of size n, say, $X_1, X_2, \ldots, X_n$, is

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

The log likelihood is

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

Now

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

and upon equating this last result to zero, we obtain

$$\hat{\lambda} = n / \sum_{i=1}^{n} X_i = 1 / \bar{X}$$

Conclusion: Thus, the maximum likelihood estimator of λ is the reciprocal of the sample mean. Notice that this is the same as the moment estimator.

It is easy to illustrate graphically just how the method of maximum likelihood works. Figure 7-9(a) plots the log of the likelihood function for the exponential parameter from Example 7-12, using the $n = 8$ observations on failure time given following Example 7-6. It is common for the log likelihood function to be negative. We found that the estimate of λ was $\hat{\lambda} = 0.0462$. From Example 7-12, we know that this is a maximum likelihood estimate. Figure 7-9(a) shows clearly that the log likelihood function is maximized at a value of λ that is approximately equal to 0.0462. Notice that the log likelihood function is relatively flat in the region of the maximum. This implies that the parameter is not estimated very precisely. If the parameter were estimated precisely, the log likelihood function would be very peaked at the maximum value. The sample size here is relatively small, and this has led to the imprecision in estimation. This is illustrated in Fig. 7-9(b) where we have plotted the difference in log likelihoods for the maximum value, assuming that the sample sizes were $n = 8$, 20, and 40 but that the sample average time to failure remained constant at

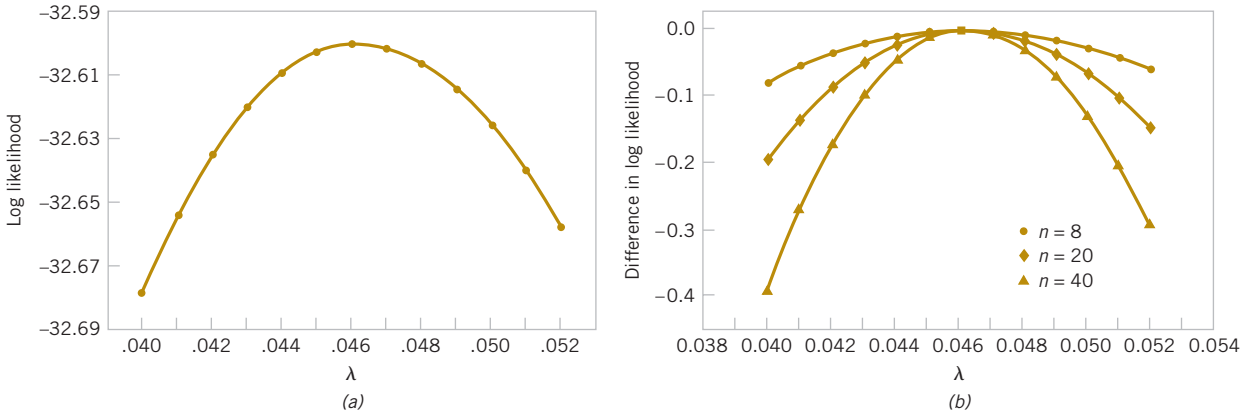

FIGURE 7-9 Log likelihood for the exponential distribution, using the failure time data. (a) Log likelihood with $n = 8$ (original data). (b) Log likelihood if $n = 8$, 20, and 40.

$\bar{x} = 21.65$. Notice how much steeper the log likelihood is for $n = 20$ in comparison to $n = 8$, and for $n = 40$ in comparison to both smaller sample sizes.

The method of maximum likelihood can be used in situations that have several unknown parameters, say, $\theta_1, \theta_2, \ldots, \theta_k$ to estimate. In such cases, the likelihood function is a function of the k unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$, and the maximum likelihood estimators $\{\hat{\Theta}_i\}$ would be found by equating the k partial derivatives $\partial L(\theta_1, \theta_2, \ldots, \theta_k)/\partial\theta_i$, $i = 1, 2, \ldots, k$ to zero and solving the resulting system of equations.

Example 7-13 **Normal Distribution MLEs For μ and σ^2** Let X be normally distributed with mean μ and variance σ^2 where both μ and σ^2 are unknown. The likelihood function for a random sample of size n is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

and

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Now

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

The solutions to these equations yield the maximum likelihood estimators

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

Conclusion: Once again, the maximum likelihood estimators are equal to the moment estimators.

Properties of the Maximum Likelihood Estimator

As noted previously, the method of maximum likelihood is often the estimation method that we prefer because it produces estimators with good statistical properties. We summarize these properties as follows.

Properties of a Maximum Likelihood Estimator

Under very general and not restrictive conditions when the sample size n is large and if $\hat{\Theta}$ is the maximum likelihood estimator of the parameter θ,

(1) $\hat{\Theta}$ is an approximately unbiased estimator for $\theta[E(\hat{\Theta})] \simeq \theta$,

(2) The variance of $\hat{\Theta}$ is nearly as small as the variance that could be obtained with any other estimator.

