Example 2.8 Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from a population with probability density function

$$f(x) = \frac{1}{\theta} e^{-x/\theta} \qquad x > 0,$$

where θ is a positive unknown parameter. Find the maximum likelihood estimator of θ.

Once again, the population distribution is recognized as an exponential distribution with population mean θ. As before, we assume that a visual inspection of the histogram and/or theoretical considerations have revealed that the exponential distribution is an appropriate probability model for the data set, so we proceed with parameter estimation. The data values are denoted by $x_1, x_2, \ldots, x_n$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} x_i/\theta}.$$

The log likelihood function is

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

The score is

$$\frac{\partial \ln L(\theta)}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i.$$

When the score is equated to zero,

$$-\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i = 0,$$

and this equation is solved for θ, the maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

For data drawn from an exponential population, the point estimate via maximum likelihood estimation for the population mean is the sample mean. To see that the maximum likelihood estimator maximizes (rather than minimizes) the log likelihood function, a second derivative is taken:

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^{n} x_i.$$

When θ is replaced with the maximum likelihood estimator $\hat{\theta}$, this expression becomes

$$\left.\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = \frac{n}{\hat{\theta}^2} - \frac{2}{\hat{\theta}^3} \sum_{i=1}^{n} x_i = \frac{n^3}{\left(\sum_{i=1}^{n} x_i\right)^2} - \frac{2n^3}{\left(\sum_{i=1}^{n} x_i\right)^2} = -\frac{n^3}{\left(\sum_{i=1}^{n} x_i\right)^2}.$$

For data values drawn from a distribution with positive support (such as the exponential population in this example), this expression is always negative. This implies that $\hat{\theta}$ maximizes the log likelihood function, and therefore, $\hat{\theta}$ also maximizes the likelihood

function. For this particular population, the maximum likelihood estimate happens to be identical to the method of moments estimate. This is not universally true, however.

As mentioned earlier, parameter estimators are random variables that have probability density functions. Switching the notation from the data values $x_1, x_2, \ldots, x_n$ to the associated random variables $X_1, X_2, \ldots, X_n$, the expected value of the maximum likelihood estimator is

$$E\left[\hat{\theta}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \theta = \frac{1}{n}(n\theta) = \theta.$$

So the maximum likelihood estimator $\hat{\theta}$ is an unbiased estimator of θ. Having the expected value of the parameter estimate equal to the parameter itself is a highly desirable property for the estimate, because the estimator is "on target" or "unbiased" on average.

In order to appreciate the geometry associated with maximum likelihood estimation, consider the tiny data set with just $n = 4$ observations:

$$x_1 = 1.3, \qquad x_2 = 0.5, \qquad x_3 = 0.3, \qquad x_4 = 1.9.$$

Now consider all of the possible probability density functions for exponential populations, that is, all of the probability density functions of the form

$$f(x) = \frac{1}{\theta}e^{-x/\theta} \qquad\qquad x > 0$$

for some θ in the parameter space $\Omega = \{\theta \,|\, \theta > 0\}$. The θ value corresponding to the maximum likelihood estimate, which is

$$\hat{\theta} = \bar{x} = \frac{1.3 + 0.5 + 0.3 + 1.9}{4} = 1,$$

has the largest (maximum) product of the lengths of the vertical lines shown in Figure 2.6. Any other choice of θ would give a lower value for this product, which is also the value of the likelihood function $L(\theta)$. The data values are plotted as $\times$s on the horizontal axis in Figure 2.6, and the particular probability density function plotted is

$$f(x) = e^{-x} \qquad\qquad x > 0,$$

which is the probability density function associated with the maximum likelihood estimate $\hat{\theta} = 1$. The vertical lines connecting the data values to the probability density function have lengths $f(x_1)$, $f(x_2)$, $f(x_3)$, and $f(x_4)$. The product of these lengths is the value of the likelihood function for $\hat{\theta}$, which is

$$L\left(\hat{\theta}\right) = f(x_1) \cdot f(x_2) \cdot f(x_3) \cdot f(x_4) = e^{-x_1}e^{-x_2}e^{-x_3}e^{-x_4} = e^{-1.3}e^{-0.5}e^{-0.3}e^{-1.9} = e^{-4}.$$

In this sense, this particular choice of θ gives the exponential distribution that is most likely to have resulted in the observed data set. Hence, the name *maximum likelihood* is used to describe this parameter estimator.

The next example shows that the procedure for finding maximum likelihood estimates is essentially the same for a discrete population as it is for a continuous population.

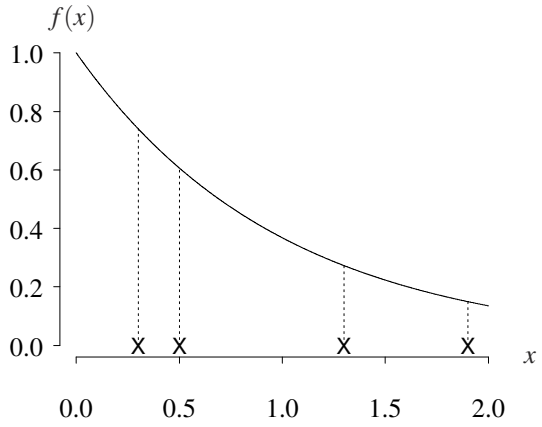

Figure 2.6: Geometry associated with the maximum likelihood estimator.

Example 2.9 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson(λ) population, where λ is a positive unknown parameter. Find the maximum likelihood estimator $\hat{\lambda}$.

