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 estimates of θ. 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):

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

$$= \log n! - \sum_{i=1}^{3} \log X_i! + (2X_1 + X_2)\log(1-\theta)$$

$$+ (2X_3 + X_2)\log\theta + X_2 \log 2$$

[We have not explicitly incorporated 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:

$$\hat{\theta} = \frac{2X_3 + X_2}{2X_1 + 2X_2 + 2X_3}$$

$$= \frac{2X_3 + X_2}{2n}$$

$$= \frac{2 \times 187 + 500}{2 \times 1029} = .4247$$

How precise is this estimate? Do we have faith in the accuracy of the first, second, third, or fourth decimal place? Do the data in the table actually fit the Hardy–Weinberg Law? We will address these questions in later sections of this chapter and in chapter 9. □

8.5.2 Large Sample Theory for Maximum Likelihood Estimates

Maximum likelihood estimates are functions of the sample values and are therefore random variables. Although we would like to know the sampling distribution in order to ascertain the precision of the estimate and to form confidence intervals for the population parameters, we cannot usually obtain the sampling distribution of the maximum likelihood estimates explicitly. We have already encountered a similar problem in chapter 7, where we wished to know the sampling distribution of the sample mean. In that situation, we used the central limit theorem to approximate the sampling distribution of $\bar{X}$; in this section, we develop some large sample theory to approximate the sampling distributions of maximum likelihood estimates by normal distributions.

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 below may be skipped without loss of continuity. Rigorous proofs may be found in Cramer (1946).]

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 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 maximize $l(\theta)$ should be close to the θ that maximize $E \log f(X|\theta)$. (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 expressions just above, 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,

$$0 = l'(\hat\theta) \approx l'(\theta_0) + (\hat\theta - \theta_0) l''(\theta_0)$$

$$(\hat\theta - \theta_0) \approx \frac{-l'(\theta_0)}{l''(\theta_0)}$$

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

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

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

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:

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

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

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

which may be interpreted as an average radius of curvature of $l(\theta)$ at θ_0. When this radius of curvature is small, the estimate is relatively well resolved and the asymptotic variance is small.

A corresponding result can be proved for 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, and the elements of the vector of estimates have variances and covariances given by the corresponding elements of the matrix $(1/n)I^{-1}(\theta_0)$, where $I(\theta)$ is a matrix with the ij component

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

The following sections will apply these results in several examples.

8.5.3 Confidence Intervals for 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 the 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 $\hat{\theta}$.

In some cases, the exact sampling distribution of a maximum likelihood estimate can be obtained, allowing the derivation of exact confidence intervals; however, this is typically not possible. For moderate to large sample sizes, confidence intervals based on the normal approximation to the sampling distribution developed in Section 8.5.2 may be used. In that section, it was shown that the asymptotic variance of a maximum likelihood estimate depends on $I(\theta_0)$. As it was given there, this result is difficult to use, since θ_0 is not known. The obvious thing to try is to substitute $\hat{\theta}$ for θ_0 and use $I(\hat{\theta})$. In fact, it can be shown that the limiting distribution of $\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)$ is also the standard normal distribution, so this procedure can be justified.

From essentially the same argument that was used to form approximate confidence intervals for the population mean (see Section 7.3.3), it follows that an approximate $100(1 - \alpha)\%$ confidence interval for θ_0 is

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

Denoting the estimated standard deviation, or standard error, of $\hat{\theta}$ by $s_{\hat{\theta}}$, we can write this confidence interval as

$$\hat{\theta} \pm z(\alpha/2)s_{\hat{\theta}}$$

This form parallels that of the $100(1-\alpha)\%$ confidence interval for μ developed in chapter 7, which was $\bar{X} \pm z(\alpha/2)s_{\bar{X}}$. In both cases, the confidence interval is the parameter estimate plus or minus a multiple of its standard error. Many, but not all, confidence intervals are of this form.

Let us consider some examples of exact and approximate confidence intervals.

