Serfling, R. J.
(1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Shorack, G. R.
(1972). "Functions of order statistics." *Ann. Math. Statist.* **43**, 412–427.

Siddiqui, M. M. and Raghunandanan, K.
(1967). "Asymptotically robust estimators of location." *J. Amer. Statist. Assoc.* **62**, 950–953.

Spjøtvoll, E.
(1968). "A note on robust estimation in analysis of variance." *Ann. Math. Statist.* **39**, 1486–1492.

Spjøtvoll, E. and Aastreit, A. H.
(1980). "Comparison of robust estimators on data from field experiments." *Scand. J. Statist.* **7**, 1–13.

Staudte, R. G., Jr.
(1980). "Robust estimation." *Queen's Papers in Pure and Applied Mathematics*. No. 53. Queen's University, Kingston, Ontario.

Stigler, S. M.
(1973a). "Laplace, Fisher, and the discovery of the concept of sufficiency." *Biometrika* **60**, 439–445.

Stigler, S. M.
(1973b). "Simon Newcomb, Percy Daniell, and the history of robust estimation." *J. Amer. Statist. Assoc.* **68**, 872–879.

Stigler, S. M.
(1973c). "The asymptotic distribution of the trimmed mean." *Ann. Statist.* **1**, 472–477.

Stigler, S. M.
(1974). "Linear functions of order statistics with smooth weight functions." *Ann. Statist.* **2**, 676–693. (Correction, *Ann. Statist.* **7**, 466).

Stigler, S. M.
(1977). "Do robust estimators work with real data?" *Ann. Statist.* **5**, 1055–1077.

Stigler, S. M.
(1980). "An Edgeworth curiosum." *Ann. Statist.* **8**, 931–934.

Stone, C. J.
(1974). "Asymptotic properties of estimators of a location parameter." *Ann. Statist.* **2**, 1127–1137.

Stone, C. J.
(1975). "Adaptive maximum likelihood estimators of a location parameter." *Ann. Statist.* **3**, 267–284.

Theil, H.
(1950). "A rank-invariant method of linear and polynomial regression analysis." *Koninkl. Ned. Weteusch. Proc.* **53**, 386–392, 521–525, 1397–1412.

Tukey, J. W.
(1960). "A survey of sampling from contaminated distributions." In *Contributions to Probability and Statistics*. Olkin, Ed. Stanford University Press, Stanford, Calif.

CHAPTER 6

Asymptotic Optimality

1. ASYMPTOTIC EFFICIENCY

The large sample approximations of the preceding chapter not only provide a convenient method for assessing the performance of an estimator and for comparing different estimators, they also permit a new approach to optimality that is much less restrictive than the theories of unbiased and equivariant estimation developed in Chapters 2 and 3.

It was seen in Sections 5.1 and 5.2 that estimators* of interest typically are consistent as the sample sizes tend to infinity and, suitably normalized, are asymptotically normally distributed about the estimand with a variance $v(\theta)$ (the asymptotic variance), which provides a reasonable measure of the accuracy of the estimator sequence. (In this connection, see Problem 1.1.) Within this class of consistent asymptotically normal estimators, it turns out that under mild additional restrictions there exist estimators that uniformly minimize $v(\theta)$. The present chapter is mainly concerned with the development of fairly explicit methods of obtaining such *asymptotically efficient* estimators.

Before embarking on this program, it may be helpful to note an important difference between the present large-sample approach and the small-sample results of Chapters 2 through 4. Both UMVU and MRE estimators tend to be unique (Theorem 1.6.5) and so are at least some of the minimax estimators derived in Chapter 4. On the other hand, it is in the nature of asymptotically optimal solutions not to be unique since asymptotic results refer to the limiting behavior of sequences, and the same limit is shared by many different sequences. More specifically, if

$$\sqrt{n}\,[\delta_n - g(\theta)] \xrightarrow{\mathcal{L}} N(0, v)$$

*As in the preceding chapter, we shall frequently use *estimator* instead of the more accurate but cumbersome term *estimator sequence*.

and $\{\delta_n\}$ is asymptotically optimal in the sense of minimizing v, then $\delta_n + R_n$ is also optimal provided

$$\sqrt{n}\, R_n \to 0 \qquad \text{in probability.}$$

As we shall see later, asymptotically equivalent optimal estimators can be obtained from quite different starting points.

A central role in the theory of asymptotic efficiency is played by an analog of the information inequality (2.6.28). If $X_1, \ldots, X_n$ are iid according to a density $f_\theta(x)$ (with respect to μ) satisfying suitable regularity conditions, this inequality states that the variance of any unbiased estimator δ of $g(\theta)$ satisfies

$$\text{(1)} \qquad \operatorname{var}_\theta(\delta) \geqslant \frac{[g'(\theta)]^2}{nI(\theta)},$$

where $I(\theta)$ is the amount of information in a single observation defined by (2.6.10).

Suppose now that $\delta_n = \delta_n(X_1, \ldots, X_n)$ is asymptotically normal, say that

$$\text{(2)} \qquad \sqrt{n}\,[\delta_n - g(\theta)] \xrightarrow{\mathcal{L}} N[0, v(\theta)], \quad v(\theta) > 0.$$

Then it turns out that under some additional restrictions one also has

$$\text{(3)} \qquad v(\theta) \geqslant \frac{[g'(\theta)]^2}{I(\theta)}.$$

However, although the lower bound (1) is attained only in exceptional circumstances (Sect. 2.6), there exist sequences $\{\delta_n\}$ that satisfy (2) with $v(\theta)$ equal to the lower bound (3) subject only to quite general regularity conditions. A sequence $\{\delta_n\}$ satisfying (2) with

$$\text{(4)} \qquad v(\theta) = \frac{[g'(\theta)]^2}{I(\theta)}$$

is said to be *asymptotically efficient*.

At first glance, (3) might be thought to be a consequence of (1). Two differences between the inequalities (1) and (3) should be noted, however. (i) The estimator δ in (1) is assumed to be unbiased, while (2) implies consistency of $\{\delta_n\}$ but not that δ_n is unbiased or even that its bias tends to zero (Problem 1.5). (ii) The quantity $v(\theta)$ in (3) is an asymptotic variance and (1) refers to the actual variance of δ. It follows from Lemma 5.1.2 that

$$\text{(5)} \qquad v(\theta) \leqslant \liminf[n \operatorname{var}_\theta \delta_n]$$

but equality need not hold. Thus, (3) is a consequence of (1) provided

$$\text{(6)} \qquad \operatorname{var}\{\sqrt{n}\,[\delta_n - g(\theta)]\} \to v(\theta)$$

and if δ_n is unbiased, but not necessarily if these requirements do not hold.

For a long time, (3) was nevertheless believed to be valid subject only to regularity conditions on the densities f_θ. This belief was exploded by the example (due to Hodges; see Le Cam, 1953) given below. Before stating the example, note that in discussing the inequality (3) under assumption (2), if θ is real-valued and $g(\theta)$ is differentiable, it is enough to consider the case $g(\theta) = \theta$, for which (3) reduces to

$$\text{(7)} \qquad v(\theta) \geqq \frac{1}{I(\theta)}.$$

For if

$$\text{(8)} \qquad \sqrt{n}\,(\delta_n - \theta) \xrightarrow{\mathcal{L}} N[0, v(\theta)]$$

and if g has derivative g', it was seen in Theorem 5.1.5 that

$$\text{(9)} \qquad \sqrt{n}\,[g(\delta_n) - g(\theta)] \xrightarrow{\mathcal{L}} N[0, v(\theta)g'^2(\theta)].$$

After the obvious change of notation, this implies (3).

