Int. Stat. Rev., Vol. 40, No. 3, 1972, pp. 329–354/*Longman Group Ltd/Printed in Great Britain*

A Survey of Maximum Likelihood Estimation

R. H. Norden
University of Bath, England

Contents

(* Part 2 following in next issue of the *Review*)

Abstract

This survey, which is in two parts, is expository in nature and gives an account of the development of the theory of Maximum Likelihood Estimation (MLE) since its introduction in the papers of Fisher (1922, 1925) up to the present day where original work in this field still continues. After a short introduction there follows a historical account in which in particular it is shown that Fisher may indeed be rightfully regarded as the originator of this method of estimation since all previous apparently similar methods depended on the method of inverse probability.

There then follows a summary of a number of fundamental definitions and results which originate from the papers of Fisher himself and the later contributions of Cramer and Rao. Consistency and Efficiency in relation to maximum likelihood estimators (MLE's) are then considered in detail and an account of Cramer's important (1946) theorem is given, though it is remarked that, with regard to consistency. Wald's (1949) proof has greater generality. Comments are then made on a number of subsequent related contributions.

The paper then continues with a discussion of the problem of comparing MLE's with other estimators. Special mention is then made of Rao's measure of "second-order efficiency" since this does provide a means of discriminating between the various best asymptotically Normal (BAN) estimators of which maximum likelihood estimators form a subclass. Subsequently there is a survey of work in the multiparameter field which begins with a brief account of the classical Cramer-Rao theory.

Finally there is a section on the theoretical difficulties which arise by reason of the existence of inconsistent maximum likelihood estimates and of estimators which are super efficient. Particular reference is made to Rao's (1962) paper in which these anomalies were considered and at least partially resolved by means of alternative criteria of consistency and efficiency which nevertheless are rooted in Fisher's original ideas.

1. Introduction

The aim of this paper and its sequel is to give an account of the theoretical aspect of Maximum Likelihood Estimation. To this end we shall consider a number of contributions in detail in which the main themes are consistency and efficiency, both with regard to single and multiparameter situations.

The scope of these papers will also include a discussion of various theoretical difficulties with special reference to examples of inconsistency of maximum likelihood estimators and superefficiency of other estimators. The related question of to what extent Maximum Likelihood Estimation has an advantage over other methods of estimation will also be considered.

An extensive bibliography of nearly 400 titles divided into four main groups is given at the end of the first paper, in which it is hoped that all, or at least almost all, contributions which are in some way related to this subject are included. In the text we refer to papers mainly in the first two groups.

2. Historical Note

In two renowned papers R. A. Fisher (1921, 1925) introduced into statistical theory the concepts of consistency, efficiency and sufficiency, and he also advocated the use of Maximum Likelihood as an estimation procedure.

His 1921 definition of consistency, although important from the theoretical standpoint, cannot easily be applied to actual situations, and he gave a further definition in his 1925 paper.

"A statistic is said to be a consistent estimate of any parameter if when calculated from an indefinitely large sample it tends to be accurately equal to that parameter."

This is the definition of consistency which is now in general use, and mathematically it may be stated thus.

The estimator T_n of a paremeter θ, based on a sample of n is consistent for θ if for an arbitrary $\varepsilon>0$,

$$P[|T_n-\theta|>\varepsilon]\to 0 \text{ as } n\to\infty.$$

We shall refer to Fisher's earlier definition in Section 9, however, where we consider examples of inconsistency of MLE's.

With regard to efficiency, he says: "The criterion of efficiency is satisfied by those statistics which when derived from large samples tend to a normal distribution with the least possible standard deviation".

Later, 1925, he says: "We shall prove that when an efficient statistic exists, it may be found by the method of Maximum Likelihood Estimation".

Evidently, therefore, he regarded efficiency as an essentially asymptotic property. Nowadays we would say that an estimator is efficient, whatever the sample size, if its variance attains the Cramer-Rao lower bound, and Fisher's property would be described as asymptotic efficiency.

His definition of sufficiency is as follows: "A statistic satisfies the criterion of sufficiency when no other sample provides any additional information to the value of the parameter to be estimated".

It is well known, of course, that there is a close relation between sufficient statistics and Maximum Likelihood Estimation.

Further, he defines the likelihood as follows: "The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed".

The method of Maximum Likelihood as expounded by Fisher consists of regarding the likelihood as a function of the unknown parameter and obtaining the value of this parameter which makes the likelihood greatest. Such a value is then said to be the Maximum Likelihood Estimate of the parameter.

A formal definition of the method would now therefore be as follows.

Let X be a random variable with probability density function $f = f(x; \theta)$, where the form of f is known but the parameter θ, which may be a vector, is unknown. Suppose $x_1, \ldots, x_n$ are the realized values of X in a sample of n independent observations.

Define the Likelihood Function (LF) by

$$L = f(x_1; \theta) \ldots f(x_n; \theta).$$

Then if $\hat{\theta}$ is such that

$$\sup_{\theta \in \Theta} L = L(\hat{\theta})$$

where Θ is the parameter space, i.e. the set of possible values of θ, then we say that $\hat{\theta}$ is the MLE[1] of θ. If there is no unique MLE then $\hat{\theta}$ will be understood to be any value of θ at which L attains its supremum.

The question as to whether Fisher may be rightly regarded as the originator of the method of MLE has arisen from time to time, so that although most statisticians would now attribute the method to him there does not appear to have been absolute agreement on this matter.

There is, however, a full discussion of the historical aspect by C. R. Rao (1962), where he surveys the work of F. Y. Edgeworth (1908*a* and *b*), and also of Gauss, Laplace and Pearson in so far as they are relevant to this question.

According to Rao, all arguments, prior to those of Fisher, which appeared to be of a ML type, were in fact dependent on the method of inverse probability, i.e. the unknown parameter is estimated by maximizing the "a posterior" probability derived from Bayes law. They could not therefore be regarded as MLE as such, for as Rao remarks, "If $\hat{\theta}$ is a maximum likelihood estimate of θ, then $\phi(\hat{\theta})$ is a maximum likelihood estimate of any one-to-one function of $\phi(\theta)$, while such a property is not true of estimates obtained by the inverse probability argument".

After further discussion he concludes by saying: "We, therefore do not have any literature supporting prior claims to the method of maximum likelihood estimation as a principle capable of wide application, and justifying its use on reasonable criteria (such as sufficiency in a sense wider than that used by Edgeworth and consistency) and not on inverse probability argument, before the fundamental contributions by Fisher in 1922 and 1925".

This agrees with the view put forward by Le Cam (1953), who also gives an outline account of developments of MLE since 1925, particularly with regard to the various attempts that were made to prove rigorously that MLE's are both consistent and efficient. Fisher himself, of course, provided a proof of efficiency, but there was no explicit statement of the restrictions on the type of estimation situation to which his conclusions would apply. He also gave no separate proof of consistency, although this could be regarded as implied by his proof at efficiency.

Subsequently these matters were considered by a number of writers. Hotelling (1931) attempted to prove the consistency and asymptotic normality of MLE's and his work was extended by J. L. Doob (1934). The question of efficiency was considered separately by D. Dugué (1936–1937), in a series of notes, but according to Le Cam all these papers are in some degree erroneous, even though they contain a number of important ideas. He does say, however, that a paper by S. S. Wilks (1938), does contain "the essential elements of a proof" that within a certain class of estimates MLE's are efficient.

A major advance was made by H. Cramer (1946), who provided a proof of the consistency and asymptotic normality of MLE's under certain regularity conditions. The consistency aspect of his proof is not, however, complete, for he only established that some root of the likelihood equation, and not necessarily the absolute maximum of the likelihood function, L,

[1] In the sequel we shall now use the abbreviation ML for Maximum Likelihood and MLE for Maximum Likelihood Estimate, Estimator or Estimation, etc., in accordance with the context.

is consistent. This result was improved by Hurzabazaar (1948), but it was not until A. Wald's (1949) paper appeared on the scene that a rigorous proof of the consistency of the MLE of a parameter within a wide class of estimation situations existed.

By this time there could, of course, be no question of MLE's being generally consistent. In 1948, J. Neyman and E. L. Scott showed that in the general case, where the number of parameters increased with the number of observations, the MLE's would be inconsistent.

The concept of efficiency suffered a similar fate, for in 1951 J. L. Hodges constructed an estimator, not the MLE, which was asymptotically normal and yet for some values of the unknown parameter had asymptotically a variance that was definitely less than that of the MLE. Thus efficiency, in Fisher's sense, could not be regarded as a property that would apply to MLE's generally, but only to a limited class of estimators.

By the end of the 1950s, examples of consistency and "super efficiency", many of them highly ingenious, were fairly numerous, and it might now be supposed that MLE cannot be accorded any special importance relative to other methods of estimation.

This, however, is an oversimplification, for many of these examples are of a very theoretical type and do not have much practical value. In any case, a reformulation of the concepts of consistency and efficiency in operational rather than mathematical terms, by Rao (1961), showed that MLE's could still be regarded as having optimal asymptotic properties within a wide class of estimators, and to this extent, at least, Fisher's original assertions have been realized.

3. Some Fundamental Properties

In this section we summarize a number of well-known results of fundamental importance.

We have already given a formal definition of MLE, and we see that it is very general and implies neither the differentiability of L nor even the uniqueness of $\hat{\theta}$.

In the single parameter case, where Θ is some subset of the set of real numbers, L is a function of a single variable θ. Clearly, if $\hat{\theta}$ exists and is not a terminal value and $\frac{\partial L}{\partial \theta}$ exists for all $\theta \in \Theta$, then $\hat{\theta}$ will be a root of $\frac{\partial L}{\partial \theta} = 0$.

In many simple situations, L is unimodal and $\hat{\theta}$ is the unique root of $\frac{\partial L}{\partial \theta} = 0$, whose solution is usually straightforward. However, $l = \log L$ is in general a simpler function to work with, and as the maxima of l and L occur at the same values of θ, it is usual to consider the equation

$$\frac{\partial l}{\partial \theta} = 0$$

which is referred to as the Likelihood Equation (LE). This is perhaps unfortunate since in some instances no root of the LE is the MLE.

As an illustration, consider the estimation of θ in the type III distribution

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

where λ is a known positive constant.

For a sample of n independent observations $x_1, x_2, \ldots, x_n$, it is easily shown that

$$l = n\lambda \log \theta - n \log \Gamma(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log x_i - \theta \sum_{i=1}^{n} x_i$$

which is a continuous differentiable function of θ for all $\theta > 0$.

Clearly l has an absolute maximum at $\theta = \frac{\lambda}{\bar{x}}$ which is therefore $\hat{\theta}$, since

$$\frac{\partial l}{\partial \theta} = \frac{n\lambda}{\theta} - \sum_{i=1}^{n} x_i$$

is zero at this value of θ, and nowhere else, and $\frac{\partial^2 l}{\partial \theta^2} < 0$ for all $\theta > 0$.

Thus $\hat{\theta}$ is unique in this case, and there is no problem of the LE having multiple roots.

It is also not difficult to establish that, in this case, $\hat{\theta}$ is not unbiased. There are in fact many instances of biased MLE's, but this is not, of course, a serious objection to the method, provided, as is usual, the bias tends to zero as n tends to infinity.

This is in fact the case here, for it is not difficult to show that

$$E(\hat{\theta}) - \theta = \frac{\theta}{n\lambda - 1} = 0\left(\frac{1}{n}\right).$$

It is in any case, often possible to correct for bias in small samples.[1]

Difficulties will, however, very often arise when L has more than one root, or is not differentiable everywhere, or when $\hat{\theta}$ is a terminal value. In the multiparameter case the situation may become extremely involved and it may not be possible to obtain the absolute maximum of L merely by setting the partial derivation of L to zero. We might for example get a "saddle" point by this method[2] so that the solution of the LE would not in fact be the MLE in accordance with the definition.

Even if this procedure does, in principle, give us $\hat{\theta}$, we may still be confronted with serious computational difficulties. The LE's can be extremely complicated, and a closed-form solution will be often out of the question. Some form of iterative procedure will then be necessary, and the computer time aspect may then be a limiting factor. We might then have to abandon MLE altogether and employ some other estimation procedure such as the method of moments which even if less efficient would lead to more straightforward calculations.

However, these problems clearly form a separate study in themselves, and as such come outside the scope of this paper, which is primarily concerned with the properties of $\hat{\theta}$ as an estimator, rather than with the question of how it may be calculated.

We pass on therefore to consider again consistency, efficiency and sufficiency. The existence of a sufficient statistic, has, in particular, a close bearing on the MLE, for in this case L can be factored as (in the usual notation)

$$L(\underset{\sim}{x} \mid \theta) = g(t \mid \theta)\, h(\underset{\sim}{x}) \tag{3.1}$$

where t is sufficient for θ, and $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$.

Thus any solution of the LE is also a solution of

$$\frac{\partial}{\partial \theta} g(t \mid \theta) = 0$$

which implies that $\hat{\theta}$ will be a function of t. Thus, if a sufficient statistic exists, then $\hat{\theta}$ will be a function of it.

[1] The problem of determining the magnitude up to various orders of $\frac{1}{n}$ of the bias in certain MLE's has been considered in a number of papers, e.g. Haldane (1953), Haldane and Smith (1956), Bowman and Shenton (1965, 1968, 1969) and Shenton and Bowman (1963, 1967). There is also a recent paper by Box (1971) which deals with the problem from a more general standpoint.

[2] See, for example, Solari (1968). Further comments about this type of difficulty will be made in section 9.

It is also known that under the regularity conditions that $\frac{\partial L}{\partial \theta}$ and $\frac{\partial^2 L}{\partial \theta^2}$ exist for all θ and all $\underset{\sim}{x}$, that a necessary and sufficient condition for there to exist an estimator t of θ, whose variance attains the Minimum Variance Bound (MVB) of Cramer-Rao,

$$\frac{1}{E\left[\frac{-\partial^2 l}{\partial \theta^2}\right]} = Vo, \text{ say}$$

is that $\frac{\partial l}{\partial \theta}$ can be written in the form

$$\frac{\partial l}{\partial \theta} = A(\theta)(t-\theta) \tag{3.2}$$

where $A(\theta)$ is a function of θ only. Further, if this condition is satisfied then no function of θ, apart from θ itself, can have an estimator with this property.

If the above identity (3.2) holds, then clearly the LE has the unique solution $\theta = t$, and since at this value of θ, $\frac{\partial^2 l}{\partial \theta^2}$ is negative, we have that l, and hence L, has an absolute maximum at $\theta = t$, which is therefore $\hat{\theta}$. We thus have the important result that in a regular estimation situation, if there is an MVB estimator, it can be obtained by the ML method.

A consideration of (3.1) and (3.2) will show further that in a regular situation the MVB cannot be attained unless a sufficient statistic exists. The converse of this is not, however, necessarily true as is illustrated by our example earlier in this section.

Here $\hat{\theta}$ is indeed sufficient for θ, but $\frac{\partial l}{\partial \theta}$ is not of the form $A(\theta)(\hat{\theta}-\theta)$. In fact, the efficiency of $\hat{\theta}$, i.e. $\frac{Vo}{V(\hat{\theta})}$[1] can be shown to be $\frac{n\lambda-2}{n\lambda}$ which is less than unity for any given n. (See Cramér (1946), Chapter 33.)

On the other hand, as this ratio tends to one as n tends to infinity, then we can say that $\hat{\theta}$ is asymptotically efficient. To show also that it is efficient in Fisher's sense we would have to establish that, asymptotically, $\hat{\theta}$ has a normal distribution, which is in fact the case.

In a limited number of cases where the MVB is not attained the MLE can be used in conjunction with the Rao-Blackwell process to obtain the MVUE.

Thus, for example, in the case of the binomial distribution with parameters n and θ, if r is the number of "successes" realised then $\hat{\theta} = \frac{r}{n}$. As this estimator is unbiased and sufficient for θ, and the distribution is complete, i.e. $E_\theta[f(r)] = 0$ for all θ implies $f(r) \equiv 0$, we have immediately that any function of $\hat{\theta}$ which is unbiased for the function of θ which it estimates is the MVUE.[2]

Thus as $E\left[\frac{r(n-r)}{n-1}\right] = n\theta(1-\theta)$ then $\frac{r(n-r)}{n-1}$ is the MVUE of $n\theta(1-\theta)$ which is the variance.

Similarly, for example, $\frac{r(r-1)}{n(n-1)}$ is the MVUE of θ^2.

[1] Since $\hat{\theta}$ is biased, $Vo = \frac{[1+b'(\theta)]^2}{E\left[\frac{-\partial^2 l}{\partial \theta^2}\right]}$ where $b'(\theta) = \frac{db(\theta)}{d\theta}$ and $b(\theta) = E(\hat{\theta})-\theta$.

[2] Note that as $\hat{\theta} = \frac{r}{n}$ is an MVB estimator, then in view of our earlier remarks no other function of $\hat{\theta}$ can be MVB for the function of θ it estimates.

It does not necessarily follow, however, that we must always first obtain the MLE in order to begin the Rao-Blackwell process. All that is required initially is a sufficient statistic and some unbiased estimator which may be obtained in any way we like.

Consistency, on the other hand, is of its very nature an asymptotic property, and so the distinction between small and large samples does not arise. It is a property which holds under fairly wide conditions. Certainly in cases such as we have considered in this section, where Cramer's (1946) conditions hold and the LE has a unique solution, we know at once that $\hat{\theta}$ is consistent for θ.

We shall consider this property in detail in the next section, but before passing on, we conclude with a brief discussion of the property of invariance which MLE's possess.

It can be shown that if $\phi\,(\theta)$ is a $1-1$ function of θ, i.e. both $\phi\,(\theta)$ and $\phi^{-1}\,(\theta)$ are single-valued, then the MLE of $\phi\,(\theta)$ is $\phi\,(\hat{\theta})$. This result can lead to considerable simplification of any necessary calculations when, if for any reason, we need not only the estimate of a parameter but also an estimate of some function of it. So long as we are dealing with MLE's, therefore, no reparametrization will be necessary.

Thus in the case of the normal density

$$\frac{1}{\sqrt{2\pi v}}\exp\left[-\frac{1}{2v}(x-\mu)^2\right]$$

with μ known but v, the variance, unknown, it is easily shown that

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

and we know therefore that the MLE of σ, the standard deviation, is

$$\hat{\sigma} = \sqrt{\frac{1}{n}\Sigma\,(x_i-\mu)^2}.$$ [1]

4. Consistency

We now give an outline of Cramer's (1946) proof that some root of the LE which is consistent for θ exists under the following regularity conditions. These are stated in full since they are now standard in statistical theory.

C.1. For almost all x, $\dfrac{\partial \log f}{\partial\theta}$, $\dfrac{\partial^2 \log f}{\partial\theta^2}$ and $\dfrac{\partial^3 \log f}{\partial\theta^3}$ [2] exist for every $\theta \in \Theta$.

C.2. For all $\theta \in \Theta$, $\left|\dfrac{\partial f}{\partial\theta}\right| < F_1\,(x)$, $\left|\dfrac{\partial^2 f}{\partial\theta^2}\right| < F_2\,(x)$ and $\left|\dfrac{\partial^3 \log f}{\partial\theta^3}\right| < H\,(x)$, the functions F_1 and F_2 being integrable over $(-\infty, \infty)$ and $\displaystyle\int_{-\infty}^{\infty} H\,(x)\,f(x,\,\theta)\,dx < M$, where M does not depend on θ.

C.3. For all $\theta \in \Theta$, the integral

$$\int_{-\infty}^{\infty}\left(\frac{\partial \log f}{\partial\theta}\right)^2 f dx$$

is finite and positive.

[1] In order to preserve the $1-1$ aspect, we regard v and σ as related by $\sigma = v^{\frac{1}{2}}$ rather than by $\sigma^2 = v$.

[2] f is written for $f(x;\ \theta)$.

By Taylor's theorem we can therefore write,

$$\frac{\partial \log f_i}{\partial \theta} = \left(\frac{\partial \log f_i}{\partial \theta}\right)_{\theta_0} + (\theta-\theta_0)\left(\frac{\partial^2 \log f_i}{\partial \theta^2}\right)_{\theta_0} + \tfrac{1}{2}\lambda\,(\theta-\theta_0)^2\, H(x)$$

where $f_i = f(x_i, \theta)$, $|\lambda| < 1$, θ_0 is the true value of the parameter and the partial derivatives with respect to θ are evaluated at $\theta = \theta_0$.

