MAXIMUM LIKELIHOOD ESTIMATION

Maximum likelihood is by far the most popular general method of estimation. Its widespread acceptance is seen on the one hand in the very large body of research dealing with its theoretical properties, and on the other in the almost unlimited list of applications.

To give a reasonably general definition of maximum likelihood estimates, let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector of observations whose joint distribution is described by a density $f_n(\mathbf{x}|\boldsymbol{\Theta})$ over the n-dimensional Euclidean space R^n. The unknown parameter vector $\boldsymbol{\Theta}$ is contained in the parameter space $\Omega \subset R^s$. For fixed $\mathbf{x}$ define the likelihood* function of $\mathbf{x}$ as $L(\boldsymbol{\Theta}) = L_{\mathbf{x}}(\boldsymbol{\Theta}) = f_n(\mathbf{x}|\boldsymbol{\Theta})$ considered as a function of $\boldsymbol{\Theta} \in \Omega$.

Definition 1. Any $\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Theta}}(\mathbf{x}) \in \Omega$ which maximizes $L(\boldsymbol{\Theta})$ over Ω is called a *maximum likelihood estimate* (MLE) of the unknown true parameter $\boldsymbol{\Theta}$.

Often it is computationally advantageous to derive MLEs by maximizing $\log L(\boldsymbol{\Theta})$ in place of $L(\boldsymbol{\Theta})$.

Example 1. Let X be the number of successes in n independent Bernoulli trials with success probability $p \in [0, 1]$; then

$$\begin{aligned} L_x(p) &= f(x|p) = P(X = x|p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n. \end{aligned}$$

Solving

$$\frac{\partial}{\partial p} \log L_x(p) = x/p - (n-x)/(1-p) = 0$$

for p, one finds that $\log L_x(p)$ and hence $L_x(p)$ has a maximum at

$$\hat{p} = \hat{p}(x) = x/n.$$

This example illustrates the considerable intuitive appeal of the MLE as that value of p for which the probability of the observed value x is the largest.

It should be pointed out that MLEs do not always exist, as illustrated in the following natural mixture example; see Kiefer and Wolfowitz [32].

Example 2. Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) with density

$$\begin{aligned} f(x|\mu, \nu, \sigma, \tau, p) = {} & \frac{p}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ & + \frac{1-p}{\sqrt{2\pi}\tau} \exp\left[-\frac{1}{2}\left(\frac{x-\nu}{\tau}\right)^2\right], \end{aligned}$$

where $0 \leqslant p \leqslant 1, \mu, \nu \in R$, and $\sigma, \tau > 0$.

The likelihood function of the observed sample $x_1, \ldots x_n$, although finite for any permissible choice of the five parameters, approaches infinity as, for example, $\mu = x_1, p > 0$ and $\sigma \to 0$. Thus the MLEs of the five unknown parameters do not exist.

Further, if an MLE exists, it is not necessarily unique as is illustrated in the following example.

Example 3. Let $X_1, \ldots, X_n$ be i.i.d. with density $f(x|\alpha) = \frac{1}{2}\exp(-|x-\alpha|)$. Maximizing $f_n(x_1, \ldots, x_n|\alpha)$ is equivalent to minimizing $\sum |x_i - \alpha|$ over α. For $n = 2m$ one finds that any $\hat{\alpha} \in [x_{(m)}, x_{(m+1)}]$ serves as MLE of α, where $x_{(i)}$ is the ith order statistic of the sample.

The method of maximum likelihood estimation is generally credited to Fisher* [17–20], although its roots date back as far as Lambert*, Daniel Bernoulli*, and Lagrange in the eighteenth century; see Edwards [12] for an historical account. Fisher introduces the method in [17] as an alternative to the method of moments* and the method of least squares*. The former method Fisher criticizes for its arbitrariness in the choice of moment equations and the latter for not being invariant under scale changes in the variables. The term *likelihood** as distinguished from (*inverse*) *probability* appears for the first time in [18]. Introducing the measure of information named after him (*see* FISHER INFORMATION) Fisher [18–20] offers several proofs for the efficiency of MLEs, namely that the asymptotic variance of asymptotically normal estimates cannot fall below the reciprocal of the information contained in the sample and, furthermore, that the MLE achieves this lower bound. Fisher's proofs, obscured by the fact that assumptions are not always clearly stated, cannot be considered completely rigorous by today's standards and should be understood in the context of his time. To some extent his work on maximum likelihood estimation was anticipated by Edgeworth [11], whose contributions are discussed by Savage [51] and Pratt [45]. However, it was Fisher's insight and advocacy that led to the prominence of maximum likelihood estimation as we know it today.

For a discussion and an extension of Definition 1 to richer (nonparametric) statistical models which preclude a model description through densities (i.e., likelihoods will be missing), see Scholz [52]. At times the primary concern is the estimation of some function g of $\mathbf{\Theta}$. It is then customary to treat $g(\hat{\mathbf{\Theta}})$ as an "MLE" of $g(\mathbf{\Theta})$, although strictly speaking, Definition 1 only justifies this when g is a one-to-one function. For arguments toward a general justification of $g(\hat{\mathbf{\Theta}})$ as MLE of $g(\mathbf{\Theta})$, see Zehna [58] and Berk [7].

CONSISTENCY

Much of maximum likelihood theory deals with the large sample (asymptotic) properties of MLEs; i.e., with the case in which it is assumed that $X_1, \ldots, X_n$ are independent and identically distributed with density $f(\cdot|\mathbf{\Theta})$ (i.e., $X_1, \ldots, X_n$ i.i.d. $\sim f(\cdot|\mathbf{\Theta})$). The joint density of $\mathbf{X} = (X_1, \ldots, X_n)$ is then $f_n(\mathbf{x}|\mathbf{\Theta}) = \prod_{i=1}^{n} f(x_i|\mathbf{\Theta})$. It further is assumed that the distributions $P_{\boldsymbol{\theta}}$ corresponding to $f(\cdot|\mathbf{\Theta})$ are identifiable, i.e., $\mathbf{\Theta} \neq \mathbf{\Theta}'$, and $\mathbf{\Theta}, \mathbf{\Theta}' \in \Omega$ implies $P_{\mathbf{\Theta}} \neq P_{\mathbf{\Theta}'}$. For future reference we state the following assumptions:

- A0: $X_1, \ldots, X_n$ i.i.d. $f(\cdot|\mathbf{\Theta})\mathbf{\Theta} \in \Omega$;
- A1: the distributions $P_{\mathbf{\Theta}}, \mathbf{\Theta} \in \Omega$, are identifiable.