Example 1.1. Let $X_1, \ldots, X_n$ be iid according to the normal distribution $N(\theta, 1)$ and let the estimand be θ. It was seen in Table 2.6.1 that in this case $I(\theta) = 1$ so that (7) reduces to $v(\theta) \geqq 1$. On the other hand, consider the sequence of estimators,

$$\delta_n = \begin{cases} \bar{X} & \text{if } |\bar{X}| \geqq 1/n^{1/4} \\ a\bar{X} & \text{if } |\bar{X}| < 1/n^{1/4}. \end{cases}$$

Then (Problem 1.2)

$$\sqrt{n}\,(\delta_n - \theta) \xrightarrow{\mathcal{L}} N[0, v(\theta)]$$

There with $v(\theta) = 1$ when $\theta \neq 0$ and $v(\theta) = a^2$ when $\theta = 0$. If $a < 1$, inequality (3) is therefore violated at $\theta = 0$.

This phenomenon is quite general (Problems 1.3–1.4). There will typically exist estimators satisfying (8) but with $v(\theta)$ violating (7) for at least some values of θ, called points of *superefficiency*. However, (7) is almost true, for it was shown by Le Cam (1953) that for any sequence δ_n satisfying (8) the set S of points of superefficiency has Lebesgue measure zero. The following version of this result which we shall not prove, is due to Bahadur (1964). The assumptions are somewhat stronger but similar to those of Theorem 2.6.4.

Remark on notation. In the present chapter we shall use X_i, X, and x_i, x for real-valued random variables and the values they take on, and $\underline{X}$, $\underline{x}$ for the vectors $(X_1, \ldots, X_n)$, $(x_1, \ldots, x_n)$. This differs from the notation in earlier chapters, for example, from that in Section 2.6.

Theorem 1.1. *Let* $\mathrm{X}_1, \ldots, \mathrm{X}_n$ *be iid, each with density* $\mathrm{f}(\mathrm{x}, \theta)$ *with respect to a* σ*-finite measure* μ*, where* θ *is real-valued, and suppose the following regularity conditions hold.*

(i) *The parameter space* Ω *is an open interval (not necessarily finite).*

(ii) *The distributions* P_θ *of the* X_i *have common support, so that the set* $\mathrm{A} = \{\mathrm{x}: \mathrm{f}(\mathrm{x}, \theta) > 0\}$ *is independent of* θ.

(iii) *For every* $\mathrm{x} \in \mathrm{A}$, *the density* $\mathrm{f}(\mathrm{x}, \theta)$ *is twice differentiable with respect to* θ, *and the second derivative is continuous in* θ.

(iv) *The integral* $\int \mathrm{f}(\mathrm{x}, \theta)\, \mathrm{d}\mu(\mathrm{x})$ *can be twice differentiated under the integral sign.*

(v) *The Fisher information* $\mathrm{I}(\theta)$ *defined by* (2.6.10) *satisfies* $0 < \mathrm{I}(\theta) < \infty$.

(vi) *For any given* $\theta_0 \in \Omega$ *there exists a positive number* c *and a function* $\mathrm{M}(\mathrm{x})$ *(both of which may depend on* θ_0*) such that*

$$|\partial^2 \log \mathrm{f}(\mathrm{x}, \theta)/\partial\theta^2| \leqslant \mathrm{M}(\mathrm{x})$$

$$\text{for all} \quad \mathrm{x} \in \mathrm{A}, \quad \theta_0 - \mathrm{c} < \theta < \theta_0 + \mathrm{c}$$

and

$$\mathrm{E}_{\theta_0}[\mathrm{M}(\mathrm{X})] < \infty.$$

Under these assumptions, if $\delta_\mathrm{n} = \delta_\mathrm{n}(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n})$ *is any estimator satisfying* (8), *then* $\mathrm{v}(\theta)$ *satisfies* (7) *except on a set of Lebesgue measure zero.*

Note that by Lemma 2.6.1, condition (iv) ensures that for all $\theta \in \Omega$

(vii) $$E\left[\partial \log f(X, \theta)/\partial\theta\right] = 0$$

and

(viii) $$E\left[-\partial^2 \log f(X, \theta)/\partial\theta^2\right] = E\left[\partial \log f(X, \theta)/\partial\theta\right]^2 = I(\theta).$$

Condition (iv) can be replaced by conditions (vii) and (viii) in the statement of the theorem.

The example makes it clear that no regularity conditions on the densities $f(x, \theta)$ can prevent estimators from violating (7). This possibility can be avoided only by placing restrictions also on the sequence of estimators. In view of the information inequality (2.6.28), an obvious sufficient condition is (6) [with $g(\theta) = \theta$] together with

(10) $$b_n'(\theta) \to 0$$

where $b_n(\theta) = E_\theta(\delta_n) - \theta$ is the bias of δ_n.

If $I(\theta)$ is continuous, as will typically be the case, a more appealing assumption is perhaps that $v(\theta)$ also be continuous. Then (7) clearly cannot be violated at any point since otherwise it would be violated in an interval around this point in contradiction to Theorem 1.1. As an alternative, which under mild assumptions on f implies continuity of $v(\theta)$, Rao (1963) and Wolfowitz (1965) require the convergence in (2) to be uniform in θ. By working with coverage probabilities rather than asymptotic variance, the latter author also removes the unpleasant assumption that the limit distribution in (2) must be normal. An analogous result is proved by Pfanzagl (1970), who requires the estimators to be asymptotically median unbiased.

The search for restrictions on the sequence $\{\delta_n\}$, which would ensure (7) for all values of θ, is motivated in part by the hope of the existence, within the restricted class, of uniformly best estimators for which $v(\theta)$ attains the lower bound. It is further justified by the fact, brought out by Le Cam (1953), Huber (1966), and Hájek (1972), that violation of (7) at a point θ_0 entails certain unpleasant properties of the risk of the estimator in the neighborhood of θ_0.

This behavior can be illustrated on the Hodges example.

Example 1.1. *(Continued).* The normalized risk function

(11) $$R_n(\theta) = nE(\delta_n - \theta)^2$$

of the Hodges estimator δ_n can be written as

$$R_n(\theta) = 1 - (1 - a^2)\int_{I_n}^{\bar{I}_n} (x + \sqrt{n}\theta)^2 \phi(x)\, dx$$

$$+ 2\theta\sqrt{n}\,(1 - a)\int_{I_n}^{\bar{I}_n} (x + \sqrt{n}\theta)\phi(x)\, dx$$

where $\bar{l}_n = \sqrt[4]{n} - \sqrt{n}\theta$, and $\underline{l}_n = -\sqrt[4]{n} - \sqrt{n}\theta$. When the integrals are broken up into their three and two terms, respectively, and the relations

$$\Phi'(x) = \phi(x) \quad \text{and} \quad \phi'(x) = -x\phi(x)$$

are used, $R_n(\theta)$ reduces to

$$R_n(\theta) = \left[1 - (1-a^2)\int_{\underline{l}_n}^{\bar{l}_n} x^2\phi(x)\,dx\right]$$

$$+ n\theta^2(1-a)^2\left[\Phi(\bar{l}_n) - \Phi(\underline{l}_n)\right] + 2\sqrt{n}\theta a(1-a)\left[\phi(\bar{l}_n) - \phi(\underline{l}_n)\right].$$

Consider now the sequence of parameter values $\theta_n = 1/\sqrt[4]{n}$, so that

$$\sqrt{n}\theta_n = n^{1/4}, \quad \underline{l}_n = -2n^{1/4}, \quad \bar{l}_n = 0.$$

