Efficient Estimates and Optimum Inference Procedures in Large Samples

By C. Radhakrishna Rao

Indian Statistical Institute, Calcutta

[Read at a Research Methods Meeting of the Society, December 6th, 1961, Professor D. R. Cox in the Chair]

Summary

The concept of efficiency in estimation is linked with closeness of approximation to the derivative of log likelihood, which plays an important role in statistical inference in large samples. Various orders of efficiency are defined depending on degrees of closeness, and properties of estimates satisfying these criteria are studied. Such measures of efficiency appear to be more appropriate than the one related to asymptotic variance of an estimate for judging the performance of an estimate, when used as a substitute for the whole sample in drawing inference about unknown parameters. It is found that, under some conditions, the maximum likelihood estimate has some optimum properties which distinguish it from all other large sample estimates.

1. Introduction

It is nearly four decades since Sir Ronald Fisher introduced the concept of likelihood which, as a function of unknown parameters given the sample, plays a fundamental role in statistical inference. He had also studied and established optimum properties of estimates obtained by maximizing likelihood, in the light of criteria of consistency and efficiency in large samples, and of sufficiency and amount of information in the case of small samples (Fisher, 1922, 1925). There has been, however, some controversy regarding the superiority of maximum likelihood (m.l.) estimates over others. For instance, it has been said that the method of m.l. is just one out of an infinity of estimation procedures yielding what are called B.A.N. (best asymptotically normal) estimates having the same optimum asymptotic properties as m.l. estimates (Neyman, 1949; Haldane, 1951), and that further criteria are necessary for establishing the superiority, if any, of m.l. estimates. It has also been thought that the existence of superefficient estimates, i.e. with asymptotic variances smaller than those of m.l. estimates, invalidates the concept of efficiency on which the use of m.l. estimates is advocated (LeCam, 1953).

The anomalies apparently arose in judging an estimate from narrow concepts, such as asymptotic variance and concentration, defined in a restricted manner round the true value, which are not by themselves well-conditioned indicators of the usefulness of an estimate in statistical inference. For instance, if T_n is an estimate of θ, the limit of

$$\Pr(|T_n-\theta|<\lambda/\sqrt{n};\ \theta)$$

as $n\to\infty$ is defined as limiting concentration, whereas the natural thing to do is to examine the probability of concentration in a *fixed* interval round the true value as

$n \to \infty$ and not in intervals tending to zero. Recent investigations at the Indian Statistical Institute on criteria for estimation and limiting properties of estimates, from the wider point of view of statistical inference, have led to some definite results regarding m.l. estimates, which I would like to present at this meeting.

The main line of investigation has been to enquire how good a given estimate is as a substitute for the whole sample in drawing inference about unknown parameters (Rao, 1960a, 1961). This approach is implicit in Fisher's work (Fisher, 1922, 1925) and is also stated by Barnard (1949) in his fundamental paper on statistical inference. Such an approach is considered by some statisticians as of limited value because it is general and not intended to answer specific questions in making decisions from observed data (Berkson, 1960). On the contrary, in many practical situations, the object of reduction of data in the form of an estimate is only to facilitate answering a variety of questions of immediate interest. Further, it would be more economical to preserve for future use an estimate instead of the whole sample; and this could be done satisfactorily only if the estimate is a good substitute for the sample. Another line of work initiated by Bahadur (1960) has been to study the concentration of an estimate in fixed intervals round the true value of a parameter, as the sample size increases. It may be observed from the approach adopted by Bahadur, or as explicitly demonstrated in the present paper, that concentration is equivalent to certain other properties of an estimate used as a substitute for the whole sample in tests of significance. Thus the approach developed in the earlier papers (Rao, 1960a, 1961) seems to provide a convenient framework for discussing the problem of estimation.

The criteria for judging the performance of an estimate compared with that of the whole sample are obtained by a suitable reformulation of consistency and efficiency introduced by Fisher (1922, 1925). This is done in section 2 and the properties of estimates satisfying these criteria are examined in sections 3, 4 and 5.

While thanking the Royal Statistical Society for giving me an opportunity to read a paper at one of its meetings, I must apologize for choosing a subject which may appear somewhat classical. But I hope this small attempt intended to state in precise terms what can be claimed about m.l. estimates, in large samples, will at least throw some light on current controversies.

2. CRITERIA OF CONSISTENCY AND EFFICIENCY

2.1. *Consistency*

It will help us in our discussion if we state in precise terms the criteria of consistency and efficiency, especially as there is some misunderstanding in the interpretations of the original definitions given by Fisher (1922, 1925). Although the concept of consistency as discussed in this section is not necessary for the development of the paper it has been included for the sake of completeness, to demonstrate how this criterion fits in with the general approach to the problem of estimation indicated in the introduction.

Let $\mathscr{X}^\infty$ be the space of infinite sequences of observations and S, the Kolmogoroff σ-field of measurable sets. Further $\{P_\theta\}$ represents a family of probability measures defined over S, and indexed by a parameter θ varying in a set $\{\theta\}$. Let $T_n(X_n)$ be a real valued statistic defined on $\mathscr{X}^n$, the space of the first n observations, X_n. The family of probability measures induced by T_n, which may be regarded as a function on X^∞, is $\{P_\theta T_n^{-1}\}$. Two probability measures μ, ν are said to be orthogonal $(\mu \perp \nu)$ if there exist disjoint sets A and B such that $\mu(A) = \nu(B) = 1$, $\mu(B) = \nu(A) = 0$.

Definition 2.1. Let $P_\theta T_n^{-1} \to \mu_\theta$ weakly. Then T_n is said to be consistent for the family $\{P_\theta\}$, if $\mu_\theta \perp \mu_\phi$ whenever $P_\theta \perp P_\phi$.

When the limiting measures μ_θ and μ_ϕ do not exist an alternative definition of consistency may be given.

Definition 2.2. The statistic T_n is said to be consistent for the family $\{P_\theta\}$ if for any given $\epsilon > 0$, there exists an $n_0(\epsilon)$ such that for each $n > n_0(\epsilon)$ disjoint sets A_n and B_n can be found in $\mathscr{X}^n$ with the property

$$P_\theta T_n^{-1}(A_n) \geqslant 1-\epsilon, \quad P_\phi T_n^{-1}(A) \leqslant \epsilon,$$

$$P_\theta T_n^{-1}(B_n) \leqslant \epsilon, \quad P_\phi T_n^{-1}(B_n) \geqslant 1-\epsilon,$$

whenever $P_\theta \perp P_\phi$.

It is easy to show that the definition 2.1 implies the definition 2.2 when μ_θ and μ_ϕ exist.

Orthogonality of measures P_θ and P_ϕ implies that complete discrimination is possible between these two members of the family as the sample size is increased indefinitely. The criterion of consistency states that the same could be achieved by considering only the estimate (as a statistic) in the place of the whole sample at each stage.

There are two definitions of consistency current in literature, one of which known as probability consistency (P.C.) requires that $T_n \to \theta$ weakly or strongly, in which case it is easy to see that the definitions 2.1 and 2.2 are satisfied. According to our wider concept, T_n would be consistent even if $T_n \to g(\theta)$, a single valued function of θ, and not necessarily to θ. The author has shown (Rao, 1960a) that a few examples of inconsistency of m.l. estimates recorded in literature relate to an estimate tending to a function of the parameter instead of the parameter with respect to which maximization of likelihood is sought.

Another definition suitable for sequences of independent and identically distributed observations is called Fisher consistency (F.C.). If S_n denotes the empirical distribution function based on n observations, Fisher (1956) considers a statistic $T_n = f(S_n)$ where f is a functional defined over the space of all distribution functions. Then T_n is said to be F.C. if $f(F_\theta) \equiv \theta$ where F_θ is the true distribution function. If the functional f is weakly continuous, F.C. implies P.C.

2.2. *Efficiency*

While consistency ensures that an estimate achieves perfect discrimination between alternative distributions as $n \to \infty$, efficiency is concerned with differences in discrimination provided by an estimate and the whole sample as $n \to \infty$. For simplicity, let us consider the case where probability densities exist and only one unknown parameter is involved. For a given θ, let $P(X_n, \theta)$ denote the probability density of the sample point X_n in $\mathscr{X}^n$ and $P(T_n, \theta)$, that for the statistic T_n. The best discriminator (discriminant function) between alternative distributions with indices θ and ϕ is the likelihood ratio $P(X_n, \theta)/P(X_n, \phi)$, while that based on T_n alone is $P(T_n, \theta)/P(T_n, \phi)$. Now T_n is equivalent to the whole sample if

$$\frac{P(X_n, \theta)}{P(X_n, \phi)} = \frac{P(T_n, \theta)}{P(T_n, \phi)} \quad (\text{all } \theta, \phi), \tag{2.1}$$

which is realized if T_n is sufficient for the unknown parameter.