The following simple result further supports the intuitive appeal of the MLE; see Bahadur [3]:

Theorem 1. Under **A0** and **A1**

$$P_{\mathbf{\Theta}'}[f_n(\mathbf{X}|\mathbf{\Theta}') > f_n(\mathbf{X}|\mathbf{\Theta})] \to 1$$

as $n \to \infty$ for any $\mathbf{\Theta}, \mathbf{\Theta}' \in \Omega$ with $\mathbf{\Theta} \neq \mathbf{\Theta}'$. If, in addition, Ω is finite, then the MLE $\hat{\mathbf{\Theta}}_n$ exists and is consistent.

The content of Theorem 1 is a cornerstone in Wald's [56] consistency proof of the MLE for the general case. Wald assumes that Ω is compact, which by a familiar compactness argument reduces the problem to the case in which Ω contains only finitely many elements. Aside from the compactness assumption on Ω, which often is not satisfied in practice, Wald's uniform integrability conditions (imposed on $\log f(\cdot|\mathbf{\Theta})$) often are not satisfied in typical examples.

Many improvements in Wald's approach toward MLE consistency were made by later researchers. For a discussion and further references, see Perlman [42]. Instead of Wald's theorem or any of its refinements, we present another theorem, due to Rao [47], which shows under what simple conditions MLE consistency may be established in a certain specific situation.

Theorem 2. Let **A0** and **A1** be satisfied and let $f(\cdot|\mathbf{\Theta})$ describe a multinomial experiment with cell probabilities $\boldsymbol{\pi}(\mathbf{\Theta}) = (\pi_1(\mathbf{\Theta}), \ldots, \pi_k(\mathbf{\Theta}))$. If the map $\mathbf{\Theta} \to \boldsymbol{\pi}(\mathbf{\Theta}), \mathbf{\Theta} \in \Omega$, has a continuous inverse (the inverse existing because of A1), then the MLE $\hat{\mathbf{\Theta}}_n$, if it exists, is a consistent estimator of $\mathbf{\Theta}$.

For a counterexample to Theorem 2 when the inverse continuity assumption is not satisfied see Kraft and LeCam [33].

A completely different approach toward proving consistency of MLEs was given by Cramér [9]. His proof is based on a Taylor expansion of $\log L(\mathbf{\Theta})$ and thus, in contrast to Wald's proof, assumes a certain amount of smoothness in $f(\cdot|\mathbf{\Theta})$ as a function of $\mathbf{\Theta}$. Cramér gave the consistency proof only for $\Omega \subset R$. Presented here are his conditions generalized to the multiparameter case, $\Omega \subset R^s$:

- C1: The distributions $P_{\mathbf{\Theta}}$ have common support for all $\mathbf{\Theta} \in \Omega$; i.e., $\{x : f(x|\mathbf{\Theta}) > 0\}$ does not change with $\mathbf{\Theta} \in \Omega$.

C2: There exists an open subset ω of Ω containing the true parameter point $\mathbf{\Theta}_0$ such that for almost all x the density $f(x|\mathbf{\Theta})$ admits all third derivatives

$$\frac{\partial^3}{\partial\Theta_i\partial\Theta_j\partial\Theta_k}f(x|\mathbf{\Theta}) \qquad \text{for all} \quad \mathbf{\Theta} \in \omega.$$

C3:

$$E_{\Theta}\left[\frac{\partial}{\partial\Theta_j}\log f(X|\mathbf{\Theta})\right] = 0, \quad j = 1, \ldots, s$$

and

$$\begin{aligned} I_{jk}(\mathbf{\Theta}) :&= E_{\mathbf{\Theta}}\left[\left(\frac{\partial}{\partial\mathbf{\Theta}_j}\log f(X|\mathbf{\Theta})\right)\right. \\ &\quad \times \left.\left(\frac{\partial}{\partial\mathbf{\Theta}_k}\log f(X|\mathbf{\Theta})\right)\right] \\ &= -E_{\mathbf{\Theta}}\left[\frac{\partial^2}{\partial\Theta_j\partial\Theta_k}\log f(X|\mathbf{\Theta})\right] \end{aligned}$$

exist and are finite for $j, k = 1, \ldots, s$ and all $\mathbf{\Theta} \in \omega$.

C4: The Fisher information matrix $\mathbf{I}(\mathbf{\Theta}) = (I_{jk}(\mathbf{\Theta}))_{j,k=1,\ldots,s}$ is positive definite for all $\mathbf{\Theta} \in \omega$.

C5: There exist functions $M_{ijk}(x)$ independent of $\mathbf{\Theta}$ such that for all $i, j, k, = 1, \ldots, s$,

$$\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\log f(x|\mathbf{\Theta})\right| \leqslant M_{ijk}(x) \qquad \text{for all} \quad \mathbf{\Theta} \in \omega,$$

where

$$E_{\mathbf{\Theta}_0}(M_{ijk}(X)) < \infty.$$

We can now state Cramér's consistency theorem.

Theorem 3. Assume **A0, A1**, and **C1–C5**. Then with probability tending to one as $n \to \infty$ there exist solutions $\tilde{\mathbf{\Theta}}_n = \tilde{\mathbf{\Theta}}_n(X_1, \ldots, X_n)$ of the likelihood equations

$$\frac{\partial}{\partial\Theta_j}\log f_n(\mathbf{X}|\mathbf{\Theta}) = \frac{\partial}{\partial\Theta_j}\sum_{i=1}^{n}\log f(X_i|\mathbf{\Theta}) = 0, \qquad j = 1, \ldots, s,$$

such that $\tilde{\mathbf{\Theta}}_n$ converges to $\mathbf{\Theta}_0$ in probability; i.e., $\tilde{\mathbf{\Theta}}_n$ is consistent.