Then

$$\sqrt{n}\theta_n\phi(\underline{l}_n) \to 0$$

so that the third term tends to infinity as $n \to \infty$. Since the second term is positive and the first term is bounded, it follows that

$$R_n(\theta_n) \to \infty \qquad \text{for} \quad \theta_n = 1/\sqrt[4]{n},$$

and hence a fortiori that

$$\sup_\theta R_n(\theta) \to \infty.$$

Let us now compare this result with the fact that (Problem 1.6) for any fixed θ

$$R_n(\theta) \to 1 \qquad \text{for} \quad \theta \neq 0, \qquad R_n(0) \to a^2.$$

(This shows that in the present case the limiting risk is equal to the asymptotic variance found in Example 1.1.) The functions $R_n(\theta)$ are continuous functions of θ with discontinuous limit function

$$L(\theta) = 1 \qquad \text{for} \quad \theta \neq 0, \qquad L(0) = a^2.$$

However, each of the functions with large values of n rises to a high above the limit value 1, at values of θ tending to the origin with n, and with the value of the peak tending to infinity. The improvement (over $\bar{X}$) from 1 to a^2 in the limiting risk at the origin and hence for large finite n also near the origin, therefore, leads to an enormous increase in risk at points slightly further away but nevertheless close to the origin. (In this connection, see Problem 1.8.)

2. EFFICIENT LIKELIHOOD ESTIMATION

Under smoothness assumptions similar to those of Theorem 1.1, we shall in the present section prove the existence of asymptotically efficient estimators and provide a method for determining such estimators which in many cases leads to an explicit solution.

We begin with the following assumptions:

(A0) The distributions P_θ of the observations are distinct (otherwise, θ cannot be estimated consistently; but see Redner (1981) for a different point of view).

(A1) The distributions P_θ have common support.

(A2) The observations are $\underline{X} = (X_1, \ldots, X_n)$, where the X_i are iid with probability density $f(x_i, \theta)$ with respect to μ.

(A3) The parameter space Ω contains an open interval ω of which the true parameter value θ_0 is an interior point.

Note: The true value of θ will be denoted by θ_0. The density $f(x_1, \theta) \cdots f(x_n, \theta)$, considered for fixed $\underline{x}$ as a function of θ, is called the *likelihood function*.

Theorem 2.1. *Under assumptions* $(A0)$–$(A2)$,

$$(1) \quad P_{\theta_0}\{f(X_1, \theta_0) \cdots f(X_n, \theta_0) > f(X_1, \theta) \cdots f(X_n, \theta)\} \to 1 \quad \text{as } n \to \infty$$

for any fixed $\theta \neq \theta_0$.

Proof. The inequality is equivalent to

$$\frac{1}{n}\Sigma \log[f(X_i, \theta)/f(X_i, \theta_0)] < 0.$$

By the law of large numbers, the left side tends in probability toward

$$E_{\theta_0}\log[f(X, \theta)/f(X, \theta_0)].$$

Since $-\log$ is strictly convex, Jensen's inequality shows that

$$E_{\theta_0}\log[f(X, \theta)/f(X, \theta_0)] < \log E_{\theta_0}[f(X, \theta)/f(X, \theta_0)] = 0,$$

and the results follows.

By (1), the density of $\underline{X}$ at the true θ_0 exceeds that at any other fixed θ with high probability when n is large. We do not know θ_0 but we can determine the value $\hat{\theta}$ of θ which maximizes the density of $\underline{X}$, that is, which

maximizes the likelihood function at the observed $\underline{X}$. If this value exists and is unique, it is the *maximum likelihood estimator* (MLE) of θ.* The MLE of $g(\theta)$ is defined to be $g(\hat{\theta})$. If g is 1 : 1 and $\xi = g(\theta)$, this agrees with the definition of $\hat{\xi}$ as the value of ξ that maximizes the likelihood, and the definition is consistent also in the case that g is not 1 : 1. (In this connection, see Berk, 1967, and Zehna, 1966.)

Theorem 2.1 suggests that if the density of $\underline{X}$ varies smoothly with θ, the MLE of θ typically should be close to the true value of θ, and hence be a reasonable estimator.

Example 2.1. Let X have the binomial distribution $b(p, n)$. Then the MLE of p is obtained by maximizing $\binom{n}{x}p^x q^{n-x}$ and hence is $\hat{p} = X/n$ (Problem 2.1).

Example 2.2. If $X_1, \ldots, X_n$ are iid as $N(\xi, \sigma^2)$, it is convenient to obtain the MLE by maximizing the logarithm of the density, $-n \log \sigma - 1/2\sigma^2 \Sigma(x_i - \xi)^2 - c$. When (ξ, σ) are both unknown, the maximizing values are $\hat{\xi} = \bar{X}$, $\hat{\sigma}^2 = \Sigma(X_i - \bar{X})^2/n$ (Problem 2.3).

As a first question regarding the MLE for iid variables, let us ask whether it is consistent. We begin with the case in which Ω is finite, so that θ can take on only a finite number of values. In this case, a sequence δ_n is consistent if and only if

$$P_\theta(\delta_n = \theta) \to 1 \qquad \text{for all} \quad \theta \in \Omega \tag{2}$$

(Problem 2.6).

Corollary 2.1. *Under assumptions (A0)–(A2) if Ω is finite, the MLE $\hat{\theta}_n$ exists, it is unique with probability tending to* 1, *and it is consistent.*

Proof. The result is an immediate consequence of Theorem 2.1 and the fact that if $P(A_{in}) \to 1$ for $i = 1, \ldots, k$, then also $P[A_{1n} \cap \cdots \cap A_{kn}] \to 1$ as $n \to \infty$.

The proof of Corollary 2.1 breaks down when Ω is not restricted to be finite. That the consistency conclusion itself can break down even if Ω is only countably infinite is shown by the following example due to Bahadur (1958) and Le Cam (1979b).

Example 2.3. Let h be a continuous function defined on $(0, 1]$, which is strictly decreasing, with $h(x) \geqslant 1$ for all $0 < x \leqslant 1$ and satisfying

$$\int_0^1 h(x)\, dx = \infty \tag{3}$$

*For a more general definition, see Scholz (1980). A discussion of the MLE as a summarizer of the data rather than an estimator is given by Efron (1982).

Given a constant $0 < c < 1$, let a_k, $k = 0, 1, \ldots$ be a sequence of constants defined inductively as follows: $a_0 = 1$; given $a_0, \ldots, a_{k-1}$, the constant a_k is defined by

$$\int_{a_k}^{a_{k-1}} [h(x) - c]\, dx = 1 - c. \tag{4}$$

It is easy to see that there exists a unique value $0 < a_k < a_{k-1}$ satisfying (4) (Problem 2.7). Since the sequence $\{a_k\}$ is decreasing, it tends to a limit $a \geqslant 0$. If a were > 0, the left side of (4) would tend to zero which is impossible. Thus $a_k \to 0$ as $k \to \infty$.

Consider now the sequence of densities

$$f_k(x) = \begin{cases} c & \text{if} \quad x \leqslant a_k \quad \text{or} \quad x > a_{k-1} \\ h(x) & \text{if} \quad a_k < x \leqslant a_{k-1} \end{cases},$$

and the problem of estimating the parameter k on the basis of independent observations $X_1, \ldots, X_n$ from f_k. We shall show that the MLE exists and that it tends to infinity in probability regardless of the true value k_0 of k and is therefore not consistent, provided $h(x) \to \infty$ sufficiently fast as $x \to 0$.

Let us denote the joint density of the X's by

$$p_k(\underline{x}) = f_k(x_1) \cdots f_k(x_n).$$

That the MLE exists follows from the fact that $p_k(\underline{x}) = c^n < 1$ for any value of k for which the interval $I_k = (a_k, a_{k-1}]$ contains none of the observations, so that the maximizing value of k must be one of the $\leqslant n$ values for which I_k contains at least one of the x's.