Summing over i we obtain for the LE

$$0 = \frac{1}{n}\frac{\partial l}{\partial \theta} = B_0 + B_1\,(\theta-\theta_0) + \tfrac{1}{2}\lambda B_2\,(\theta-\theta_0)^2 \tag{4.1}$$

where

$$B_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \log f_i}{\partial \theta}\right)_{\theta_0}, \quad B_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \log f_i}{\partial \theta^2}\right)_{\theta_0}$$

and

$$B_2 = \frac{1}{n}\sum_{i=1}^{n} H(x_i).$$

From the conditions (1), (2) and (3) it follows that

$$E\left(\frac{\partial \log f}{\partial \theta}\right)_{\theta=\theta_0} = 0 \quad \text{and} \quad E\left(\frac{\partial^2 \log f}{\partial \theta^2}\right)$$

is a strictly negative constant, $-k^2$ say.

Thus by Khintchine's theorem[1] B_0 and B_1 converge in probability to zero and $-k^2$ respectively, while B_2 converges in probability to $E[H(x)]$ which is positive and less than a fixed quantity independent of θ, by condition 2.

Therefore if we put $\theta = \theta_0 + \delta$ so that

$$\frac{1}{n}\frac{\partial l}{\partial \theta} = B_0 - B_1\delta + \tfrac{1}{2}\lambda B_2 \delta^2$$

then it follows that for all sufficiently small δ, $\frac{1}{n}\frac{\partial l}{\partial \theta}$ and $-k^2\delta$ have the same sign in probability as n tends to infinity. Thus the probability that there is a root of the LE in the interval $(\theta_0 - \delta, \theta_0 + \delta)$ tends to one as n tends to infinity. Since δ may be arbitrarily small it follows that a consistent root of the LE exists.

V. S. Hurzabazaar (1948) argues along similar lines, and from Cramer's conditions proves that if $\hat{\theta}$ is a consistent solution of the LE then

$$\lim_{n\to\infty} P\left[\left(\frac{\partial^2 L}{\partial \theta^2}\right)_{\hat{\theta}} < 0\right] = 1$$

which implies that the probability that L has a local maximum at $\theta = \hat{\theta}$ approaches unity as n approaches infinity. A straightforward application of Rolle's theorem then shows that there cannot be more than one consistent solution.

Thus Cramer's and Hurzabazaar's results together prove that under certain regularity conditions the LE has a unique consistent solution, and that at this value of θ, the LF has a local maximum with probability one.

L. Weiss (1963, 1966) develops these ideas further. In the first of these two papers he proves Cramer's result on consistency, but under far less restrictive conditions. The LF is not even assumed continuous.

[1] The conditions of Khintchine's theorem, which is a form of the law of large numbers, require identically distributed variables with finite expectations.

This result is improved further in the second paper, where it is shown that there exists a sequence $\{\delta_n\}$, converging to zero, such that the probability that the LF has a relative maximum in the open interval $(\theta_0-\delta_n, \theta_0+\delta_n)$ tends to one as n tends to infinity.

This result has an interesting application from the computational point of view. Suppose that

$$f(x, \theta) = \frac{1}{\pi\,[1+(x-\theta)^2]}\;(-\infty<x<\infty)$$

then it is well known that $\sqrt{n}\,(\hat{\theta}_n-\theta_0)$ is asymptotically a normal variate. In practice, this may not be very helpful for the LE is a polynomial equation of degree $2n-1$, and a closed-form solution will not, in general, be possible.

However, if $\bar{\theta}_n$ is defined by

$$\bar{\theta}_n = Z_n - \frac{\left[\dfrac{\partial l}{\partial \theta}\right]_{\theta = Z_n}}{\left[\dfrac{\partial^2 l}{\partial \theta^2}\right]_{\theta = Z_n}}$$

where Z_n is the median of the observed values $x_1, x_2, \ldots, x_n$, then from Weiss' result it can be shown that $\bar{\theta}_n$ is asymptotically identical to $\hat{\theta}$. Since, by definition, there is an explicit form for $\bar{\theta}_n$ which can be easily computed for a given sample, it is therefore a relatively straightforward matter to obtain, for example, an "asymptotic confidence interval" for θ_0.

The results of Cramer and Hurzabazaar together do not imply the consistency of $\hat{\theta}$ in the general case. It might, however, be inferred that this would follow if we restrict Θ to be a sufficiently small neighbourhood of θ_0.

That this is not the case is shown by the following counter-example due to Kraft and Le Cam (1956). Here Cramer's conditions are satisfied and θ is identifiable. The LE has roots and with probability tending to one, as n tends to infinity, the MLE exists and is a unique root of the LE. Nevertheless, the MLE is inconsistent while at the same time consistent estimates do exist.

Suppose $\Theta = \bigcup_{k=1}^{\infty} A_k$, where A_k is the open interval $(2k, 2k+1)$ and that $\{\alpha_k\}$ is some defined order of the rationals in $(0, 1)$. Let $\rho\,(\theta) = \alpha_k$ if $\theta \in A_k$ and let (X_i, Y_i, Z_i) be multinomially distributed with probabilities $p_1 = \rho\,(\theta) \cos^2 2\pi\theta$, $p_2 = \rho\,(\theta) \sin^2 2\pi\theta$ and $p_3 = 1-\rho\,(\theta)$.

For n independent observations we have

$$l = \log p_n = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + f\,(n_1, n_2, n_3)$$

where

$$n_1 = \sum_{i=1}^{n} X_i, \quad n_2 = \sum_{i=1}^{n} Y_i, \quad n_3 = \sum_{i=1}^{n} Z_i,$$

and $f\,(n_1, n_2, n_3)$ is a function of n_1, n_2, n_3 independent of θ.

The solutions of the LE can be shown to be of the form $\theta = \frac{1}{2}m + \dfrac{\alpha}{2\pi}$, where α is the acute angle $\tan^{-1}\sqrt{\dfrac{n_2}{n_1}}$, and m is any integer. Since, however, l is maximized by taking $p_i = n_i/n$ then only one of these solutions for θ can be $\hat{\theta}$.

Also we must have $\lim\limits_{n\to\infty} P\left[1-\rho\,(\hat{\theta}) = \dfrac{n_3}{n}\right] = 1$, so that if $\hat{\theta}$ is consistent it must eventually remain in some fixed interval A_k so that $\dfrac{n_3}{n} = 1-\alpha_k$. The probability of this occurring, however, tends to zero as n tends to infinity.

We pass on therefore to consider Wald's proof of consistency which is more satisfactory, for it does establish the consistency of $\hat{\theta}$ and not merely some root of the LE. His approach is quite different, for his regularity conditions do not include any differentiability assumptions. We give only a brief outline of the proof which is set in a multiparameter context.

If the random variable $u = \log f(x, \theta) - \log f(x, \theta_0)$ is considered, then it can be established by a generalized form of the theorem of means that when $\theta \neq \theta_0$

$$E_{\theta_0}[\log f(x_i, \theta)] < E_{\theta_0}[\log f(x_i, \theta_0)].$$

Summing over i and letting n tend to infinity, it follows from the strong law of large numbers that

$$\lim_{n\to\infty} P[l(\underset{\sim}{x}, \theta) < l(\underset{\sim}{x}, \theta_0)] = 1$$ [1]

On the other hand, by definition,

$$l(\underset{\sim}{x}, \hat{\theta}) \geqq l(\underset{\sim}{x}, \theta_0)$$

so that we must have $\lim_{n\to\infty} P(\hat{\theta} = \theta_0) = 1$.

There is an interesting comment by Wolfowitz (1949) which follows immediately afterwards. He makes the point that by appealing to the strong law of large numbers Wald actually proves strong convergence which is more than consistency. He shows that in fact it is only necessary to appeal to the weak law of large numbers to establish the consistency of $\hat{\theta}$. This, according to Wolfowitz, extends the class of dependent random variables to which the result can be applied.

Kraft (1955) also establishes a theorem giving conditions under which $\hat{\theta}$ will be consistent, strongly consistent and uniformly consistent. He observes that in fact strong convergence is of more than theoretical importance, as Wald (1951) has shown that there is a class of asymptotic minimax estimates whose construction depends on the existence of a strong consistent estimate. Kraft's final theorem proves strong consistency for the case where the observations are independent and identically distributed, though he himself remarks that this can be deduced from a trivial generalization of Wald's (1949) results.

We conclude by remarking that since there are cases of inconsistent MLE's, then some conditions must be imposed in order to prove consistency. Conditions such as Wald's form a sufficient set. A more difficult problem would be to establish a set of necessary and sufficient conditions, as for example Kolmogorov has done for a form of the law of large numbers. This problem appears not to have been considered by anyone.

5. Efficiency

We have already given a definition of efficiency in section 3, which is now a standard measure of estimator performance. However, it is to some extent arbitrary, as will be explained subsequently, and we shall therefore begin by considering the possibility of more general criteria.

The "natural" definition of optimality is the following. The estimator T of θ is optimum if

$$P[|T-\theta| < \varepsilon \mid \theta] \geqq P[|y-\theta| < \varepsilon \mid \theta] \qquad \text{(A)}$$

for all other estimators y of θ.

Some development of this idea has been achieved by Bahadur (1960), who introduces the concept of the effective standard deviation, $\tau = \tau_g(T_n, \varepsilon, \theta)$ defined by the equation

$$P\left[|N| \geqq \frac{\varepsilon}{\tau}\right] = P[|T_n - g(\theta)| \geqq \varepsilon \mid \theta] \qquad \text{(B)}$$

where T_n is an estimator, based on n independent observations, of $g(\theta)$ [2] a real-valued function of an unknown parameter θ, and N is the standard normal variate.

[1] $l(\underset{\sim}{x}, \theta) = \log f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta)$, and similarly $l(\underset{\sim}{x}, \theta_0)$.

[2] $g(\theta)$ could, of course, be simply θ.

The motivation for definition (B) arises in the first place from a consideration of the approximation

$$P\left[\,|N| \geqq \frac{\varepsilon}{\sigma_n}\right] \simeq P\left[|\,T_n - g(\theta)| \geqq \varepsilon \mid \theta\right] \tag{5.1}$$

where σ_n^2 is the asymptotic variance of T_n. Such an approximation is unsatisfactory, however, since the error involved is unknown. Thus σ_n might differ considerably from τ which is a true measure of the performance of T_n. It would in any case be permissible to use (5.1) only within the class of CAN estimators.

Definition (B), however, is quite general and clearly leads to a criterion of optimality in agreement with (A), for if U_n is any other estimator of $g\,(\theta)$ then

$$P\left[|T_n - g\,(\theta)| \geqq \varepsilon \mid \theta\right] \geqq P\left[|\,U_n - g\,(\theta)| \geqq \varepsilon \mid \theta\right]$$

if and only if $\tau_g\,(T_n, \varepsilon, \theta) \geqq \tau_g\,(U_n, \varepsilon, \theta)$.

In fact, without restriction to the class of CAN estimators Bahadur establishes the following results.

If T_n is consistent for $g\,(\theta)$ then

$$\underline{\lim}_{\varepsilon\to 0}\ \underline{\lim}_{n\to\infty}\ \{n\tau_g^2\,(T_n, \varepsilon, \theta)\} \geqq \frac{[g'\,(\theta)]^2}{I\,(\theta)} \tag{5.2}$$

where $I\,(\theta)$ is the classical "information contained in x".

This result is clearly analogous to Fisher's result

$$\lim_{n\to\infty}\,(n\sigma_n^2) \geqq \frac{[g'\,(\theta)]^2}{I\,(\theta)} \tag{5.3}$$

for a consistent estimator. The appearance of $I\,(\theta)$ in Bahadur's result again bears witness to its general importance as a statistical measure.

He also proves that under certain regularity conditions which are a combination of Wald's (1949) conditions, Cramer's (1946) conditions, plus certain other assumptions, that if U_n is the MLE of $g\,(\theta)$ based on a sample of n independent observations, then

$$\overline{\lim}_{\varepsilon\to 0}\ \overline{\lim}_{n\to\infty}\ \{n\tau_g^2\,(U_n, \varepsilon, \theta)\} \leqq \frac{[g'\,(\theta)]^2}{I\,(\theta)}. \tag{5.4}$$

Clearly (5.2) and (5.4) imply that MLE's are asymptotically efficient, i.e. "best", in the sense of definition (A) or (B).

This is a very satisfactory result, although limited to a relatively small class of estimators, and it might now be hoped that it would be possible to go on and using the parameter τ, establish the asymptotic efficiency of any estimator of whatever type. Thus if T_n is asymptotically efficient, the asymptotic efficiency of T_n', some other estimator, would be measured by an expression such as

$$\lim_{\varepsilon\to 0}\ \lim_{n\to\infty} \frac{P\left[|\,T_n - g\,(\theta)| > \varepsilon\right]}{P\left[|\,T_n' - g\,(\theta)| > \varepsilon\right]}.$$

Unfortunately the analysis necessary to follow up such an approach appears to be of extreme difficulty.

Another contribution along similar lines is provided by Wolfowitz (1965), who is strongly critical of Fisher's conclusions about the efficiency of MLE's, partly because it does not take non-normal estimators into account, and partly because super efficient estimators do exist.

Wolfowitz's argument requires heavy regularity conditions on $f\,(x, \theta)$ which nevertheless he claims are operationally sensible, but no restrictions on a general estimator sequence $\{T_n\}$, except that the distribution of $\sqrt{n}\,(T_n - \theta)$ should approach some limiting distribution $L\,(x, \theta)$

Z

uniformly in both x and θ. There is therefore no requirement, as in the case of Fisher's argument, that T_n should be asymptotically normal.

He then proves that if θ_0 is a "smooth" point for L, i.e. any point of Θ except possibly for an at most denumerable set of turning points, then for any positive b and c,

$$\lim_{n\to\infty} P\left[-c+w(\theta_0)<\sqrt{n}\,(\hat{\theta}-\theta_0)<W(\theta_0)+b \mid \theta_0\right]$$
$$\geqq \lim_{n\to\infty} P\left[-c+w(\theta_0)<\sqrt{n}\,(T_n-\theta_0)<W(\theta_0)+b \mid \theta_0\right]$$

where $w(\theta_0)$ and $W(\theta_0)$ are defined as follows.

Let
$$l(\theta) = \min\left[x \mid L(x,\theta) = \tfrac{1}{2}\right]$$
$$u(\theta) = \max\left[x \mid L(x,\theta) = \tfrac{1}{2}\right]$$

and
$$W(\theta_0) = \limsup_{\delta\to 0+}\left[u(\theta)\middle|\, \theta_0<\theta<\theta_0+\delta\right]$$
$$w(\theta_0) = \limsup_{\delta\to 0+}\left[l(\theta)\middle|\, \theta_0-\delta<\theta<\theta_0\right].$$ [1]

Clearly all this is equivalent to saying that the MLE is efficient in a sense which is essentially that of definition (A).

Thus, the results of Bahadur and Wolfowitz point to the optimality of MLE's within a wide class of estimators whether normal or non-normal. Nevertheless, since no immediately useful mathematical development of their ideas appears possible then we must continue to measure estimator performance in terms of its variance. That efficiency in this sense does not imply efficiency in the sense of definition (A), however, is demonstrated by the following example due to Basu (1956).

If each of $x, \ldots, x_n$ are identically and normally distributed about an unknown mean μ with a variance of 1, then it is well known that

$$\bar{X}_n = \sum_{i=1}^{n} \frac{x_i}{n} \quad \text{and} \quad S_n = \sum_{i=1}^{n} (x_i-\bar{x}_n)^2$$

are independently distributed, and that the distribution of S_n does not depend on μ.

Let a_n be such that

$$P(S_n>a_n) = \frac{1}{n}$$

and define H_n by

$$H_n = 0 \text{ if } S_n \leqq a_n$$
$$H_n = 1 \text{ if } S_n > a_n.$$

Further, let $t_n = (1-H_n)\bar{X}_n+nH_n$, and $t_n' = \bar{X}_{[\sqrt{n}]}$ where, as usual, $[x]$ denotes the largest integer not exceeding x. Then as, $\sqrt{n}\,(t_n-\mu) = \sqrt{n}\,(\bar{X}_n-\mu)+\sqrt{n}H_n\,(n-\bar{X}_n)$ and

$$P(H_n=0) = 1-\frac{1}{n}$$

tends to 1 as n tends to infinity, then it follows that $n\,(t_n-\mu)$ converges in distribution to a standard normal variate. Thus the asymptotic variance of t_n is $\frac{1}{n}$, whereas the asymptotic variance of t_n' is clearly $1/\sqrt{n}$.

[1] $[l(\theta), u(\theta)]$ is referred to as the median interval. If there is a unique median then $l(\theta) = u(\theta) = w(\theta) = W(\theta)$.

On the other hand as $\bar{X}_n$ and H_n are independent, then for all $n > \mu + \varepsilon$

$$P[|t_n - \mu| > \varepsilon \mid \mu] = P(H_n = o)\, P[|\bar{X}_n - \mu| > \varepsilon \mid \mu] + P(H_n = 1) = \frac{1}{n} + o\left(\frac{1}{n}\right)$$

since

$$P[|\bar{X}_n - \mu| > \varepsilon \mid \mu] = o\left(\frac{1}{n}\right).$$

Also

$$P[|t'_n - \mu| > \varepsilon \mid \mu] = o\left(\frac{1}{n}\right).$$

Judged therefore by criterion (A), t'_n is better, one might say infinitely better, than t_n, but on the basis of the usual definition of efficiency t_n is better than t'_n.

Nevertheless, this type of situation is probably very exceptional, and indeed the CAN estimators form a large class. We can therefore continue our discussion of efficiency in the usual sense without undue restriction.

We have already remarked that Fisher never gave what we would now describe as regularity conditions in the proof of the asymptotic efficiency of MLE's. The first rigorous proof is due, in fact, to Cramer (1946). This forms part of the theorem to which reference has already been made in connection with consistency. Under the conditions stated there, it is proved that the solution of the LE which is consistent is also asymptotically efficient. The whole theorem can thus be summarized by saying that there is a root of the LE, $\hat{\theta}$, such that the variate $\sqrt{n}\,(\hat{\theta} - \theta_0)$ converges in distribution to that of a $N\left(0, \dfrac{1}{I(\theta_0)}\right)$ variate.

Cramer's proof of efficiency follows on immediately from the argument for consistency. In the notation of section 4 it is shown that if $\hat{\theta}$ is the solution of the LE now known to exist then from the LE we obtain,

$$k\sqrt{n}\,(\hat{\theta} - \theta_0) = \frac{\dfrac{1}{k\sqrt{n}} \displaystyle\sum_{i=1}^{n} \frac{\partial \log f_i}{\partial \theta}}{\dfrac{-B_1}{k^2} - \dfrac{\lambda B_2}{2k^2}(\hat{\theta} - \theta_0)}.$$

Since $\hat{\theta}$ has now been proved consistent for θ_0, then it follows that the denominator of the right-hand side of the above converges in probability to 1. Further, $\left(\dfrac{\partial \log f_i}{\partial \theta}\right)_{\theta = \theta_0}$ for each i, is a variate with zero mean and variance k^2, so that a straightforward application of the Lindeberg-Lèvy theorem shows that

$$\sum_{i=1}^{n} \left(\frac{\partial \log f_i}{\partial \theta}\right)_{\theta = \theta_0}$$

is asymptotically $N(O, k^2 n)$, so that the numerator above is asymptotically $N(0, 1)$. Thus $\sqrt{n}\,(\hat{\theta} - \theta_0)$ is asymptotically $N\left(0, \dfrac{1}{k^2}\right)$ or $N\left(0, \dfrac{1}{I(\theta_0)}\right)$ where $I(\theta_0) = E\left(\dfrac{-\partial^2 \log f}{\partial \theta^2}\right)_{\theta = \theta_0}$.

We remark in passing that the quantity $I(\theta_0)$ forms part of the lower bound $\dfrac{1}{nI(\theta_0)}$ for the asymptotic variance of any estimator of θ, when θ_0 is the true value of θ. Fisher himself showed in a non-rigorous way that $\hat{\theta}$ attains this bound, and may therefore be said to be efficient. He therefore described such an estimator as one which utilizes all the available information in the sample, in the sense that no other statistic based on the same sample could in probability give a more precise estimate of θ_0. For this reason $I(\theta_0)$ is often described as the information contained in a single observation, or simply as "Fisher's information". Similarly, the quantity $E\left(\frac{-\partial^2 l}{\partial\theta^2}\right)_{\theta=\theta_0}$ which can be shown to be equal to $nI\ (\theta_0)$, is the information available in a sample of n independent observations.

Returning now to Cramer's proof of efficiency we remark that his conditions, as stated at the beginning of section 4, are sufficient rather than necessary, as is clearly shown by a consideration of the density $f(x, \theta) = \frac{1}{2}e^{-|x-\theta|}$. Here $\hat{\theta}$ is the median which asymptotically is normally distributed and is efficient. Yet $\frac{\partial}{\partial\theta} \log f(x, \theta)$ is discontinuous and $\frac{\partial^2}{\partial\theta^2} \log f(x, \theta)$ is zero almost everywhere, hence contradicting condition 1.