In large samples it is perhaps relevant to consider alternatives close to one another. In such a case, if the first derivative $P'(X_n, \theta)$ of $P(X, \theta)$ with respect to θ exists, the discriminator may be written as $P'(X_n, \theta)/P(X_n, \theta)$. The corresponding expression for T_n is $P'(T_n, \theta)/P(T_n, \theta)$ and the difference

$$d(\theta, n) = \frac{P'(X_n, \theta)}{P(X_n, \theta)} - \frac{P'(T_n, \theta)}{P(T_n, \theta)} \tag{2.2}$$

plays a key role in studying the performance of T_n. The statistic T_n is equivalent in some sense to the whole sample, when n is large, if

$$n^{\alpha} d(\theta, n) \to 0 \quad \text{in probability,}$$

where α is chosen so that $n^{\alpha} P'(X_n, \theta)/P(X_n, \theta)$ does not itself converge to zero in probability. Usually $\alpha = -\frac{1}{2}$ serves the purpose.

Definition 2.3. A statistic T_n is said to be efficient (to the first order) if, for a suitable choice of α, such that the statistic $n^{\alpha} P'(X_n, \theta)/P(X_n, \theta)$ does not converge to zero, $n^{\alpha} d(\theta, n) \to 0$ in probability.

There may be a very wide class of statistics satisfying the criterion of first-order efficiency, in which case a further criterion may be necessary for restricting the choice of statistics. This should depend on the rapidity of convergence of $n^{\alpha} d(\theta, n)$ or the asymptotic behaviour of $d(\theta, n)$ itself. Since $E\{d(\theta, n)\} = 0$, $V\{d(\theta, n)\}$ may provide a satisfactory measure.

Definition 2.4. The second-order efficiency of T_n is

$$\lim_{n \to \infty} V\{d(\theta, n)\} = \lim_{n \to \infty} \{I(X_n) - I(T_n)\},$$

where $I(X_n)$ and $I(T_n)$ stand for the amounts of information (in Fisher's sense) contained in the sample and in the statistic respectively.

Second-order efficiency, as given in definition 2.4, examines the amount of information lost in using a statistic instead of the whole sample. This aspect was first examined by Fisher (1925) when he conjectured that m.l. estimates have the least limiting loss.

The criteria of efficiency given in definitions 2.3 and 2.4 are extremely difficult to verify in practice. They are, therefore, replaced by simpler definitions 2.5 and 2.6, which are formulated so as to incorporate the essential features of definitions 2.3 and 2.4 and to be equivalent to them under some restrictive conditions on the probability densities.

Definition 2.5. The statistic T_n is said to be efficient (first order) if

$$n^{\frac{1}{2}} \left\{ \frac{P'(X_n, \theta)}{nP(X_n, \theta)} - \beta(\theta)(T_n - \theta) \right\} \to 0 \quad \text{in probability,} \tag{2.3}$$

where $\beta(\theta)$ is a function of θ only.

Note that if (2.3) holds, T_n is automatically consistent, because it tends to θ in probability. If we replace $\beta(\theta)(T_n - \theta)$ by $\alpha(\theta) + \beta(\theta) T_n$, we have a more general situation, but this is not of direct interest in the context of the present investigation.

Definition 2.6. The second-order efficiency is the minimum asymptotic variance of

$$\frac{P'(X_n, \theta)}{P(X_n, \theta)} - n\beta(\theta)(T_n - \theta) - n\lambda(\theta)(T_n - \theta)^2, \tag{2.4}$$

when minimized with respect to λ.

In the rest of the paper we shall use first- and second-order efficiencies only in the sense of definitions 2.5 and 2.6 respectively.

2.3. *Second-order Efficiency of m.l. Estimates*

It has been shown by Rao (1961) that, under some regularity conditions, the first-order efficiency in the sense of definition 2.5 ensures that as $n \to \infty$

$$\frac{I(T_n)}{n} \to \frac{I(X_n)}{n} = i, \tag{2.5}$$

where $I(X_n)$ and $I(T_n)$ are the amounts of information contained in the sample and in the statistic respectively. There may be an infinity of estimation procedures for which (2.5) is true, in which case the second-order efficiency will be of use in restricting the choice to a subset.

It is of some interest to examine the conditions under which definitions 2.4 and 2.6 of second-order efficiency are equivalent. In such a case the results already proved regarding second-order efficiency of estimates have direct significance as statements concerning the actual amount of information lost.

Now, for a multinomial distribution, it has been shown by Rao (1961) that m.l. is the only method with an optimum second-order efficiency under the following conditions of which the first one is purely a restriction on the choice of a parameter:

(i) The parameter under consideration is a continuous functional of the distribution function;

(ii) The cell probabilities represented by $\pi_1(\theta), \ldots, \pi_k(\theta)$ admit continuous derivatives up to the second order;

(iii) The estimating equation

$$f(\theta, n_1/n, \ldots, n_k/n) = 0, \tag{2.6}$$

where $n_1, \ldots, n_k$ are observed frequencies in the k classes, is consistent, i.e., $f\{\theta, \pi_1(\theta), \ldots, \pi_k(\theta)\} = 0$ and has continuous derivatives up to the second order in θ, as well as in n_i/n considered as variables. A similar result may be proved in the case of continuous distributions but this would involve some further conditions.

In the light of the above result it may be of interest to examine the differences in the second-order efficiency of some common methods of estimation suggested as alternatives to m.l.

Some of these methods and E_2, the second-order efficiency computed on the basis of definition 2.6, are given in Table 1, where

$$\mu = \frac{\mu_{02} - 2\mu_{21} + \mu_{40}}{i} - i - \frac{\mu_{11}^2 + \mu_{30}^2 - 2\mu_{11}\mu_{30}}{i^2},$$

$$\Delta = \tfrac{1}{4}\Sigma\left(\frac{\pi_r'}{\pi_r}\right)^2 - \frac{\mu_{40}}{i} + \frac{\mu_{30}^2}{2i^2},$$

$$\mu_{rs} = \Sigma\pi_j(\pi_j'/\pi_j)^r(\pi_j''/\pi_j)^s, \quad i = \mu_{20}$$

and π_i' and π_i'' are the first and second derivatives of $\pi_i(\theta)$.

Although all the six methods considered in Table 1 provide first-order efficient estimates, they are clearly distinguishable by their second-order efficiencies. Both μ and Δ are non-negative which shows that m.l. is the best and minimum Hellinger's distance is better than the others. The modification made in the minimum chi-square method by substituting the sample frequency in the denominator for computational simplicity seems to bring in additional loss of information.

TABLE 1

Second-order efficiencies of different methods of estimation applicable to a multinomial distribution

Method of estimation	*Formula*	E_2
max. likelihood	$\Sigma n_i \log \pi_i$	μ
min. chi-square	$\sum \frac{(n_i - n\pi_i)^2}{n\pi_i}$	$\mu+\Delta$
min. modified chi-square†	$\sum \frac{(n_i - n\pi_i)^2}{n_i}$	$\mu+4\Delta$
min. discrepancy†	$\sum \frac{\pi_i^{k+1}}{n_i^k}$	$\mu+(k+1)^2\Delta$
min. Kullback–Leibler separator	$\Sigma \pi_i \log (\pi_i/n_i)$	$\mu+\Delta$
min. Hellinger's distance	$\Sigma \sqrt{(n_i \pi_i)}$	$\mu+\frac{\Delta}{4}$

† The actual formulae are slightly different from those given in Table 1, but the changes made do not affect the large sample properties of the estimates.

2.4. *First-order Efficiency and Asymptotic Variance*

The concept of efficiency has been generally linked with the attainment of the least asymptotic variance for a consistent estimate. It is also believed that, without any restriction on the class of estimates as functions of observations, the asymptotic variance of $(T_n - \theta)\sqrt{n}$ cannot be less than $1/i$, the reciprocal of information, provided only T_n is consistent and asymptotically normally distributed. Unfortunately, the last result is not true as shown by Hodges in an example quoted by LeCam (1953). Kallianpur and Rao (1955) have given some sufficient conditions for the existence of a lower bound. The main restrictions imposed on T_n are:

(i) $T_n = f(S_n)$ where f is a functional defined over the space of distribution functions, and

(ii) T_n is F.C. and f is Frechet differentiable.

It can be shown (the proof is omitted as it is straightforward), under the same conditions, that the attainment of the lower bound $1/i$ implies that the statistic under consideration has first-order efficiency. It is of some interest to examine whether this result could be established withdrawing some of the severe restrictions imposed on T_n.

