- The abbreviation MLE is often used for the maximum likelihood estimator.
- In practice it is often easier to maximize the *log likelihood function*, $\ln L(\theta)$, than the likelihood function $L(\theta)$. The maximum likelihood estimators will be the same for maximizing either function because a function and its increasing monotonic transformation are maximized at the same value.
- The derivative of the log likelihood function $\frac{\partial \ln L(\theta)}{\partial \theta}$ is often called the *score*, which generalizes to the *score vector* when there are several unknown parameters.
- When the second partial derivative of the likelihood function evaluated at the maximum likelihood estimator $\frac{\partial^2 L(\hat{\theta})}{\partial \theta^2} < 0$ or when the second partial derivative of the log likelihood function evaluated at the maximum likelihood estimator $\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta^2} < 0$, the maximum likelihood estimate $\hat{\theta}$ maximizes $L(\theta)$ or $\ln L(\theta)$ by the second derivative test from calculus.
- As will be seen in a subsequent example, maximum likelihood estimators are not necessarily unique.
- Maximum likelihood estimates have the *invariance property*, which states that if $\hat{\theta}$ is the maximum likelihood estimator for θ that exists uniquely and h is a one-to-one function, then $h(\hat{\theta})$ is the maximum likelihood estimator for $h(\theta)$.
- Although the mathematics required to find the maximum likelihood estimator for a parameter θ is often a straightforward calculus problem, the derivation is mathematically intractable for some distributions and numerical methods must be used to calculate $\hat{\theta}$.

There is an important property of point estimators that will be discussed in the examples that follow. When the expected value of a point estimate $\hat{\theta}$ is θ, that is when $E\left[\hat{\theta}\right] = \theta$, the point estimator is known as an *unbiased estimator*. This is a good property for a point estimator to possess because it means that the estimator is, in some sense, aiming at the right target. A formal definition and more details associated with unbiased estimators are given in the next section. The unbiased property can be applied to any point estimator, including a method of moments estimator.

Example 2.7 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 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

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

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 for 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 estimator happens to be identical to the method of moments estimator. This is not universally true.

As mentioned earlier, parameter estimates 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 estimate of θ. Having the expected value of the parameter estimate equal to the parameter itself is a highly desirable property for the estimate, because it is "on target" on average.

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

$$1.3,\ 0.5,\ 0.3,\ 1.9.$$

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 estimator, 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 2.5. 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 2.5, and the particular probability density function plotted is

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

which is the probability density function associated with the maximum likelihood estimator $\hat{\theta} = 1$. 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."

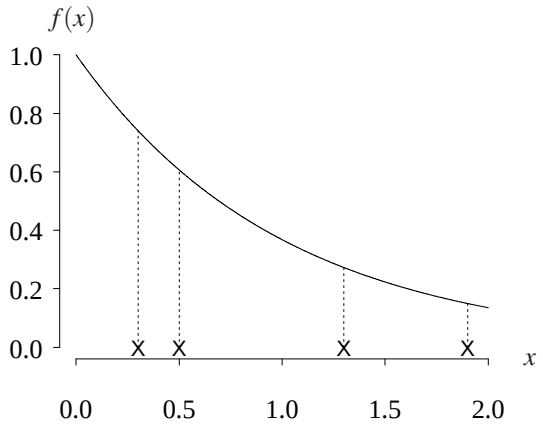

Figure 2.5: Geometry associated with the maximum likelihood estimator.

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.

Example 2.8 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson(λ) population, where λ is an unknown positive parameter. Find the maximum likelihood estimator $\hat{\lambda}$.

As before, a visual inspection of a histogram reveals that the Poisson distribution is an appropriate probability model for the data set. The probability mass function for the Poisson distribution is

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