For a proof see Lehmann [37, Sect. 6.4]. The theorem needs several comments for clarification:

(a) If the likelihood function $L(\mathbf{\Theta})$ attains its maximum at an interior point of Ω then the MLE is a solution to the likelihood equation. If in addition the likelihood equations only have one root, then Theorem 3 proves the consistency of the MLE($\hat{\mathbf{\Theta}}_n = \tilde{\mathbf{\Theta}}_n$).

(b) Theorem 3 does not state how to identify the consistent root among possibly many roots of the likelihood equations. One could take the root $\tilde{\mathbf{\Theta}}_n$ which is closest to $\mathbf{\Theta}_0$, but then $\tilde{\mathbf{\Theta}}_n$ is no longer an estimator since its construction assumes knowledge of the unknown value of $\mathbf{\Theta}_0$. This problem may be overcome by taking that root which is closest to a (known) consistent estimator of $\mathbf{\Theta}_0$. The utility of this approach becomes clear in the section on efficiency.

(c) The MLE does not necessarily coincide with the consistent root guaranteed by Theorem 3. Kraft and LeCam [33] give an example in which Cramér's conditions are satisfied, the MLE exists, is unique, and satisfies the likelihood equations, yet is not consistent.

In view of these comments, it is advantageous to establish the uniqueness of the likelihood equation roots whenever possible. For example, if $f(x|\mathbf{\Theta})$ is of nondegenerate multiparameter exponential family type, then $\log L(\mathbf{\Theta})$ is strictly concave. Thus the likelihood equations have at most one solution. Sufficient conditions for the existence of such a solution may be found in Barndorff-Nielsen [4]. In a more general context Mäkeläinen et al. [38] give sufficient conditions for the existence and uniqueness of roots of the likelihood equations.

EFFICIENCY

The main reason for presenting Cramér's consistency theorem and not Wald's is the following theorem which specifically addresses consistent roots of the likelihood equations and not necessarily MLEs.

Theorem 4. Assume **A0, A1**, and **C1–C5**. If $\tilde{\mathbf{\Theta}}_n$ is a consistent sequence of roots of the likelihood equations, then as $n \to \infty$,

$$\sqrt{n}(\mathbf{\Theta}_n - \mathbf{\Theta}_0) \xRightarrow{L} N_s(\mathbf{0}, \mathbf{I}(\mathbf{\Theta}_0)^{-1});$$

i.e., in large samples the distribution of $\tilde{\mathbf{\Theta}}_n$ is approximately s-variate normal with mean $\mathbf{\Theta}_0$ and covariance matrix $\mathbf{I}(\mathbf{\Theta}_0)^{-1}/n$. This theorem is due to Cramér [9], who gave a proof for $s = 1$. A proof for $s \geqslant 1$ may be found in Lehmann [37].

Because of the form, $\mathbf{I}(\mathbf{\Theta})^{-1}$, of the asymptotic covariance matrix for $\sqrt{n}(\tilde{\mathbf{\Theta}}_n - \mathbf{\Theta})$, one generally regards $\tilde{\mathbf{\Theta}}_n$ as an efficient estimator for $\mathbf{\Theta}$. The reasons for this are now discussed. Under regularity conditions (weaker than those of Theorem 4) the Cramér—Rao* lower bound (CRLB) states that

$$\operatorname{var}(T_{jn}) \geqslant (\mathbf{I}(\mathbf{\Theta})^{-1})_{jj}/n$$

for any unbiased estimator T_{jn} of Θ_j which is based on n observations. Here $(\mathbf{I}(\mathbf{\Theta})^{-1})_{jj}$ refers to the jth diagonal element of $\mathbf{I}(\mathbf{\Theta})^{-1}$. Note, however, that the CRLB refers to the actual variance of an (unbiased) estimator and not to the asymptotic variance of such estimator. The relationship between these two variance concepts is clarified by the following inequality. If as $n \to \infty$

$$\sqrt{n}(T_{jn} - \Theta_j) \xrightarrow{L} N(0, \upsilon_j(\mathbf{\Theta})), \qquad (1)$$

then

$$\lim_{n\to\infty} \left[n \operatorname{var}(T_{jn}) \right] \geqslant \upsilon_j(\mathbf{\Theta}),$$

where equality need not hold; see Lehmann [37].

Thus for unbiased estimators T_{jn} which are asymptotically normal, i.e., satisfy (1), and for which

$$\lim_{n\to\infty} [n \operatorname{var}(T_{jn})] = \upsilon_j(\mathbf{\Theta}),$$

the CRLB implies that $\upsilon_j(\mathbf{\Theta}) \geqslant (\mathbf{I}(\mathbf{\Theta})^{-1})_{jj}$. It was therefore thought that $(\mathbf{I}(\mathbf{\Theta})^{-1})_{jj}$ is a lower bound for the asymptotic variance of any asymptotically normal unbiased estimate of Θ_j. Since the estimators $\tilde{\Theta}_{jn}$ of Theorem 4 have asymptotic variances equal to this lower bound, they were called *efficient estimators*. In fact, Lehmann [36] refers to $\tilde{\mathbf{\Theta}}_n$ as an *efficient likelihood estimator* (ELE) in contrast to the MLE, although the two often will coincide. For a discussion of the usage of ELE versus MLE refer to his paper. As it turns out, $(\mathbf{I}(\mathbf{\Theta})^{-1})_{jj}$ will not serve as a true lower bound on the asymptotic variance of asymptotically normal estimators of Θ_j unless one places some restrictions on the behavior of such estimators. Without such restrictions Hodges (see LeCam [35]) was able to construct so called *superefficient estimators* (*see* SUPEREFFICIENCY, HODGES for a simple example). It was shown by LeCam [35] (see also Bahadur [2]) that the set of superefficiency points must have Lebesgue measure zero for any particular sequence of estimators. LeCam (see also Hájek [28]) further showed that falling below the lower bound at a value $\mathbf{\Theta}_0$ entails certain unpleasant properties for the mean squared error* (or other risk functions) of such superefficient estimators in the vicinity of $\mathbf{\Theta}_0$. Thus it appears not advisable to use superefficient estimators.