This example is due to H. E. Daniels (1961), who proves asymptotic efficiency under more general conditions which do not involve $\frac{\partial^2}{\partial\theta^2} \log f(x, \theta)$.

In the case of finite samples we have already seen that if the MVB is attained then efficient estimators can be obtained by the method of ML.

What happens in a regular estimation problem when the MVB is not attained? The following elementary example is enough to show that the MLE does not necessarily have the least possible variance, in this case.

Let $f(x, \theta) = \theta^2 x e^{-\theta x}$, $(\theta > 0)$, then for a sample of n independent observations $x_1, x_2, \ldots, x_n$, it is easily shown that

$$\hat{\theta} = \frac{2n}{\sum_{i=1}^{n} x_i}$$

whereas an MVUE is

$$\tilde{\theta} = \frac{2n-1}{\sum_{i=1}^{n} x_i}$$

so that

$$\frac{\text{var}\,(\tilde{\theta})}{\text{var}\,(\hat{\theta})} = \left(\frac{2n-1}{2n}\right)^2 < 1.$$

Nevertheless this is clearly a regular estimation situation, since

$$\frac{\partial l}{\partial\theta} = \frac{2n}{\theta} - \Sigma x_i \quad \text{and} \quad \frac{\partial^2 l}{\partial\theta^2} = \frac{-2n}{\theta^2}$$

both exist for $\theta > 0$ and all x.

Another possibility might be to consider the variance bounds derived by Bhattacharyya (1946–7–8),[1] since they are a generalization of the Cramer-Rao MVB. A possible result for the MLE would be that in a regular estimation situation the MLE attains the kth order B-bound, if this is the minimum variance which is attained. Fend (1959), however, has shown that this is not necessarily true. This is done by considering

$$f(x, \theta) = \theta^{-c} e^{-x\theta^{-c}} \quad (c > 0, \theta > 0)$$

where c is supposed known and a single observation is taken by which to estimate θ. It is shown that if $\frac{1}{c}$ is an integer, k, then the kth B-bound attained is attained by the estimator $\frac{x^k}{k!}$ which is unbiased for θ. The MLE, on the other hand, is x^k, which clearly therefore does not attain this bound for $k > 1$.

[1] A full account of these bounds is also given, for example, in *The Advanced Theory of Statistics*, Vol. 2, chapter 17, by Kendall and Stuart.

There does not, in fact, appear to be any particular bound such as the MVB of Cramer-Rao which the MLE must attain in such cases.

In the non-regular case the MVB may not even be defined so that we cannot talk of efficiency in this sense. Neither is it possible, as an alternative, to define efficiency relative to the MVUE, for Basu (1952) has shown that such estimators do not exist in a large number of situations.

Some progress towards defining non-trivial variance bounds for certain well-defined classes of non-regular situations has, however, been made by, for example, Chapman and Robins (1951) and Fraser and Guttman (1952). There is also a set of three papers by Blischke *et al.* (1965, 1968*a* and *b*) on the estimation of the parameters of the Pearson type III distribution

$$f(x) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x-\alpha}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x-\alpha}{\beta}\right)}, \quad x>\alpha\,(\alpha>0,\ \beta>0)$$

$$= 0 \text{ otherwise,}$$

which is a good example of a non-regular situation.

They also derive a new bound which is applied to the Pearson type I and IV distributions, and to the Weibull distribution.

There is, however, certainly nothing approaching a unified theory of non-regular estimation, and there are no simple measures such as the MVB to provide a yardstick of efficiency. Thus no general conclusions about the MLE, or indeed about any estimator, can be made for such situations, and most non-regular estimation problems must therefore be considered *ad hoc.*

Neither is it necessarily true that MLE's have asymptotically the least variance for large samples in non-regular situations, as the following example due to Basu (1952) shows.

The estimator $T = \frac{2g+l}{5}$, of θ, where g and l are the greatest and least values of a sample of n independent observations drawn from a population with a rectangular distribution on $[\theta, 2\theta]$, has an asymptotic variance of $\frac{1}{5n^2}$. The MLE, on the other hand, has a variance of $\frac{1}{4n^2}$ so that $\frac{v(T)}{v(\hat{\theta})} = 0{\cdot}8$.

Note on Arrangement of Bibliography

Those papers which are in the first group are of importance in estimation theory generally, and for various reasons come within the scope of this paper. Some of them contain results about MLE in particular. The papers in the second group are again of a theoretical type but are mainly or even exclusively concerned with MLE.

In the third group we have papers containing particular applications of MLE. In the final group are those which consider simultaneously various estimators, among them MLE's, with regard to some particular problem of estimation. They therefore often provide useful comparisons of MLE's with other estimates.

Papers marked with a single asterisk have some numerical content, and those with a double asterisk are almost wholly numerical or are concerned with the computational aspect of deriving MLE's.

Note. An additional forty entries were added at proof stage and are printed at the end of this Bibliography. (Editor.)

Bibliography

Group I

Aitken, A. C. and Silverstone, H. (1942). On the estimation of statistical parameters. *Proc. Royal Soc. Edinburgh*, A, **61**, 186.

Barankin, E. W. and Gurland, J. (1951). On asymptotically normal and efficient estimates. *Univ. California Pub. in Statist.* **1**, 89.

Bartlett, M. S. (1936). The information available in small samples. *Proc. Cambridge Phil. Soc.* **32**, 560–566.

Basu, D. (1952). An example of non-existence of a minimum variance estimator. *Sankhyā*, **12**, 43.

Basu, D. (1954). Choosing between two simple hypotheses and the criterion of consistency. From a Ph.D. thesis submitted to Calcutta University.

Basu, D. (1956). The concept of asymptotic efficiency. *Sankhyā*, **17**, 193.

→ Basu, D. (1964). Recovery of ancillary information. *Sankhyā*, A, **26**, 3–16.

Bhattacharyya, A. (1946–7–8). On some analogues of the amount of information and their use in statistical estimation. *Sankhyā*, **8**, 1, 201, 315.

Blischke, W. R., Mundle, P. B. and Truelove, A. J. (1969). On non-regular estimation I. Variance bounds for estimation of location parameters. *J. Amer. Statist. Assoc.* **64**, 1056–1072.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc.*, B, **26**, 211–252.

Chapman, D. G. and Robbins, H. E. (1951). Minimum variance estimation without regularity assumptions. *Ann. Math. Statist.* **22**, 581–586.

Edgeworth, F. Y. (1908*a*). On the probable errors of frequency constants. *J. R. Statist. Soc.* **LXXI**, 381.

Edgeworth, F. Y. (1908*b*). On the probable errors of frequency constants. *J. R. Statist. Soc.* **LXXI**, 499.

Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transaction of the Royal Society*, A, **222**, 308–358.

Fisher, R. A. (1925). Theory of Statistical Estimation. *Proc. Cambridge Phil. Soc.* **22**, 700–725.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Royal Soc.*, A, **144**, 285.

Ford, A. V. (1959). On the attainment of C-R and Bhattacharyya bounds for the variance of an estimate. *Ann. Math. Statist.* **30**, 381–388.

Fraser, D. A. S. and Guttman, I. (1952). Bhattacharyya Bounds without regularity assumptions. *Ann. Math. Statist.* **23**, 629–632.

Fraser, D. A. S. (1964). On local inference and information. *J. R. Statist. Soc.*, B, **26**, 253–260.

Hodges, J. L.(Jr.) and Lehmann, E. L. (1950). Some applications of the Cramer-Rao inequality. *Proc. 2nd Berkeley Symp. on Math. Statist. Prob.* 13–22.

Kalbfleisch, J. D. and Sprott, D. A. (1969). Application of likelihood and fiducial probability to sampling finite populations. *New Developments in Survey Sampling*, 358–389.

Kallianpur, G. and Rao, C. R. (1955). On Fisher's lower bound to asymptotic variance of a constant estimate. *Sankhyā*, **15**, 331–342.

Koopman, B. O. (1936). On distributions admitting sufficient statistics. *Trans. Ann. Math. Soc.* **39**, 399–409.

Kraft, C. (1955). Some conditions for consistency and uniform consistency of statistical procedures. *Univ. California Pub. in Statist.* **2**, 125.

Kullback, S. and Liebler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79.

Le Cam, L. (1953). On the asymptotic theory of estimation and testing hypotheses. *Proc. 3rd Berkeley Symp. on Math. Statist. Prob.* **1**, 129–156.

Lehmann, E. and Scheffé, H. (1950). Completeness, similar regions and unbiased estimation. *Sankhyā*, **10**, 305.

Neyman, J. (1949). Contributions to the theory of χ^2 test. *Proc. 1st Berkeley Symp. on Math. Statist. Prob.* 239–273.

Pearson, K. (1896). Mathematical contributions to the theory of evolution IV. Regression, Heredity, Panmixia. *Phil. Trans. Royal Soc. London, Ser.*, A, **187**, 253–318.

Pearson, K. and Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Phil. Trans. Royal Soc. London*, **191**, 229–311.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika*, **14**, 249–272.

Pitman, E. J. C. (1938). The estimation location and scale parameters of a continuous population of any given form. *Biometrika*, **30**, 391.

Rao, C. R. (1945). Information and accuracy attainable in estimation of statistical parameters. *Bull. Cal. Math. Soc.* **37**, 81–91.

Rao, C. R. (1949). Sufficient statistics and minimum variance estimates. *Proc. Cambridge Phil. Soc.* **45**, 213–218.

Rao, C. R. (1952*a*). Some theorems on minimum variance estimations. *Sankhyā*, **12**, 27–42.

Rao, C. R. (1952*b*). Minimum variance estimations in distributions admitting ancillary statistics. *Sankhyā*, **12**, 53–56.

Rao, C. R. (1952*c*). On statistics with uniformly minimum variance. *Science and Culture*, **17**, 483–484.

Rao, C. R. (1955). Theory of the method of estimation by minimum chi-square. *Bull. Inter. Statist. Inst.* **35**, 25–32.

→ Rao, C. R. (1961*b*). A study of large sample test criteria through properties of efficient estimates. *Sankhyā*, **23**, 25–40.

Rao, C. R. and Poti, S. J. (1946). On locally most powerful tests when alternatives are one-sided. *Sankhyā*, **7**, 439.

Roy, Jogabratha, Mitra and Sujitkumar (1957). Unbiased minimum variance estimation in a class of discrete distributions. *Sankhyā*, **18**, 371–378.

Smith, J. H. (1947). Estimation of linear functions of cell proportions. *Ann. Math. Statist.* **18**, 231–254.

Wald, A. (1941). Asymptotically most powerful tests of statistical hypotheses. *Ann. Math. Statist.* **12**, 1, 1–19.

Wald, A. (1943). Test of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54**, 426–482.

Wald, A. (1948). Estimation of a parameter when the number of unknown parameters increases indefinitely with the number of observations. *Ann. Math. Statist.* **19**, 220–227.

Wald, A. (1951). Asymptotic minimax solutions of sequential point estimation problems. *Proc. 2nd Berkeley Symp. on Math. Statist. Prob.* 1–11.

Walker, A. M. (1963). A note on asymptotic efficiency of an asymptotically normal estimator sequence. *J. R. Statist. Soc.*, B, **25**, 195–200.

Group II

Bahadur, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhyā*, **20**, 207–210.

Bahadur, R. R. (1960). On the asymptotic efficiency of tests and estimates. *Sankhyā*, **22**, 229.

Barnard, G. A. (1967). The use of the likelihood function in statistical practice. *Proc. 5th Berkeley Symp. Math. Statist. Prob.* 27–40.

Barnard, G. A., Jenkins, G. M. and Winsten, C. B. (1962). Likelihood inference and time series. *J. R. Statist. Soc.*, A, **125**, 321–372.

Bar-Shalom (1971). Asymptotic properties of maximum likelihood estimates. *J. R. Statist. Soc.*, B, Vol. 33, No. 1.

Barton, D. E. (1968). The solution of stochastic integral relations for strongly consistent estimators at an unknown distribution function from a sample subject to variable censoring and truncation. *Trab. Estadist.* **19**, III, 51–73.

Basu, D. (1955). An inconsistency of the method of maximum likelihood. *Ann. Math. Statist.* **26**, 144.

Berkson, J. (1955). Estimation by least squares and by maximum likelihood. *Proc. 3rd Berkeley Symp. Math. Statist. Prob.* **1**.

Birnbaum, A. (1961). Generalised M.L. methods with exact justifications on two levels. *Bull. Inter. Statist. Inst.* **38**, IV, 457–462.

Bradley, R. A. and Grant, J. J. (1962). The asymptotic properties of ML estimators when sampling for associated populations. *Biometrika*, **49**, 205–214.

Brillinger, D. R. (1962). A note on the rate of convergence of a mean. *Biometrika*, **49**, 574–576.

Chanda, K. C. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika*, **41**, 56.

Daniels, H. E. (1961). The asymptotic efficiency of the maximum likelihood estimator. *Proc. 4th Berkeley Symp. Math. Statist. Prob.* **1**, 151–163.

Dharmadhikari, S. W. (1967). A note on a theorem of H. Cramer. *Skand. Aktuarietidskr.* **50**, 153–154.

Doob, J. L. (1934). Probability and statistics. *Trans. Amer. Math. Soc.* **36**, 766–775.

Doob, J. L. (1936). Statistical estimation. *Trans. Amer. Math. Soc.* **39**, 410.

Doss, S. A. D. C. (1962). A note on consistency and asymptotic efficiency of maximum likelihood estimates in multi-parametric problems. *Bull. Cal. Statist. Ass.* **11**, 85.

Doss, S. A. D. C. (1963). On consistency and asymptotic efficiency of maximum likelihood estimates. *J. Indian Soc. Agric. Statist.* **15**, 232–241.

Dugué, D. (1936*a*). Sur le maximum de précision des estimation gaussiennes a la limite. *Comptes Rendus, Paris*, **202**, 193–196.

Dugué, D. (1936*b*) Sur le maximum de précision des limites d'estimations. *Comptes Rendus, Paris*, **202**, 452–454.

Dugué, D. (1937). Application des propriétés de la limite au sens de calcul des probabilités à l'étude de diverses questions d'estimation. *J. Ecole Polytechnique*, **3**, 305–374.

→ Dutta, M. (1966). On maximum (information-theoretic) entropy estimation. *Sankhyā*, A, **28**, 319-328.

Fraser, D. A. S. (1963). On sufficiency and the exponential family. *J. R. Statist. Soc.*, B, **25**, 115–123.

Fraser, D. A. S. (1964). Local conditional sufficiency. *J. R. Statist. Soc.*, B, **26**, 52–62.

Geary, R. C. (1942). The estimation of many parameters. *J. R. Statist. Soc.* **105**, 213–17.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1212.

Gurland, J. (1954). On regularity conditions for M.L.E.'s *Skand. Aktuarietidskrift*, **37**, 71–76.

Hanson, M. A. (1965). Irregularity constrained maximum likelihood estimation. *Ann. Inst. Statist. Math., Tokyo*, **17**, 311–321.

Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, **55**, 547–557.

Hotelling, M. (1930). The consistency and ultimate distribution of optimum statistics. *Trans. Amer. Math. Soc.* **32**, 847.

Huber, P. J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. *Proc. 5th Berkeley Symp. Math. Statist. Prob.* **1**, 221–233.

Huzurbazar, V. S. (1948). The likelihood equation, consistency and the maxima of the likelihood function. *Annals of Eugenics*, **14**, 185–200.

Huzurbazar, V. S. (1949). On a property of distributions admitting sufficient statistics. *Biometrika*, **36**, 71.

→ Kale, B. K. (1963). Some remarks on a method of maximum likelihood estimation proposed by Richards. *J. R. Satist. Soc.*, B, **25**, 209–212.

Kale, B. K. (1966). Approximations to the maximum likelihood estimator using grouped data. *Biometrika*, **53**, 282–285.

Kallianpur, G. (1963). Von Mises functionals and maximum likelihood estimation. *Sankhyā*, A, **25**, 148–149.

Kaufman, S. (1966). Asymptotic efficiency of the maximum likelihood estimator. *Ann. Inst. Statist. Math., Tokyo*, **18**, 155–178.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887–906.

Konijn, H. S. (1963). Note on the non-existence of a maximum likelihood estimate. *Aust. J. Statist.* **5**, 143–146.

Kraft, C. and Le Cam, L. (1956). A remark on the roots of the maximum likelihood equation. *Ann. Math. Statist.* **27**, 1174.

Kriz, T. A. and Talacko, J. V. (1968). Equivalence of the M.L.E. to a minimum entropy estimator. *Trb. Estadist.* **19**, I/II, 55–65.

Kulldorf, G. (1957). On the conditions for consistency and asymptotic efficiency of maximum likelihood estimates. *Skand. Aktuarietidskrift*, **40**, 129.

Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of the M.L.E.'s. *Ann. Math. Statist.* **41** (3), 802–828.

→ Linnik, Yu V. and Mitroafanova, N. M. (1965). Some asymptotic expansions for the distribution of the maximum likelihood estimate. *Sankhyā*, A, **27**, 73–82.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1.

Pakshirajan, R. P. (1963). On the solution of the maximum likelihood equation. *J. Indian Statist. Ass.* **1**, 196–201.

Raja Rao, B. (1960). A formula for the curvature of the likelihood surface of a sample drawn from a distribution admitting sufficient statistics. *Biometrika*, **47**, 203–207.

Rao, C. R. (1947). Minimum variance and the estimation of several parameters. *Proc. Cambridge Phil. Soc.* **43**, 280–283.

Rao, C. R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, **18**, 139–148.

Rao, C. R. (1958). Maximum likelihood estimation for the multinomial distribution with an infinite number of cells. *Sankhyā*, **20**, 211–218.

Rao, C. R. (1961*a*). Asymptotic efficiency and limiting information. *Proc. 4th Berkeley Symp. Math. Statist. Prob.* 531–545.

→ Rao, C. R. (1962*a*). Apparent anomalies and irregularities in maximum likelihood estimation (with discussion). *Sankhyā*, A, **24**, 73-101.

Rao, C. R. (1962*b*). Efficient estimates and optimum inference procedures in large samples. *J. R. Statist. Soc.*, B, **24**, 46–72.

→ Rao, C. R. (1963). Criteria of estimation in large samples. *Sankhyā*, **25**, 189–206.

*Richards, F. S. G. (1961). A method of M.L.E. *J. R. Statist. Soc.*, B, **23**, 469–475.

Roussas, G. G. (1965). Extension to Markov processes of a result by A. Wald about the consistency of the maximum likelihood estimate. *Zeit. Wahrscheinlichkeitsth.* **4**, 69–73.

Silvey, S. D. (1961). A note on M.L. in the case of dependent random variables. *J. R. Statist. Soc.*, **B**, **23**, 444–452.

Sprott, D. A. and Kalbfleisch, J. D. (1965). Use of the likelihood function in inference. *Psychological Bulletin*, **64**, 15–22.

*Sprott, D. A. and Kalbfleisch, J. D. (1969). Examples of likelihoods and comparisons with point estimates and large sample approximations. *J. Amer. Statist. Ass.* **64**, 468–484.

Wald, A. (1949). Note on the consistency of maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595–601.

Wedgeman, E. J. (1970). M.L.E. of a unimodal density function. *Ann. Math. Statist.* **41**, (2), 457–471.

Weiss, L. (1963). The relative maxima of the likelihood function. *Skand. Aktuartidskr.* **46**, 162–166.

Weiss, L. (1966). The relative maxima of the likelihood function II. *Skand. Aktuartidskr.* **49**, 119–121.

Weiss, L. (1968). Approximating maximum likelihood estimators based on bounded random variables. *Naval Res. Logist. Quart.* **15**, 168–177.

Weiss, L. and Wolfowitz, J. (1967). Estimation of a density function at a point. *Zeit. Wahrscheinlichkeitsth.* **7**, 327–335.

Welch, B. L. (1965). On comparisons between confidence point procedures in the case of a single parameter. *J. R. Statist. Soc.*, B, **27**, 1–8.

Wilks, S. S. (1938). Shortest average confidence intervals from large samples. *Ann. Math. Statist.* **9**, 166–175.

Wolfowitz, J. (1949). On Wald's proof of the consistency of the maximum likelihood function. *Ann. Math. Statist.* **20**, 601–602.

Wolfowitz, J. (1953). The method of maximum likelihood and the Wald theory of decision functions. *Proc. Royal Dutch Acad. Sciences*, Series A, **56**, 114–119.

Wolfowitz, J. (1965). Asymptotic efficiency of the maximum likelihood estimator. *Teor. Vesoyatnost Primen.* **10**, 267–281.

Zehna, P. W. (1966). Invariance of maximum likelihood estimates. *Ann. Math. Statist.* **37**, 744.