For $n = 1$, the MLE is the value of k for which $X_1 \in I_k$, and for $n = 2$ the MLE is the value of k for which $X_{(1)} \in I_k$. For $n = 3$, it may happen that one observation lies in I_k and two in I_l $(k < l)$, and whether the MLE is k or l then depends on whether $c \cdot h(x_{(1)})$ is greater than or less than $h(x_{(2)})h(x_{(3)})$.

We shall now prove that the MLE $\hat{K}_n$ (which is unique with probability 1) tends to infinity in probability, that is, that

$$P(\hat{K}_n > k) \to 1 \qquad \text{for every} \quad k, \tag{5}$$

provided h satisfies

$$h(x) \geqslant e^{1/x^2} \tag{6}$$

for all sufficiently small values of x.

To prove (5), we will show that for any fixed j

$$P\left[p_{K_n^*}(\underline{X}) > p_j(\underline{X})\right] \to 1 \qquad \text{as} \quad n \to \infty \tag{7}$$

where K_n^* is the value of k for which $X_{(1)} \in I_k$. Since $p_{\hat{K}_n}(\underline{X}) \geqslant p_{K_n^*}(\underline{X})$, it then follows that for any fixed k

$$P\left[\hat{K}_n > k\right] \geqslant P\left[p_{\hat{K}_n}(\underline{X}) > p_j(\underline{X}) \qquad \text{for} \quad j = 1, \ldots, k\right] \to 1.$$

To prove (7), consider

$$L_{jk} = \log\frac{f_k(x_1)\cdots f_k(x_n)}{f_j(x_1)\cdots f_j(x_n)} = \Sigma_i^{(1)}\log\frac{h(x_i)}{c} - \Sigma_i^{(2)}\log\frac{h(x_i)}{c}$$

where $\Sigma^{(1)}$ and $\Sigma^{(2)}$ extend over all i for which $x_i \in I_k$ and $x_i \in I_j$, respectively. Now $x_i \in I_j$ implies that $h(x_i) < h(a_j)$, so that

$$\Sigma^{(2)}\log[h(x_i)/c] < \nu_{jn}\log[h(a_j)/c]$$

where ν_{jn} is the number of x's in I_j. Similarly, for $k = K_n^*$,

$$\Sigma^{(1)}\log[h(x_i)/c] \geqslant \log[h(x_{(1)})/c]$$

since $\log[h(x)/c] \geqslant 0$ for all x. Thus,

$$\frac{1}{n}L_{j,K_n^*} \geqslant \frac{1}{n}\log\frac{h(x_{(1)})}{c} - \frac{1}{n}\nu_{jn}\log\frac{h(a_j)}{c}.$$

Since ν_{jn}/n tends in probability to $P(X_1 \in I_j) < 1$, it only remains to show that

$$\frac{1}{n}\log h(X_{(1)}) \to \infty \quad \text{in probability.} \tag{8}$$

Instead of $X_1,\ldots, X_n$ consider a sample $Y_1,\ldots, Y_n$ from the uniform distribution $U(0, 1/c)$. Then for any x, $P(Y_i > x) \geqslant P(X_i > x)$ and hence

$$P[h(Y_{(1)}) > x] \leqslant P[h(X_{(1)}) > x]$$

and it is therefore enough to prove that $(1/n)\log h(Y_{(1)}) \to \infty$ in probability. If h satisfies (6), $(1/n)\log h(Y_{(1)}) \geqslant 1/nY_{(1)}^2$, and the right side tends to infinity in probability since $nY_{(1)}$ tends to a limit distribution (Problem 5.2.16). This completes the proof.

For later reference, note that the proof has established not only (7) but the fact that for any fixed A (Problem 2.8)

$$P[p_{K_n^*}(\underline{X}) > A^n p_j(\underline{X})] \to 1. \tag{9}$$

The example suggests (and this suggestion will be verified in the next section) that also for densities depending smoothly on a continuously varying parameter θ, the MLE need not be consistent. We shall now show, however, that a slightly weaker conclusion is possible under relatively mild conditions. Throughout the present section we shall assume θ to be real-valued. The case of several parameters will be taken up in Section 6.4.

The logarithm of the joint density

$$p(\underline{x}, \theta) = f(x_1, \theta)\cdots f(x_n, \theta) \tag{10}$$

when considered for fixed $\underline{x}$ as a function of θ,

$$L(\theta, \underline{x}) = \Sigma \log f(x_i, \theta) \tag{11}$$

is called the *log-likelihood*. We shall frequently use the shorthand notation $L(\theta)$ for $L(\theta, \underline{x})$ and $L'(\theta)$, $L''(\theta)$,... for its derivatives with respect to θ.

Theorem 2.2. *Let* $\mathrm{X}_1,\ldots, \mathrm{X}_n$ *satisfy* (A0)–(A3) *and suppose that for almost all* x, $\mathrm{f}(\mathrm{x}, \theta)$ *is differentiable with respect to* θ *in* ω, *with derivative* $\mathrm{f}'(\mathrm{x}, \theta)$. *Then with probability tending to* 1 *as* $\mathrm{n} \to \infty$, *the* likelihood equation

$$\frac{\partial}{\partial\theta}[\mathrm{f}(\mathrm{x}_1, \theta)\cdots \mathrm{f}(\mathrm{x}_n, \theta)] = 0 \tag{12}$$

or, equivalently, the equation

$$\mathrm{L}'(\theta, \underline{\mathrm{x}}) = \Sigma\frac{\mathrm{f}'(\mathrm{x}_i, \theta)}{\mathrm{f}(\mathrm{x}_i, \theta)} = 0 \tag{13}$$

has a root $\hat{\theta}_n = \hat{\theta}_n(\mathrm{x}_1,\ldots, \mathrm{x}_n)$ *such that* $\hat{\theta}_n(\mathrm{X}_1,\ldots, \mathrm{X}_n)$ *tends to the true value* θ_0 *in probability.*

Proof. Let a be small enough so that $(\theta_0 - a, \theta_0 + a) \subset \omega$, and let

$$S_n = \{\underline{x}: L(\theta_0, \underline{x}) > L(\theta_0 - a, \underline{x}) \quad \text{and} \quad L(\theta_0, \underline{x}) > L(\theta_0 + a, \underline{x})\}. \tag{14}$$

By Theorem 2.1, $P_{\theta_0}(S_n) \to 1$. For any $\underline{x} \in S_n$ there thus exists a value $\theta_0 - a < \hat{\theta}_n < \theta_0 + a$ at which $L(\theta)$ has a local maximum, so that $L'(\hat{\theta}_n) = 0$. Hence for any $a > 0$ sufficiently small, there exists a sequence $\hat{\theta}_n = \hat{\theta}_n(a)$ of roots such that

$$P_{\theta_0}(|\hat{\theta}_n - \theta_0| < a) \to 1.$$

It remains to show that we can determine such a sequence, which does not depend on a.

Let θ_n^* be the root closest to θ_0, [This exists because the limit of a sequence of roots is again a root by the continuity of $L(\theta)$.] Then clearly $P_{\theta_0}(|\theta_n^* - \theta_0| < a) \to 1$ and this completes the proof.

In connection with this theorem, the following comments may be helpful.