(3) $\hat{\Theta}$ has an approximate normal distribution.

Properties 1 and 2 essentially state that the maximum likelihood estimator is approximately an MVUE. This is a very desirable result and, coupled with the facts that it is fairly easy to obtain in many situations and has an asymptotic normal distribution (the "asymptotic" means "when n is large"), explains why the maximum likelihood estimation technique is widely used. To use maximum likelihood estimation, remember that the distribution of the population must be either known or assumed.

To illustrate the "large-sample" or asymptotic nature of these properties, consider the maximum likelihood estimator for σ^2, the variance of the normal distribution, in Example 7-13. It is easy to show that

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\,\sigma^2$$

The bias is

$$E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\,\sigma^2 - \sigma^2 = \frac{-\sigma^2}{n}$$

Because the bias is negative, $\hat{\sigma}^2$ tends to underestimate the true variance σ^2. Note that the bias approaches zero as n increases. Therefore, $\hat{\sigma}^2$ is an asymptotically unbiased estimator for σ^2.

We now give another very important and useful property of maximum likelihood estimators.

Invariance Property

Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$ be the maximum likelihood estimators of the parameters $\theta_1, \theta_2, \ldots, \theta_k$. Then the maximum likelihood estimator of any function $h(\theta_1, \theta_2, \ldots, \theta_k)$ of these parameters is the same function $h(\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k)$ of the estimators $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$.

Example 7-14

In the normal distribution case, the maximum likelihood estimators of μ and σ^2 were $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \Sigma_{i=1}^{n}(X_i - \bar{X})^2/n$, respectively. To obtain the maximum likelihood estimator of the function $h(\mu, \sigma^2) = \sqrt{\sigma^2} = \sigma$, substitute the estimators $\hat{\mu}$ and $\hat{\sigma}^2$ into the function h, which yields

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]^{1/2}$$

Conclusion: The maximum likelihood estimator of the standard deviation σ is *not* the sample standard deviation S.

Complications in Using Maximum Likelihood Estimation

Although the method of maximum likelihood is an excellent technique, sometimes complications arise in its use. For example, it is not always easy to maximize the likelihood function because the equation(s) obtained from $dL(\theta)/d\theta = 0$ may be difficult to solve. Furthermore, it may not always be possible to use calculus methods directly to determine the maximum of $L(\theta)$. These points are illustrated in the following two examples.

Example 7-15 **Uniform Distribution MLE** Let X be uniformly distributed on the interval 0 to a. Because the density function is $f(x) = 1/a$ for $0 \le x \le a$ and zero otherwise, the likelihood function of a random sample of size n is

$$L(a) = \prod_{i=1}^{n} \frac{1}{a} = \frac{1}{a^n}$$

for

$$0 \le x_1 \le a, 0 \le x_2 \le a, \ldots, 0 \le x_n \le a$$

Note that the slope of this function is not zero anywhere. That is, as long as $\max(x_i) \le a$, the likelihood is $1/a^n$, which is positive, but when $a < \max(x_i)$, the likelihood goes to zero as illustrated in Fig. 7-10. Therefore, calculus methods cannot be used directly because the maximum value of the likelihood function occurs at a point of discontinuity. However, because $d/da(a^{-n}) = -n/a^{n+1}$ is less than zero for all values of $a > 0$, a^{-n} is a decreasing function of a. This implies that the maximum of the likelihood function $L(a)$ occurs at the lower boundary point. The figure clearly shows that we could maximize $L(a)$ by setting $\hat{a}$ equal to the smallest value that it could logically take on, which is $\max(x_i)$. Clearly, a cannot be smaller than the largest sample observation, so setting $\hat{a}$ equal to the largest sample value is reasonable.

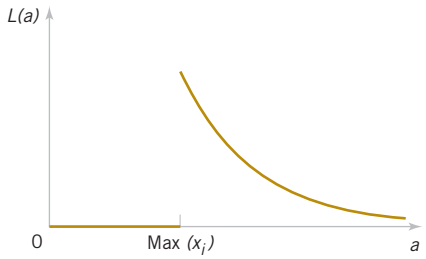

FIGURE 7-10 The likelihood function for the uniform distribution in Example 7-15.