Group III

Alma, D. (1965). An efficiency comparison between two different estimators for the fraction p of a normal distribution falling below a fixed limit (in Dutch). *Statist. Neerlandica*, **19**, 81–91.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Ass.* **52**, 200–203.

Anscombe, F. J. (1964). Normal likelihood functions. *Ann. Inst. Statist. Math.* **26**, 1–19.

Arato, M. (1962). Some remarks to the notion of I-divergence. *Magy. Tud. Akad. III Oszt. Kozl.* **12**, 325–327.

*Arato, M. (1968). Unbiased estimates of the complex stationary Gauss-Markov process. Approximate distribution functions. *Studia Sci. Math. Hung.* **3**, 153–158.

Barnett, V. D. (1966). Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, **53, 151–166.

Barnett, V. D. (1967). A note on linear structural relationships when both residual variances are known. *Biometrika*, **54**, 670–672.

Batholomew, D. J. (1957). A problem in life testing. *J. Amer. Statist. Ass.* **52**, 350–357.

Berger, Agnes and Gold, Ruth Z. (1967). On estimating recessive frequencies from truncated samples. *Biometrics*, **23**, 356–360.

*Bhattacharya, N. (1959). An extension to Hold's table for the one-sided censored normal distribution. *Sankhyā*, **21**, 377–380.

Blischke, W. R., Glinski, A. M., Johns, M. V., Mundle, P. B. and Truelove, A. J. (1965). On non-regular estimation, minimum variance bounds and the Pearson Type III distribution. *Arl. Tech. Rep.* Aerospace Res. Lab. Wright-Patterson A.F. Base No. 65-177.

*Blischke, W. R., Mundle, P. B., Johns, M. V. and Truelove, A. J. (1968*a*). Further results on estimation of the parameters of the Pearson Type III and Weibull distributions in the non-regular case. *Arl. Tech. Rep.* Aerospace Res. Lab. Wright-Patterson A.F. Base.

*Bowman, K. O. and Shenton, L. R. (1970). Properties of M.L.E. for the parameter of the log series distribution. *Random Count in Scientific Work*, **1**, 127–150.

→ Brillinger, D. R. (1964). The asymptotic behaviour of Tukey's general method of setting approximate confidence limits ("The Jacknife") when applied to M.L.E. *Rev. Int. Statist. Inst.* **32**, 202–206.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26**, 607–616.

Brunk, H. D., Franck, W. E., Hanson, D. L. and Hogg, R. V. (1966). Maximum likelihood estimation of the distribution of two stochastically ordered random variables. *J. Amer. Statist. Ass.* **61**, 1067–1080.

Brunk, H. D., Ewiing, G. M. and Reid, W. T. (1954). The minimum of a certain definite integral suggested by the M.L.E. *Bull. Amer. Math. Soc.* **60**, 535.

Cannings, C. J. and Smith, C. A. B. (1966). A note on the estimation of gene frequencies by maximum likelihood in systems similar to Rhesus. *Ann. Hum. Genet.* **29**, 277–279.

Chan, L. K. (1967). Remark on the linearised M.L.E. *Ann. Math. Statist.* **38**, 1876–1881.

Chapman, D. G. and Douglas, G. (1956). Estimating the parameters of a truncated gamma distribution. *Ann. Math. Statist.* **27**, 498–506.

Chernoff, H. L. and Lehmann, E. L. (1954). The use of the maximum likelihood in the χ^2 test for goodness of fit. *Ann. Math. Statist.* **25**, 579–586.

*Choi, S. C. and Wette, R. (1969). M.L.E. of the parameters of the gamma distribution and their bias. *Technometrics*, **11**, 683–690.

Clutton-brock, M. (1965). Using the observation to estimate the prior distribution. *J. R. Statist. Soc.*, **B**, **27**, 17–27.

Cohen, A. C. (1950*a*). Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *Ann. Math. Statist.* **21**, 557–569.

Cohen, A. C. (1950*b*). Estimating parameters of Pearson Type III populations from truncated samples. *J. Amer. Statist. Ass.* **45**, 411–423.

Cohen, A. C. (1959). Simplified estimators for the normal distributions when samples are singly censored or truncated. *Technometric*, **1**, 217–237.

Cohen, A. C. (1960*a*). Estimating the parameter in a conditional Poisson distribution. *Biometrics*, **16**, 203–211.

→ Cohen, A. C. (1960*b*). Estimating the parameters of a modified Poisson distribution. *J. Amer. Statist. Assn.* **55**, 139–143.

Cohen, A. C. (1960*c*). Estimation in the truncated Poisson distribution when zeros and some ones are missing. *J. Amer. Statist. Ass.* **55**, 342–348.

Cohen, A. C. (1960*d*). Simplified estimators for the normal distribution when samples are singly censored or truncated. *Bull. Int. Statist. Inst.* **37**, III, 251–269.

Cohen, A. C. (1960*e*). Estimation in the Poisson distribution when sample values of C+1 are sometimes erroneously reported as C. *Ann. Inst. Statist. Math. Tokyo*, **11**, 189–193.

*Cohen, A. C. (1965). M.L.E. in the Weibull distribution based on complete and on censored samples. *Technometrics*, **7**, 579–588.

Curnow, R. N. (1961). The estimation of repeatability and heritability from records subject to culling. *Biometrics*, **17**, 553–566.

*Das Gupta, P. (1964). On the estimation of the total number of events and of the probabilities of deleting an event from information supplied to several agencies. Bull. *Calcutta Statist. Ass.* **13**, 89–100.

David, F. N. and Johnson, N. L. (1952). The truncated Poisson. *Biometrics*, **8, 175–185.

*Davidson, R. R. and Bradley, R. A. (1969). Multivariate paired comparisons: the extension of a univariate model and associated estimation and test procedures. *Biometrika*, **46**, 81–95.

*Day, N. E. (1969). Estimating the components of a mixture of normal populations. *Biometrika*, **56**, 463–474.

*Day, N. E. and Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics*, **23**, 313–325.

*Dorfman, D. D. and Alf, E. Jr. (1968). Maximum likelihood estimation of parameters of signal detection theory. *Psychometrika*, **33**, 117–124.

Dorogovtzev, A. I. (1959). Confidence intervals in estimation of parameters. *Dopov. Acad. Nank Ukrain. S.S.R.* 355–358.

Dubey, S. D. (1963). On some statistical inferences for Weibull laws. *J. Amer. Statist. Ass.* **58**, 549.

Dubey, S. D. (1967). Monte Carlo study of the moment and maximum likelihood estimators of Weibull parameters. *Trab. Estadist.* **18**, II/III, 131–141.

Edgett, G. (1956). Multiple regression with missing observations among the independent variables. *J. Amer. Statist. Ass.* **51**, 122–31.

Feigl, Polly and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**, 826–838.

Ferguson, T. (1958). A method of generating BAN estimates with application to the estimation of bacterial densities. *Ann. Math. Statist.* **29**, 1046–1062.

Fields, R. I., Kramer, C. Y. and Clunies-Ross, C. W. (1962). Joint estimation of the parameters of two normal populations. *J. Amer. Statist. Ass.* **57, 446–454.

Finney, D. J. (1949). The truncated binomial distribution. *Ann. Eugen.* **14, 319–328.

Finney, D. J. and Varley, G. C. (1955). An example of the truncated Poisson distribution. *Biometrics*, **11, 387–394.

Fisher, F. M. (1967). Approximate specification and the choice of a *k*-class estimator. *J. Amer. Statist. Ass.* **62**, 1265–1276.

Fisk, P. R. (1961). Estimation of location and scale parameters in a truncated grouped sech square distribution. *J. Amer. Statist. Ass.* **56, 692–702.

Frome, E. L. and Beauchamp, J. J. (1968). Maximum likelihood estimation of survival curve parameters. *Biometrics*, **24**, 595–605.

*Gajjar, A. V. and Khatri, C. G. (1969). Progressively censored samples from log-normal and logistic distributions. *Technometrics*, **11**, 793–803.

Ghosh, J. K. and Rajinder Singh (1966). Unbiased estimation of location and scale parameters. *Ann. Math. Statist.* **37**, 1671–1675.

*Glaser, M. (1967). Exponential survival with covariance. *J. Amer. Statist. Ass.* **62**, 561–68.

*Gnadadesikan, R., Pinkham, R. S. and Hughes, Laura P. (1967). M.L.E. of the parameters of the beta distribution from smallest order statistics. *Technometrics*, **9**, 607–620.

*Greenwood, J. and Durand, D. (1960). Aids for fitting the gamma distribution by M.L. *Technometrics*, **2**, 55–65.

Hald, A. (1949). M.L.E. of the parameters of a normal distribution which is truncated at a known point. *Skand. Aktuariedskrift.* **32**, 119–134.

Haldane, J. B. S. (1941). The fitting of the Binomial distribution. *Ann. Eng.* **11, 179–181.

Halperin, M. (1952). M.L.E. in truncation samples. *Ann. Math. Statist.* **23**, 226–238.

Hannan, J. (1960). Consistency of M.L.E. of discrete distributions. *Contributions to Prob. and Statist.*, pp. 249–257. Stanford U.P.

Hannan, J. F. and Tate, R. F. (1965). Estimation of the parameters of a multivariate normal distribution when one variable is dichotomised. *Biometrika*, **52**, 664–668.

*Harter, H. L. (1966). Asymptotic variances and covariances of maximum likelihood estimators, from censored samples, of the parameters of a four parameter generalised gamma population. *Arl. Tech. Rep.* Aerospace Res. Lab. Wright-Patterson A.F. Base, No. 66-0158, iv + 147 pages.

*Harter, H. L. (1967). Maximum likelihood estimation of the parameters of a four parameter generalised gamma population from complete and censored examples. *Technometrics*, **9**, 159–165.

*Harter, H. L. and Moore, A. H. (1965). Maximum likelihood estimation of the parameters of gamma and Weibull populations from complete and censored samples. *Technometrics*, **7**, 639-643.

Harter, H. L. and Moore, A. H. (1966*a*). Iterative maximum likelihood estimation of the parameters of normal populations from singly and doubly censored samples. *Biometrika*, **53**, 205–213.

*Harter, H. L. and Moore, A. H. (1966*b*). Local maximum likelihood estimation of the parameters of three parameter lognormal populations from complete and censored samples. *J. Amer. Statist. Ass.* **61**, 842–851.

Harter, H. L. and Moore, A. H. (1967*a*). Asymptotic variances and covariances of maximum likelihood estimators, from censored samples, of the parameters of Weibull and gamma populations. *Ann. Math. Statist.* **38**, 557–571.

Harter, H. L. and Moore, A. H. (1967*b*). A note on estimation from a type I extreme-value distribution. *Technometrics*, **9**, 325–331.

Harter, H. L. and Moore, A. H. (1967*c*). Maximum likelihood estimation, from censored samples, of the parameters of a logistic distribution. *J. Amer. Statist. Ass.* **62**, 675–684.

*Harter, H. L. and Moore, A. H. (1968*a*). M.L.E. from doubly censored samples, of the parameters of the first asymptotic distribution of extreme values. *J. Amer. Statist. Ass.* **63**, 889–901.

*Harter, H. L. and Moore, A. H. (1968*b*). Conditional M.L.E. from singly censored samples, of the scale parameters of type II extreme-value distributions. *Technometrics*, **10**, 349–359.

Hartley, H. O. and Rao, J. W. K. (1967). M.L.E. for the mixed analysis of variance model. *Biometrika*, **54**, 93–108.

Hasselblad, V. (1969). Estimation of finite mixtures of distribution from the exponential family. *J. Amer. Statist. Ass.* **64**, 1459–1471.

Hildreth, C. (1969). Asymptotic distribution of M.L.E.'s in a linear model with autoregression disturbances. *Ann. Math. Statist.* **40** (2), 583-594.

Hill, B. M. (1963). The three parameter lognormal distribution and Bayesian analysis of a point source epidemic. *J. Amer. Statist. Ass.* **58**, 72–84.

*Hocking, R. R. and Smith, W. B. (1968). Estimation of parameters in the multivariate normal distribution with missing observations. *J. Amer. Statist. Ass.* **63**, 159–173.

*Holgate, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrika*, **51**, 241–245.

*Irwin, J. O. (1959). On the estimation of the mean of a Poisson distribution from a sample with the zero class missing. *Biometrics*, **15**, 324–326.

*Isida, M. and Tagami, S. (1959). The bias and precision in the M.L.E. of the parameters of normal populations from singly truncated samples. *Rep. Statist. Appl. Res.* (*JUSE*), **6**, 105–110.

Iyer, P. V. K. and Sing, N. (1962). Estimation of parameters from generalised censored normal samples. *J. Indian Soc. Agric. Statist.* **14**, 165–176.

*Jacquez, J. A., Mather, F. J. and Crawford, C. R. (1968). Linear regression with non-constant, unknown error variances, sampling experiments with least squares, weighted least squares and maximum likelihood estimators. *Biometrics*, **24**, 607–626.

Jogdeo, K. (1967). Monotone convergence of binomial probabilities with an application to maximum likelihood estimation. *Ann. Math. Statist.* **38**, 1583–1586.

Jolly, G. M. (1963). Estimates of population parameters from multiple recapture data with both death and dilution – deterministic model. *Biometrika*, **50**, 113–128.

Kale, B. K. (1963). M.L.E. for a truncated exponential family. *J. Indian Statist. Ass.* **1**, 86–90.

Katti, S. K. and Gurland, J. (1962). Efficiency of certain methods of estimation for the negative binomial and the Neyman type A distribution. *Biometrika*, **49**, 215–226.

→ Khan, R. A. (1969). M.L. estimation in sequential experiments. *Sankhyā*, A, **31**, 49–56.

Khatri, C. G. (1962*a*). A simplified method of fitting the doubly and singly truncated negative binomial distribution. *J. University Baroda*, **11, 35–38.

Khatri, C. G. (1962*b*). A method of estimating approximately the parameters of a certain class of distributions. *Ann. Inst. Statist. Math., Tokyo*, **14, 57–62.

*Khatri, C. G. (1963). Joint estimation of the parameters of multivariate normal populations. *J. Indian Statist. Ass.* **1**, 125–133.

Khatri, C. G. and Jaiswall, M. C. (1963). Estimation of parameters of a truncated bivariate normal distribution. *J. Amer. Statist. Ass.* **58**, 519–526.

Klotz, J. and Putter, J. (1969). M.L.E. of multivariate covariance components for the balanced one-way lay-out. *Ann. Math. Statist.* **40** (3), 1100–1105.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, **28**, 221–238.

Kulldorf, G. (1959). A problem of M.L.E. from a grouped sample. *Metrika*, **2**, 94–99.

Lee, T. C., Judge, G. G. and Zellner, A. (1968). M.L. and Bayesian estimation of transition probabilities. *J. Amer. Statist. Ass.* **63**, 1162–1179.

Lejeune, J. (1958). On an "a priori" solution of Haldane's "a posteriori" method. *Biometrics*, **14**, 513–520.

Lellough, J. and Wambersie, A. (1966). Maximum likelihood estimation of survival curves of irradiated microorganisms (in French). *Biometrics*, **22**, 673–683.

Li, C. C. and Mantel, N. (1968). A simple method of estimating the segregation ratio under complete ascertainment. *Amer. J. Hum. Genet.* **20**, 61–81.

Lloyd, D. E. (1959). Note on a problem of estimation. *Biometrika*, **46**, 231–235.

Lord, F. M. (1955*a*). Estimation of parameters from incomplete data. *J. Amer. Statist. Ass.* **50**, 870–876.

Lord, F. M. (1955*b*). Equating test scores a M.L. solution. *Psychometrika*, **20**, 193–200.

Malik, H. J. (1968). Estimation of the parameters of the power function population. *Metron*, **27**, 196–203.

Mantel, N. (1967). Adaption of Karber's method for estimating the exponential parameter from quantal data, and its relationship to birth, death and branching processes. *Biometrics*, **23**, 739–746.

Maritz, J. S. (1966). Smooth empirical Bayes estimation for one-parameter discrete distributions. *Biometrika*, **53**, 417–429.

Marshall, A. W. and Proschan, F. (1965). Maximum likelihood estimation for distributions with monotone failure rate. *Ann. Math. Statist.* **36**, 69–77.

Matthai, A. (1951). Estimate of parameters from incomplete data with applications to design of sample surveys. *Sankhyā*, **11**, 145–152.

Mead, R. (1967). A mathematical model for the estimation of interplant competition. *Biometrics*, **23**, 189–206.

*Mendenhall, W. and Hader, R. J. (1958). Estimation of parameters of mixed exponentially distributed failure-time distribution from censored life test data. *Biometrika*, **45**, 504–520.

Mitchell, Ann F. S. (1968). Exponential regression with correlated observations. *Biometrika*, **55**, 149–162.

Moore, A. H. and Harter, H. L. (1969). Conditional M.L.E. from singly censored samples, of the shape parameters of Pareto and limited distributions. *IEEE Trans. Rel.* **R-18**, 76–78.

Morgan, R. W. (1965). The estimation of parameters from the spread of a disease by considering households of two. *Biometrika*, **52**, 271–274.

→ Mukherji, V. (1967). A note on maximum likelihood – a generalisation. *Sankhyā*, A, **29**, 105–106.

*Mustafi, C. K. (1963). Estimation of parameters of the extreme value distribution with limited type of primary probability distribution. *Bull. Calcutta Statist. Ass.* **12**, 47–54.

Nelson, A. C. Jr., Williams, J. S. and Fletcher, N. T. (1963). Estimation of the probability of defection failure from destructive tests. *Technometrics*, **5**, 459–468.

Ohlsen, Sally (1964). On estimating edpidemic parameters from household data. *Biometrika*, **51**, 511–512.

Pal, L. and Kiddi Eva. (1965). On some statistical problems connected with the measuring of the total cross section of thermal neutrons. *Pub. Res. Inst. Physics Hungarian Acad. Sci.* **11**, 3–19.

*Parker, R. A. (1963). On the estimation of population size, mortality and recruitment. *Biometrics*, **19**, 318–328.

*Patil, G. P. (1962). M.L.E. for generalised power series distribution and its application to truncated binomial distribution. *Biometrika*, **49**, 227–238.

Paulik, G. J. (1963). Estimates of mortality rates from tag recoveries. *Biometrics*, **19**, 28–57.

*Pielou, E. C. (1963). The distribution of diseased trees with respect to healthy ones in a patchily infected forest. *Biometrics*, **19**, 450–459.

Pike, M. C. (1966). A suggested method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, **22**, 142–161.

Prakasa Rao, B. L. S. (1968). Estimation of the location of the cusp of a continuous density. *Ann. Math. Statist.* **39**, 76–87.

→ Prakasa Rao, B. L. S. (1969). Estimation of unimodal density. *Sankhyā*, A, **31**, 23–36.

Press, S. J. (1968). Estimating from mis-classified data. *J. Amer. Statist. Ass.* **63**, 123–133.

Prince, B. M. and Tate, R. F. (1966). Accuracy of M.L.E.'s of correlation for a biserial model. *Psychometrika*, **31**, 85–92.

Quandt, R. E. (1966). Old and new methods of estimation and the Pareto distribution. *Metrika*, **10**, 55–82.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cambridge Phil. Soc.* **44**, 50–57.

Remage, R. Jr. and Thompson, W. A. Jr. (1966). Maximum likelihood paired comparison rankings. *Biometrika*, **53**, 143–149.

Riffenburgh, R. H. (1966). On growth parameter estimation for early life stages. *Biometrics*, **22**, 162–178.

*Rutemiller, H. C. (1967). Estimation of the probability of zero failures in *m* binomial trials. *J. Amer. Statist. Ass.* **62**, 272–277.

Samuel, Ester (1968). Sequential maximum likelihood estimation of the size of a population. *Ann. Math. Statist.* **39**, 1057–1068.

Samuel, Ester (1969). Comparison for sequential rules for estimation of the size of a population. *Biometrics*, **25**, 517–527.

*Saw, J. G. (1961). The bias of the M.L.E.'s of the location and scale parameters given a type II censored sample. *Biometrika*, **48**, 448–451.

*Saw, J. G. (1962). Efficient moment estimators when the variables are dependent with special reference to type II censored normal data. *Biometrika*, **49**, 155–161.

Seber, G. A. F. and Le Cren, E. D. (1967). Estimating population parameters from catches large relative to the population. *J. Animal Ecol.* **36**, 631–643.

Shah, S. M. and Jaiswall, M. C. (1966). Estimation of parameters of doubly truncated normal distribution from first four sample moments. *Ann. Inst. Statist. Math., Tokyo*, **18**, 107–111.

Shenton, L. R. (1958). Moment estimators and M.L. *Biometrika*, **45**, 411–420.