1. The proof yields the additional fact that with probability tending to 1, the roots $\hat{\theta}_n(a)$ can be chosen to be local maxima and so, therefore, can the θ_n^* if we let θ_n^* be the closest root corresponding to a maximum.
2. On the other hand, the theorem does not establish the existence of a consistent estimator sequence since, with the true value θ_0 unknown, the data do not tell us which root to choose so as to obtain a consistent sequence. An exception, of course, is the case in which the root is unique.
3. It should also be emphasized that existence of a root $\hat{\theta}_n$ is not asserted for all $\underline{x}$ (or for a given n even for any $\underline{x}$). This does not affect consistency, which only requires $\hat{\theta}_n$ to be defined on a set S_n', the probability of which tends to 1 as $n \to \infty$.
4. Although the likelihood equation can have many roots, the consistent sequence of roots generated by Theorem 2.2 is essentially unique. For a more precise statement of this result which is due to Huzurbazar (1948), see Problem 2.27.

Corollary 2.2. *Under the assumptions of Theorem* 2.2, *if the likelihood equation has a unique root* δ_n *for each* n *and all* $\underline{x}$, *then* $\{\delta_n\}$ *is a consistent sequence of estimators of* θ. *If, in addition, the parameter space is an open interval* $(\underline{\theta}, \bar{\theta})$ (*not necessarily finite*), *then with probability tending to* 1, δ_n *maximizes the likelihood, that is,* δ_n *is the MLE, which is therefore consistent.*

Proof. The first statement is obvious. To prove the second, suppose the probability of δ_n being the MLE does not tend to 1. Then for sufficiently large values of n, the likelihood must tend to a supremum as θ tends toward $\underline{\theta}$ or $\bar{\theta}$ with positive probability. Now with probability tending to 1, δ_n is a local maximum of the likelihood, which must then also possess a local minimum. This contradicts the assumed uniqueness of the root.

The conclusion of Corollary 2.2 holds, of course, not only when the root of the likelihood equation is unique but also when the probability of multiple roots tends to zero as $n \to \infty$. On the other hand, even when the root is unique, the corollary says nothing about its properties for finite n.

Example 2.4. Let X take on the values 0, 1, 2 with probabilities $6\theta^2 - 4\theta + 1$, $\theta - 2\theta^2, 3\theta - 4\theta^2$ $(0 < \theta < 1/2)$. Then the likelihood equation has a unique root for all x, which is a minimum for $x = 0$ and a maximum for $x = 1$ and 2 (Problem 2.10).

Theorem 2.2 establishes the existence of a consistent root of the likelihood equation. The next theorem asserts that any such sequence is asymptotically normal and efficient.

Theorem 2.3. *Suppose that* $X_1, \ldots, X_n$ *are iid and satisfy the assumptions of Theorem* 1.1, *with* (*iii*) *and* (*vi*) *replaced by the corresponding assumptions on the third* (*rather than the second*) *derivative, that is, by the existence of a third derivative satisfying*

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x,\theta)\right| \leqslant M(x) \tag{15}$$

$$\text{for all } x \in A, \qquad \theta_0 - c < \theta < \theta_0 + c$$

with

$$E_{\theta_0}[M(x)] < \infty.$$

Then any consistent sequence $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ *of roots of the likelihood equation satisfies*

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{I(\theta)}\right). \tag{17}$$

We shall call such a sequence $\hat{\theta}_n$ of roots an *efficient likelihood estimator* (ELE) of θ. It is typically (but need not be, see Example 3.1) provided by the MLE. More generally, any sequence $\hat{\theta}_n^*$ of estimators satisfying (17) will be said to be *asymptotically efficient.*

Proof. For any fixed x, expand $L'(\hat{\theta}_n)$ about θ_0

$$L'(\hat{\theta}_n) = L'(\theta_0) + (\hat{\theta}_n - \theta_0)L''(\theta_0) + \tfrac{1}{2}(\hat{\theta}_n - \theta_0)^2 L'''(\theta_n^*)$$

where θ_n^* lies between θ_0 and $\hat{\theta}_n$. By assumption, the left side is zero, so that

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) = \frac{(1/\sqrt{n})L'(\theta_0)}{-(1/n)L''(\theta_0) - (1/2n)(\hat{\theta}_n - \theta_0)L'''(\theta_n^*)}$$

where it should be remembered that $L(\theta)$, $L'(\theta)$, and so on are functions of

$\underline{X}$ as well as θ. We shall show that

$$\frac{1}{\sqrt{n}}L'(\theta_0) \xrightarrow{\mathcal{L}} N[0, I(\theta_0)], \tag{18}$$

that

$$-\frac{1}{n}L''(\theta_0) \xrightarrow{P} I(\theta_0) \tag{19}$$

and that

$$\frac{1}{n}L'''(\theta_n^*) \quad \text{is bounded in probability.} \tag{20}$$

The desired result then follows from Theorem 5.1.4.

Of the above statements, (18) follows from the fact that

$$\frac{1}{\sqrt{n}}L'(\theta_0) = \sqrt{n}\,\frac{1}{n}\Sigma\left[\frac{f'(X_i, \theta_0)}{f(X_i, \theta_0)} - E_{\theta_0}\frac{f'(X_i, \theta_0)}{f(X_i, \theta_0)}\right]$$

since the expectation term is zero, and then from the CLT and the definition of $I(\theta)$.

Next,

$$-\frac{1}{n}L''(\theta_0) = \frac{1}{n}\Sigma\frac{f'^2(X_i, \theta_0) - f(X_i, \theta_0)f''(X_i, \theta_0)}{f^2(X_i, \theta_0)}.$$

By the law of large numbers, this tends in probability to

$$I(\theta_0) - E_{\theta_0}\frac{f''(X_i, \theta_0)}{f(X_i, \theta_0)} = I(\theta_0).$$

Finally,

$$\frac{1}{n}L'''(\theta) = \frac{1}{n}\Sigma\frac{\partial^3}{\partial\theta^3}\log f(X_i, \theta)$$

so that by (15)

$$\left|\frac{1}{n}L'''(\theta_n^*)\right| < \frac{1}{n}[M(X_1) + \cdots + M(X_n)]$$

with probability tending to 1. The right side tends in probability to $E_{\theta_0}[M(X)]$, and this completes the proof.

Although the conclusions of Theorem 2.3 are quite far-reaching, the proof is remarkably easy. The reason is that Theorem 2.2 already puts $\hat{\theta}_n$ into the neighborhood of the true value θ_0, so that an expansion about θ_0 essentially linearizes the problem and thereby prepares the way for application of the central limit theorem.

Corollary 2.3. *Under the assumptions of Theorem* 2.3, *if the likelihood equation has a unique root for all* n *and* $\underline{x}$, *and more generally if the probability of multiple roots tends to zero as* $n \to \infty$, *the MLE is asymptotically efficient.*

Before giving some applications of Theorem 2.3 and Corollary 2.3, let us briefly consider what is involved in checking its assumptions. The following are two conditions that may not be obvious:

1. Differentiability twice of $\int f(x, \theta)\,d\mu(x)$ with respect to θ by differentiating under the integral sign.
2. Condition (15), which states that the third derivative is uniformly bounded by an integrable function.

Conditions for 1. are given in books on calculus, although it is often easier simply to calculate the difference quotient and pass to the limit.

Condition 2 is usually easy to check after realizing that it is not necessary for (15) to hold for all θ, but that it is enough if there exist $\theta_1 < \theta_0 < \theta_2$ such that (15) holds for all $\theta_1 \leqq \theta \leqq \theta_2$.