Unpleasant as such superefficient estimators are, their existence led to a reassessment of large sample properties of estimators. In particular, a case was made to require a certain amount of regularity not only in the distributions but also in the estimators. For example, the simple requirement that the asymptotic variance $\upsilon_j(\mathbf{\Theta})$ in (1) be a continuous function in $\mathbf{\Theta}$ would preclude such estimator from being superefficient since, as remarked above, such phenomenon may occur only on sets of Lebesgue measure zero. For estimators satisfying (1) with continuous $\upsilon_j(\mathbf{\Theta})$ one thus has $\upsilon_j(\mathbf{\Theta}) \geqslant (\mathbf{I}(\mathbf{\Theta})^{-1})_{jj}$. Rao [49] requires the weak convergence in (1) to be uniform on compact subsets of Ω, which under mild assumption on $f(\cdot|\mathbf{\Theta})$ implies the continuity of the asymptotic variance $\upsilon_j(\mathbf{\Theta})$.

Hájek [27] proved a very general theorem which gives a succinct description of the asymptotic distribution of regular estimators. His theorem will be described in a somewhat less general form below. Let $\mathbf{\Theta}(n) = \mathbf{\Theta}_0 + \mathbf{h}/\sqrt{n}, \mathbf{h} \in R^s$ and denote by $\xrightarrow{L_{\mathbf{\Theta}(n)}}$ convergence in law when $\mathbf{\Theta}(n)$ is the true parameter.

Definition 2. An estimator sequence $\{\mathbf{T}_n\}$ is called *regular* if for all $\mathbf{h} \in R^s$

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\Theta}(n)) \xrightarrow{L_{\boldsymbol{\Theta}(n)}} \mathbf{T}$$

as $n \to \infty$, where the distribution of the random vector $\mathbf{T}$ is independent of $\mathbf{h}$.

The regularity conditions on $f(\cdot|\boldsymbol{\Theta})$ are formulated as follows.

Definition 3. $f(\cdot|\boldsymbol{\Theta})$ is called *locally asymptotically normal* (LAN) if for all $\mathbf{h} \in R^s$

$$\begin{aligned}&\log[f_n(\mathbf{X}|\boldsymbol{\Theta}(n))/f_n(\mathbf{X}|\boldsymbol{\Theta}_0)]\\ &\quad = \mathbf{h}'\boldsymbol{\Delta}_n(\boldsymbol{\Theta}_0) - \tfrac{1}{2}\mathbf{h}'\mathbf{I}(\boldsymbol{\Theta}_0)\mathbf{h} + Z_n(\mathbf{h}, \boldsymbol{\Theta}_0)\end{aligned}$$

with

$$\boldsymbol{\Delta}_n(\boldsymbol{\Theta}_0) \xrightarrow{L_{\boldsymbol{\Theta}_0}} N_s(\mathbf{0}, \mathbf{I}(\boldsymbol{\Theta}_0))$$

and

$$Z_n(\mathbf{h}, \boldsymbol{\Theta}_0) \xrightarrow{P_{\boldsymbol{\Theta}_0}} 0 \qquad \text{as} \quad n \to \infty.$$

Comment. Under the conditions **A0, A1**, and **C1–C5**, one may show that $f(\cdot|\boldsymbol{\Theta})$ is LAN and in that case

$$\begin{aligned}\boldsymbol{\Delta}'{}_n(\boldsymbol{\Theta}_0) = \Bigg(&\frac{\partial}{\partial\Theta_1}\log f_n(\mathbf{X}|\boldsymbol{\Theta}), \ldots,\\ &\times\frac{\partial}{\partial\Theta_s}\log f_n(\mathbf{X}|\boldsymbol{\Theta})\Bigg)\Big/\sqrt{n}.\end{aligned}$$

Theorem 5. (Hájek). If $\{\mathbf{T}_n\}$ is regular and $f(\cdot|\boldsymbol{\Theta})$ is LAN then $\mathbf{T} = \mathbf{Y} + \mathbf{W}$, where $\mathbf{Y} \sim N_s(\mathbf{0}, \mathbf{I}(\boldsymbol{\Theta}_0)^{-1})$ and $\mathbf{W}$ is a random vector independent of $\mathbf{Y}$. The distribution of $\mathbf{W}$ is determined by the estimator sequence.

Comment. Estimators for which $P_{\boldsymbol{\Theta}_0}(\mathbf{W} = \mathbf{0}) = 1$ are most concentrated around $\boldsymbol{\Theta}_0$ (see Hájek [27]) and may thus be considered efficient among regular estimators. Note that this optimality claim is possible because competing estimators are required to be regular; on the other hand, it is no longer required that the asymptotic distribution of the estimator be normal.

As remarked in the comments to Theorem 3 the ELE $\tilde{\boldsymbol{\Theta}}_n$ of Theorem 4 may be chosen by taking that root of the likelihood equations which is closest to some known consistent estimate of $\boldsymbol{\Theta}$. The latter estimate need not be efficient. In this context we note that the consistent sequence of roots generated by Theorem 3 is essentially unique. For a more precise statement of this result see Huzurbazar [31] and Perlman [43]. Roots of the likelihood equations may be found by one of various iterative procedures offered in several statistical computer packages; *see* ITERATED MAXIMUM LIKELIHOOD ESTIMATES. Alternatively on may just take a one-step iteration estimator in place of a root. Such one step estimators use as starting point $\sqrt{n}$-consistent estimators (not necessarily efficient) and are efficient. A precise statement is given in Theorem 6. First note the following definition.