Shenton, L. R. (1963). A note on bounds for the asymptotic sampling variance of the maximum likelihood estimator of a parameter in the negative binomial distribution. *Ann. Inst. Statist. Math. Tokyo*, **15**, 145–151.

Shenton, L. R. and Wallington, P. A. (1962). The bias of moment estimators with an application to the negative binomial distribution. *Biometrika*, **49**, 193–204.

Sheps, M. C. (1966). Characteristics of a ratio used to estimate failure rates: occurrence per person year of exposure. *Biometrics*, **22**, 310–321.

Sibuya, M. (1963). Randomised unbiased estimation of restricted parameters. *Ann. Inst. Statist. Math., Tokyo*, **15**, 61–66.

Singh, S. N. (1963). A note on the inflated Poisson distribution. *J. Indian Statist. Ass.* **1**, 140–144.

Skibinsky, M. and Cote, L. (1963). On the inadmissability of some standard estimates in the presence of prior information. *Ann. Math. Statist.* **34**, 539–548.

*Smith, B. W. (1966). Parameter estimation in the presence of measurement noise. *Int. J. Control*, **3**, 297–312.

Smith, W. E. (1966). An "a posteriori" probability method for solving an over-determined system of equations. *Technometrics*, **8**, 675–686.

Sprott, D. A. (1965*a*). Statistical estimation – some approaches and controversies. *Statist. Hefte.* **6**, 97–111.

Sprott, D. A. (1965*b*). A class of contagious distributions; and maximum likelihood estimation. *Sankyā*, A, **27**, 369–382.

Sverdrup, E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. *Skand. Aktuarietidskr.* **48**, 184–211.

*Swamy, P. S. (1960). Estimating the mean variance of a normal distribution from singly and doubly truncated samples of grouped observations. *Bull. Calcutta Statist. Ass.* **9**, 145–156.

→ *Swamy, P. S. and Doss, S. A. D. (1961). On a problem of Bartholomew in life testing. *Sankhyā*, A, **23**, 225–230.

*Swan, A. V. (1969). Computing M.L.E.'s for parameters of the normal distribution from grouped and censored data. *Appl. Statist.* **18**, 65–69.

*Tallis, G. M. and Light, R. (1968). The use of fractional moments for estimating the parameters of a mixed exponential distribution. *Technometrics*, **10**, 161–175.

*Tallis, G. M. and Young, S. S. Y. (1962). M.L.E. of parameters of the normal, log normal, truncated normal and bivariate normal distribution from grouped data. *Aust. J. Statist.* **4**, 49–54.

Teicher, H. (1961). Maximum likelihood characterisation of distributions. *Ann. Math. Statist.* **32**, 1214–1222.

Thomason, R. L. and Kapadia, C. H. (1968). On estimating the parameter of a truncated geometric distribution. *Ann. Inst. Statist. Math., Tokyo*, **20**, 519–523.

*Thoni, H. (1969). A table for estimating the mean of a lognormal distribution. *J. Amer. Statist. Ass.* **64**, 632–636.

Tiku, M. L. (1967*a*). Estimating the mean and standard deviation from a censored normal sample. *Biometrika*, **54**, 155–165.

Tiku, M. L. (1967*b*). A note on estimating the location and scale parameters of the exponential distribution from a censored sample. *Aust. J. Statist.* **9**, 49–54.

*Tiku, M. L. (1968). Estimating the mean and standard deviation from progressively censored normal samples. *J. Indian Soc. Agric. Statist.* **20**, 1, 20–25.

Trawinski, Irene M. and Bargmann, R. E. (1964). Maximum likelihood estimation with incomplete multivariate data. *Ann. Math. Statist.* **35**, 647–657.

Villegas, C. (1961). Maximum likelihood estimation of a linear functional relationship. *Ann. Math. Statist.* **32**, 1048–1062.

*Villegas, C. (1963). On the L.S.E. of a linear relation. *Publ. Inst. Mat. Estadist. Fac. Ingen. Agric., Montivideo*, **3**, 189–204.

Wald, A. (1948). Asymptotic properties of the M.L.E. of an unknown parameter of a discrete stochastic process. *Ann. Math. Statist.* **19**, 40–46.

→ Watson, G. S. (1964). A note on maximum likelihood. *Sankhyā*, A, **26**, 303–304.

Weibull, C. (1963*a*). Maximum likelihood estimation from truncated, censored and grouped samples. *Skand. Aktuartiedskr.* **46**, 70–77.

Wilk, M. B., Gnanadesikan, R. and Hyett, Marilyn J. (1962). Estimation of parameters of the gamma distribution using order statistics. *Biometrika*, **49, 525–545.

Wilk, M. B., Gnanadesikan, R. and Hyett, Marilyn J. (1963). Separate M.L.E. of scale or shape parameters of the gamma distribution using order statistics. *Biometrika*, **50**, 217–221.

Wilk, M. B., Gnanadesikan, R. and Laugh, Elizabeth (1966). Scale parameter estimation from the order statistics of unequal gamma components. *Ann. Math. Statist.* **37**, 152–176.

Wilks, S. S. (1932). Moments and distributions of estimating of population parameters from fragmentary samples. *Ann. Math. Statist.* **3**, 163–195.

Wyshak, G. and White, C. (1968). Estimation of parameters of a truncated Poissonian binomial. *Biometrics*, **24**, 377–388.

Zacks, S. (1966). Sequential estimation of the mean of a log-normal distribution having a prescribed proportional closeness. *Ann. Math. Statist.* **37**, 1688–1696.

Zacks, S. and Even, M. (1966*a*). The efficiencies in small samples of the maximum likelihood and best unbiased estimation of reliability function. *J. Amer. Statist. Ass.* **61**, 1033.

Zacks, S. and Even, M. (1966*b*). Minimum variance unbiased and M.L.E.'s of reliability functions for systems in series and in parallel. *J. Amer. Statist. Ass.* **61**, 1052.

Zippin, C. and Armitage, P. (1966). Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics*, **22**, 665–672.

Group IV

*Amemiya, T. and Fuller, W. A. (1967). A comparative study of alternative estimators in a distributed log model. *Econometrica*, **35**, 509–529.

*Arnold, B. C. (1968). Parameter estimation for a multivariate exponential distribution. *J. Amer. Statist. Ass.* **63**, 848–852.

*Bain, L. J. and Antle, C. E. (1967). Estimation of parameters in the Weibull distribution. *Technometrics*, **9**, 621–627.

Berkson, J. (1955). M.L. and minimum χ^2 estimates of the logistic function. *J. Amer. Statist. Ass.* **50**, 130–162.

Berkson, J. (1960). Problems recently discussed regarding estimating the logistic curve. *Bull. Int. Statist. Inst.* **37**, III, 207–211.

Berkson, J. and Elveback, Lillian (1960). Computing exponential risks with particular reference to the study of smoking and lung cancer. *J. Amer. Statist. Ass.* **55**, 415–428.

Birnbaum, A. (1961). A unified theory of estimation. *Ann. Math. Statist.* **32**, 112–135.

Birnbaum, A. (1962). Intrinsic confidence methods. *Bull. Int. Statist. Inst.* **39**, II, 376–383.

*Blischke, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. *J. Amer. Statist. Ass.* **59**, 510–528.

*Blischke, W. R., Mundle, P. B. and Johns, P. B. (1968*b*). Further results on estimation of the parameters of the Pearson type III and Weibull distributions in the non-regular case. *Arl. Tech. Rep.* Aerospace Res. Lab. Wright-Patterson A.F. Base, No. 68-0207.

*Blumenthal, S. and Cohen, A. (1968). Estimation of the larger of two normal means. *J. Amer. Statist. Ass.* **63**, 861–876.

Chakraborty, P. N. (1963). On a method of estimating birth and death rates from several agencies. *Bull. Calcutta Statist. Ass.* **12**, 106–112.

*Cohen, A. (1967). Estimation in mixtures of two normal distributions. *Technometrics*, **9**, 15–28.

*Cornell, R. G. and Speckman, J. A. (1967). Estimation for a simple exponential model. *Biometrics*, **23**, 717–737.

*Cragg, J. G. (1967). On the relative small-sample properties of several structural-equation estimations. *Econometrica*, **35**, 89–110.

Doss, S. A. D. (1962*a*). On the efficiency of BAN estimates of the parameters of normal populations based on singly censored samples. *Biometrika*, **49**, 570–573.

→ Doss, S. A. D. (1962*b*). On uniqueness and maxima of the roots of likelihood equations under truncated and censored sampling from normal populations. *Sankhyā*, A, **24**, 355–362.

Dubey, S. D. (1965*a*). Asymptotic properties of several estimators of Weibull parameters. *Technometrics*, **7**, 423–434.

Dubey, S. D. (1965*b*). M.L. and Bayes solutions for the true score. *J. Indian Statist. Ass.* **3**, 65–81.

Dubey, S. D. (1966). Comparative performance of several estimators of Weibull parameters. *Ann. Tech. Conf. Trans., Amer. Soc. Qual. Control*, **20**, 723–735.

Epstein, B. (1961). Estimates of bounded relative error for the mean life of an exponential distribution. *Technometrics*, **1**, 107–109.

Fraser, D. A. S. (1964). On local unbiased estimation. *J. R. Statist. Soc.*, B, **26**, 46–51.

Ghirtis, G. C. (1967). Some problems of statistical inference relating to the double-gamma distribution. *Trab. Estadist.* **18**, II/III, 67–87.

Glasser, G. J. (1962). Minimum variance unbiased estimators for Poisson probabilities. *Technometrics*, **4**, 409–418.

*Graybill, F. A. and Cornell, T. L. (1964). Sample size required to estimate the parameter in the uniform density within *d* units of the true value. *J. Amer. Statist. Ass.* **59**, 550–556.

Gumbell, E. J. (1965). A quick estimation of the parameters in Frechets distributions. *Rev. Int. Statist. Inst.* **33**, 349–363.

Haldane, J. B. S. (1951). A class of efficient estimates of a parameter. *Bull. Int. Statist. Inst.* **33**, 231.

*Haldane, J. B. S. and Smith, C. A. B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Ann. of Eugenics*, **14**, 10–31.

Halperin, M. (1961). Almost linearly-optimum combination of unbiased estimates. *J. Amer. Statist. Ass.* **56**, 36–43.

Harter, H. L. (1968). The use of order statistics in estimation. *Oper. Res.* **16**, 783–798.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Huber, P. J. (1968). Robust estimation. *Math. Centre Tracts. Selected Statistical Papers*, **27**, 3–25.

Jaiswal, M. C. and Khatri, C. G. (1967). Estimation of parameters for the selected samples from bivariate normal populations. *Metron*, **26**, 326–333.

Jorgenson, D. W. (1961). Multiple regression analysis of a Poisson process. *J. Amer. Statist. Ass.* **56**, 235–245.

Katti, S. K. (1962). Use of some "a priori" knowledge in the estimation of means from double samples. *Biometrics*, **18**, 139–147.

Katti, S. K. and Gurland, J. (1962). Some methods of estimation for the Poisson binomial distribution. *Biometrics*, **18**, 42–51.

Kharshikar, A. V. (1968). On information from folded distributions. *J. Indian Soc. Agric. Statist.* **20**, I, 49–54.

Lindley, D. V. and El-Sayyad, G. M. (1968). The Bayesian estimation of a linear functional relationship. *J. R. Statist. Soc.*, B, **30**, 190–202.

Miller, R. G. (1960). Early failures in life testing. *J. Amer. Statist. Ass.* **55**, 491–502.

*Molenaar, W. (1965). Survey of estimation methods for a mixture of two normal distributions. *Statist. Neerlandica*, **19**, 249–263.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–1076.

Posner, E. C., Rodemich, E. R., Ashlock, J. C. and Lurie, Sandra (1969). Application of an estimator of high efficiency in bivariate extreme-value theory. *J. Amer. Statist. Ass.* **64**, 1403–1414.

Reiersøl, O. (1961). Linear and non-linear multiple comparisons in logit analysis. *Biometrika*, **48**, 359–365.

Roberts, H. V. (1967). Information stopping rules and inference about population size. *J. Amer. Statist. Ass.* **62**, 763–775.

Rothenberg, T. J., Fisher, F. M. and Tilanus, C. B. (1964). A note on estimation from a Cauchy sample. *J. Amer. Statist. Ass.* **59**, 460–463.

Saw, J. G. (1961). Estimation of the normal population parameters given a type I censored sample. *Biometrika*, **48**, 367–377.

Silverstone, H. (1957). Estimating the logistic curve. *J. Amer. Statist. Ass.* **52**, 567.

Smith, J. H. (1947). Estimation of linear functions of cell proportions. *Ann. Math. Statist.* **18**, 231–254.

Snoeck, N. (1961). The construction of asymptotically exhaustive estimators. *C. R. Acad. Sci. Paris*, **253**, 777–779.

*Soong, T. T. (1969). An extension of the moment method in statistical estimation. *Siam. J. Appl. Math.* **17**, 560–568.

Stormer, H. (1961). On a test for the parameters of a life distribution. *Metrika*, **4**, 63-77.

Theil, H. and Schweitzer, A. (1961). The best quadratic estimator of the residual variance in regression analysis. *Statist. Neerlandica*, **15**, 19–23.

*Tiku, M. L. (1968*a*). Estimating the parameters of log-normal distributions from censored samples. *J. Amer. Statist. Ass.* **63**, 134–140.

Tiku, M. L. (1968*b*). Estimating the parameters of normal and logistic distributions from censored samples. *Aust. J. Statist.* **10**, 64–74.

Varde, S. D. (1969). Life testing and reliability estimation for the two parameter exponential distribution. *J. Amer. Statist. Ass.* **64**, 621–631.

Books

Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications,* Volume 2. New York, Wiley.

Kendall, M. G. and Stuart, A. (1961). *The Advanced Theory of Statistics,* Volume 2. London, Griffin.

Additions to Bibliography

Group II

Anderson, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. R. Statist. Soc.,* B., **32**, 283–301.

Barton, D. E. (1956). A class of distributions for which the maximum likelihood estimator is unbiased and of maximum variance for all sample sizes. *Biometrika,* **43**, 200–202.

Bodin, N. A. (1968). On the theory of grouped samples. *Dokl. Akad. Nank. SSSR,* **178**, 17–20; *Soviet Math. Dokl.* **9**, 10–13.

Box, M. J. (1971). Bias in Nonlinear Estimation. *J. R. Statist. Soc.,* B, **33**, 171–201.

Herman, R. J. and Patel, R. K. N. (1971). Maximum likelihood estimation for multi-risk model. *Technometrics,* **13**, 385–396.

Hocking, R. R. and Oxspring, H. H. (1971). Maximum likelihood estimation with incomplete multinomial data. *J. Amer. Statist. Ass.* **66**, 65–70.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. R. Statist. Soc.,* B, **32**, 175–208.

Kale, B. K. (1970). Inadmissibility of the maximum likelihood estimator in the presence of prior information. *Canad. Math. Bull.* **13**, 391–393.

Group III

Antle, C., Klimko, L. and Harkness, W. (1970). Confidence intervals for the parameters of the logistic distribution. *Biometrika,* **57**, 397–402.

Armitage, P. (1970). The combination of assay results. *Biometrika,* **57**, 665–666.

Bairagi, R. (1969). Estimation of the covariance between two sets of values of a variable given at two periods. *Bull. Inst. Statist. Res.* Tv **3, 42–47.

Blight, B. J. N. (1970). Estimation from a censored sample for the exponential family. *Biometrika,* **57**, 389–395.

Bowman, K. and Shenton, L. R. (1965). AEC Research and Development Report K1633, Union Carbide Corporation.

Bowman, K. and Shenton, L. R. (1968). Properties of estimates for the gamma distribution. *Statist. Abstr.* **10**, 401.

Bowman, K. and Shenton, L. R. (1969). Remarks on maximum likelihood estimates for the gamma distribution. ICQC 69 - Tokyo, TS 10–05, 519–522.

Bowman, K. and Shenton, L. R. (1970). Small sample properties of estimators for the gamma distribution. Union Carbide Corporation, Oakridge, Tennessee, Report No. CTC-28.

Easterling, R. G. and Prairie, R. R. (1971). Combining component and system information. *Technometrics,* **13**, 271–280.

Fryer, J. G. and Holt, D. (1970). On the robustness of the standard estimates of the exponential mean to contamination. *Biometrika,* **57**, 641–648.

Garg, M. L., Lao, B. R. and Redmond, Carol K. (1970). Maximum likelihood estimation of the parameters of the Gompertz survival function. *Appl. Statist.* **19**, 152–159.

Haldane, J. B. S. (1953). The estimation of two parameters from a sample. *Sankhyā,* **12**, 313–320.

Haldane, J. B. S. and Smith, S. M. (1956). The sampling distribution of a maximum likelihood estimate. *Biometrika,* **43**, 96–103.

Hartley, H. O. and Rao, J. H. K. (1969). A new estimation theory for sample surveys, II. *New Developments in Survey Sampling,* 147–169.

Heiny, R. L. and Siddiqui, M. M. (1970). Estimation of the parameters of a normal distribution when the mean is restricted to an interval. *Aust. J. Statist.* **12**, 112–117.

Lambert, J. A. (1970). Estimation of parameters in the four-parameter lognormal distribution. *Aust. J. Statist.* **12**, 33–43.

Malik, M. J. (1970). Estimation of the parameters of the Pareto distribution. *Metrika,* **15**, 126–132.

Meyer, P. L. (1967). The maximum likelihood estimate of the non-centrality parameters of a non-central χ^2 variate. *J. Amer. Statist. Ass.* **62**, 1258–1264.

*Oberhofer, W. (1970). Maximum likelihood estimation of the parameters for the multisector models. *Operat. Res. Verfahren,* **8**, 211–219.

Seber, G. A. F. and Whale, J. F. (1970). The removal method for two and three samples. *Biometrics,* **26**, 393–400.

Shenton, L. R. and Bowman, K. (1963). Higher moments of a maximum likelihood estimate. *J. R. Statist. Soc.,* B, **25**, 305–317.

Shenton, L. R. and Bowman, K. (1967). Remarks on large sample estimators for some discrete distributions. *Technometrics*, **9**, 587–598.

→ Shenton, L. R. and Bowman, K. (1969). Maximum likelihood estimator moments for the two parameter gamma distribution. *Sankhyā*, B, **31**, 379–396.

Ta-Chung, Lu and Breen, W. J. (1969). The covariance matrix of the limited information estimator and the identification test. *Econometrica*, **37**, 222–227.

Group IV

Blischke, W. R., Brady, F. J. and Mundle, P. B. (1970). Further results on estimation of the parameters of the Pearson type III distribution in the regular and non-regular cases. *ARL Tech. Rep.* Aerospace Res. Labs., Wright-Patterson Air Force Base, No. 70–0017.

*Griffin, B. S. and Krutchkoff, R. G. (1971). Optimal linear estimates: an empirical Bayses version with application to the binomial distribution. *Biometrika*, **58**, 195–201.

Hamdan, M. A. (1971). Estimation of class boundaries in fitting a normal distribution to a qualitative multinominal distribution. *Biometrics*, **27**, 457–459.

John, S. (1970). On identifying the population of origin of each observation in a mixture of observations from two gamma populations. *Technometrics*, **12**, 565–568.

*Mann, Nancy R. (1970). Estimation of location and scale parameters under various models of censoring and truncation. *ARL Tech. Rep.* Aerospace Res. Labs. Wright-Patterson Air Force Base No. 70-0026.

Oliver, F. R. (1970). Some asymptotic properties of Colquhoun's estimates of a rectangular hyperbola. *Appl. Statist.* **19**, 269–273.

*Rowe, I. M. (1970). A bootstrap method for the statistical estimation of model parameters. *Int. J. Control*, **12**, 721–738.

Rustagi, J. S. and Laitnen, R. (1970). Moment estimation in a Markov-dependent firing distribution. *Operat. Res.* **18**, 918–923.

Book

Silvey, S. D. (1970). *Statistical Inference.* London, Penguin Books, 1970, especially chapter 2.

Résumé

Présenté sous forme d'exposition, ce mémoire (en deux parties) traite du développement de la théorie du Maximum de Vraisemblance (m.l.), depuis son introduction dans les exposés de Fisher (1922 et 1925), jusqu'à l'époque actuelle, où des recherches originales se poursuivent toujours. Après un court préambule, l'auteur donne un aperçu historique dans lequel il démontre notamment que c'est à juste titre que l'on considère Fisher comme l'originateur de cette méthode d'estimation, puisque toutes les méthodes antérieures, bien que similaires en apparence, dépendaient en effet de la méthode de la probabilité inverse.