Example 2.5. One-Parameter Exponential Family. Let $X_1, \ldots, X_n$ be iid according to a one-parameter exponential family with density

$$f(x_i, \eta) = e^{\eta T(x_i) - A(\eta)} \tag{21}$$

with respect to a σ-finite measure μ, and let the estimand be η. The likelihood equation is

$$\frac{1}{n}\Sigma T(x_i) = A'(\eta)$$

and hence by (1.4.11)

$$E_\eta[T(X_j)] = \frac{1}{n}\Sigma T(x_i). \tag{22}$$

The left side of (22) is a strictly increasing function of η since by (1.4.12)

$$\frac{d}{d\eta}[E_\eta T(X_j)] = \mathrm{var}_\eta T(X_j) > 0.$$

It follows that equation (22) has at most one solution. The conditions of Theorem 2.3 are easily checked in the present case. In particular condition 1 follows from Theorem 1.4.1, and 2 from the fact that the third derivative of $\log f(x, \eta)$ is independent of x and a continuous function of η. With probability tending to 1, (22) therefore has a solution $\hat{\eta}$. This solution is unique, consistent, and asymptotically efficient, so that

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{\mathcal{L}} N(0, 1/\operatorname{var} T) \tag{23}$$

where $T = T(X_i)$ and the asymptotic variance follows from (2.7.13).

Example 2.6. Truncated Normal. As an illustration of the preceding example, consider a sample of n observations from a normal distribution $N(\xi, 1)$, truncated at two fixed points $a < b$. The density of a single X is then

$$\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x - \xi)^2\right] / [\Phi(b - \xi) - \Phi(a - \xi)], \qquad a < x < b,$$

which satisfies (21) with $\eta = \xi, T(x) = x$. An ELE will therefore be the unique solution of $E_\xi(X) = \bar{x}$ if it exists. To see that this equation has a solution for any value $a < \bar{x} < b$, note that as $\xi \to -\infty$ or $+\infty$, X tends in probability to a or b, respectively (Problem 2.11). Since X is bounded, this implies that also $E_\xi(X)$ tends to a or b. Since $E_\xi(X)$ is continuous, existence of $\hat{\xi}$ follows.

As a second class of examples, consider some location problems. Since the MLE or ELE is equivariant (see Section 3.1), its risk cannot be smaller than that of the MRE estimator. One would thus typically expect the MRE (Pitman) estimator also to be asymptotically efficient. This was shown to be the case for a large class of loss functions by Stone (1974). Given a specific loss function L and finite n, the Pitman estimator δ_{Ln} has a risk advantage over the ELE $\hat{\xi}_n$. However, if one is vague about the loss function, the ELE may be viewed as an alternative of δ_{Ln}, with the two asymptotically equivalent in the sense that $\sqrt{n}(\hat{\xi}_n - \delta_{Ln}) \to 0$ in probability [Stone (1974)], but where for finite n and $L^* \neq L$, $\hat{\xi}_n$ may have smaller or larger risk than δ_{Ln}.

Example 2.7. Logistic. If $X_1, \ldots, X_n$ are iid according to the logistic density of Table 1.3.1 with $a = \theta, b = 1$, the likelihood equation after some simplification becomes (Problem 2.13)

$$\sum \frac{1}{1 + \exp(x_i - \theta)} = \frac{n}{2}. \tag{24}$$

The left side is an increasing function of θ which is zero at $-\infty$ and n at $+\infty$. (24) therefore has a unique root $\hat{\theta}$, which is the MLE since $L'(\theta)$ is > 0 or < 0 as θ is $< \hat{\theta}$ or $> \hat{\theta}$.

Consider now the general case of $X_1, \ldots, X_n$ iid each with density $f(x - \theta)$, where f is differentiable and $f(x) > 0$ for all x. The likelihood equation reduces to

$$\sum \frac{f'(x_i - \theta)}{f(x_i - \theta)} = 0. \tag{25}$$

If f is strictly strongly unimodal so that $f'(x)/f(x)$ is strictly decreasing, this equation has at most one root. On the other hand, the likelihood tends to zero as $\theta \to \pm\infty$ and therefore has a maximum $\hat{\theta}$ in the interior which must satisfy (25). When f is not strongly unimodal, (25) may have several roots. The determination of an ELE in this case will be considered in the next section.

Example 2.8. Double Exponential. For the double exponential density $DE(\theta, 1)$ given in Table 1.3.1, it is not true that for all (or almost all) x, $f(x - \theta)$ is differentiable with respect to θ, since for every x there exists a value $(\theta = x)$ at which the derivative does not exist. Despite this failure, the MLE (which is the median of the X's) satisfies the conclusion of Theorem 2.3.

In the following example, $f(x, \theta)$ is neither an exponential, location, or scale family.

Example 2.9. Weibull. Let $X_1, \ldots, X_n$ be iid according to a *Weibull distribution* with density

$$f_\theta(x) = \theta x^{\theta - 1} \exp(-x^\theta), \qquad x > 0, \qquad \theta > 0. \tag{26}$$

The log likelihood of $x_1, \ldots, x_n$ is

$$n \log \theta + (\theta - 1)\Sigma \log x_i - \Sigma \exp(\theta \log x_i)$$

so that

$$L'(\theta) = \frac{n}{\theta} + \Sigma \log x_i - \Sigma \log x_i \exp(\theta \log x_i)$$

and

$$L''(\theta) = -\frac{n}{\theta^2} - \Sigma (\log x_i)^2 \exp(\theta \log x_i) < 0.$$

Thus, if the likelihood equation has a root, it is unique and equals the MLE. As $\theta \to 0$, $L'(\theta) \to \infty$ while for θ sufficiently large, $L'(\theta)$ is negative. This proves the existence of a root.

3. LIKELIHOOD ESTIMATION: MULTIPLE ROOTS

When the likelihood equation has multiple roots, the assumptions of Theorem 2.3 are no longer sufficient to guarantee consistency of the MLE, even when it exists for all n. This is shown by the following example due to Le Cam (1979b), which is obtained by embedding the sequence $\{f_k\}$ of Example 2.3 in a sufficiently smooth continuous-parameter family.

Example 3.1. For $k \leqslant \theta < k+1$, $k = 1, 2, \ldots$, let

$$f(x,\theta) = [1 - u(\theta - k)]f_k(x) + u(\theta - k)f_{k+1}(x), \tag{1}$$

with f_k defined as in Example 2.3 and u defined on $(-\infty, \infty)$ such that $u(x) = 0$ for $x \leqslant 0$, $u(x) = 1$ for $x \geqslant 1$, is strictly increasing on $(0,1)$ and infinitely differentiable on $(-\infty, \infty)$ (Problem 3.1). Let $X_1, \ldots, X_n$ be iid, each with density $f(x,\theta)$, and let $p(\underset{\sim}{x}, \theta) = \Pi f(x_i, \theta)$.

Since for any given $\underset{\sim}{x}$, the density $p(\underset{\sim}{x}, \theta)$ is bounded and continuous in θ and is equal to c^n for all sufficiently large θ and greater than c^n for some θ, it takes on its maximum for some finite θ and the MLE $\hat{\theta}_n$ therefore exists.

To see that $\hat{\theta}_n \to \infty$ in probability, note that for $k \leqslant \theta < k+1$

$$p(\underset{\sim}{x}, \theta) \leqslant \Pi \max[f_k(x_i), f_{k+1}(x_i)] = \bar{p}_k(\underset{\sim}{x}). \tag{2}$$

If $\hat{K}_n$ and K_n^* are defined as in Example 2.3, the argument of that example shows that it is enough to prove that for any fixed j

$$P\left[p_{K_n^*}(\underset{\sim}{X}) > \bar{p}_j(\underset{\sim}{X})\right] \to 1 \qquad \text{as} \qquad n \to \infty, \tag{3}$$

where $p_k(\underset{\sim}{x}) = p(\underset{\sim}{x}, k)$. Now

$$\bar{L}_{jk} = \frac{p_k(\underset{\sim}{x})}{\bar{p}_j(\underset{\sim}{x})} = \Sigma^{(1)} \log \frac{h(x_i)}{c} - \Sigma^{(2)} \log \frac{h(x_i)}{c} - \Sigma^{(3)} \log \frac{h(x_i)}{c}$$

where $\Sigma^{(1)}$, $\Sigma^{(2)}$, and $\Sigma^{(3)}$ extend over all i for which $x_i \in I_k$, $x_i \in I_j$ and $x_i \in I_{j+1}$, respectively. The argument is now completed as before to show that $\hat{\theta}_n \to \infty$ in probability regardless of the true value of θ and is therefore not consistent.