Definition 4. An estimator $\mathbf{T}_n$ is $\sqrt{n}$-*consistent* for estimating the true $\boldsymbol{\Theta}_0$ if for every $\epsilon > 0$ there is a K_ϵ and an N_ϵ such that

$$P_{\boldsymbol{\Theta}_0}(\| \sqrt{n}(\mathbf{T}_n - \boldsymbol{\Theta}_0) \| \leqslant K_\epsilon) \geqslant 1 - \epsilon$$

for all $n \geqslant N_\epsilon$, where $\| \cdot \|$ denotes the Euclidean norm in R^s.

Theorem 6. Suppose the assumptions of Theorem 4 hold and that $\boldsymbol{\Theta}_n^*$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\Theta}$. Let $\boldsymbol{\delta}_n' = (\delta_{1n}, \ldots, \delta_{sn})$ be the solution of the linear equations

$$\sum_{k=1}^{s}(\delta_{kn} - \Theta_{kn}^*)R_{jk}''(\boldsymbol{\Theta}_n^*) = -R_j'(\boldsymbol{\Theta}_n^*),$$

$j = 1, \ldots, s$, where

$$R_j'(\boldsymbol{\Theta}) = \frac{\partial}{\partial\Theta_j}\log L(\boldsymbol{\Theta}),$$

and

$$R_{jk}''(\boldsymbol{\Theta}) = \frac{\partial^2}{\partial\Theta_j\partial\Theta_k}\log L(\boldsymbol{\Theta}).$$

Then

$$\sqrt{n}(\boldsymbol{\delta}_n - \boldsymbol{\Theta}_0) \xrightarrow{L} N_s(\mathbf{0}, \mathbf{I}(\boldsymbol{\Theta}_0)^{-1})$$

as $n \to \infty$, i.e., $\boldsymbol{\delta}_n$ is asymptotically efficient.

For a proof of Theorem 6 see Lehmann [37].

The application of Theorem 6 is particularly useful in that it does not require the solution of the likelihood equations. The two estimators $\tilde{\boldsymbol{\Theta}}_n$ (Theorem 4) and $\boldsymbol{\delta}_n$ (Theorem 6) are asymptotically efficient. Among other estimators that share this property are the Bayes estimators. Because of this multitude of asymptotically efficient estimators one has tried to discriminate between them by considering higher order terms in the asymptotic analysis. Several measures of second-order efficiency* have been examined (*see* EFFICIENCY, SECOND-ORDER) and it appears that with some qualifications MLEs are second-order efficient, provided the MLE is first corrected for bias of order $1/n$. Without such bias correction one seems to be faced with a problem similar to the nonexistence of estimators with uniformly smallest mean squared error in the finite sample case (*see* ESTIMATION, CLASSICAL). For investigations along these lines see Rao [48], Ghosh and Subramanyam [23], Efron [13], Pfanzagl and Wefelmeyer [44] and Ghosh and Sinha [22]. For a lively discussion of the issues involved see also Berkson [8].

Other types of optimality results for MLEs, such as the local asymptotic admissibility* and minimax* property, were developed by LeCam and Hájek. For an exposition of this work and some perspective on the work of others, see Hájek [28]. For a simplified introduction to these results, see also Lehmann [37].

The following example (see Lehmann [37]) illustrates a different kind of behavior of the MLE when the regularity condition **C1** is not satisfied.

Example 4. Let $X_1,\ldots,X_n$ be i.i.d. $\sim U(0, \Theta)$ (uniform on $(0,\Theta)$). Then the MLE is $\hat{\Theta}_n = \max(X_1,\ldots,X_n)$ and its large sample behavior is described by

$$n(\Theta - \hat{\Theta}_n) \xrightarrow{L} \Theta \cdot E$$

as $n \to \infty$, where E is an exponential random variable with mean 1. Note that the normalizing factor is n and not $\sqrt{n}$. Also the asymptotic distribution ΘE is not normal and is not centered at zero. Considering instead $\delta_n = (n+1)\hat{\Theta}_n/n$ one finds as $n \to \infty$

$$n(\Theta - \delta_n) \xrightarrow{L} \Theta(E - 1).$$

Further

$$E[n(\hat{\Theta}_n - \Theta)]^2 \to 2\Theta^2$$

and

$$E[n(\delta_n - \Theta)]^2 \to \Theta^2.$$

Hence the MLE, although consistent, may no longer be asymptotically optimal.

For a different approach which covers regular as well as nonregular problems, see Weiss and Wolfowitz [57] on maximum probability estimators*.

MLES IN MORE GENERAL MODELS

The results concerning MLEs or ELEs discussed so far all assumed **A0**. In many statistical applications the sampling structure gives rise to independent observations which are not identically distributed. For example one may be sampling several different populations or with each observation one or more covariates may be recorded. For the former situation Theorems 3 and 4 are easily extended, see Lehmann [37]. In the case of independent but not identically distributed observations, results along the lines of Theorem 3 and 4 were given by Hoadley [29] and Nordberg [40]. Nordberg deals specifically with exponential family* models presenting the binary logit model and the log-linear Poisson model as examples. See also Haberman [26], who treats maximum likelihood theory for diverse parametric structures in multinomial and Poisson counting data models.

The maximum likelihood theory for incomplete data* from an exponential family is treated in Sundberg [55]. Incomplete data models include situations with grouped*, censored*, or missing data and finite mixtures. See Sundberg [55] and Dempster et al. [10] for more examples. The latter authors present the EM algorithm for iterative computation of the MLE from incomplete data.

Relaxing the independence assumption of the observations opens the way to stochastic process applications. Statistical inference problems concerning stochastic processes* only recently have been treated with vigor. The maximum likelihood approach to parameter estimation here plays a prominent role. Consistency and asymptotic normality* of MLEs (or ELEs) may again be established by using appropriate martingale* limit theorems. Care has to be taken to define the concept of likelihood function when the observations consist of a continuous time stochastic process. As a start into the growing literature on this subject see Feigin [16], Basawa and Prakasa Rao [5,6].