Le mémoire donne ensuite, en résumé, quelques définitions et résultats fondamentaux, ayant leurs origines dans les exposés de Fisher lui-même et dans les contributions apportées plus tard par MM. Cramer et Rao. La consistance et l'efficience, par rapport au m.l. sont alors considérées dans le détail. Le théorème important de Cramer (1946) est également mentionné, mais avec la remarque que, en ce qui concerne la consistance, la preuve de Wald (1949) est d'une application plus générale.

Viennent ensuite quelques observations sur des exposés subséquents qui s'y rattachent. Le mémoire se poursuit par la discussion du problème de la comparaison des m.l. avec d'autres estimateurs. L'auteur attache une importance toute spéciale au concept de Rao – "l'efficience de second ordre" – étant donné que celle-ci fournit en réalité le moyen de différencier les divers estimateurs BAN ("best asymptotic normal"), dont les m.l. forment une sous-classe. L'auteur continue en résumant le travail fait dans le domaine des multi-paramètres, où il commence par un bref résumé de la théorie Cramer-Rao bien connue. Pour finir, il y a une section traitant des difficultés théoriques que suscite l'existence de m.l. inconsistants et d'estimateurs super-efficients. L'auteur fait une allusion particulière à l'exposé de Rao (1962), dans lequel ces anomalies ont été considérées et résolues, au moins partiellement, par l'utilisation d'autres critères de consistance et d'efficience, lesquels critères ont néanmoins comme source les idées originales de Fisher.

Dans la bibliographie, à la fin du mémoire, se trouvent les titres d'environ quatre cents exposés. Quelques définitions importantes sont données séparément en annexe.

Int. Stat. Rev., Vol. 41, No. 1, 1973, pp. 39–58/*Longman Group Ltd/Printed in Great Britain*

A Survey of Maximum Likelihood Estimation

Part 2

R. H. Norden

University of Bath, England

Contents

Introduction

In Part 2 of this survey the discussion is mainly confined to the problem of comparing maximum likelihood estimation with other estimation procedures and the related problems that arise by reason of the existence of inconsistent maximum likelihood estimates, and of other estimators which are superefficient, with regard to which special reference is made to the work of C. R. Rao (1961*a* and 1962*a*).

Initially, however, there is a further discussion of Cramer's conditions as they relate to consistency and efficiency, and there is also a section on the general multiparameter situation.

All references are contained within the bibliography at the end of Part 1 (*Int. Statist. Rev.*, Vol. 40, No. 3, 1972, pp. 329–354), but a number of important definitions are given separately in the Appendix at the end of this paper.

6. Consistency and Efficiency continued

A very thorough analysis of Cramer's conditions has been carried out by G. Kulldorf (1957), who gives them in the following equivalent way:

C.1. $\dfrac{\partial \log f}{\partial \theta}$[1] exists for every $\theta \in \Theta$ and for almost all x.

C.2. $\dfrac{\partial^2 \log f}{\partial \theta^2}$ exists for every $\theta \in \Theta$ and for almost all x.

C.3. $\dfrac{\partial^3 \log f}{\partial \theta^3}$ exists for every $\theta \in \Theta$ and for almost all x.

C.4. $\displaystyle\int_{-\infty}^{\infty} \frac{\partial f}{\partial \theta}\, dx = 0$ for every $\theta \in \Theta$.

[1] f is an abbreviation for $f(x, \theta)$.

C.5. $\int_{-\infty}^{\infty} \frac{\partial^2 f}{\partial \theta^2} dx = 0$ for every $\theta \in \Theta$.

C.6. $-\infty < \int_{-\infty}^{\infty} \frac{\partial^2 \log f}{\partial \theta^2} f dx < 0$ for every $\theta \in \Theta$.

C.7. There exists a function $H(x)$ such that for all

$$\theta \in \Theta, \left| \frac{\partial^3 \log f}{\partial \theta^3} \right| < H(x) \quad \text{and} \quad \int_{-\infty}^{\infty} H(x) f(x, \theta)\, dx < \infty.$$

He then defines any root of the LE as a MLE in the loose sense, and $\hat{\theta}_n$, as defined in this paper, as the MLE in the strict sense. Also, following Wald, Kulldorf defines an estimator t_n as asymptotically efficient in the strict sense if the limit of the distribution of $\sqrt{n}(t_n - \theta)$ as n tends to infinity, is $N(0, v)$ where

$$v = \left[E \left(\frac{\partial \log f}{\partial \theta} \right)^2 \right]^{-1}.$$

He then shows that conditions 1 to 4, 6 and 7 are sufficient for proving consistency in the loose sense, which is Cramer's result, and that conditions 1 to 7 imply asymptotic efficiency in the strict sense.

He also gives an example of an MLE which is consistent in the loose sense, is asymptotically efficient in the strict sense, and yet, Cramer's conditions are not satisfied. It is

$$f(x, \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-x^2/2\theta}$$

for which

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

but

$$\frac{\partial^3 \log f}{\partial \theta^3} = -\frac{1}{\theta^3} + \frac{3x^2}{\theta^4} \to \infty \text{ as } \theta \to 0,$$

and is not bounded therefore in the open interval $0 < \theta < \infty$. Thus condition 7 is not satisfied.

A new condition, C.8, is then introduced. There exists a function $g(\theta)$ which is positive and twice differentiable for every $\theta \in \Theta$, and a function $H(x)$ such that for every $\theta \in \Theta$

$$\left| \frac{\partial^2}{\partial \theta^2} \left\{ g(\theta) \frac{\partial \log f}{\partial \theta} \right\} \right| < H(x)$$

and

$$\int_{-\infty}^{\infty} H(x) f(x, \theta)\, dx < \infty.$$

Kulldorf's Theorem 2 (Theorem 1 is Cramer's theorem) is then as follows.

(i) Conditions 1 to 4, 6 and 8 imply that there is a MLE in the loose sense which is consistent.
(ii) Conditions 1 to 6 and 8 imply that this root is asymptotically efficient in the strict sense.

Further conditions are introduced as follows:

C.9. $\frac{\partial \log f}{\partial \theta}$ is a continuous function of θ for almost all x.

C.10. There exists a positive function $h(\theta)$ such that

$$\frac{1}{\theta_1-\theta_2}\int_{-\infty}^{\infty}\left(\frac{\partial \log f}{\partial\theta}\right)^2_{\theta=\theta_1} f(x,\theta_2)\,dx<0$$

for every pair θ_1 and θ_2 belonging to Θ and satisfying the inequality

$$0<|\theta_1-\theta_2|<h(\theta_2).$$

C.11. There exists a function $g(\theta)$ which is positive and differentiable for every $\theta\in\Theta$ such that

$$\frac{\partial}{\partial\theta}\left[g(\theta)\frac{\partial f}{\partial\theta}\right]$$

is a continuous function of θ uniformly in x. This leads to Theorem 3. Conditions 1, 9 and 10 imply that consistency in the loose sense holds, and conditions 1, 2, 4, 5, 6, 10 and 11 imply that the corresponding root of the LE is asymptotically efficient in the strict sense.

The importance of this last result in contrast to Theorems 1 and 2 is that $\dfrac{\partial^3 f(x,\theta)}{\partial\theta^3}$ is not assumed to exist, and the consistency property does not even assume the existence of $\dfrac{\partial^2 f(x,\theta)}{\partial\theta^2}$.

To illustrate this theorem consider the density

$$f(x,\theta)=\frac{1}{\sqrt{2\pi\theta}}\,e^{-x^2}/2\theta$$

for which conditions 9 and 10 are clearly satisfied. Also condition 11 is satisfied by taking $g(\theta)=\theta^2$ since then $\dfrac{\partial}{\partial\theta}\left[g(\theta)\dfrac{\partial \log f}{\partial\theta}\right]=-\tfrac{1}{2}$.

Previously Gurland (1954) had attempted to establish similar results, but according to Kulldorf there are a number of errors which invalidate the proofs.

Huber (1967) has also established consistency and asymptotic normality of the MLE under fairly weak conditions. He does not assume that the true distribution belongs to the parametric family which defines the MLE, and the regularity conditions themselves do not involve the second and higher derivatives of f. This is of importance in relation to questions of robustness where it is necessary to establish asymptotic normality under such conditions.

A further question of importance is to what extent does any dependence between the observations affect the asymptotic properties of $\hat{\theta}$. This problem has been studied by Silvey (1961), by Hartley and Rao (1967), for a particular case, and very recently by Bar-Shalom (1971).

Silvey's results which are closely related to Martingale theory are not expressed as explicit conditions on the distribution of the observations, and therefore it is difficult to see how they can be applied directly to actual situations.

More useful results, on the other hand, have been obtained by Bar-Shalom (1971), since he does derive a set of relatively straightforward conditions which imply weak consistency and asymptotic normality. In the first place this requires a reformulation of Cramer's conditions to take dependence into account. Thus the observations $x_1, x_2, \ldots, x_n$ are assumed to have a known joint probability density function $p(x_1, x_2, \ldots, x_n \mid \theta)$ with respect to a σ-finite product measure μ^n $(n=1, 2, 3, \ldots)$. θ_0, the unknown true value of θ, is estimated by $\hat{\theta}=\hat{\theta}_n(x^n)$ obtained by maximizing

$$L_n=p(x^n\mid\theta)=\prod_{i=1}^{n} p_i$$

where $\hat{\theta}_n(x^n) = \hat{\theta}_n(x_1, x_2, \ldots, x_n)$, $p(x^n \mid \theta) = p(x_1, \ldots, x_n \mid \theta)$ and $p_i = p(x_i \mid x^{i-1}, \theta) = \dfrac{L_i}{L_{i-1}}$ i.e. p_i is the conditional probability density function of x_i, given $x_1, x_2, \ldots, x_{i-1}$ and θ.

It is then supposed that the following regularity conditions are satisfied for all i. They are quoted in full.

C.1. $\dfrac{\partial^s \log p_i}{d\theta^s}$ $(s = 1, 2, 3)$ exist for all i and all $\theta \in \Theta$.

C.2. $E\left[\dfrac{\partial}{\partial\theta} \log p_i\right]_{\theta=\theta_0} = 0.$

C.3. $$I_i(\theta_0) = E\left[\left(\frac{\partial}{\partial\theta} \log p_i\right)^2\right]_{\theta=\theta_0}$$
$$= \int_{\mathscr{E} i} \left(\frac{\partial}{\partial\theta} \log p_i\right)^2 p(x^i \mid \theta)\, \mu^i(dx^i)|_{\theta=\theta_0} \leqq C_1 < \infty.$$

C.4. $E\left[\dfrac{\partial^2}{\partial\theta^2} \log p_i\right]_{\theta=\theta_0} = -I_i(\theta_0).$

C.5. There exists a (μ^i)-measurable function $H_i(x^i)$ such that $\left|\dfrac{\partial^3}{\partial\theta^3} \log p_i\right| < H_i(x^i)$ for all $\theta \in \Theta$, and $H_i(x^i)$ is such that given an arbitrary $\varepsilon > 0$, there exists a finite positive M such that $P[H_i > M] < \varepsilon$ where M is independent of θ and i.

C.6. $\lim\limits_{|k-j| \to \infty} E\left[\dfrac{\partial}{\partial\theta} \log p_k \dfrac{\partial}{\partial\theta} \log p_j\right]_{\theta=\theta_0} = 0$

or

C.6′. $E\left[\dfrac{\partial}{\partial\theta} \log p_k \dfrac{\partial}{\partial\theta} \log p_j\right]_{\theta=\theta_0} = 0$, for all j $\neq k$ which is a stronger version of C.6, and

C.7. $\operatorname{var}\left[\dfrac{\partial^2}{\partial\theta^2} \log p_k\right]_{\theta=\theta_0} \leqq A < \infty$ where A is independent of k, and

$$\lim_{|k-j| \to \infty} \operatorname{cov}\left[\frac{\partial^2}{\partial\theta^2} \log p_k, \frac{\partial^2}{\partial\theta^2} \log p_j\right]_{\theta=\theta_0} = 0.$$

The form of the argument then follows that of Cramer, except that when we get to equation (4.1) it is not possible to apply Khintchine's form of the law of large numbers since the observations are not independent.

Instead a form of the weak law of large numbers for dependent random variables is applied. This states that if for all i the random variables $\{u_i\}$ are such that

$$|E(u_i)| < \infty,\ \operatorname{var}(u_i) \leqq A < \infty \text{ and } \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E(u_i) < \infty$$

and also

$$\lim_{n\to\infty} \left[\frac{1}{n} \sum_{i=1}^{n} \operatorname{cov}(u_i, u_n)\right] = 0$$

then

$$\lim_{n\to\infty} \left[\frac{1}{n} \sum_{i=1}^{n} u_i - \frac{1}{n} \sum_{i=1}^{n} E(u_i)\right] = 0$$

in probability.

Clearly the final conditions of this theorem motivates the "asymptotic uncorrelatedness" conditions C.6 and C.7.

Bar-Shalom's then proves that under the regularity conditions 1 to 7,

$$\lim_{n\to\infty} \hat{\theta}_n (x^n) = \theta_0 \text{ in probability}$$

and if also C.6′ is satisfied, then

$$E(\hat{\theta}_n - \theta_0)^2 = \left[\sum_{i=1}^{n} I_i(\theta_0)\right]^{-1}$$

for sufficiently large n, i.e. $\hat{\theta}_n$ is asymptotically efficient.

7. Comparison of MLE's with other Estimators

It is well known that J. Neyman (1949), in developing the theory of minimum chi-square estimators, formulated the class of Regular Best Asymptotically Normal estimates, RBAN, or simply BAN estimates. This idea was further developed by Barankin and Gurland (1951), who derived a general method for constructing such estimators.

Although Neyman defined BAN estimates in terms of the multinomial distribution, they now have a completely general definition which is that $\hat{\theta}_n$, an estimate from a sample of n observations, is BAN if $\sqrt{n}\,\hat{\theta}_n$ is asymptotically distributed as $N\left(\sqrt{n}\,\theta, \frac{1}{I(\theta)}\right)$ where $I(\theta)$ is Fisher's information for a single observation.

He also obtained a set of necessary and sufficient conditions for an estimator to be BAN and showed that minimum chi-square (MCS) as well as ML estimators belong to this class. The same is also true of modified MCS and other estimators, so that MLE's are certainly not alone in having optimal asymptotic properties.

Thus in actual and hence finite samples, asymptotic efficiency as we have understood that term so far does not enable us to distinguish between these estimators. Some finite sample studies have, of course, been carried out e.g. Berkson (1955*a* and *b*) and Silverstone (1957) for particular cases, but we do not have and indeed it does not appear possible to have any clear-cut conclusions so long as we keep to minimum variance as a yardstick of efficiency.

Nevertheless, a reformulation of efficiency by Rao (1961*a*) has provided us with a statistical measure which does distinguish between the various BAN estimators, even when the sample itself is not very large. His scheme begins with the idea that if a statistic T is efficient, in the old sense, then I_T, the information provided by T per observation, in the Fisherian sense, tends to I the information for a single observation. This will be the case if

$$\lim_{n\to\infty} P\left[\left| n^{-\frac{1}{2}} \frac{\partial l}{\partial \theta} - \alpha - n(T_n - \theta)\right| > \varepsilon\right] = 0 \qquad (A)$$

for any $\varepsilon > 0$.

Taking (A) as a definition of efficiency does not lead us immediately to any new conclusions, for since it is equivalent to the old definition we find as previously that ML, MCS and other estimators are all equally efficient. However, it can be extended to formulate a criterion of second order efficiency defined in the following way

Let E_2 be the variance of

$$\frac{\partial l}{\partial \theta} - \sqrt{n}\,\alpha - n\beta(T_n - \theta) - \lambda n (T_n - \theta)^2$$

minimized with respect to λ. Then T_n is said to have second-order efficiency if no other estimator has a smaller value of E_2.

This approach is in agreement with Rao's earlier contention (1960) that the efficiency of an estimator T_n should be judged by the extent to which $\frac{\partial l}{\partial \theta}$ can be approximated by a function of T_n. In this (1961) paper, Rao then goes on to consider various estimators based on a random sample from a multinomial distribution [1] where the cell probabilities are functions of a single parameter. It is shown that, in this case, E_2 for the MLE is lower than that for the MCS, the minimum discrepancy (Haldane, 1951) the minimum Hellinger distance, and the minimum Kullback-Liebler separator estimators. Thus within this class, at least, the MLE has second order efficiency.

We can also approach the problem of comparing MLE with other estimation procedures, by investigating whether it can be regarded as a special case of some more general process. In this direction we have, for example, a paper by Wolfowitz (1953) who shows in a non-rigorous way that the theory of decision functions can explain why the MLE is asymptotically efficient.

The paper by Berkson (1955), to which we have just referred, is also of interest here. He defines an extended class of Least Squares Estimates to which belong, in particular, the MLE, minimum χ^2, minimum reduced χ^2, minimum transform χ^2 and two other estimators described by Smith (1947) as "ideal" least squares estimates. All these estimates are efficient since the extended class is itself BAN.

Such approaches, however, demand essentially the acceptance of a particular loss function and are to that extent arbitrary. Different loss functions can give different results so that it is possible to have two estimators $\tilde{\theta}_1$ and $\tilde{\theta}_2$ of θ, and a function $f(\theta)$, such that, although

$$E\,[(\tilde{\theta}_1-\theta)^2]>E\,[(\tilde{\theta}_2-\theta)^2]$$

yet

$$E\,[\{f(\tilde{\theta}_1)-f(\theta)\}^2]<E\,[\{f(\tilde{\theta}_2)-f(\theta)\}^2].$$ [2]

We can mention too, other incidental difficulties in connection with the "loss function" approach. Thus, for example, Silverstone (1957) points out that for the binomial distribution with probability θ, when $n = 3$ the estimator $T \equiv \frac{1}{2}$ of θ has a smaller mean square error when $\frac{1}{4} \leqq \theta \leqq \frac{3}{4}$, than the estimator $T' = \frac{r}{n}$, yet clearly T' is to be preferred.

It seems doubtful, therefore, whether Decision Theory can claim any greater generality than the method of MLE, as an estimation procedure. The only way open to us, therefore in order to assess the ML process generally is by direct comparison of results, and in order to achieve this it is necessary to search continuously for measures, such as efficiency and second-order efficiency which do indeed discriminate between the various estimators.

8. Multiparameter Situations

We have already seen that consistency for the multiparameter case has been established by Wald.

It is, however, possible to generalize the work of Cramer (1946) and Hurzabazaar (1948), and the first step in this direction appears to have been taken by Chanda (1954). He modifies the conditions of Cramer in an obvious way, so that condition I for example, now becomes that for $\theta = (\theta_1, \ldots, \theta_k)$

$$\frac{\partial \log f}{\partial \theta_r}, \frac{\partial^2 \log f}{\partial \theta_r \partial \theta_s}, \frac{\partial^3 \log f}{\partial \theta_r \partial \theta_s \partial \theta_t} (r, s, t = 1, \ldots, k)$$

[1] Rao's papers (1957, 1958) have established a complete theory of MLE for the multinomial distribution.

[2] This would be the case, for example if θ is distributed as a normal variate with a mean θ_0 and a unit variance, and $\tilde{\theta}_1 = \theta_0$, $\tilde{\theta}_2 = \sqrt{1+\theta_0^2}$ with $f(\theta) \equiv \theta^2$.

all exist for all $\theta \in \Theta$ and almost all x, and similarly for the other conditions. The proof that there is a unique consistent root of the LE, and that with probability tending to one the matrix $\left\| \dfrac{\partial^2 l}{\partial\theta_r \partial\theta_s} \right\|$ is negative definite is analogous to the single parameter situation. This result, however, suffers from the defect of proving consistency at a local maximum and not necessarily at an absolute maximum, as in the case of Wald's proof.

When, however, we have a set of k jointly sufficient statistics, $t_1, \ldots, t_k$ for k parameters $\theta_1, \ldots, \theta_k$, then as in the single parameter case the MLE's will be functions of them. It was established by Hurzabazaar (1949) that in this case the LE's have a unique solution at which the LF has a maximum.

The multiparameter form of the Cramer-Rao MVB (Rao, 1945) in which Fisher's information is replaced by the "information" matrix $\left\| -E \dfrac{\partial^2 \log f}{\partial\theta_r \partial\theta_s} \right\|$ enables us, in a regular estimation problem, to fix bounds for the variance of the estimators of several unknown parameters, whatever estimation process is used. Thus if we have k estimators $t_1, \ldots, t_k$ for k parameters $\theta_1, \ldots, \theta_k$, then defining $V_{rr} = v\,(t_r)$ we have simultaneously that

$$nV_{rr} \geqq I^{rr}$$

where

$$\| I^{rs} \| = \| I_{rs} \|^{-1} \quad (r, s = 1, \ldots, k)$$

and $\| I_{rs} \|$ is the information matrix, I say, defined above.

Since it can be shown that $I^{rr} \geqq I_{rr}^{-1}$ then we have an improved result relative to the normal Cramer-Rao MVB for the single parameter case.