But the concept of first-order efficiency is applicable to a wider class of estimates which may not have any lower bound to their variance. In fact it need not bear any relationship to asymptotic variance, and estimates with a lower asymptotic variance may be less efficient in the present sense in certain situations. To give an obvious example, consider the statistic T_n based on a sample of n observations from a normal distribution with variance unity and unknown mean θ. Let

$$T_n = \begin{matrix} \bar{x} & (|\bar{x}| \geqslant n^{-\frac{1}{4}}), \\ \alpha x_m & (|\bar{x}| \leqslant n^{-\frac{1}{4}}), \end{matrix} \qquad (2.7)$$

where $\bar{x}$ and x_m are the mean and median respectively. Obviously T_n is consistent for the mean of the normal distribution and its asymptotic variance is $\frac{1}{2}\alpha^2\pi/n$ when $\theta = 0$ and $1/n$ otherwise. The asymptotic variance of T_n is the same as that of $\bar{x}$ when $\theta \neq 0$ and can be made smaller than that of $\bar{x}$ at $\theta = 0$ by choosing α arbitrarily small. Now T_n, being stochastically equivalent to the median when $\theta = 0$, is, obviously, not a more satisfactory estimate than $\bar{x}$ from any point of view. Indeed one can construct, by an extension of Hodges's method suggested by LeCam (1953), T_n so as to be asymptotically equivalent to the median for a countable set of values of θ and possessing a smaller asymptotic variance at these points. That would be making the position worse. What in effect first-order efficiency demands is not that the asymptotic variance of an estimate is a minimum but its asymptotic correlation with the derivative of log likelihood be unity. For T_n constructed in (2.7), the asymptotic correlation is $\sqrt{(2/\pi)}$ at $\theta = 0$ and unity elsewhere, whereas for $\bar{x}$, it is unity for all θ. If the deficiency in an estimate is measured by $1-r^2$, where r is the asymptotic correlation, then the deficiency in T_n defined in (2.7) is about 0·363 at $\theta = 0$.

2.5. *Assumptions and Notations*

In the rest of the paper we shall consider only sequences of independent and identically distributed variables with probability density $p(x, \theta)$ with respect to a σ-finite measure v_1 and distribution function $F_\theta(x)$. The probability density of n observations is denoted by $P(X_n, \theta)$ with respect to v_n, and that of T_n with respect to a σ-finite measure u by $P(T_n, \theta)$. The following assumptions are referred to in the various propositions proved.

Assumption 1. The derivative of $p(x, \theta)$ exists and $i = E(d\log p/d\theta)^2$ is finite.

Assumption 2. If E_n is any Lebesgue measurable set in $\mathfrak{X}^n$,

$$\frac{d}{d\theta}\int_{E_n}\!\!\cdots\!\int P(X_n, \theta)\,dv_n = \int_{E_n}\!\!\cdots\!\int \frac{dP(X_n, \theta)}{d\theta}\,dv_n,$$

$$\frac{d}{d\theta}\int_{E_1} P(T_n, \theta)\,du = \int_{E_1} \frac{dP(T_n, \theta)}{d\theta}\,du.$$

It may be noted that, as a consequence of assumption 1, the statistic

$$Z_n = \{P'(X_n, \theta)/P(X_n, \theta)\}/\sqrt{n}$$

is asymptotically normally distributed with mean zero and variance i. Further $E(Z_n) = 0$, $V(Z_n) = i$ for each n.

If we define

$$Y_n = \{P'(T_n, \theta)/P(T_n, \theta)\}/\sqrt{n},$$

the information contained in T_n is

$$V(Y_n\sqrt{n}) = ni_T.$$

If T_n has first-order efficiency in the sense of definition 2.5, it has been shown by Rao (1961) that both $(T_n - \theta)\sqrt{n}$ and Y_n are normally distributed in the limit.

3. Efficient Estimates and Tests of Significance

In this section, we establish some optimum properties of tests of significance based on first-order efficient estimates. Since optimum tests provide a basis for interval estimation, a justification of the choice of efficient estimates in two important methodological problems is provided. The main results are given in Theorems 1 and 2.

Lemma 3.1. Let $r_n(\theta)$ be the power function of a test of the hypothesis $H_0:\theta = \theta_0$ (a specified value), at a given level of significance α. If $r_n'(\theta_0)$ is the derivative of $r_n(\theta)$ at θ_0, then, under Assumptions 1 and 2,

$$\overline{\lim_{n\to\infty}}\,\{n^{-\frac{1}{2}}r_n'(\theta_0)\} \leqslant (\tfrac{1}{2}i/\pi)^{\frac{1}{2}}e^{-\frac{1}{2}a^2} \qquad (3.1$$

where a is the α per cent point of the standard normal distribution.

For any n it is known (Rao and Poti, 1946) that the test based on the critical region

$$w_n = \{X_n : P'(X_n, \theta_0)/P(X_n, \theta_0) \geqslant a_n \sqrt{(ni)}\},$$

where a_n is chosen such that the size of w_n is α, has a maximum slope for the power function at θ_0. Hence

$$r_n'(\theta_0) \leqslant \int_{w_n} P'(X_n, \theta_0)\,dv_n = \sqrt{n}\int_{w_n} Z_n P(X_n, \theta_0)\,dv_n.$$

Dividing by $\sqrt{n}$ and taking limits as $n\to\infty$, we have that

$$\overline{\lim}\, r_n'(\theta_0)/\sqrt{n} \leqslant \lim\int_{Z_n > a_n\sqrt{i}} Z_n P(Z_n, \theta_0)\,dv_n$$

$$= \int_{Z>a\sqrt{i}} (2\pi i)^{-\frac{1}{2}} Z e^{-\frac{1}{2}Z^2/i}\,dZ = (\tfrac{1}{2}i/\pi)^{\frac{1}{2}}e^{-\frac{1}{2}a^2}. \qquad (3.2)$$

Lemma 3.2. If T_n has first-order efficiency and Assumptions 1 and 2 are satisfied, then $i_T \to i$.

The proof is given by Rao (1961). We shall now prove Theorem 1, which shows that a test based on an efficient estimate has maximum local power asymptotically.

Theorem 1. Let T_n be a first-order efficient estimate and $P'(T_n, \theta_0) > \lambda_n P_n(T_n, \theta_0)$, a test of the hypothesis $H_0 : \theta = \theta_0$ at a given level of significance α. If $r_n(\theta)$ is the power function of this test then, under Assumptions 1 and 2,

$$\lim_{n\to\infty} r_n'(\theta_0)/\sqrt{n} = (\tfrac{1}{2}i/\pi)^{\frac{1}{2}}e^{-\frac{1}{2}a^2}, \qquad (3.3)$$

the upper bound derived in Lemma 3.1.

Following the notation of section 2.5, let

$$w_n = \{X_n : Z_n > c_n\} \quad \text{and} \quad w_T = \{X_n : Y_n \geqslant b_n\} \qquad (3.4)$$

be two critical regions of the same size α. Since $i - i_T = E(Z_n - Y_n)^2$ and $i_T \to i$, as a result of Lemma 3.2,

$$\lim_{n\to\infty}\int_{w_n}(Z_n - Y_n)^2 P(X_n, \theta_0)\,dv_n = 0.$$

Hence $$\lim_{n\to\infty}\int_{w_n} Z_n P(X_n, \theta_0)\,dv_n = \lim_{n\to\infty}\int_{w_n} Y_n P(X_n, \theta_0)\,dv_n. \qquad (3.5)$$

But $$\int_{w_T} Y_n P(T_n, \theta_0)\,du \geqslant \int_{w_n} Y_n P(X_n, \theta_0)\,dv_n$$

because $Y_n \geqslant b_n$ is the locally powerful test based on T_n. Hence

$$\varliminf_{n\to\infty} \int_{w_T} Y_n P(T_n, \theta_0)\, du = \varliminf_{n\to\infty} r_n'(\theta_0)/\sqrt{n}$$

$$\geqslant \lim_{n\to\infty} \int_{w_n} Y_n P(X_n, \theta_0)\, dv_n = \lim_{n\to\infty} \int Z_n P(X_n, \theta_0)\, dv_n$$

$$= (\tfrac{1}{2}i/\pi)^{\frac{1}{2}}\, e^{-\frac{1}{2}a^2}, \tag{3.6}$$

using (3.2). The result (3.3) of Theorem 1 follows from (3.1) and (3.6).

It is important to consider the property of the test based on T_n in the form $(T_n - \theta_0)\sqrt{n} \geqslant \lambda$, which is usually used, and not as considered in Theorem 1 using the log derivative of the density function of T_n. Theorem 2 shows that the same local property is true of such a test.

Theorem 2. Let the assumptions on the probability density $p(x, \theta)$ and T_n be as in Theorem 1. If the test criterion is $U_n = (T_n - \theta_0)\sqrt{n} \geqslant \lambda$ where λ is chosen such that the limiting level of significance is α, then $r_n'(\theta_0)/\sqrt{n}$ has the same limit as in Theorem 1.

By definition
$$r_n'(\theta_0)/\sqrt{n} = \int_{U_n > \lambda} P'(X_n, \theta_0)\, dv_n/\sqrt{n}$$

$$= \int_{U_n > \lambda} Z_n P(X_n, \theta_0)\, dv_n. \tag{3.7}$$

Since $|Z_n - \beta U_n| \to 0$ in probability because of the first-order efficiency of T_n, the joint asymptotic distribution of Z_n and U_n exists and is, in fact, degenerate. If $F(Z, U)$ represents this asymptotic distribution, (3.7) tends to

$$\int_{U > \lambda} Z dF(Z, U) = \int_{Z > \lambda/\beta} Z dF(Z) = \int_{Z > a\sqrt{i}} Z dF(Z),$$

since λ/β is to be chosen so that $\Pr(Z > \lambda/\beta) = \alpha$, which proves the required result.