The example is not yet completely satisfactory since $\partial f(x,\theta)/\partial\theta = 0$ and hence $I(\theta) = 0$ for $\theta = 1, 2, \ldots$. [The remaining conditions of Theorem 2.3 are easily checked (Problem 3.2).] To remove this difficulty, define

$$g(x,\theta) = \tfrac{1}{2}\left[f(x,\theta) + f(x, \theta + \alpha e^{-\theta^2})\right], \qquad \theta \geqslant 1 \tag{4}$$

for some fixed $\alpha < 1$.

If $X_1, \ldots, X_n$ are iid according to $g(x,\theta)$, we shall now show that the MLE $\hat{\theta}_n$ continues to tend to infinity for any fixed θ. We have, as before

$$P\left[\hat{\theta}_n > k\right] \geqslant P\left[\Pi g(x_i, K_n^*) > \Pi g(x_i, \theta) \qquad \text{for all} \quad \theta \leqslant k\right]$$

$$\geqslant P\left\{\frac{1}{2^n}\Pi f(x_i, K_n^*) > \Pi\left(\frac{1}{2}\left[f(x_i,\theta) + f\left(x_i, \theta + \alpha e^{-\theta^2}\right)\right]\right)\right\}.$$

For $j \leqslant \theta < j+1$ it is seen from (1) that $[f(x_i,\theta) + f(x_i, \theta + \alpha e^{-\theta^2})]/2$ is a weighted average of $f_j(x_i), f_{j+1}(x_i)$ and possibly $f_{j+2}(x_i)$. By using $\bar{p}_j(\underset{\sim}{x}) = \Pi \max[f_j(x_i), f_{j+1}(x_i), f_{j+2}(x_i)]$ in place of $p_j(\underset{\sim}{x})$, the proof can now be completed as before. Since the densities $g(x_i,\theta)$ satisfy the conditions of Theorem 2.3 (Problem 3.3) these conditions therefore are not enough to ensure the consistency of the MLE. (For another example, see Ferguson, 1982.)

Even under the assumptions of Theorem 2.3 one is thus, in the case of multiple roots, still faced with the problem of identifying a consistent sequence of roots. Following are three possible approaches.

(i) In many cases, the maximum likelihood estimator is consistent. Conditions which ensure this were given by among others Wald (1949), Le Cam (1953, 1970), Kiefer and Wolfowitz (1956), Bahadur (1967), and Perlman (1972). A survey of the literature can be found in Perlman (1983). This material is technically difficult, and even when the conditions are satisfied, the determination of the MLE may present problems [see Barnett (1966)]. We shall therefore turn to somewhat simpler alternatives.

The following two methods require that some sequence of consistent (but not necessarily efficient) estimators be available. The existence of such a sequence is guaranteed under very weak assumptions by a theorem of Le Cam (1956); see also Kraft and Le Cam (1956). In any given situation, it is usually easy to construct a consistent sequence, as will be illustrated below and in the next section.

(ii) Suppose that δ_n is any consistent estimator of θ and that the assumptions of Theorem 2.3 hold. Then the root $\hat{\theta}_n$ of the likelihood equation closest to δ_n (which exists by the proof of Theorem 2.2) is also consistent, and hence is efficient by Theorem 2.3.

To see this, note that by Theorem 2.3 there exists a consistent sequence of roots, say $\hat{\theta}_n^*$. Since $\hat{\theta}_n^* - \delta_n \to 0$ in probability, so does $\hat{\theta}_n - \delta_n$.

The following approach, which does not require the determination of the closest root, and in which the estimators are no longer exact roots of the likelihood equation, is often more convenient.

(iii) The usual iterative methods for solving the likelihood equation

$$L'(\theta) = 0 \tag{5}$$

are based on replacing the left side by the linear terms of its Taylor expansion about an approximate solution $\tilde{\theta}$. If $\hat{\theta}$ denotes a root of (5), this leads to the approximation

$$0 = L'(\hat{\theta}) \doteq L'(\tilde{\theta}) + (\hat{\theta} - \tilde{\theta})L''(\tilde{\theta}), \tag{6}$$

and hence to

$$\hat{\theta} = \tilde{\theta} - \frac{L'(\tilde{\theta})}{L''(\tilde{\theta})}. \tag{7}$$

The procedure is then iterated by replacing $\tilde{\theta}$ by the value $\tilde{\tilde{\theta}}$ of the right side of (7), and so on. (For a discussion of the performance of this procedure, see for example Barnett, 1966.)

Here we are concerned only with the first step and with the performance of the one-step approximation (7) as an estimator of θ. The following result gives conditions on $\tilde{\theta}$ under which the resulting sequence of estimators is consistent, asymptotically normal, and efficient.

Theorem 3.1. *Suppose that the assumptions of Theorem* 2.3 *hold and that* $\tilde{\theta}_n$ *is not only a consistent but a* $\sqrt{n}$*-consistent* estimator of* θ*, that is, that* $\sqrt{n}(\tilde{\theta}_n - \theta)$ *is bounded in probability so that* $\tilde{\theta}_n$ *tends to* θ *at least at the rate of* $1/\sqrt{n}$*. Then the estimator sequence*

$$\delta_n = \tilde{\theta}_n - \frac{L'(\tilde{\theta}_n)}{L''(\tilde{\theta}_n)} \tag{8}$$

is asymptotically efficient, that is, it satisfies (2.17) *with* δ_n *in place of* $\hat{\theta}_n$.

Proof. As in the proof of Theorem 2.3, expand $L'(\tilde{\theta}_n)$ about θ_0 as

$$L'(\tilde{\theta}_n) = L'(\theta_0) + (\tilde{\theta}_n - \theta_0)L''(\theta_0) + \tfrac{1}{2}(\tilde{\theta}_n - \theta_0)^2 L'''(\theta_n^*)$$

where θ_n^* lies between θ_0 and $\tilde{\theta}_n$. Substituting this expression into (8) and simplifying, we find

$$\sqrt{n}(\delta_n - \theta_0) = \frac{(1/\sqrt{n})L'(\theta_0)}{-(1/n)L''(\tilde{\theta}_n)} + \sqrt{n}(\tilde{\theta}_n - \theta_0) \times \left[1 - \frac{L''(\theta_0)}{L''(\tilde{\theta}_n)} - \frac{1}{2}(\tilde{\theta}_n - \theta_0)\frac{L'''(\theta_n^*)}{L''(\tilde{\theta}_n)}\right]. \tag{9}$$

To prove the result, we shall use (2.18) and (2.19), which are consequences of the assumptions on $f(x, \theta)$, and (2.20) together with the fact that

*A general method for constructing $\sqrt{n}$-consistent estimators is given by Le Cam (1969, p. 103).

$\theta_n^* \to \theta_0$ in probability. We shall also need that

$$\frac{L''(\tilde{\theta}_n)/n}{L''(\theta_0)/n} \to 1 \quad \text{in probability.} \tag{10}$$

This follows from (2.19) and (2.20), and the expansion

$$\frac{1}{n}L''(\tilde{\theta}_n) = \frac{1}{n}L''(\theta_0) + \frac{1}{n}(\tilde{\theta}_n - \theta_0)L'''(\theta_n^{**})$$

where θ_n^{**} is between θ_0 and $\tilde{\theta}_n$.