Thus, for example, in the estimation of π_1 and π_2 in the trinomial distribution

$$f(\underset{\sim}{x}, \theta) = \frac{n!}{x_1!x_2!x_3!}\pi_1^{x_1}\pi_2^{x_2}(1-\pi_1-\pi_2)^{x_3}$$

where $x_1+x_2+x_3 = n$, it is easily shown that

$$\hat{\pi}_1 = \frac{x_1}{n}, \quad \hat{\pi}_2 = \frac{x_2}{n}$$

and that

$$I = \begin{bmatrix} \dfrac{1-\pi_2}{\pi_1(1-\pi_1-\pi_2)} & \dfrac{1}{1-\pi_1-\pi_2} \\ \dfrac{1}{1-\pi_1-\pi_2} & \dfrac{1-\pi_1}{\pi_2(1-\pi_1-\pi_2)} \end{bmatrix}$$

and that

$$I^{-1} = \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{bmatrix}.$$

Since it can be shown independently that $V\,(\hat{\pi}_1) = \dfrac{\pi_1(1-\pi_1)}{n}$ and that $V\,(\hat{\pi}_2) = \dfrac{\pi_2(1-\pi_2)}{n}$ then it follows that in this case the MLE's of π_1 and π_2 do attain the lower bound. [See Silvey (1970), chapter 2.]

We note also that if π_1 only is unknown, then the MVB for $\hat{\pi}_1$ is $\dfrac{\pi_1(1-\pi_1-\pi_2)}{n(1-\pi_2)}$ which is less than $\dfrac{1}{n}\pi_1(1-\pi_1)$.

D

However, as in the single parameter case, there are many instances where these bounds are not attained in finite samples and it is again to large samples that we must look for the optimal properties of MLE to become more generally apparent. An important paper in this respect is again by Rao (1947) who proves the following.

If L is the dispersion matrix of $\hat{\theta}_1, \ldots, \hat{\theta}_k$, the MLE's of $\theta_1, \ldots, \theta_k$, and K that of any other set of estimators then under certain conditions:

(i) $\lim\limits_{n\to\infty} nK - \lim\limits_{n\to\infty} nL = K_\infty - L_\infty$, say,

is positive semi-definite

(ii) $L_\infty^{-1} = I$, the information matrix.

Actually part (i) of the theorem follows from an earlier result of Wald (1943), who by means of general limit theorems showed that the joint distribution of $\hat{\tilde{\theta}}_1, \ldots, \hat{\theta}_k$ tends to that of the multivariate normal with mean $(\theta_1, \ldots, \theta_k)$ and dispersion matrix $D^{-1} = \frac{1}{n} I^{-1}$.

An immediate consequence of Rao's result is that $\det K \geqq \det L$, which is Geary's (1942) result that MLE's have the least "generalized" variance. A second corollary of this theorem is that $K_{rr}^\infty \geqq L_{rr}^\infty$ showing that each MLE has least possible variance in large samples.

There are also two important papers by Doss (1962, 1963), in the first of which he establishes the properties of consistency, uniqueness and asymptotic normality, and joint asymptotic efficiency of the MLE for the multiparameter case. The conditions there are weaker than those given by Chanda in the case of consistency, and Doss gives as an example of the case of the normal distribution with mean α and variance β which satisfy his conditions but not those of Chanda, yet the MLE's of α and β do have all the properties under consideration. In the second paper a different set of conditions is given to establish the same results. The two sets of conditions therefore widen the class of situations in which MLE's have the optimal asymptotic properties.

One might say that Doss investigated the conditions of Chanda in the same way as Kulldorf examined Cramer's conditions. All this points again to the fact that there are a large number of possible sets of sufficient conditions and that, as we have already remarked, a much more difficult problem would be to establish a set of necessary and sufficient conditions under which desirable asymptotic properties would hold.

A different approach to the question of efficiency similar to Wolfowitz's (1965) investigations in the single parameter case is provided by Kaufman (1966). Wolfowitz's results, as we have already seen in section 5, imply that subject to certain conditions which do not demand normality, that among all estimators of θ_0, $\hat{\theta}$ is such that $\sqrt{n}\,(\hat{\theta}-\theta_0)$ is most concentrated about its median.

Kaufman's main theorem is a symmetrized multi-parameter version of this result. He proves, in fact, that if it is assumed that Θ is a closed subset of k-dimensional space R^k, and $\{f_\theta;\, \theta \in \Theta\}$ a family of probability densities over a σ-finite measure space $(\mathscr{X}, \mathscr{F}, \mu)$, be such that a MLE sequence $\{\hat{\theta}_n\}$ exists, that $n^{\frac{1}{2}}\,(\theta_n - \theta)$ is asymptotically joint normal $N\,[0, I\,(\theta)^{-1}]$ uniformly in $R^k \times C$, where $I\,(\theta)$ is the information matrix and that $\{\hat{\theta}_n\}$ is asymptotically sufficient for Θ^0, the interior of Θ, in the sense defined by Le Cam (1956), then if $\{T_n\}$ is any estimator sequence, which need not be asymptotically normal, such that the distribution of $n^{\frac{1}{2}}\,(T-\theta)$ converges uniformly in $R^k \times C$, and S is a convex symmetric set in R^k, we have

$$\lim_{n\to\infty} P\,[n^{\frac{1}{2}}\,(T_n-\theta) \in S] \leqq \lim_{n\to\infty} P\,[n^{\frac{1}{2}}\,(\theta_n-\theta) \in S]$$

for all $\theta \in \Theta^0$.

He then gives a counter-example to show that this result is not in general true for the unsymmetrized multidimensional form.

In conclusion, we comment briefly on a paper by Richards (1961), who devises a computational method which can greatly reduce the calculations necessary to derive a set of MLE's. It could also be of great value in regression problems.

From k unknown parameters $\theta_1, \ldots, \theta_k$ choose r, $\theta_1, \ldots, \theta_r$, in some convenient way, and obtain conditional MLE's of $\theta_{r+1}, \ldots, \theta_k$ given $\theta_1, \ldots, \theta_r$. Denoting these conditional estimates by $\tilde{\theta}_{r+1}, \ldots, \tilde{\theta}_k$ we have, by implication, that there exists functional relationships

$$\tilde{\theta}_{r+j} = \psi_j(\theta_1, \ldots, \theta_r) \quad j = 1, 2, \ldots, k-r. \tag{8.1}$$

Now substitute $\tilde{\theta}_{r+j}$ for θ_{r+j} in $l(x \mid \theta_1, \ldots, \theta_k)$ to obtain $l'(x \mid \theta_1, \ldots, \theta_r)$, say, and hence new modified LE's,

$$\frac{\partial l'}{\partial \theta_i} = 0 \quad i = 1, 2, \ldots, r. \tag{8.2}$$

If $\hat{\theta}_1, \ldots, \hat{\theta}_r$ are the solutions of (8.2) then the MLE's are given by

$$\hat{\theta}_1, \ldots, \hat{\theta}_r, \psi_1(\hat{\theta}_1, \ldots, \hat{\theta}_r), \ldots, \psi_{k-r}(\hat{\theta}_1, \ldots, \hat{\theta}_r).$$

There are some theoretical gaps in this argument, but these have been filled in by Kale (1963), who establishes that the method is essentially valid provided correct use is made of the implicit function theorem.

9. Theoretical Difficulties

We have already seen that MLE as a procedure appears to be deficient in the following ways:

(*a*) MLE's are not always consistent.

(*b*) There are other estimators which have a smaller asymptotic variance, at least for some values of the unknown parameter, and are therefore, in this sense, more efficient than the MLE.

(*c*) Nothing very much can be said about the MLE small sample performance.

With regard to (*c*), it can be said that as there is no complete theory of estimation for small samples, no simple conclusions capable of a wide application about MLE's, or any other estimator, can be made. It cannot therefore be advanced as a reason for preferring another estimation procedure. Certainly Fisher's claims for MLE's cannot be criticized in this way, as he clearly regarded the optimal properties of MLE as essentially asymptotic ones.

The rest of this discussion will therefore be confined to (*a*) and (*b*), and we begin with the well-known example of inconsistency of Neyman and Scott (1948), who also at the same time introduced the important concepts of structural and incidental parameters.

Suppose that we have observations x_{ij} $(i = 1, 2, \ldots, s;\ j = 1, \ldots, n)$ from s populations which are characterized by the following probability density functions,

$$p_i(x_{ij} \mid \mu_i, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\tfrac{1}{2}\left(\frac{x_{ij}-\mu_i}{\sigma}\right)^2\right].$$

The parameter σ is common to each population so that if s is indefinitely large then σ appears in the probability law of an infinite number of observations, and therefore, in accordance with the definition given, is a structural parameter. On the other hand, each μ_i relates only to a finite number of observations, however large s may be, and is therefore an incidental parameter.

If these parameters are estimated by the method of ML then it is easily found that

$$\hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} = \bar{x}_i, \text{ say,}$$

and

$$\hat{\sigma}^2 = \frac{1}{ns} \sum_{i=1}^{s} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2.$$

Since $E\left[\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2\right] = (n-1)\,\sigma^2$, for each i, it follows that $\hat{\sigma}^2$ is the arithmetic mean of s quantities each of which has an expected value $\frac{(n-1)\,\sigma^2}{n}$. Thus $\hat{\sigma}^2$ converges in probability as s tends to infinity to $\frac{n-1}{n}\,\sigma^2$, so that it is not consistent for σ^2.

It is, of course, possible to correct for this by taking $\frac{n}{n-1}\,\hat{\sigma}^2$ as our estimator of σ^2, but this is irrelevant since this new estimator would not then be the MLE. We note, too, that when $n = 2$, which is the form in which this example is most frequently quoted, $\hat{\sigma}^2$ converges in probability to $\frac{1}{2}\sigma^2$, so that there is a substantial error.

Kiefer and Wolfowitz (1956), on the other hand, have shown that when the μ_i are independently but identically distributed random variables then the MLE $\hat{\sigma}^2$ is strongly consistent. Evidently therefore, this situation is more regular than the one in which the incidental parameters are unknown constants.

In general, however, the problem of eliminating incidental or nuisance parameters so as to be able to make statements of uncertainty about the parameters of interest is now recognized in statistics as a major one. A "likelihood" as distinct from a Bayesian approach to this problem is provided in a very recent paper by Kalbfleisch and Sprott (1970). This is an important as well as a highly controversial paper. It is here that integrated likelihoods, marginal likelihoods, maximum relative likelihoods, conditional likelihoods and second-order likelihoods are introduced. Definitions of these terms are given in the Appendix.

An example of a different type is due to Basu (1952). Let $X = 0$ or 1 be a random variable whose distribution depends on a single unknown parameter θ, $0 \leqq \theta \leqq 1$, as follows

$$\begin{aligned} P(X = 1) &= \theta \quad \text{when } \theta \in A \\ &= 1 - \theta \text{ when } \theta \in B \end{aligned}$$

where A is the set of rationals and B any countable set of irrationals in $[0, 1]$. If in a sample of n independent observations we obtain r 1's, then $\hat{\theta} = \frac{r}{n}$ which converges in probability to θ or $1-\theta$ according as $\theta \in A$ or $\theta \in B$, and it therefore cannot be consistent for θ.

Another more complicated example due to Kraft and Le Cam (1956) was discussed earlier in section 4. Two further examples of inconsistency of a different type are provided by Bahadur (1958). Here we are concerned with the estimation of the population distribution function. Bahadur's viewpoint is that where "the (ML) estimated distribution converges to the actual distribution in the strong sense that, with probability one, the estimated probability of any event in the sample space of a single observation tends, uniformly in events, to the correct probability" it is unreasonable to speak of the MLE as being inconsistent. He claims, therefore, that the importance of his examples lies in the fact that unlike earlier examples of inconsistency, the estimated distribution function does not converge to the true one.

Actually, in the first of his examples, the MLE does not, in any way, indicate the true form of the distribution function. In the second which we now summarize the results obtained by the ML process are misleading.

(i) The random variable X is a non-negative integer $j = 0, 1, 2, \ldots$.

(ii) The set of possible distributions of X is a countable set $\{P_1, P_2, \ldots\}$ in which $\lim_{k\to\infty} P_k$ exists $= P_\infty$ say.

(iii) For all sets of sample values $\{X_1, X_2, \ldots, X_n\}$ the MLE exists.

(iv) Nevertheless, no matter what the actual distribution of X, any sequence of MLE distributions converges to P_∞. Thus if $h_n = h_n(X_1, \ldots, X_n)$ is the least index h such that P_h is a MLE based on the first n observations then

$$P_k^{(\infty)}\left[\lim_{n\to\infty} h_n = \infty\right] = 1$$

for each $k = 1, 2, \ldots$ when $P_k^{(\infty)}$ denotes the probability measure on the space of infinite sequences $X_1, X_2, \ldots$, where X is distributed as P_k.

At first sight, therefore, it would appear that we are confronted with serious theoretical difficulties with regard to MLE. The situation is, however, not as bad as it seems, for consistency as a property is artificial, and indeed it is possible for there to be situations where it is quite valueless. For example, if μ_n is the mean of n independent observations from a normal population with unknown mean μ and known variance σ^2, then $\hat{\mu} = \mu_n$ and is consistent for μ. Now define, T_n is arbitrary for $n \leqq 10^6$ and is equal to μ_n for $n > 10^6$. Then T_n is also a consistent estimator of μ but in practical terms it is clearly useless.

Of greater intrinsic value in estimation theory is Fisher's first (1922) definition of consistency, now referred to as Fisher consistency (FC). It is defined in the following way.

"A statistic satisfies the criterion of consistency if when it is calculated from the whole population it is equal to the required parameter" (page 309). Later in the paper he says "That when applied to the whole population the derived statistic should be equal to the parameter".

It is also apparent that these definitions and the examples given in the paper imply that he was considering the statistic as a function of the observed distribution function. We therefore have the following formal definition of FC due to Rao (1962).

A statistic is said to be FC for a parameter θ if

(1) it is an explicit function of the sample distribution function S_n (or the observed proportions $[p_1, \ldots, p_k]$ in the case of the multinomial), and

(2) the value of the function reduces to θ identically when S_n is replaced by the true distribution function $F(\theta)$ (or true proportions $[\pi_1(\theta), \ldots, \pi_k(\theta)]$ in the case of the multinomial).

This definition places restrictions on the functions of observations which may be considered, whereas the definition of consistency used heretofore, probability consistency (PC), does not. In the case where the statistic is a continuous functional (see Appendix) of the distribution function then PC$\Leftrightarrow$FC, and no doubt Fisher himself had in mind this kind of situation when instead he put forward PC as a definition of consistency.

With regard to the general problem of the consistency of MLE's Rao (1962) then makes the following observations by which the importance of FC, in particular, will readily become apparent. In the first place it can easily be shown that the MLE is always FC. In addition, however, we must also require that when S_n is close to $F(\theta_0)$ then $F(\hat{\theta})$ should be close to $F(\theta_0)$. This is always the case for the multinomial distribution (Rao, 1957). It is also the case for the infinite multinomial distribution provided $\Sigma\pi_i \log \pi_i$ is convergent, where π_i are the cell probabilities (Kiefer and Wolfowitz, 1956; Rao, 1958), and under slightly more restrictive conditions, for any continuous distribution (Kraft, 1955). The example due to Bahadur, which we have just considered, is of a very special type.

Thus it can be said that the MLE of the distribution function is consistent under very non-restrictive conditions, and from this will follow the corresponding consistency of the MLE of a parameter provided we also have the continuity condition $F(\theta)\rightarrow F(\phi)$ implies $\theta\rightarrow\phi$. This seems a reasonable requirement, and in fact, it is not satisfied in the examples of Basu (1955) and Kraft and Le Cam (1956).

Rao then shows how it is possible to accommodate some, at least, of these anomolous cases, in the following way. Consider first the likelihood ratio $\frac{P(T,\theta_1)}{P(T,\theta_2)}$ where $P(T,\theta_i)$ is the probability density function of T when $\theta=\theta_i\,(i=1,2)$, so that the more T discriminates between the two possible values θ_1 and θ_2 of θ, the more this ratio will differ from unity. As a measure of discrimination, over all samples, and hence over all possible values of T, the quantity [1]

$$J_T(\theta_1,\theta_2)=\mathop{E}_{\theta_1}\log\left[\frac{P(T,\theta_1)}{P(T,\theta_2)}\right]+\mathop{E}_{\theta_2}\log\left[\frac{P(T,\theta_2)}{P(T,\theta_1)}\right]$$

is put forward.

Obviously the maximum value of this quantity is obtained by replacing T by the sample itself, when we would obtain

$$J_s(\theta_1,\theta_2)=n\left\{\mathop{E}_{\theta_1}\log\frac{P(X,\theta_1)}{P(X,\theta_2)}+\mathop{E}_{\theta_2}\log\frac{P(X,\theta_2)}{P(X,\theta_1)}\right\}$$

since the observations are independent. A measure of the effectiveness of T as a discriminator, therefore, would be the ratio $\frac{J_T(\theta_1,\theta_2)}{J_s(\theta_1,\theta_2)}$ which would be unity if T is a sufficient statistic.

Now Basu (1954) has shown that as n tends to infinity $J_s(\theta_1,\theta_2)\rightarrow\infty$, so that we have perfect discrimination between θ_1 and θ_2, as we should expect. We must also demand therefore that for T similarly, $J_T(\theta_1,\theta_2)\rightarrow\infty$ as $n\rightarrow\infty$, in which case $P(T,\theta_1)$ and $P(T,\theta_2)$ would be said to be orthogonal in the limit. This will be the case if T converges in probability to a $1-1$ function of θ, which is precisely Fisher's criterion of consistency.

The problem of inconsistent MLE's can therefore be partly resolved by the following wider definition of consistency. A statistic T will be said to be consistent in the wide sense if the two distributions of T corresponding to the two possible values of θ are orthogonal. Such a definition does not necessarily demand the convergence of T to a unique limit and in fact Basu's estimate is consistent in this sense. So also is the example due to Neyman and Scott.

Bahadur's example, however, is not even consistent in the wide sense though it would be reasonable to regard this as very exceptional, and only very weak conditions are necessary for its exclusion.

Examples of apparent inconsistency arise in some cases of estimation of parameters of Functional and Structural relationships. This is an extensive subject and reference will be made to only two simple examples.

In the case of the functional relationship

$$Y=\alpha+\beta X$$

when α and β are unknown constants, it is supposed that we observe $x=X+u$ and $y=Y+v$ where $u\sim N(0,\sigma_u^2)$ and $v\sim N(0,\sigma_v^2)$. If we have n unreplicated observations

$$(x_i, y_i),\ i=1,2,\ldots,n$$

then

$$L=(2\pi\sigma_u\sigma_v)^{-n}\exp\left[-\frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(x_i-X_i)^2-\frac{1}{2\sigma_v^2}\sum_{i=1}^{n}(y_i-\alpha-\beta X_i)^2\right].$$

[1] This originates from the work → Kullback and Liebler (1951) in connection with information theory.

From the LE's it follows that $\hat{\sigma}_v^2 = \hat{\beta}^2\hat{\sigma}_u^2$ which is an inconsistency in any sense, since no corresponding relationship between the true values of σ_u^2, σ_v^2 and β necessarily exists.

This anomaly, which was first discovered by Lindley [1] (1947) has, however, been resolved by Solari (1968) who has shown that the solutions of the likelihood equations for the three structural parameters β, σ_u^2, σ_v^2 and the n incidental parameters X_i do not give rise to even a local maximum of L, but in fact to a saddle point. Thus they are not MLE's, and so no inconsistency exists here.

Solari goes on, however, to show further that the function

$$g = 2\log\sigma_u\sigma_v + \frac{1}{n\sigma_u^2}\sum_{i=1}^{n}(x_i - X_i)^2 + \frac{1}{n\sigma_v^2}\sum_{i=1}^{n}(y_i - \beta X_i)^2$$

which is such that

$$L = (2\pi)^{-n}e^{-ng/2}$$

so that L is a decreasing function of g, has an essential singularity at any points of the parameter space for which

$$\sigma_u = 0 \text{ and } \sum_{i=1}^{n}(x_i - X_i)^2 = 0,$$

and/or

$$\sigma_v = 0 \text{ and } \sum_{i=1}^{n}(y_i - \beta X_i)^2 = 0.$$

Thus, in the neighbourhood of such points, g can take any value, so that the minimum of g obtained by first minimizing with respect to one parameter and then with respect to another, and so on, would depend on the order of the parameters chosen. Thus unless prior information was available to indicate a particular order no unique minimum of g and hence no MLE could exist.