The foregoing analysis suggests the following definition of limiting efficiency of a test.

Definition. Let $r_n(\theta)$ be the power function of a test of the hypothesis H_0; $\theta = \theta_0$ at a fixed level of significance α. The limiting efficiency of a test is then

$$\left\{\varliminf_{n\to\infty} r_n'(\theta_0)/\sqrt{n}\right\} \Big/ J(\theta_0), \tag{3.8}$$

or its square, where $J(\theta_0)$ is the upper bound derived in lemma 3.1.

An explicit evaluation of the limiting efficiency is provided by Lemma 3.3.

Lemma 3.3. Let the test criterion be $U_n = (T_n - \theta)\sqrt{n} \geqslant \lambda$ and let the joint limiting distribution of Z_n and U_n be bivariate normal with correlation coefficient ρ. Then under the assumptions of Theorem 2, the limiting efficiency of the test is ρ.

For a first-order efficient estimate the asymptotic correlation between the estimate and Z_n is unity, in which case the efficiency of any other estimate may be measured by its asymptotic correlation with Z_n. Lemma 3.3 shows that this measure is directly related to the asymptotic slope of the power function of the test based on the estimate.

To prove Lemma 3.3, we have that

$$\lim_{n\to\infty}\int_{U_n>\lambda} Z_n P(X_n,\theta_0)\,dv_n = \int_{U>\lambda} Z\,dF(Z,U)$$

$$= \int_{U>\lambda} Z\,dF(Z|U)\,dF(U) = \int_{U>\lambda} \eta U\,dF(U), \qquad (3.9)$$

where ηU is the regression of Z on U, so that $\eta = \rho\sigma_Z/\sigma_U$. The last expression in (3.9), after simplification, reduces to

$$(\tfrac{1}{2}i\rho^2/\pi)^{\frac{1}{2}}e^{-\frac{1}{2}\lambda^2},$$

and dividing by the upper bound (3.1) we obtain the efficiency of the test as ρ.

Sundrum (1954) observed that in examining linkage in inheritance of two factors, the test based on the m.l. estimate of the recombination fraction (linkage parameter) is locally less powerful than an alternative large sample test. This cannot, obviously, be true in view of what has been established in Theorem 2, since the m.l. estimate in this particular situation has first-order efficiency. The alternative test is based on a statistic which happens to be *efficient* when linkage does not exist, i.e. for a particular value of the linkage parameter. Hence it is expected to have the same *local property* as the test based on the m.l. estimate for this particular value of the parameter. Sundrum's result is, therefore, misleading, especially as he claims to provide a justification for a statement attributed to Fisher (1950, pp. 314–315) that good tests may be based on inefficient estimators. The particular inefficient estimator referred to by Fisher happens to be efficient at the point specified by the null hypothesis, and it provides a test as good as the m.l. estimate, but not better. The results of section 4 of this paper will show that efficiency in an interval of the unknown parameter ensures some other desirable properties.

4. STRONGER FIRST-ORDER EFFICIENCY AND TESTS OF SIGNIFICANCE

The limiting properties proved in Theorems 1 and 2 state that, compared to any other given test, the power of a one-sided test based on a first-order efficient estimate is not less (and is perhaps better) in a neighbourhood of θ_0, for each sufficiently large n, but the neighbourhood may depend on n. It would, indeed, be better if it could be claimed of any test that its power cannot be less than that of any other given test in a specified interval of θ for all sufficiently large n. It appears that such a statement can be made if the test is based on an estimate that bears a relation to Z_n stronger than that implied by first-order efficiency.

Lemma 4.1. Let w_n be a critical region of size α_n in $\mathscr{X}^n$ for testing the hypothesis, $H_0:\theta=\theta_0$ and $\beta_n(\theta)$ the second kind of error. If α_n is bounded away from unity, then

$$\varliminf_{n\to\infty} n^{-1}\log\beta_n(\theta) \geqslant \int p(x,\theta_0)\log\frac{p(x,\theta)}{p(x,\theta_0)}\,dv_1 = \mu(\theta,\theta_0). \qquad (4.1)$$

If $\mu(\theta,\theta_0)=-\infty$, there is nothing to prove. Let $\mu(\theta,\theta_0)$ be finite and define $w_n' = R_n - w_n$. Then

$$\int_{w_n'} P(X_n,\theta_0)\,dv_n = 1-\alpha_n, \quad \beta_n = \int_{w_n'} P(X_n,\theta)\,dv_n$$

and

$$n^{-1}\log\left\{\frac{\beta_n}{1-\alpha_n}\right\} = n^{-1}\log\left\{\int_{w_n'}\frac{P(X_n,\theta)}{1-\alpha_n}dv_n\right\} \geqslant n^{-1}\int_{w_n'}\log\left\{\frac{P(X_n,\theta)}{P(X_n,\theta_0)}\right\}\frac{P(X_n,\theta_0)}{1-\alpha_n}dv_n$$

$$= \mu(\theta,\theta_0)+\int_{w_n'}\left[n^{-1}\log\left\{\frac{P(X_n,\theta)}{P(X_n,\theta_0)}\right\}-\mu(\theta,\theta_0)\right]\frac{P(X_n,\theta_0)}{1-\alpha_n}dv_n. \quad (4.2)$$

Hence
$$\underline{\lim}_{n\to\infty} n^{-1}\log\beta_n \geqslant \mu(\theta,\theta_0),$$

since, by the mean ergodic theorem (Doob, 1953, p. 469), the second expression in (4.2) tends to zero.

Lemma 4.2. For the likelihood ratio test $P(X_n,\theta)\geqslant\lambda_n P(X_n,\theta_0)$ such that the size of the critical region $\alpha_n\to\alpha, 0<\alpha<1$,

$$\lim_{n\to\infty} n^{-1}\log\beta_n(\theta) = \mu(\theta,\theta_0) \quad (4.3)$$

where $\mu(\theta,\theta_0)$ is as defined in Lemma 4.1.

By definition

$$\beta_n = \int_{w_n'} P(X_n,\theta)\,dv_n \leqslant \lambda_n\int_{w_n'} P(X_n,\theta_0)\,dv_n = \lambda_n(1-\alpha_n).$$

Since
$$n^{-1}\log\left\{\frac{P(X_n,\theta)}{P(X_n,\theta_0)}\right\}\to\mu(\theta,\theta_0)$$

in probability, we have, as observed by Basu (1953),

$$\lambda_n = \exp\{n\mu(\theta,\theta_0)+o(n)\},$$

provided $\mu(\theta,\theta_0)$ is finite. Hence

$$\lim_{n\to\infty} n^{-1}\log\beta_n \leqslant \lim_{n\to\infty} n^{-1}\log(1-\alpha_n)+\lim_{n\to\infty} n^{-1}\log\lambda_n = \mu(\theta,\theta_0). \quad (4.4)$$

If $\mu(\theta,\theta_0)=-\infty$, then $n^{-1}\log\lambda_n$ can be made less than $-k$ for any given positive k by choosing n sufficiently large. Hence the limit of the right-hand side of (4.4) is $<-k$. Since k is arbitrary,

$$\overline{\lim}_{n\to\infty} n^{-1}\log\beta_n \leqslant -\infty.$$

On combining (4.4) with (4.1) of Lemma 4.1, the required result is obtained.

It is stated by Chernoff (1956) that the result of Lemma 4.2, which is independent of the limit of α_n, is contained in an unpublished paper by Stein. The author is not aware of the conditions imposed by Stein, but it is interesting that the result needs no assumption whatsoever to be made on the likelihood ratios.

Corollary. If β_n and $\mu(\theta,\theta_0)$ are as defined in Lemma 4.1 and

$$\mu(\theta,\theta_0) = -\tfrac{1}{2}(\theta-\theta_0)^2\, i(\theta_0)+o\{(\theta-\theta_0)^2\}$$

then
$$\underline{\lim}_{\theta\to\theta_0}\ \underline{\lim}_{n\to\infty}\frac{n^{-1}\log\beta_n(\theta)}{(\theta-\theta_0)^2}\geqslant-\frac{i(\theta_0)}{2}. \quad (4.5)$$

Bahadur (1960) proved the same inequality when β_n is replaced by the probability of concentration of a consistent estimate in the neighbourhood of the true value θ_0. We shall in fact show in section 5 that the probability of concentration of a consistent estimate round the true value is related to the power of a test based on the estimate in a simple way.