Another assumption made so far is that the parameter space Ω has dimension s, where s remains fixed as the sample size n increases. For the problems one encounters as s grows with n or when Ω is so rich that it cannot be embedded into any finite dimensional Euclidean space, see Kiefer and Wolfowitz [32] and Grenander [25] respectively. Huber [30] examines the behavior of MLEs derived from a particular parametric model when the sampled distribution is different from this parametric model. These results are useful in studying the robustness properties of MLEs (*see* ROBUST ESTIMATION).

MISCELLANEOUS REMARKS

Not much can be said about the small sample properties of MLEs. If the MLE exists it is generally a function of the minimal sufficient statistic*, namely the likelihood function $L(\cdot)$. However, the MLE by itself is not necessarily sufficient. Thus in small samples some information may be lost by considering the MLE by itself. Fisher [19] proposed the use of ancillary statistics* to recover this information loss, by viewing the distribution of the MLE conditional on the ancillary statistics. For a recent debate on this subject, see Efron and Hinkley [15], who make an argument for using the observed and not the expected Fisher information* in assessing the accuracy of MLEs. Note, however, the comments to their paper. See also Sprott [53,54], who suggests using parameter transformations to achieve better results in small samples when appealing to large sample maximum likelihood theory. Efron [14], in discussing the relationship between maximum likelihood and decision theory*, highlights the role of maximum likelihood as a summary principle in contrast to its role as an estimation principle.

Although MLEs are asymptotically unbiased in the regular case, this will not generally be the case in finite samples. It is not necessarily clear whether the removal of bias from an MLE will result in a better estimator, as the following example shows (*see also* UNBIASEDNESS).

Example 5. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$; then the MLE of σ^2 is

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n,$$

which has mean value $\sigma^2(n-1)/n$, i.e., $\hat{\sigma}^2$ is biased. Taking instead $\tilde{\sigma}^2 = \hat{\sigma}^2 n/(n-1)$ we have an unbiased estimator of σ^2 which has uniformly smallest variance among all unbiased estimators; however,

$$E(\hat{\sigma}^2 - \sigma^2)^2 < E(\tilde{\sigma}^2 - \sigma^2)^2$$

for all σ^2.

The question of bias removal should therefore be decided on a case by case basis with consideration of the estimation objective. See, however, the argument for bias correction of order $1/n$ in the context of second-order efficiency of MLEs as given in Rao [48] and Ghosh and Subramanyam [23].

Gong and Samaniego [24] consider the concepts of pseudo maximum likelihood estimation which consists of replacing all nuisance parameters* in a multiparameter model by suitable estimates and then solving a reduced system of likelihood equations for the remaining structural parameters. They present consistency and asymptotic normality results and illustrate them by example.

AN EXAMPLE

As an illustration of some of the issues and problems encountered, consider the Weibull* model. Aside from offering much flexibility in its shape, there are theoretical extreme value type arguments (Galambos [21]), recommending the Weibull distribution as an appropriate model for the breaking strength of certain materials.

Let $X_1, \ldots, X_n$ be i.i.d. $W(t, \alpha, \beta)$, the Weibull distribution with density

$$f(x|t,\alpha,\beta) = \frac{\beta}{\alpha}\left(\frac{x-t}{\alpha}\right)^{\beta-1} \exp\left(-(\frac{x-t}{\alpha})^{\beta}\right),$$
$$x > t, \quad t \in R, \quad \alpha, \beta > 0.$$

When the threshold parameter t is known, say $t = 0$ (otherwise subtract t from X_i) the likelihood equations for the scale and shape parameters α and β have a unique solution. In fact, the likelihood equations may be rewritten

$$\hat{\alpha} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{\beta}} \right)^{1/\hat{\beta}}, \qquad (2)$$

$$\sum_{i=1}^{n} x_i^{\hat{\beta}} \log x_i \left(\sum_{i=1}^{n} x_i^{\hat{\beta}} \right)^{-1} - \hat{\beta} - \frac{1}{n} \sum_{i=1}^{n} \log x_i = 0, \qquad (3)$$

i.e., $\hat{\alpha}$ is given explicitly in terms of $\hat{\beta}$, which in turn can be found from the second equation by an iterative numerical procedure such as Newton's method. The regularity conditions of Theorems 3 and 4 are satisfied. Thus we conclude that

$$\sqrt{n} \left(\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right) \xrightarrow{L} N_2(\mathbf{0}, \mathbf{I}^{-1}(\alpha, \beta)),$$

where

$$\mathbf{I}^{-1}(\alpha, \beta) = \begin{bmatrix} 1.109 \left(\frac{\alpha}{\beta} \right)^2 & .257\alpha \\ .257\alpha & .608\beta^2 \end{bmatrix}.$$

From these asymptotic results it follows that, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta} - \beta)/\beta \xrightarrow{L} N(0, .608),$$
$$\sqrt{n}\hat{\beta} \log(\hat{\alpha}/\alpha) \xrightarrow{L} N(0, 1.109),$$

from which large sample confidence intervals* for β and α may be obtained. However, the large sample approximations are good only for very large samples. Even for $n = 100$ the approximations leave much to be desired. For small to medium sample sizes a large collection of tables is available (see Bain [1]) to facilitate various types of inference. These tables are based on extensive Monte Carlo* investigations. For example, instead of appealing to the asymptotic $N(0, .608)$ distribution of the pivot $\sqrt{n}(\hat{\beta} - \beta)/\beta$, the distribution of this pivot was simulated for various samples sizes and the percentage points of the simulated distributions were tabulated. Another approach to finite sample size inference is offered by Lawless [34]. His method is based on the conditional distribution of the MLEs given certain ancillary statistics. It turns out that this conditional distribution is analytically manageable; however, computer programs are ultimately required for the implementation of this method.