To begin, a visual inspection of a histogram associated with a data set of values from a discrete population suggests that the Poisson distribution is an appropriate probability model for the data set in order to proceed with fitting via maximum likelihood. The probability mass function for the Poisson distribution is

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots,$$

where λ is an unknown parameter in the parameter space $\Omega = \{\lambda \,|\, \lambda > 0\}$. The likelihood function is

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

The log likelihood function is

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

and the score is

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

When the score is equated to zero,

$$-n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0,$$

and when this equation is solved for λ, the maximum likelihood estimate of λ is

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

As in the previous example, the sample mean is the point estimate of the population mean. The second partial derivative of the log likelihood function is

$$\frac{\partial^2 \ln L(\lambda)}{\partial \lambda^2} = -\frac{1}{\lambda^2} \sum_{i=1}^{n} x_i.$$

When λ is replaced by the maximum likelihood estimate, this expression becomes

$$\left.\frac{\partial^2 \ln L(\lambda)}{\partial \lambda^2}\right|_{\lambda=\hat{\lambda}} = -\frac{1}{\hat{\lambda}^2} \sum_{i=1}^{n} x_i = -\frac{n^2}{\sum_{i=1}^{n} x_i}.$$

Since this expression is always negative for data drawn from a Poisson population (except in the rare case in which all of the data values are zeros), the maximum likelihood estimate is associated with a local maximum of the log likelihood function.

Again switching the data values $x_1, x_2, \ldots, x_n$ to their associated random variables $X_1, X_2, \ldots, X_n$, the expected value of the maximum likelihood estimator is

$$E\left[\hat{\lambda}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \lambda = \frac{1}{n}\,(n\lambda) = \lambda.$$

As was the case with the maximum likelihood estimator of the population mean for a random sample from an exponential population, the maximum likelihood estimator $\hat{\lambda}$ is an unbiased estimator of the unknown parameter λ.

Now consider fitting a Poisson distribution to a data set. The famous horse kick data set consists of annual deaths of Prussian cavalry soldiers due to horse kicks. There are $n = 200$ corps-years of data given in Table 2.1. This table shows that there are $n = 200$ data values, and they consist of 109 zeros, 65 ones, 22 twos, 3 threes, and 1 four. Since the maximum likelihood estimate of the Poisson distribution is the sample mean,

$$\hat{\lambda} = \frac{(109)(0)+(65)(1)+(22)(2)+(3)(3)+(1)(4)}{200} = \frac{122}{200} = 0.61$$

fatalities per corps-year. The R statements

```
x = 0:10
200 * dpois(x, 0.61)
```

can be used to add a third row to the table containing the data. Table 2.2 includes the fitted Poisson probabilities, rounded to the nearest tenth, that are predicted by the Poisson model. The spectacular agreement between the data and the fitted Poisson probability model allows us to conclude that the Poisson distribution is an appropriate model—horse kick deaths were likely to have been a Poisson process over time. The deaths appear to be occurring randomly over time, consistent with the assumptions

number of deaths per corps per year	0	1	2	3	4
number of observed values	109	65	22	3	1

Table 2.1: Observed horse kick deaths.

number of deaths per corps per year	0	1	2	3	4 or more
number of observed values	109	65	22	3	1
number of predicted values	108.7	66.3	20.2	4.1	0.7

Table 2.2: Observed and predicted horse kick deaths.

associated with a Poisson process concerning random events. This is sensible because the horses would be unlikely to have conspired against the soldiers in a systematic fashion.

Another geometric aspect of the maximum likelihood estimation technique can be seen by plotting the log likelihood function in a vicinity of the maximum likelihood estimate. Figure 2.7 shows the log likelihood function

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

plotted in the vicinity of $\hat{\lambda} = 0.61$. Such a plot is generated by the R code given below.

```
x     = c(rep(0, 109), rep(1, 65), rep(2, 22), rep(3, 3), 4)
n     = length(x)
lam   = mean(x)
xlam  = seq(lam - 0.02, lam + 0.02, length = 200)
logl  = -n * xlam + sum(x) * log(xlam) - sum(log(factorial(x)))
plot(xlam, logl, type = "l")
```

Figure 2.7 shows that the log likelihood function achieves a local maximum at the point $\left(\hat{\lambda}, \ln L(\hat{\lambda})\right) = (0.61, -206.1067)$. The second coordinate is negative because it is the natural logarithm of a joint probability mass function, which typically assumes values between 0 and 1.

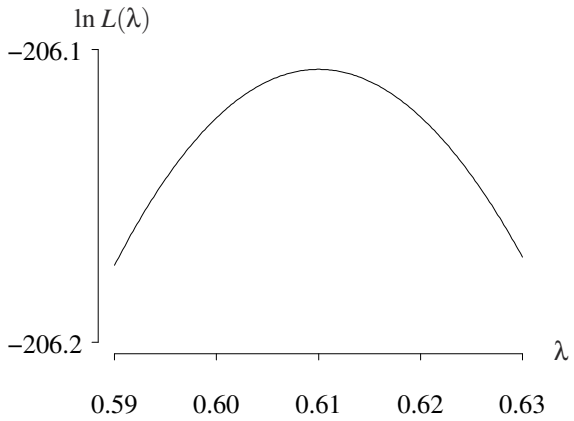

Figure 2.7: The log likelihood function near $\hat{\lambda} = 0.61$ for the horse kick data.