The inequality (4.1) is, however, quite general, and if

$$\frac{\mu(\theta, \theta_0)}{\delta(\theta, \theta_0)} \to k(\theta_0) \quad \text{as} \quad \theta \to \theta_0,$$

where $\delta(\theta, \theta_0)$ is a non-negative function such that $\delta(\theta, \theta_0) \to 0$ as $\theta \to \theta_0$, then

$$\varliminf_{\theta \to \theta_0} \varliminf_{n \to \infty} \frac{n^{-1} \log \beta_n(\theta)}{\delta(\theta, \theta_0)} \geqslant k(\theta_0),$$

which is a general form of the type of inequality given in (4.5).

Lemma 4.3. Besides Assumptions 1 and 2, let (i) the moment generating function of $z(\theta_0) = p'(x, \theta_0)/p(x, \theta_0)$ exist for each true value of the parameter θ in the neighbourhood of θ_0, and (ii)

$$\int z(\theta_0)\, p(x, \theta)\, dv_1 = (\theta - \theta_0)\, i(\theta_0) + o(\theta - \theta_0) \quad \text{as} \quad \theta \to \theta_0,$$

$$\int z^2(\theta_0)\, p(x, \theta)\, dv_1 = i(\theta_0) + o(1) \quad \text{as} \quad \theta \to \theta_0,$$

and $$\int z^3(\theta_0)\, p(x, \theta)\, dv_1$$

be bounded as a function of θ in the neighbourhood of θ_0.

Then for the test $Z_n(\theta_0) \geqslant \lambda$ (fixed)

$$\lim_{\theta \to \theta_0 +} \lim_{n \to \infty} \frac{n^{-1} \log \beta_n(\theta)}{(\theta - \theta_0)^2} = -\frac{i(\theta_0)}{2}, \tag{4.6}$$

where $\beta_n(\theta)$ is the second kind of error.

Lemma 4.3 shows that for the test $Z_n(\theta_0) \geqslant \lambda$, which is known to be locally most powerful on one side, the stronger result (4.6), which ensures its superiority over any other test in a neighbourhood of θ_0 as $n \to \infty$, is true under the additional conditions (i) and (ii).

Under the conditions (i) and (ii) it is easy to show that

$$\begin{aligned} \log G(\theta, \theta_0) &= \log \int \exp\{-(\theta - \theta_0)\, z(\theta_0)\}\, p(x, \theta)\, dv_1 \\ &= \log \left[1 - \frac{(\theta - \theta_0)^2}{2}\, i(\theta_0) + o\{(\theta - \theta_0)^2\}\right]. \end{aligned}$$

Therefore

$$\varliminf_{\theta \to \theta_0} (\theta - \theta_0)^{-2} \log \int \exp\{-(\theta - \theta_0)\, z(\theta_0)\}\, p(x, \theta)\, dv_1 = -\frac{i(\theta_0)}{2}. \tag{4.7}$$

Consider $$\begin{aligned} &\int \exp[-(\theta - \theta_0)\{Z_n(\theta_0) - \lambda\} \surd n]\, P_n(x, \theta)\, dv_n \\ &\qquad = \exp\{\lambda(\theta - \theta_0) \surd n\}\{G(\theta, \theta_0)\}^n. \end{aligned} \tag{4.8}$$

By Tchebycheff's inequality, (4.8) is not less than

$$\int_{Z_n(\theta_0)<\lambda} P(X_n, \theta)\, dv_n = \beta_n(\theta)$$

provided $(\theta-\theta_0) \geqslant 0$, and therefore

$$n^{-1}\log\beta_n(\theta) \leqslant \frac{(\theta-\theta_0)\lambda\sqrt{n}}{n} + \log G(\theta, \theta_0)$$

and

$$\overline{\lim_{\theta\to\theta_0+}}\ \overline{\lim_{n\to\infty}}\ \frac{n^{-1}\log\beta_n}{(\theta-\theta_0)^2} \leqslant \frac{-i(\theta_0)}{2}, \tag{4.9}$$

using (4.7). Combining (4.9) with (4.5), we obtain the result (4.6).

Corollary. If the test is of the form $|Z_n(\theta_0)| \geqslant \lambda$, and $\xi_n(\theta) = \Pr(|Z_n(\theta_0)| < \lambda|\theta)$, then, under the conditions assumed in Lemma 4.3,

$$\lim_{\theta\to\theta_0}\ \lim_{n\to\infty} \frac{n^{-1}\log\xi_n(\theta)}{(\theta-\theta_0)^2} = -\frac{i(\theta_0)}{2}, \tag{4.10}$$

where θ can approach θ_0 from either side.

It is easy to see that $\xi_n(\theta) \leqslant \beta_n(\theta)$ of Lemma 4.3. Hence (4.9) is true with $\beta_n(\theta)$ replaced by $\xi_n(\theta)$. Similarly it can be proved that

$$\lim_{\theta\to\theta_0-}\ \lim_{n\to\infty} \frac{n^{-1}\log\xi_n(\theta)}{(\theta-\theta_0)^2} = -\frac{i(\theta_0)}{2}. \tag{4.11}$$

Hence (4.10) follows.

Lemma 4.4. Let x be a random variable such that $E(x) = 0$ and $0 < E(x^2) < \infty$. Also let $\phi(t) = E(e^{tx}) < \infty$ for all t with $|t| < \delta > 0$. If $\bar{x}$ is the mean of n independent observations on x, then

$$\Pr(|\bar{x}| > \epsilon) = \exp[-\tfrac{1}{2}n\epsilon^2\{1+\delta_n(\epsilon)\}/E(x^2)], \tag{4.12}$$

where

$$\lim_{\epsilon\to 0}\ \lim_{n\to\infty} \delta_n(\epsilon) = 0.$$

This lemma is due to Bahadur (1960), who uses the earlier work of Cramer (1938) and Chernoff (1952).

What we need for proving further results is the inequality (4.14) given below, which may be deduced independently or by applying (4.12) to the random variable $p'(x, \theta)/p(x, \theta)$. In our notation,

$$\log\Pr(|Z_n/\sqrt{n}| > \epsilon) = -\frac{n\epsilon^2}{2i(\theta)}\{1+\delta_n(\epsilon)\},$$

$$\lim_{n\to\infty} n^{-1}\log\Pr(|Z_n/\sqrt{n}| > \epsilon) = -\frac{\epsilon^2}{2i(\theta)}, \tag{4.13}$$

which implies that, considering Z_n/i instead of Z_n, for sufficiently large n,

$$\log\Pr(|Z_n/(i\sqrt{n})| > \epsilon) < \exp\{-\tfrac{1}{2}n\epsilon'^2 i(\theta)\}, \tag{4.14}$$

where $\epsilon' < \epsilon$.

As observed earlier, the property proved in Lemma 4.3 is expected to be true of a test based on an estimate T_n provided there is a strong stochastic relationship between T_n and Z_n, which may be called stronger first-order efficiency of an estimate. This is stated in Theorem 3 which establishes the corresponding result for a test based on an estimate.

Theorem 3. Let T_n be an estimate such that for a given ϵ, either

$$\Pr(|T_n-\theta-Z_n(\theta)/(i\sqrt{n})|\geqslant\epsilon|T_n-\theta|;\ \theta)<\rho_\epsilon^n; \tag{4.15}$$

or

$$\Pr(|T_n-\theta-Z_n(\theta)/(i\sqrt{n})|\geqslant\epsilon|Z_n(\theta)/(i\sqrt{n})|;\ \theta)<\rho_\epsilon^n,$$

where $\rho_\epsilon<1$ and is independent of θ for some interval of θ enclosing θ_0. Then for the test $(T_n-\theta_0)\sqrt{n}\geqslant\lambda$ of the hypothesis $H_0:\theta=\theta_0$

$$\lim_{\theta\to\theta_0+}\ \lim_{n\to\infty}\frac{n^{-1}\log\beta_n}{(\theta-\theta_0)^2}=-\frac{i(\theta_0)}{2}. \tag{4.16}$$

By definition

$$\beta_n=\Pr[(T_n-\theta_0)\sqrt{n}<\lambda;\ \theta]$$
$$=\Pr\left\{T_n-\theta\leqslant\frac{\lambda}{\sqrt{n}}-(\theta-\theta_0);\ \theta\right\}.$$

Let us define the events

$$A=\{X_n:|T_n-\theta-Z_n(\theta)/(i\sqrt{n})|<\epsilon|T_n-\theta|\};$$
$$B=\{X_n:|Z_n(\theta)/(i\sqrt{n})|<(h-\lambda/\sqrt{n})(1-\epsilon)\}\quad\text{for}\quad h=(\theta-\theta_0)>0;$$
$$C=\{X_n:|T_n-\theta|<h-\lambda/\sqrt{n}\}.$$