Consider now the right side of (9). By (2.18), (2.19), and (10), the first term has the limit distribution asserted for the left side. It only remains to show that the expression in square brackets on the right side of (9) tends to zero in probability. This follows from (2.20) and the assumptions made about $\tilde{\theta}_n$.

Corollary 3.1. *Suppose that the assumptions of Theorem* 3.1 *hold and that* $I(\theta)$ *is a continuous function of* θ*. Then the estimator*

$$\delta_n' = \tilde{\theta}_n + \frac{L'(\tilde{\theta}_n)}{nI(\tilde{\theta}_n)} \tag{11}$$

is also asymptotically efficient.

Proof. By (10), condition (viii) of Theorem 1.1, and the law of large numbers, $-(1/n)L''(\tilde{\theta}_n) \to I(\theta_0)$ in probability. Since $I(\theta)$ is continuous, also $I(\tilde{\theta}_n) \to I(\theta_0)$ in probability, so that $-(1/n)L''(\tilde{\theta}_n)/I(\tilde{\theta}_n) \to 1$ in probability, and this completes the proof.

The estimators (8) and (11) are compared by Stuart (1958), who gives a heuristic argument why (11) might be expected to be closer to the ELE than (8) and provides a numerical example supporting this argument.

Example 3.2. Location Parameter. Consider the case of a symmetric location family in which the likelihood equation (2.25) has multiple roots. [For the Cauchy distribution, for example, it has been shown by Reeds (1980) that if (2.25) has $K + 1$ roots, then as $n \to \infty$, K tends in law to a Poisson distribution with expectation $1/\pi$. The Cauchy case has also been considered by Barnett (1966).] If $\text{var}(X) < \infty$, it follows from the CLT that the sample mean $\bar{X}_n$ is $\sqrt{n}$-consistent and that an asymptotically efficient estimator of θ is therefore provided by (8) or (11) with $\tilde{\theta}_n = \bar{X}$ if $f(x, \theta)$ satisfies the conditions of Theorem 2.3. For distributions such as the Cauchy for which $E(X^2) = \infty$, Theorem 5.3.2 shows that one can, instead, take for $\tilde{\theta}_n$ the sample median provided $f(0) > 0$; other L, M, or R-estimators provide still further possibilities.

Example 3.3. Grouped or Censored Observations. Suppose that $X_1, \ldots, X_n$ are iid according to a location family with cdf $F(x-\theta)$, with F known and with $0 < F(x) < 1$ for all x, but that it is only observed whether each X_i falls below a, between a and b, or above b where $a < b$ are two given constants. The n observations constitute n trinomial trials with probabilities $p_1 = p_1(\theta) = F(a-\theta)$, $p_2(\theta) = F(b-\theta) - F(a-\theta)$, $p_3(\theta) = 1 - F(b-\theta)$ for the three outcomes. If V denotes the number of observations less than a, then

$$\sqrt{n}\left[\frac{V}{n} - p_1\right] \xrightarrow{\mathcal{L}} N[0, p_1(1-p_1)] \tag{12}$$

and, by Theorem 5.1.5,

$$\tilde{V}_n = a - F^{-1}\left(\frac{V}{n}\right) \tag{13}$$

is a $\sqrt{n}$-consistent estimator of θ. Since the estimator is not defined when $V = 0$ or $V = n$, some special definition has to be adopted in these cases whose probability however tends to zero as $n \to \infty$.

If the trinomial distribution for a single trial satisfies the assumptions of Theorem 2.3, as will be the case under mild assumptions on F, the estimator (8) is asymptotically efficient (but see the comment following Example 3.5). The approach applies, of course, equally to the case of more than three groups.

A very similar situation arises when the X's are *censored*, say at a fixed point a. For example, they might be lengths of life of light bulbs or patients, with observation discontinued at time a. The observations can then be represented as

$$Y_i = \begin{cases} X_i & \text{if } X_i < a \\ a & \text{if } X_i \geqslant a. \end{cases}$$

Here the value a of Y_i when $X_i \geqslant a$ has no significance; it simply indicates that the value of X_i is $\geqslant a$. The Y's are then iid with density

$$g(y,\theta) = \begin{cases} f(y-\theta) & \text{if } y < a \\ 1 - F(a-\theta) & \text{if } y = a \end{cases}$$

with respect to the measure μ which is Lebesgue measure on $(-\infty, a)$ and assigns measure 1 to the point $y = a$.

The estimator (13) continues to be $\sqrt{n}$-consistent in the present situation. An alternative starting point is, for example, the best linear combination of the ordered X's less than a (see, for example, Chan, 1967).

Estimation from grouped or censored observations arises not only for location families but also for any other family of distributions and has been treated in particular for exponential families (see, for example, Sundberg 1976). An algorithm for computing MLEs in such problems is discussed by Dempster, Laird, and Rubin (1977).

Example 3.4. Mixtures. Let $X_1, \ldots, X_n$ be a sample from a distribution $\theta G + (1-\theta)H$, $0 < \theta < 1$, where G and H are two specified distributions with densities g and h. The log likelihood of a single observation is a concave function of θ, and so therefore is the log likelihood of a sample (Problem 3.5). It follows that the likelihood equation has at most one solution. (The asymptotic performance of the estimator is studied by Hill, 1963.)

Even when the root is unique, Theorem 3.1 provides an alternative, which may be more convenient than the MLE. In the mixture problem, as in many other cases, a $\sqrt{n}$-consistent estimator can be obtained by the method of moments, which consists in equating the first k moments of X to the corresponding sample moments, say

$$E_\theta(X_i^r) = \frac{1}{n}\sum_{j=1}^{n} X_j^r, \qquad r = 1, \ldots, k,$$

where k is the number of unknown parameters. [For further discussion, see, for example, Cramér (1946, Sect. 33.1) and Serfling (1980, Sect. 4.3.1),] In the present case, suppose that $E(X_i) = \xi$ or η when X is distributed as G or H where $\eta \neq \xi$ and G and H have finite variance. Since $k = 1$, the method of moments estimates θ as the solution of the equation

$$\xi\theta + \eta(1-\theta) = \bar{X}_n$$

and hence by

$$\tilde{\theta}_n = \frac{\bar{X}_n - \eta}{\xi - \eta}.$$

[If $\eta = \xi$ but the second moments of X_i under H and G differ, one can instead equate $E(X_i^2)$ with $\Sigma X_j^2/n$ (Problem 3.6).] An asymptotically efficient estimator is then provided by (8).

In any application of Theorem 3.1, the question naturally arises which $\sqrt{n}$-consistent sequence $\{\tilde{\theta}_n\}$ to use as the starting point. It seems plausible that one would do best with a highly efficient starting sequence. More generally, this is a special case of the problem of choosing among the many alternative sequences of asymptotically efficient estimators. It had been pointed out earlier that an efficient estimator sequence can always be modified by terms of order $1/n$ without affecting the asymptotic efficiency. To distinguish among them requires taking into account the terms of the next order.

A number of authors (among them Rao, 1963; Pfanzagl, 1973; Ghosh and Subramanyam, 1974; Efron, 1975, 1978; and Akahira and Takeuchi, 1981), have investigated estimators that are "second-order efficient," that is, efficient and among efficient estimators have the greatest accuracy to terms of the next order, and in particular they have tried to determine to what extent the MLE is second-order efficient. Their findings, for sufficiently regular situations, can be summarized roughly as follows.