EXAMPLE A. (Poisson Distribution) The mle of λ from a sample of size n from a Poisson distribution is

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

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 $I(\lambda)$. 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

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

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 $\bar{X}$ 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 95% confidence interval for λ is

$$\hat{\lambda} \pm 1.96 s_{\hat{\lambda}}$$

or (22.9, 26.9). 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. □

EXAMPLE B. (Normal Distribution) The mle's of μ and σ^2 are

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

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

Exact theory may be used; from Section 6.3,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{s} \sim t_{n-1}$$

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

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

Since the t distribution is symmetric, an exact confidence interval for μ is

α	*Efficiency*
0.0	1.0
.1	.997
.2	.989
.3	.975
.4	.953
.5	.931
.6	.878
.7	.817
.8	.727
.9	.582
.95	.464

As α tends to 1, the efficiency tends to 0. Thus, the mle is not much better than the method of moments estimate for α close to 0 but does increasingly better as α tends to 1.

It must be kept in mind that we have used the asymptotic variance of the mle, so we have actually calculated an asymptotic relative efficiency. To gain more precise information for a given sample size, a simulation of the sampling distribution of the mle could be conducted. This might be especially interesting for $\alpha = 1$, a case for which the formula for the asymptotic variance given above does not appear to make much sense. With a simulation study, the behavior of the bias as n and α vary could be analyzed (we showed that the mle is asymptotically unbiased, but there may be bias for a finite sample size), and the actual distribution could be compared to the approximating normal. □

In searching for an optimal estimate, we might ask whether there is a lower bound for the MSE of *any* estimate. If such a lower bound existed, it would function as a benchmark against which estimates could be compared. If an estimate achieved this lower bound, we would know that it could not be improved upon. In the case in which the estimate is unbiased, the Cramer–Rao Inequality provides such a lower bound. We now state and prove the Cramer–Rao Inequality.

THEOREM A. (Cramer–Rao Inequality) Let $X_1, \ldots, X_n$ be i.i.d. with density function $f(x|\theta)$. Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimate of θ. Then, under smoothness assumptions on $f(x|\theta)$,

$$\operatorname{Var}(T) \geq \frac{1}{nI(\theta)}$$

Proof. Let

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

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

In Section 8.5.2, we showed that $E(Z) = 0$. Since the correlation coefficient of Z and T is less than or equal to 1 in absolute value

$$\operatorname{Cov}^2(Z, T) \leq \operatorname{Var}(Z) \operatorname{Var}(T)$$

It was also shown in Section 8.5.2 that

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

Therefore,

$$\operatorname{Var}(Z) = nI(\theta)$$

The proof will be completed by showing that $\operatorname{Cov}(Z, T) = 1$. Since Z has mean 0,

$$\operatorname{Cov}(Z, T) = E(ZT)$$

$$= \int \cdots \int t(x_1, \ldots, x_n) \left[\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i|\theta)}{f(x_i|\theta)}\right] \prod_{j=1}^{n} f(x_j|\theta)\, dx_j$$

Noting that

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i|\theta)}{f(x_i|\theta)} \prod_{j=1}^{n} f(x_j|\theta) = \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i|\theta)$$

we rewrite the expression for the covariance of Z and T as

$$\operatorname{Cov}(Z, T) = \int \cdots \int t(x_1, \ldots, x_n) \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i|\theta)\, dx_i$$

$$= \frac{\partial}{\partial\theta} \int \cdots \int t(x_1, \ldots, x_n) \prod_{i=1}^{n} f(x_i|\theta)\, dx_i$$

$$= \frac{\partial}{\partial\theta} E(T) = \frac{\partial}{\partial\theta}(\theta) = 1$$

which proves the inequality. [Note the interchange of differentiation and integration that must be justified by the smoothness assumptions on $f(x|\theta)$.] □

Theorem A gives a lower bound on the variance of *any* unbiased estimate. An unbiased estimate whose variance achieves this lower bound is said to be **efficient**. Since the asymptotic variance of a maximum likelihood estimate is equal to the