It is easy to see that events A and B together imply C. Hence

$$\Pr(C)\geqslant\Pr(AB)$$

or

$$1-\Pr(C)\leqslant1-\Pr(AB)\leqslant1-\Pr(A)+1-\Pr(B)$$
$$\leqslant\rho_\epsilon^n+\exp\{-nh^2i(\theta)(1-\epsilon')^2\}, \tag{4.17}$$

for large n, where $\epsilon'>\epsilon$, using the result (4.14) for $1-\Pr(B)$. Keeping ϵ fixed we can decrease h such that the second term in (4.17) dominates over ρ_ϵ^n. Hence for sufficiently small h

$$\overline{\lim_{n\to\infty}}\ n^{-1}\log\{1-\Pr(C)\}\leqslant\lim_{n\to\infty}\ n^{-1}\log[\rho_\epsilon^n+\exp\{-nh^2i(\theta)(1-\epsilon')^2\}],$$
$$=-\frac{h^2(1-\epsilon')^2i(\theta)}{2}.$$

Dividing by h^2 and taking limits as $h\to0$,

$$\overline{\lim_{\lambda\to0}}\ \overline{\lim_{n\to\infty}}\ \frac{\log\{1-\Pr(C)\}}{nh^2}\leqslant\frac{-i(\theta_0)(1-\epsilon')^2}{2}.$$

Since ϵ' is arbitrary,

$$\overline{\lim_{\lambda\to0}}\ \overline{\lim_{n\to\infty}}\ \frac{\log\{1-\Pr(C)\}}{nh^2}\leqslant\frac{-i(\theta_0)}{2}. \tag{4.18}$$

Further, since $\beta_n\leqslant1-\Pr(C)$, the result (4.18) remains true with β_n in the place $1-\Pr(C)$. Now using (4.5), the result (4.16) of Theorem 3 is established.

Corollary. For the two-sided test $|T_n-\theta_0|\sqrt{n}\geqslant\lambda$,

$$\lim_{\theta\to\theta_0}\ \lim_{n\to\infty}\frac{n^{-1}\log\beta_n}{(\theta-\theta_0)^2}=-\frac{i(\theta_0)}{2}.$$

The proof is on the same lines as that of the corollary of the Theorem 2.

5. STRONGER FIRST-ORDER EFFICIENCY AND LIMITING CONCENTRATION

In this section, we shall consider the problem of limiting concentration as developed by Bahadur (1960). As observed earlier, the main result is a consequence of Lemma 4.1 which provides a lower bound to the second kind of error. The same bound holds for the probability of deviation of an estimate from the true value by a given amount, under a mild restriction on the estimate. The connexion between the second kind of error, which plays a fundamental role in the Neyman–Pearson theory of testing of hypotheses, and the probability of deviation in an estimate has been exploited by Birnbaum (1961) in proving some optimum properties of m.l. estimates in small samples. The attempt has not been completely successful in the sense that no general statements could be made about m.l. estimates, which shows that the small sample problem may have to be viewed from an entirely different angle free from the concept of long-run frequency of errors in the estimate (Barnard, 1949).

Lemma 5.1. Let T_n be a statistic such that $\Pr(T_n \geqslant \theta;\, \theta)$ is bounded away from unity as $n \to \infty$ for each $\theta < \theta_0$, then for $h > 0$

$$\underline{\lim}_{n\to\infty}\, n^{-1} \log \Pr(T_n - \theta_0 < -h;\, \theta_0) \geqslant \mu(\theta_0, \theta_0 - h). \tag{5.1}$$

If $\mu(\theta_0, \theta_0 - h) = -\tfrac{1}{2}h^2 i(\theta_0) + o(h^2)$, then

$$\underline{\lim}_{h\to 0}\, \underline{\lim}_{n\to\infty}\, n^{-1} h^{-2} \Pr(T_n - \theta_0 < -h;\, \theta_0) \geqslant -\frac{i(\theta_0)}{2}. \tag{5.2}$$

Consider the test $T_n \geqslant \theta_0 - h$ of the null hypothesis, H_0: $\theta = \theta_0 - h$. The second kind of error when θ_0 is the true value is $\Pr(T_n < \theta_0 - h;\, \theta_0)$. Hence an application of Lemma 4.1 gives the result (5.1). Equation (5.2) follows from (5.1) by considering the expansion of $\mu(\theta_0, \theta_0 - h)$.

Lemma 5.2, due to Bahadur (1960), follows from the results of Lemma 5.1. The conditions imposed by Bahadur are, however, slightly different.

Lemma 5.2. If T_n is a statistic such that one or both of $\Pr(T_n \geqslant \theta;\, \theta)$, $\theta < \theta_0$ and $\Pr(T_n \leqslant \theta;\, \theta)$, $\theta > \theta_0$ are bounded away from unity as $n \to \infty$, then

$$\underline{\lim}_{h\to 0}\, \underline{\lim}_{n\to\infty}\, \frac{\log \Pr(|T_n - \theta_0| > h)}{nh^2} \geqslant -\frac{i(\theta_0)}{2}. \tag{5.3}$$

Let $\Pr(T_n \geqslant \theta;\, \theta)$ $(\theta < \theta_0)$ be bounded away from unity. Then (5.3) follows from (5.2) by observing that

$$\Pr(|T_n - \theta_0| > h) \geqslant \Pr(T_n - \theta_0 < -h).$$

The same result can be established if $\Pr(T_n \leqslant \theta;\, \theta)$ is bounded away from unity for $\theta > \theta_0$.

Theorem 4. Let T_n be such that for given ϵ, when θ_0 obtains

$$\Pr\{|T_n - \theta_0 - Z_n(\theta_0)/(i\sqrt{n})| \geqslant \epsilon |T_n - \theta_0|\} < \rho^n \tag{5.4}$$

or

$$\Pr\{|T_n - \theta_0 - Z_n(\theta_0)/(i\sqrt{n})| \geqslant \epsilon |Z_n(\theta_0)/(i\sqrt{n})|\} < \rho^n, \tag{5.5}$$

where $\rho < 1$, may depend on θ_0 and ϵ. If the conditions of Lemma 5.2 are true, then

$$\lim_{h\to 0}\, \lim_{n\to\infty}\, \frac{\log \Pr(|T_n - \theta| > h)}{nh^2} = -\frac{i(\theta_0)}{2} \tag{5.6}$$

The proof of Theorem 4 is analogous to and somewhat simpler than that of Theorem 3. Theorem 4 establishes an important property of estimates strongly related to the derivative of log likelihood. It says that when the stronger relation holds the m.l. estimate has a maximum concentration round the true value.

6. Concluding Remarks on Efficiency

It is observed that the derivative of log likelihood as a function of the observations and the unknown parameter, denoted in the paper by $Z_n(\theta)\sqrt{n}$, plays a fundamental role in problems of statistical inference, especially when attention is concentrated on discrimination between alternative values of the parameter close to each other (Rao and Poti, 1946; Wald, 1942). We enquire whether $Z_n(\theta)$ can be replaced by a statistic T_n, which is a function of the observations only, for purposes of statistical inference. If we demand that T_n should have the same property as $Z_n(\theta)$ in finite samples, it is necessary to postulate one to one correspondence between the two for each θ, a situation which is not usually obtainable. In a recent paper Birnbaum (1961) makes use of a strong dependence of T_n on $Z_n(\theta)$ such as

$$\{X_n : Z_n(\theta) \leqslant 0\} = \{X_n : T_n \leqslant \theta\}, \quad \{X_n : Z_n(\theta) > 0\} = \{X_n : T_n > \theta\},$$

where T_n is assumed to be the unique root of $Z_n(\theta) = 0$, to deduce *admissibility* and local *bestness* of T_n as an estimate of θ in finite samples, from the corresponding properties of admissibility and local bestness of the test based on $Z_n(\theta)$. These are extremely special cases and it appears that one has to consider large samples in order to say something definite, in general.

The various types of dependence in large samples considered are

(i) $\{Z_n(\theta) - \beta(\theta)(T_n - \theta)\sqrt{n}\} \to 0$ in probability;

(iia) $\Pr\{|Z_n(\theta)/(i\sqrt{n}) - (T_n - \theta)| \geqslant \epsilon |T_n - \theta|\} < \rho^n \quad (\rho < 1)$;

(iib) $\Pr\{|Z_n(\theta)/(i\sqrt{n}) - (T_n - \theta)| \geqslant \epsilon |Z_n(\theta)/(i\sqrt{n})|\} < \rho^n \quad (\rho < 1)$.

(iii) The asymptotic variance of

$$Z_n(\theta)\sqrt{n} - n\beta(\theta)(T_n - \theta) - n\lambda(\theta)(T_n - \theta)^2$$

is a minimum for a suitable choice of λ.

The last one, (iii), called second-order efficiency, ensures that the equivalence between T_n and $Z_n(\theta)/\sqrt{n}$ takes place at the fastest rate as $n \to \infty$. The first two imply that T_n, in large samples, has the same properties as $Z_n/\sqrt{n}$. Condition (i), called first-order efficiency, ensures the same local properties as $Z_n/\sqrt{n}$ for T_n while (iia, iib), which are stronger forms of first-order efficiency, imply that T_n is, in some sense, sufficient for parametric values in small specified intervals in which supremum of ρ is less than unity. The existence of ρ for each θ in conditions (iia, iib) ensures that T_n as a consistent estimate has the maximum concentration round the true value as $n \to \infty$.