Apparently therefore we have an estimation problem—there are in fact several problems of this type—in which the method of MLE fails to give any results at all. One might ask, however, whether in the problem just described it is logical to admit points for which $\sigma_u < \varepsilon$ or $\sigma_v < \varepsilon$ into the parameter space, for if either error variance is zero the model is changed fundamentally, and if both are zero we do not have a stochastic problem at all.

With this in mind, therefore, it would seem reasonable to restrict the parameter space initially by

$$\sigma_u^2 \geqq \sigma_1^2 > 0 \text{ and } \sigma_v^2 \geqq \sigma_2^2 > 0$$

where σ_1^2 and σ_2^2 are small but fixed.[2]

The function g would then be bounded below and continuous (and differentiable) throughout the entire parameter space and would therefore have a unique minimum in Θ, so that in fact the MLE would exist, even though we would know as a result of Solari's analysis, that it could not be obtained from the likelihood equation.

The important results of Kiefer and Wolfowitz (1956) to which reference was made earlier in this section, would lead us to expect a more straightforward situation in the structural relationship case, since here the incidental parameters are themselves random variables. We have here the model

$$\begin{aligned} Y &= \alpha + \beta X \\ x &= X + u \\ y &= Y + v \quad (u \text{ and } v \text{ independent}) \end{aligned}$$

[1] Lindley and El-Sayyad (1967) have however provided a Bayesian analysis of the problem of estimating the parameters of the functional relationship considered here.

[2] This is merely another way of saying that both variables *are* subject to error, which belief is implicit in the formulation of the model itself.

where

$$E(u) = E(v) = 0,\ V(u) = \sigma_u^2,\ V(v) = \sigma_v^2,\ E(x) = E(X) = \mu_X, \text{ say,}$$

and hence $E(y) = E(Y) = \alpha + \beta\mu_X$. Also, if $V(X) = \sigma_X^2$, then

$$V(x) = \sigma_X^2 + \sigma_u^2 \text{ and } V(y) = \beta^2\sigma_X^2 + \sigma_v^2.$$

If it is further assumed that the distributions involved are normal, then the joint distribution of x and y is bivariate normal and is therefore defined by parameters μ_1, μ_2, σ_1^2, σ_2^2 and ρ, say.

In the case of unreplicated observations, therefore, it is impossible to estimate the six parameters in the original model. This is essentially a problem of identifiability [1] and in this case it can only be resolved by our making certain assumptions about the error variances. Once these assumptions are made, however, the MLE gives perfectly sensible results.

When we turn to the problem of super-efficiency it is natural to begin with the now classical example of J. L. Hodges, since previous to its appearance it had always been thought that in large samples no estimator could have a smaller variance than that of the MLE.[2]

If μ_n is the mean of n independent observations from a normal population with an unknown mean μ and a unit variance, then it is well known that μ_n is the MLE of μ and that it is normally distributed about μ with a normal variance $\frac{1}{n}$, for all n. Define T_n by

$$\begin{aligned} T_n &= \mu_n \text{ if } |\mu_n| \geqq n^{-\frac{1}{4}} \\ &= \alpha\mu_n \text{ if } |\mu_n| < n^{-\frac{1}{4}}. \end{aligned}$$

Clearly the asymptotic distribution of T_n is $N\left(\mu, \frac{1}{n}\right)$ when $\mu \neq 0$ and $N\left(\mu, \frac{\alpha^2}{n}\right)$ when $\mu = 0$. Thus by taking $\alpha < 1$ we have when $\mu = 0$ an asymptotically normal estimator with a variance less than that of the MLE.

This example was quoted and then generalized by Le Cam (1953). He showed that it was possible to construct an estimator which would be super-efficient at a countable set of values of the unknown parameter θ. Further, it was shown that the set of all values of θ at which an estimator could be super-efficient has Lebesgue measure zero. Evidently, therefore, Hodges' example is not as drastic as at first sight might be supposed, though as it stands it does refute Fisher's assertion about the asymptotic variance of MLE's in general.

One possible remedy is to impose regularity conditions so as to exclude such estimators. This has been done, for example, by A. M. Walker (1963). His conditions are as follows:

1. Θ is an open interval of the real live R for each $\theta \in \Theta$.
2. $\frac{\partial f(x, \theta)}{\partial \theta}$ exists for almost all (μ) x.
3. $\int_R \left(\frac{\partial f}{\partial \theta}\right) d\mu = \frac{\partial}{\partial \theta} \int_R f d\mu = 0.$
4. $J(\theta) = E_\theta \left[\frac{\partial}{\partial \theta} \log f(x, \theta)\right]^2 = \int_R f^{-1} \left(\frac{\partial f}{\partial \theta}\right)^2 d\mu$ is finite and positive.
5. $\frac{\partial}{\partial \theta} E_\theta(T_n) = \int_{R^n} \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i, \theta)\, d\mu^{(n)}$ which exists for each n where $\mu^{(n)}$ is the product measure $\mu \times \mu \times \ldots \times \mu$ in the Euclidean space R^n.

[1] Identifiability of the structural parameters is one of the conditions of Kiefer and Wolfowitz's (1956) results.

[2] Actually this view persisted for a long time afterwards.

Thus $\{f(x, \theta);\ \theta \in \Theta\}$ is said to be a regular family of probability densities if conditions 1 to 4 hold, and (T_n) is a regular estimator sequence if condition 5 holds. His first theorem is then as follows.

Theorem 1. *Let $\{f(x, \theta);\ \theta \in \Theta\}$ be a regular family of probability densities with respect to σ-finite measure μ, and $X^{(n)} = (X_1, X_2, \ldots, X_n)$ be, for each n, a random sample of size n from a population with probability density $f(x, \theta)$. Suppose also that $T_n(X^{(n)})$ is a regular estimator sequence with asymptotic variance $n^{-1}v(\theta)$. Then if* (a) *for each θ*

$$E_\theta \mid n^{\frac{1}{2}}(T_n - \theta)|^{2+\delta} \lambda_{2+\delta,\, n}(\theta) < \infty$$

for each n, and the sequence $\{\lambda_{2+\delta,\, n}(\theta)\}$ is bounded for some positive δ, and

(b) $b'_n(\theta) = \dfrac{\partial}{\partial\theta} E_\theta(T_n - \theta) \to 0$ *as* $n \to \infty$

then $$v(\theta) \geqq \{J(\theta)\}^{-1}$$

or $$\mathscr{E}_T(\theta) \leqq 1 \quad \theta \in \Theta$$

where $$\mathscr{E}_T(\theta) = \{v_T(\theta)\, J(\theta)\}^{-1}$$

is the asymptotic efficiency of T.

His second theorem is merely a multiparameter generalization of this.

It is claimed that the conditions on T_n are not unduly restrictive and would not, in practice, exclude any estimator that a statistician would be likely to want to use. In fact, the estimators of Hodges and Le Cam are excluded by condition (b) of Theorem 1.

Alternatively we can reconsider other ways of defining efficiency. We have already seen that attempts to achieve this in terms of "concentration" are mathematically complicated, and no straightforward development of this idea appears to be possible. Thus following Fisher efficiency has been defined heretofore essentially in terms of ratio of variances. We have remarked too, that since Fisher himself regarded a statistic as efficient when it utilizes all the available information, and that this can be said to be the case when the asymptotic variance has the least possible value $\dfrac{1}{nI(\theta)}$ then the quantity $I(\theta)$ has come to be defined as the amount of information per observation in the sample. Fisher's proof that $\dfrac{1}{nI(\theta)}$ does provide a lower bound for the asymptotic variance of any estimator does not, however, contain the necessary regularity conditions to exclude estimators such as those of Hodges and Le Cam.

Nevertheless, the appearance of $I(\theta)$ in many distinct statistical situations shows it to be of fundamental importance. We have already seen that it forms part of Bahadur's inequalities in connection with the "effective standard deviation". It is also implicit in the discrimination measures $J_s(\theta_1, \theta_2)$ and $J_T(\theta_1, \theta_2)$, which were described earlier, for it is not difficult to show that

$$J_s(\theta_1, \theta_1 + \delta\theta_1) \sim \frac{n}{2} I(\theta_1)(\delta\theta_1)^2$$

and

$$J_T(\theta_1, \theta_1 + \delta\theta_1) \sim \frac{n}{2} I_T(\theta_1)(\delta\theta_1)^2$$

when $I_T(\theta_1)$ is the corresponding information per observation derived from the statistic T.

T is not fully efficient, therefore, if $I(\theta) > I_T(\theta)$. In the multiparameter case where $I(\theta)$ and $I_T(\theta)$ are matrices it is well known that $I(\theta) - I_T(\theta)$ is positive semi-definite, Rao (1945).

Efficiency here therefore could be measured by a scalar function of $I(\theta)-I_T(\theta)$ having the value zero when all lower bounds were simultaneously attained.

Thus efficiency in the sense of perfect discrimination, or orthogonality as defined earlier, is equivalent to efficiency in the information sense, and hence to efficiency in the minimum variance sense.

As a basis for all possible definitions of efficiency, therefore, is the criterion that T is efficient if $I(\theta)-I_T(\theta)\rightarrow 0$ as $n\rightarrow\infty$. This approach is capable of development to first- and second-order efficiency discussed earlier by which it is possible to discriminate between the various BAN estimators. It also leads to the following equivalent formulation proposed by Rao (1962).

"A statistic is said to be efficient if its asymptotic correlation with the derivative of the log-likelihood (l) is unity. The efficiency of any statistic may be measured by ρ^2, when ρ is the asymptotic correlation with $Z_n (= n^{-\frac{1}{2}}l')$", and there is an analogous definition for multiparameter estimation.

The motivation for these definitions lies, again, in the search for effective discrimination. A paper by Rao and Poti (1946) establishes that a consequence of the Neyman-Pearson lemma is that a test which best discriminates small variations $\delta\theta$ from the true value of θ for any given n, is of the form reject if and only if $Z_n \geqq \lambda$ (λ a constant). The definition just given then follows naturally from the fact that a statistic which has an asymptotic correlation with Z_n of 1 will be as good as Z_n as a discriminator.

We now pass on to consider super-efficient estimators in the light of this new definition. In the first place it should be noted that such statistics are not FC since they cannot be explicit functions of the sample distribution. In fact, Kallianpur and Rao (1955) have shown that $\dfrac{1}{nI(\theta)}$ is a lower bound for the asymptotic variance of any estimator which is FC. This again illustrates the importance, from the theoretical standpoint, of taking FC rather than PC as the criterion of consistency.

They also show that in large samples, under suitable regularity conditions, any estimator whose variance does attain the lower bound $\dfrac{1}{nI(\theta)}$ also has an asymptotic correlation of unity with Z_n, thus reconciling Fisher's and Rao's definition of efficiency.

When FC and other regularity conditions do not hold, then the estimators of Hodges and Le Cam are not excluded, but on the basis of the new definition of efficiency they are at best as good as, but never better than, the MLE. They are only as effective for hypothesis testing and interval estimation, etc., when their asymptotic correlation with the MLE is unity. When it is less than unity they are definitely worse than the MLE.

We can even have an estimator of the Hodges-Le Cam type which is super-efficient for some values of θ, but is sub-efficient in the new sense, i.e. $-1<\rho<1$. For such cases the usual definition of efficiency would be misleading, for it would cause us to choose estimators which are operationally inferior.

Consider, for example, a sample of n independent observations $x_1, x_2, \ldots, x_n$ from a normal population with unknown mean μ, and $\sigma^2 = 1$. If $\bar{x}$ and x_m are respectively the sample mean and median, then define

$$\begin{aligned} T &= \alpha x_m \quad \text{if } \bar{x} < n^{-\frac{1}{4}} \\ &= \bar{x} \quad\;\; \text{if } \bar{x} \geqq n^{-\frac{1}{4}}. \end{aligned}$$

Thus T is asymptotically distributed as $N\left(\mu, \dfrac{\alpha^2\pi}{2}\right)$ when $\mu = 0$, and as $N(\mu, 1)$ when $\mu \neq 0$.

Since, therefore, α can be arbitrarily small, then T is super-efficient. It would be wrong, however, to infer that T should be used as a means of testing the hypothesis $\mu = 0$. For if T is used, then our test statistic is essentially x_m and we have a less powerful test than if we had used $\bar{x}$.

Returning now to Basu's example as summarized at the end of section 5, we see at once that if minimum variance is our criterion then we must reject the MLE in favour of the super-efficient estimator $T = \frac{2g+l}{5}$. A more "operational" approach to this problem, however, shows that the choice of estimator is not so straightforward as might at first sight appear, since for small departures from the true value of θ the statistic T does not provide such a powerful test of a hypothesis $\theta = \theta_0$ as does T', say, $= \frac{g}{2}$, the MLE.

10. Summary

It now remains to summarize the foregoing discussion. We have seen that there are a large number of papers concerned with establishing consistency and/or efficiency under conditions which are unrestrictive as possible. The aim is therefore to obtain the widest possible class of estimation situations in which MLE's exist and have the optimal properties which Fisher claimed for them.

On the other hand, the difficulties that arise by reason of the existence of inconsistent MLE's and super-efficient estimators can be partly resolved not so much by imposing further conditions as by reformulating efficiency and consistency in operational as distinct from purely mathematical terms.

Finally, it is clear that MLE's do have certain advantages over other methods of estimation which may be summarized as follows:

1. MLE's are functions of such sufficient statistics as exist, and hence of the minimal sufficient statistic. This important property does not appertain to other estimators.

2. Although there are other BAN estimators, yet only MLE's possess second order efficiency.

To these could be added that MLE has great intuitive appeal and that it can be applied to a wide class of situations.

Appendix

(A) The following definitions are due to Kalbfleisch and Sprott (1970).

1. Integrated Likelihoods and Integrated Relative Likelihoods

If $x_1, x_2, \ldots, x_n$ are the realized values of a sample of n from a population whose probability density function is $f(x;\ \theta,\ \phi)$, then the joint distribution of the observations, assuming independence, is

$$f(x_1, x_2, \ldots, x_n;\ \theta,\ \phi) = \prod_{i=1}^{n} f(x_i;\ \theta,\ \phi).$$

If, further, we are interested in making inferences about θ, and ϕ is merely a nuisance parameter with a known prior distribution $g\ (\phi)$, then we integrate out ϕ to obtain

$$I\ (\theta;\ x_1, \ldots, x_n) = \int_{\Phi} f(x_1, \ldots, x_n;\ \theta \mid \phi)\, g\ (\phi)\, d\phi$$

where Φ is the entire ϕ-space. The quantity $I(\theta;\ x_1, \ldots, x_n)$ is thus known as the integrated likelihood function of θ and the standardized quantity

$$\frac{I(\theta;\ x_1, \ldots, x_n)}{\sup\limits_{\theta \in \Theta} I(\theta;\ x_1, \ldots, x_n)}$$

the integrated relative likelihood of θ.

2. Maximum Relative Likelihood

Let $\hat{\phi}(\theta)$ be the MLE of ϕ, for a given θ, based on a sample $x_1, \ldots, x_n$. Substituting this for ϕ in the joint density of the observations and standardizing appropriately we obtain

$$R_M(\theta) = \frac{f[x_1, \ldots, x_n;\ \theta, \hat{\phi}(\theta)]}{\sup\limits_{\theta \in \Theta} f[x_1, \ldots, x_n;\ \theta, \hat{\phi}(\theta)]}$$

a quantity which is called the maximum relative likelihood.

3. Marginal Likelihoods

The probability that the point $(x_1, \ldots, x_n)$ in the n-dimensional sample space falls in the elemental cuboid whose "edges" are of lengths $dx_1, \ldots, dx_n$ is $f(x_1, \ldots, x_n;\ \theta, \phi)\, dx_1 \ldots dx_n$. Suppose that the following conditions are satisfied in relation to this quantity.

(i) There is a non-singular transformation taking $x_1, \ldots, x_n$, into $y_1, \ldots, y_{n-r}, a_1, \ldots, a_r$ such that a factorization

$$f(x_1, \ldots, x_n;\ \theta, \phi)\, dx_1, \ldots, dx_n$$
$$= \{g(a_1, \ldots, a_r;\ \theta)\, da_1, \ldots, da_r\} \times \{h(y_1, \ldots, y_{n-r};\ \theta, \phi \mid a_1, \ldots, a_r)\, dy_1, \ldots, dy_{n-r}\}$$

where g and h are also density functions, exists.

(ii) If ϕ is unknown, then h contains "no available information" about θ.

Then the marginal likelihood function of θ is some function of θ proportional to

$$M(\theta) = g(a_1, \ldots, a_r;\ \theta)\, da_1, \ldots, da_r$$

and the corresponding standardized function

$$MR(\theta) = \frac{M(\theta)}{\sup\limits_{\theta \in \Theta} M(\theta)}$$

the marginal relative likelihood function of θ.

4. Conditional Likelihoods

These are defined in a way which is wholly analogous to the above. It is supposed that the following two conditions hold.

(i) That there is a factorization in the form

$$f(x_1, \ldots, x_n;\ \theta, \phi) = g(x_1, \ldots, x_n;\ \theta \mid T_1, \ldots, T_k)\frac{dx_1, \ldots, dx_n}{dT_1, \ldots, dT_k}$$
$$\times h(T_1, \ldots, T_k;\ \theta, \phi)\, dT_1, \ldots, dT_k$$

where the statistics T_i which by implication are jointly sufficient for ϕ, may or may not be functions of θ.

(ii) The factor $h(T_1, \ldots, T_k; \theta, \phi)\, dT_1, \ldots, dT_k$ contains no "available information" about θ if nothing is known about ϕ.

The "conditional likelihood" of θ is then defined as a function of θ proportional to

$$C(\theta) = g(x_1, \ldots, x_n;\ \theta \mid T_1, \ldots, T_k)\frac{dx_1, \ldots, dx_n}{dT_1, \ldots, dT_k}$$

which when standardized as

$$CR(\theta) = \frac{C(\theta)}{\sup_{\theta} C(\theta)}$$

is called the "conditional relative likelihood function" of θ.

5. Second-order Likelihoods

This concept is introduced by considering the situation where we have pairs of observations (x_i, y_i), $i = 1, \ldots, n$, such that x_i and y_i are independently distributed as $N(\xi_i, \sigma)$ and $N(U_i, \lambda\sigma)$ respectively, and a linear functional relationship $U_i = \theta\xi_i + \delta$ exists. Thus σ is a structural parameter and the ξ_i are incidental parameters.

The parameters of interest here are therefore θ and δ, since they define the functional relationship. Now if θ is known then it is shown that the marginal relative likelihood of δ, assuming λ also known, is

$$MR_\lambda(\delta \mid \theta) = (1+u^2)^{-\frac{1}{2}(n-1)}$$

where

$$u^2 = \frac{n(\theta\bar{x} - \bar{y} + \delta)^2}{\Sigma\{\theta(x_i - \bar{x}) - (y_i - \bar{y})\}^2}.$$

Since, however, $MR_\lambda(\delta \mid \theta)$ is independent of λ, then λ has been eliminated without being estimated.[1] Thus although λ is assumed known in the derivation of the result above, it is now apparent that if λ is unknown we can by means of $MR_\lambda(\delta \mid \theta)$ make inferences about δ for a given θ.

A logical deduction from this is that we can therefore make inferences about an unknown θ for a given δ. This is apparent by observing that if, for example, we know $\delta = 0$, then a low value of $MR_\lambda(0 \mid \theta)$ arising from a value of θ under test would suggest that such a value is implausible, and similarly a large value of $MR_\lambda(0 \mid \theta)$ would suggest it is a likely value.

Viewed in this way, $MR_\lambda(0 \mid \theta)$ is essentially a likelihood function, and since it bears the same relation to the likelihood as the likelihood does to the probability distribution it is described as a second order likelihood.

(B) The following "basic notions" are summarized in a paper by Kallianpur and Rao (1955).

(1) Any transformation $f(x)$ which has for its domain a subset D of a normed, linear space E and the set of real numbers for its range is called a (real-valued) functional.

(2) A functional f is continuous at $x_0 \in D$ if $\lim f(x) = f(x_0)$ as $\| x - x_0 \| \to 0$.

(3) A functional f having E for its domain is said to be additive if for any two elements x_1 and x_2

$$f(x_1 + x_2) = f(x_1) + f(x_2).$$

[1] It is well known that if λ is unknown then not all the parameters are identifiable. This means that a direct ML method cannot be applied. We have already alluded to this point in section 9.

(4) An additive, continuous functional f (defined over E) is called linear.

(5) Let f be any functional with domain $D \subseteq E$.

Then f is said to possess a Frechêt differential (F-differential) at x_0, if for $h \in E$ so that $x_0 + h \in D$ there exists a functional $L(x_0, h)$ linear in h such that

$$\lim_{\| h \| \to 0} \frac{[f(x_0+h)-f(x_0)-L(x_0, h)]}{\| h \|} = 0.$$