Returning to the three-parameter Weibull problem, so that the threshold is also unknown, we find that the likelihood function tends to infinity as $t \to T = \min(X_1, \ldots, X_n)$ and $\beta < 1$, i.e. the MLEs of t, α, and β do not exist. It has been suggested that the parameter β be restricted a priori to $\beta \geqslant 1$ so that the likelihood function remains bounded. In that case the MLEs will always exist, but with positive probability will not be a solution to the likelihood equations. It is not clear what the large sample properties of the MLEs are in this case.

Appealing to Theorems 3 and 4 one may attempt to find efficient roots of the likelihood equations or appeal to Theorem 6, since $\sqrt{n}$-consistent estimates are easily found (e.g., method of moments, method of quantiles*). However, the stated regularity conditions, notably **C1**, are not satisfied. In addition to that the likelihood equations will, with positive probability, have no solution at all and if they have a solution they have at least two, one yielding a saddle point and one yielding a local maximum of the likelihood function (Rockette et al. [50]). A further problem in the three-parameter Weibull model concerns identifiability for large shape parameters β. Namely, if $X \sim W(t, \alpha, \beta)$ then uniformly in α and t, as $\beta \to \infty$

$$\frac{\beta}{\alpha}(X - t - \alpha) \xrightarrow{L} E_0$$

where E_0 is a standard extreme value* random variable with distribution function $F(y) = 1 - \exp(-\exp(y))$. Hence for large β,

$$X \overset{L}{\simeq} t + \alpha + \frac{\alpha}{\beta} E_0 = u + bE_0$$

and it is clear that (t, α, β) cannot be recovered from u and b. This phenomenon is similar to the one experienced for the generalized gamma distribution*; see Prentice [46]. In the Weibull problem the identifiability problem may be remedied by proper reparametrization. For example, instead of (t, α, β) one can easily use three quantiles. However, because of the one-to-one relationship between these two sets of parameters the abovementioned problems concerning maximum likelihood and likelihood equation estimators still persist.

It is conceivable that the conditions of Theorem 6 may be weakened to accommodate the three-parameter Weibull distribution. Alternatively one may use the approach of Gong and Samaniego [24] described above, in that one estimates the threshold t by other means. Mann and Fertig [39] offer one such estimate; *see* WEIBULL DISTRIBUTION, MANN–FERTIG TEST STATISTIC FOR. Neither of these two approaches seems to have been explored rigorously so far. For an extensive account of maximum likelihood estimation as well as other methods for complete and censored samples from a Weibull distribution see Bain [1] and Lawless [34]. Both authors treat maximum likelihood estimation in the context of many other models.

For the very reason that applications employing the method of maximum likelihood are so numerous no attempt is made here to list them beyond the few references and examples given above. For a guide to the literature see the survey article on MLEs by Norden [41]. Also see Lehmann [37] for a rich selection of interesting examples and for a more thorough treatment.

Acknowledgments

In writing this article I benefited greatly from pertinent chapters of Erich Lehmann's book *Theory of Point Estimation* [37]. I would like to thank him for this privilege and for his numerous helpful comments.

I would also like to thank Jon Wellner for the use of his notes on maximum likelihood estimation and Michael Perlman and Paul Sampson for helpful comments. This work was supported in part by The Boeing Company.

REFERENCES

1. Bain, L. J. (1978). *Statistical Analysis of Reliability and Life-Testing Models*. Dekker, New York.
2. Bahadur, R. R. (1964). *Ann. Math. Statist.*, **35**, 1545–1552.
3. Bahadur, R. R. (1971). *Some Limit Theorems in Statistics*, SIAM, Philadelphia.
4. Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.
5. Basawa, I. V. and Prakasa Rao, B. L. S. (1980). *Stoch. Proc. Appl.*, **10**, 221–254.
6. Basawa, I. V. and Prakasa Rao, B. L. S. (1980). *Statistical Inference for Stochastic Processes*. Academic Press, London.
7. Berk, R. H. (1967). *Math. Rev.*, **33**, No. 1922.
8. Berkson, J. (1980). *Ann. Statist.*, **8**, 457–469.
9. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, N.J.
10. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). *J. R. Statist. Soc. B*, **39**, 1–22.
11. Edgeworth, F. Y. (1908/09). *J. R. Statist. Soc.*, **71**, 381–397, 499–512, and *J. R. Statist. Soc.*, **72**, 81–90.
12. Edwards, A. W. F. (1974). *Internat. Statist. Rev.*, **42**, 4–15.
13. Efron, B. (1975). *Ann. Statist.*, **3**, 1189–1242.
14. Efron, B. (1982). *Ann. Statist.*, **10**, 340–356.
15. Efron, B. and Hinkley, D. V. (1978). *Biometrika*, **65**, 457–487.
16. Feigin, P. D. (1976). *Adv. Appl. Prob.*, **8**, 712–736.
17. Fisher, R. A. (1912). *Messenger of Mathematics*, **41**, 155–160.
18. Fisher, R. A. (1922). *Philos. Trans. R. Soc. London A*, **222**, 309–368.
19. Fisher, R. A. (1925). *Proc. Camb. Phil. Soc.*, **22**, 700–725.
20. Fisher, R. A. (1935). *J. R. Statist. Soc.*, **98**, 39–54.
21. Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*. Wiley, New York.
22. Ghosh, J. K. and Sinha, B. K. (1981). *Ann. Statist.*, **9**, 1334–1338.
23. Ghosh, J. K. and Subramanyam, K. (1974). *Sankhya (A)*, **36**, 325–358.
24. Gong, G. and Samaniego, F. J. (1981). *Ann. Statist.*, **9**, 861–869.
25. Grenander, U. V. (1980). *Abstract Inference*. Wiley, New York.
26. Haberman, S. J. (1974). *The Analysis of Frequency Data*. The University of Chicago Press, Chicago.
27. Hájek, J. (1970). *Zeit. Wahrscheinlichkeitsth. Verw. Geb.*, **14**, 323–330.