It has not been possible in the present paper to examine the concept of *asymptotic sufficiency* of estimates as introduced in relation to tests of significance by Wald (1942) or as implicit in Barnard's work (Barnard, 1949), where he tried to approximate the likelihood by a quadratic function of the m.l. estimate, or as rigorously formulated by LeCam (1956). There are obvious connexions with the results of the present paper.

We may now ask whether m.l. estimates satisfy any of the conditions (i), (ii) and (iii). In the case of the multinomial distribution, which has been studied somewhat thoroughly, it is known that estimates exist for which condition (i) is realized when the cell probabilities as functions of the unknown parameter admit continuous first derivatives (Rao, 1960b) and conditions (ii) and (iii) are realized when continuous second derivatives exist (Rao, 1957, 1958, 1961). If the parameter chosen is a continuous functional of the distribution function, such estimates may be identified as m.l. estimates. In the continuous case no comprehensive treatment is available to answer all the questions relating to conditions (i), (ii) and (iii), except for a recent contribution due to Bahadur (1960), who imposes rather severe restrictions on the probability density. Partial answers exist in the work of Cramér (1946), Daniels (1961), Doob (1934, 1936), LeCam (1953, 1956) and others.

Some may object to the wider concept of an estimate adopted in the present paper, maintaining that estimation procedures should be discussed in terms of point-estimates and their *closeness* to the true values of the parameter. The latter approach brings in an extraneous element requiring as a datum of the problem a measure of closeness in some average sense. Often what is adopted is not a measure strictly applicable to a given practical situation, but some function which is justified as a close approximation to the true one, or which is commonly adopted and which provides algebraic convenience in the derivation of estimates. According to the wider concept, what is made available is a suitable summary of data which offers some convenience in drawing inferences about unknown parameters. There could be a valid criticism that the wider point of view offers only a partial solution and no guidance is provided in reaching decisions from estimates.

Fortunately, there is no scope for any argument in the case of large samples, for the "summary of data" in the form of an m.l. estimate appears to be satisfactory from all points of view. It is a point-estimate which has an equal or a greater chance of being closer to the true value than any given estimate as the sample size increases. It enables a ready and a reasonable test to be made for any simple hypothesis or for obtaining an interval estimate. Above all, a good approximation to the likelihood function can be obtained as an explicit function of the m.l. estimate at least in a small interval of the parameter round the true value, under suitable conditions where the m.l. estimate has maximum second-order efficiency. I hope it would be possible to provide an equally satisfactory answer, though not on the same lines, in the case of small samples.

ACKNOWLEDGEMENT

I wish to thank my colleague Mr J. Sethuraman for reading the manuscript and making some useful suggestions.

REFERENCES

BAHADUR, R. R. (1960), "On asymptotic efficiency of tests and estimates", *Sankhyā*, 22, 229-252.

BARNARD, G. A. (1949), "Statistical inference", *J. R. statist. Soc.* B, 11, 115–150.

BASU, D. (1953), "Choosing between two simple hypotheses and the criterion of consistency", *Proc. Nat. Int. Sc. (India)*, 19, 841–849.

BERKSON, J. (1960), Discussion on the paper by C. R. Rao, 32*nd Session of the Int. Stat. Institute*, Tokyo.

BIRNBAUM, A. (1961), "A unified theory of estimation, I", *Ann. math. Statist.*, 32, 112–135.

CHERNOFF, H. (1952), "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations", *Ann. math. Statist.*, 23, 493–507.

—— (1956), "Large sample theory—Parametric case", *Ann. math. Statist.*, 27, 1–22.

CRAMER, H. (1938), "Sur un nouveau théorème limité de la théorie des probabilités", *Actualites Scientifiques et Industrielles*, No. 736. Paris.

—— (1946), *Mathematical Methods of Statistics*. Princeton University Press.

Daniels, H. E. (1961), "The asymptotic efficiency of a maximum likelihood estimator", *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 151–164.

Doob, J. (1934), "Probability and statistics", *Trans. Amer. math. Soc.*, **36**, 759–772.

—— (1936), "Statistical estimation", *Trans. Amer. math. Soc.*, **39**, 410–421.

—— (1953), *Stochastic Processes*. New York: Wiley.

Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics", *Philos. Trans. Roy. Soc.*, A, **222**, 309–365.

—— (1925), "Theory of statistical estimation", *Proc. Camb. phil. Soc.*, **22**, 700–725.

—— (1950), *Statistical Methods for Research Workers*, 11th edition. Edinburgh: Oliver & Boyd.

—— (1956), *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd.

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

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

Lecam, L. (1953), "On some asymptotic properties of maximum likelihood estimates and related Bayes's estimates", *University of California Publications in Statistics*, **1**, 277–330.

—— (1956), "On the asymptotic theory of estimation and testing of hypotheses", *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 129–156.

Neyman, J. (1949), "Contribution to the theory of the χ^2 test", *Proc. Second Berkeley Symposium on Mathematical Statistics and Probability*, 239–273.

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

—— (1957), "Maximum likelihood estimation for the multinomial distribution", *Sankhyā*, **18**, 139–148.

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

—— (1960a), "Apparent anomalies and irregularities in maximum likelihood estimation", *32nd Session of the Int. Stat. Institute*, Tokyo.

—— (1960b), "A study of large sample test criteria through properties of efficient estimates", *Sankhyā*, **23**, 25–40.

—— (1961), "Asymptotic efficiency and limiting information", *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 531–546.

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

Sundrum, R. M. (1954), "On the relation between estimating efficiency and the power of tests", *Biometrika*, **41**, 542–544.

Wald, A. (1942), "Asymptotically shortest confidence intervals", *Ann. math. Statist.*, **13**, 127–137.

—— (1943), "Tests of statistical hypotheses concerning several parameters when the number of observations is large", *Trans. Amer. math. Soc.*, **54**, 426–482.

—— (1949), "A note on the consistency of maximum likelihood estimation", *Ann. math. Statist.*, **20**, 595–601.

Discussion on Paper by Professor C. R. Rao

Professor M. S. Bartlett: Many of us have in the last few weeks been listening to quite a few lectures by Professor Rao, including three lectures which he gave at the invitation of the University of London, various seminar talks, and now tonight's paper to our Society. I am sure that in proposing the vote of thanks for his paper you will allow me to congratulate Professor Rao on all his lectures, and welcome him back to England on what I trust he is finding a pleasant visit.

The first of his University lectures was not unconnected with his present paper, for he was discussing large-sample tests, and in particular the chi-square test, suggesting at one point that the log likelihood ratio test should be more sensitive than the alternative quadratic chi-square expression for small discrepancies, and the quadratic form more sensitive to large discrepancies. The comment I made to Professor Rao afterwards was that to classify the two procedures correctly in this way would require rather careful attention to their distributions, for, although we know that they are asymptotically the same on the null hypothesis in large samples, the question of the second-order departure from the chi-square distribution would be very relevant. Actually I have shown that to this next order of approximation the two expressions are equivalent on the null hypothesis, at least in distribution (*Biometrika*, 40 (1953), 306).

The same kind of remark seems to me relevant to his present paper. Professor Rao has given some very interesting and carefully discussed properties of large-sample estimates, and maximum likelihood ones in particular, but he has still tended to concentrate on their inference properties rather than their use simply as estimates. Now to do this comprehensively does, to me at least, require a correct knowledge of their sampling properties as well as of how asymptotically "sufficient" they are. Thus in definition (2.6) Professor Rao defines second-order efficiency in terms of the variance of the difference between the log likelihood derivative and the best quadratic function of $T-\theta$. But it is not too clear to me how, if at all, he proposes to make use of this quadratic function. He does not say what its distribution is, nor does he seem to stress that it involves optimizing coefficients which are functions of the unknown parameter θ. If his purpose, as he himself rather implies in his paper, is to use a function of θ which approximates to the log likelihood derivative, why not, as was my own systematic purpose in the paper that I have just referred to, use this function itself?

We are often told to study the likelihood function directly; the advantages of the logarithmic derivative are that it is usually very well behaved (especially if we choose the right function of the unknown parameter) and its sampling properties, including normality properties, are much more easily and exactly investigated, to the extent of our making second-order corrections. I believe in discussing concrete problems, so let me refer to two examples to illustrate these points. First, let me remind you that if we plot $Z_n\sqrt{n} = L'$, say, as a function of θ we hope to get a straight line near the point $Z_n = 0$. If we do, and the slope is also $-I_n$, a constant independent of the sample, then we can specify $\hat{\theta}$ and I_n in place of L'_0, say, and I_n, this is the first-order situation and to this order we cannot distinguish the two alternatives, as Professor Rao notes when referring to the linkage estimation problem discussed by Sundrum.