Example 7-16 **Gamma Distribution MLE** Let $X_1, X_2, \ldots, X_n$ be a random sample from the gamma distribution. The log of the likelihood function is

$$\ln L(r, \lambda) = \ln \left(\prod_{i=1}^{n} \frac{\lambda^r x_i^{r-1} e^{-\lambda x_i}}{\Gamma(r)} \right)$$

$$= nr \ln(\lambda) + (r-1) \sum_{i=1}^{n} \ln(x_i) - n \ln[\Gamma(r)] - \lambda \sum_{i=1}^{n} x_i$$

The derivatives of the log likelihood are

$$\frac{\partial \ln L(r, \lambda)}{\partial r} = n \ln(\lambda) + \sum_{i=1}^{n} \ln(x_i) - n \frac{\Gamma'(r)}{\Gamma(r)}$$

$$\frac{\partial \ln L(r, \lambda)}{\partial \lambda} = \frac{nr}{\lambda} - \sum_{i=1}^{n} x_i$$

When the derivatives are equated to zero, we obtain the equations that must be solved to find the maximum likelihood estimators of r and λ:

$$\hat{\lambda} = \frac{\hat{r}}{\bar{x}}$$

$$n \ln(\hat{\lambda}) + \sum_{i=1}^{n} \ln(x_i) = n \frac{\Gamma'(\hat{r})}{\Gamma(\hat{r})}$$

There is no closed form solution to these equations.

Figure 7-11 is a graph of the log likelihood for the gamma distribution using the $n = 8$ observations on failure time introduced previously. Figure 7-11a is the **log likelihood surface** as a function of r and λ, and Figure 7-11b is a **contour plot**. These plots reveal that the log likelihood is maximized at approximately $\hat{r} = 1.75$ and $\hat{\lambda} = 0.08$. Many statistics computer programs use numerical techniques to solve for the maximum likelihood estimates when no simple solution exists.

7-4.3 Bayesian Estimation of Parameters

This book uses methods of statistical inference based on the information in the sample data. In effect, these methods interpret probabilities as relative frequencies. Sometimes we call probabilities that are interpreted in this manner **objective probabilities**. Another approach to statistical inference, called the **Bayesian** approach, combines sample information with other information that may be available prior to collecting the sample. In this section, we briefly illustrate how this approach may be used in parameter estimation.

Suppose that the random variable X has a probability distribution that is a function of one parameter θ. We will write this probability distribution as $f(x|\theta)$. This notation implies that the exact form of the distribution of X is conditional on the value assigned to θ. The classical approach to estimation would consist of taking a random sample of size n from this distribution and then substituting the sample values x_i into the estimator for θ. This estimator could have been developed using the maximum likelihood approach, for example.

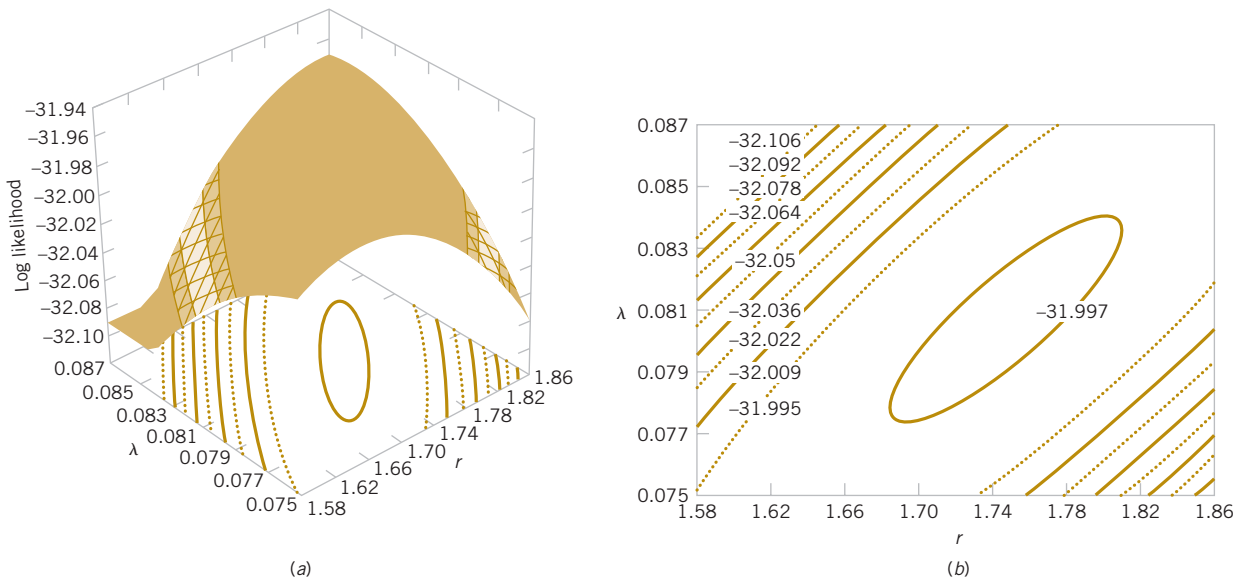

FIGURE 7-11 Log likelihood for the gamma distribution using the failure time data. (*a*) Log likelihood surface. (*b*) Contour plot.