28. Hájek, J. (1972). *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, **1**, 175–194.

29. Hoadley, B. (1971). *Ann. Math. Statist.*, **42**, 1977–1991.

30. Huber, P. J. (1967). *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, **1**, 221–233.

31. Huzurbazar, V. S. (1948). *Ann. Eugen.*, **14**, 185–200.

32. Kiefer, J. and Wolfowitz, J. (1956). *Ann. Math. Statist.*, **27**, 887–906.

33. Kraft, C. and LeCam, L. (1956). *Ann. Math. Statist.*, **27**, 1174–1177.

34. Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

35. LeCam, L. (1953). *Univ. of Calif. Publ. Statist.*, **1**, 277–330.

36. Lehmann, E. L. (1980). *Amer. Statist.*, **34**, 233–235.

37. Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York, Chaps. 5 and 6.

38. Mäkeläinen, T., Schmidt, K., and Styan, G. (1981). *Ann. Statist.*, **9**, 758–767.

39. Mann, N. R. and Fertig, K. W. (1975). *Technometrics*, **17**, 237–245.

40. Nordberg, L. (1980). *Scand. J. Statist.*, **7**, 27–32.

41. Norden, R. H. (1972/73). *Internat. Statist. Rev.*, **40**, 329–354; **41**, 39–58.

42. Perlman, M. (1972). *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, **1**, 263–281.

43. Perlman, M. D. (1983). In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his 60th Birthday*, M. H. Rizvi, J. S. Rustagi, and D. Siegmund, eds. Academic Press, New York, pp. 339–370.

44. Pfanzagl, J. and Wefelmeyer, W. (1978/79). *J. Multivariate Anal.*, **8**, 1–29; **9**, 179–182.

45. Pratt, J. W. (1976), *Ann. Statist.*, **4**, 501–514.

46. Prentice, R. L. (1973), *Biometrika*, **60**, 279–288.

47. Rao, C. R. (1957). *Sankhya*, **18**, 139–148.

48. Rao, C. R. (1961). *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **1**, 531–546.

49. Rao, C. R. (1963). *Sankhya*, **25**, 189–206.

50. Rockette, H., Antle, C., and Klimko, L. (1974). *J. Amer. Statist. Ass.*, **69**, 246–249.

51. Savage, L. J. (1976). *Ann. Statist.*, **4**, 441–500.

52. Scholz, F. -W. (1980). *Canad. J. Statist.*, **8**, 193–203.

53. Sprott, D. A. (1973). *Biometrika*, **60**, 457–465.

54. Sprott, D. A. (1980). *Biometrika*, **67**, 515–523.

55. Sundberg, R. (1974). *Scand. J. Statist.*, **1**, 49–58.

56. Wald, A. (1949). *Ann. Math. Statist.* **20**, 595–601.

57. Weiss, L. and Wolfowitz, J. (1974). *Maximum Probability Estimators and Related Topics*. Springer-Verlag, New York. (Lect. Notes in Math., No. 424.)

58. Zehna, P. W. (1966). *Ann. Math. Statist.*, **37**, 744.

BIBLIOGRAPHY

Akahira, M. and Takeuchi, K. (1981). *Asymptotic Efficiency of Statistical Estimators: Concepts of Higher Order Asymptotic Efficiency*. Springer-Verlag, New York. Lecture Notes in Statistics 7. (Technical Monograph on higher order efficiency with an approach different from the references cited in the text.)

Barndorff-Nielsen, O. (1983). *Biometrika*, **70**, 343–365. (Discusses a simple approximation formula for the conditional density of the maximum likelihood estimator given a maximal ancillary statistic. The formula is generally accurate (in relative error) to order $O(n^{-1})$ or even $O(n^{-3/2})$, and for many important models, including arbitrary transformation models, it is in fact, exact. The level of the paper is quite mathematical. With its many references it should serve as a good entry point into an area of research of much current interest although its roots date back to R. A. Fisher.)

Fienberg, S. E. and Hinkley, D. V. (eds.) (1980). *R. A. Fisher: An Appreciation*. Springer-Verlag, New York. (Lecture Notes in Statistics 1. A collection of articles by different authors highlighting Fisher's contributions in statistics.)

Ibragimov, I. A. and Has'minskii (1981). *Statistical Estimation, Asymptotic Theory*. Springer-Verlag, New York. (Technical monograph on the asymptotic behavior of estimators (MLEs and Bayes estimators) for regular as well as irregular problems, i.i.d. and non-i.i.d. cases.)

LeCam, L. (1970). *Ann. Math. Statist.*, **41**, 802–828. (Highly mathematical, weakens Cramér's third-order differentiability conditions to first-order differentiability in quadratic mean.)

LeCam, L. (1979). *Maximum Likelihood, an Introduction*. Lecture Notes No. 18, Statistics Branch, Department of Mathematics, University of Maryland. (A readable and humorous account of the pitfalls of MLEs illustrated by numerous examples.)

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, London. (This monograph deals with a class of statistical models that generalizes classical linear models in two ways: (i) the response variable may be of exponential family type (not just normal), and (ii) a monotonic smooth transform of the mean response is a linear function in the predictor variables. The parameters are estimated by the method of maximum likelihood through an iterated weight least-squares algorithm. The monograph emphasizes applications over theoretical concerns and represents a rich illustration of the versatility of maximum likelihood methods.)

Rao, C. R. (1962). *Sankhya A*, **24**, 73–101. (A readable survey and discussion of problems encountered in maximum likelihood estimation.)

See also Ancillary Statistics—I; Asymptotic Normality; Cramér–Rao Lower Bound; Efficiency, Second-Order; Efficient Score; Estimating Equations, Theory of; Estimating Functions; Estimation, Classical; Estimation: Method of Sieves; Fisher Information; Full-Information Estimators; Generalized Maximum Likelihood Estimation; Inference, Statistical; Iterated Maximum Likelihood Estimates; Large-Sample Theory; Likelihood; Likelihood Principle; Likelihood Ratio Tests; *M*-Estimators; Minimum Chi-Square; Maximum Penalized Likelihood Estimation; Partial Likelihood; Pseudo-Likelihood; Restricted Maximum Likelihood (REML); Superefficiency, Hodges; and Unbiasedness.

F. W. Scholz