Let me, however, write down the next order equations in the developments of $\delta\theta = \hat{\theta}-\theta$ and $\Delta\theta = \theta_0-\theta$. They are

$$\sqrt{I_n}\,\delta\theta \sim \frac{L'}{\sqrt{I_n}} + \left\{\frac{L'(L''+I_n)}{I_n^{\frac{3}{2}}} + \frac{(L')^2 L'''}{2I_n^{\frac{3}{2}}}\right\}, \tag{1}$$

which is equation (13) of my paper and

$$\sqrt{I_n}\,\Delta\theta + \frac{L'_0}{\sqrt{I_n}} \sim \frac{L'}{\sqrt{I_n}} + \left\{\frac{\Delta\theta(L''+I_n)}{\sqrt{I_n}} + \frac{(\Delta\theta)^2 L'''}{2\sqrt{I_n}}\right\}. \tag{2}$$

It is not clear to me off-hand which will provide the better test of $\theta = \theta_0$, but it is not too difficult to investigate this question, although easier for (2) than for (1). Sundrum concluded in the linkage problem that in certain circumstances the test based on L'_0 is better; though he did not altogether justify his neglect of bias and non-normality, aspects which it is possible to allow for when necessary. If my algebra is correct, the bias in $\hat{\theta}$ in the linkage problem is zero to the next order, and the non-normality correction the same for both statistics, so that Sundrum's conclusion appears unaffected.

As a second example, let me recall that in the physical estimation problem (*Phil. Mag.*, 44 (1953), 249) which stimulated these higher approximation methods based on L', was the estimation of the mean lifetime θ of a certain fundamental type of particle from observations of decays over a limited (and variable) track length. I noted that $L'/\sqrt{I}$ in the neighbourhood of $1/\theta = 0$ was of the form $B/\theta - A$, and hence linear in $1/\theta$. It was thus much more sensible to estimate $1/\theta$ than θ, the maximum likelihood estimate of $1/\theta$ being $0{\cdot}256 \pm 0{\cdot}198 \times 10^{10}\,\text{sec}^{-1}$, and the upper 0·95 confidence limit for $1/\theta$, inferred most accurately directly from L', being $0{\cdot}621 \times 10^{10}$.

Thus, while I agree with Professor Rao that it is most important to investigate second-order approximations in large-sample tests and estimation problems, I am less convinced of the practical value of some of the proposed definitions and results in his paper.

I am very pleased to move the vote of thanks to Professor Rao.

Professor H. E. DANIELS: Confronted by a paper of such excellence and so full of new ideas, I find it difficult to follow the tradition among voters of thanks that praise shall be to some extent salted with blame. So perhaps I may be allowed to fall back on another well-established tradition that discussion at our meetings may be broad to the point of irrelevance.

Professor Rao measures the first-order efficiency of an estimate in terms of its correlation with $\partial L/\partial\theta$, the derivative of the log likelihood $L(X_n, \theta)$. This is the best discriminator between neighbouring hypotheses under the assumed regularity conditions. It is interesting to see how far the idea may still be appropriate when the regularity conditions are relaxed. For densities like $p(x, \theta) = \frac{1}{2} e^{-|x-\theta|}$, $\partial L/\partial\theta$ has finite discontinuities in θ, but it is still a good discriminator because L is a convex polygon in θ whose flat portions are of order n^{-1}. In such cases I have shown (*Proceedings of 4th Berkeley Symposium*) that the m.l. estimate $\hat{\theta}$ is still asymptotically efficient in the old sense. On the other hand, for

$$p(x, \theta) = c \exp\{-|x-\theta|^{\alpha}\} \quad (\tfrac{1}{2} < \alpha < 1),$$

$\partial L/\partial\theta$ keeps varying between $-\infty$ and ∞ over intervals of order n^{-1} near the true θ and so is useless for discriminating between neighbouring hypotheses. Nevertheless, the information function $I(\theta) = \alpha\Gamma\{2-(1/\alpha)\}/\Gamma\{1+(1/\alpha)\}$ is well defined. The function L is not now monotonic in $|\theta-\theta_0|$ but exhibits a series of cusps as θ passes through the observation points. However, its fluctuations are relatively small when n is large so that L can still be considered approximately monotonic in $|\theta-\theta_0|$. In cases like this it is still possible to obtain an asymptotically efficient estimate $\bar{\theta}$ by maximizing L over a discrete mesh of values of θ, provided the intervals of the mesh decrease with n faster than $n^{-\frac{1}{2}}$ but slower than n^{-1}. It seems possible that by working with differences of L over such intervals, rather than with $\partial L/\partial\theta$, Professor Rao might extend his results to some extent to cover less regular situations. The definition of second-order efficiency would have to be modified because the remainders in these cases are typically worse than $O(n^{-\frac{1}{2}})$.

There is another aspect of the example $c \exp\{-|x-\theta|^{\alpha}\}$ which is worth considering. As an estimate of θ the sample median $x_{\frac{1}{2}}$ has efficiency $\sin[\pi\{(1/\alpha)-1\}]/[\pi\{(1/\alpha)-1\}]$. This is unity when $\alpha = 1$ but falls to zero as α decreases from 1 to $\frac{1}{2}$. To improve on the median, consider an estimate $T = \lambda x_{\frac{1}{4}} + \mu x_{\frac{1}{2}} + \lambda x_{\frac{3}{4}}$ incorporating the quartiles $x_{\frac{1}{4}}, x_{\frac{3}{4}}$ also. It is unbiased when $2\lambda+\mu = 1$. If λ and μ are chosen to minimize its variance it will be found that when $\frac{1}{2} < \alpha < 1$, λ turns out to be negative and μ exceeds unity. For example, when $\alpha = \frac{2}{3}$ it is found that $T = -0{\cdot}1x_{\frac{1}{4}} + 1{\cdot}2x_{\frac{1}{2}} - 0{\cdot}1x_{\frac{3}{4}}$ approximately, at any rate for large samples, and T has efficiency 63 per cent. as against 57 per cent. for the median. This estimate has the curious property that a high value of the upper quartile, for example, results in a low value of T. The "tails" of the sample provide what might appear to be misleading information about θ which has to be contradicted in the estimate, a fact which is related to the non-monotonic character of the likelihood ratio.

The estimation of location and scale parameters by linear functions of the order statistics has been studied extensively by Lloyd, Plackett, Mosteller and many others. Particularly relevant to the present discussion are the book by Blom, *Statistical Estimates and Transformed Beta Variables* (New York: Wiley, 1958) and the apparently independent work by Ogawa in *Osaka Mathematical Journal*, 1951–52. It is shown that if all the order statistics $x_{(r)}$ are used, the estimate $T = \Sigma\lambda_r x_{(r)}$ is asymptotically fully efficient for a location or scale parameter θ if λ_r is proportional'to $-\partial^2 \log p(\xi_{(r)}, \theta)/\partial\theta^2$, where $\xi_{(r)}$ is the population quantile, or to an analogous quantity when the derivative does not exist. The λ_r do not involve θ. The result holds under wider regularity conditions than are normally assumed for m.l. estimates. Moreover, in small samples the estimates have "best linear unbiased" properties. For location and scale parameters this methôd of estimation is a serious competitor to maximum likelihood and I was a little surprised that Professor Rao did not include some comparison in his paper. To encourage him to examine these estimates further I will hazard a conjecture that they have the same second-order efficiency as m.l. estimates.

It is a pleasure to second the vote of thanks to Professor Rao for his excellent paper.

The vote of thanks was put to the meeting and carried unanimously.

Professor H. CHERNOFF: I would like to say a few words in defence of the relevance of the asymptotic variance as a measure of the efficiency of an estimator.

In small-sample theory it is generally difficult to specify optimal statistical procedures without resorting to arbitrary criteria or assuming some *a priori* distributions. It is usually possible to improve the performance of a procedure for some values of θ at the expense of harming the performance at other values of θ. Consequently it is difficult to compare the procedures.

The classical phrasing of the efficiency of maximum likelihood estimators seemed to indicate that via considerations of large-sample theory one could find asymptotically optimal procedures. It seemed to indicate that performance could be improved at one point only at the expense of completely ruining it elsewhere (lack of consistency). Examples of super-efficiency proved that this was not so if *uniform excellence* of the asymptotic variance is the criterion of asymptotic optimality. However, Stein's variation of the Cramér–Rao theorem (Chernoff, 1956), which does not employ regularity conditions on the estimators, indicates that all we need do is relax our criterion of asymptotic optimality slightly. This theorem shows that it is impossible to do substantially better than the Cramér–Rao bound everywhere in *any open interval*. From this point of view one can argue that super-efficiency and the existence of B.A.N. estimates do not force one to abandon asymptotic variance as a measure of optimality but merely point out that asymptotically optimal procedures are not unique.

In his concluding remarks Professor Rao criticizes the use of approximations to some measures of closeness. I would like to mention some results in the dissertation of W. D. Commins, an abstract of which has appeared (*Ann. math. Statist.*, 31 (1960), 532). If one takes $l(t, \theta)$ as a measure of the loss of utility associated with estimating t when θ is the true value of the parameter, then the expected loss derived from the use of an estimator T_n would be $E_\theta\, l(T_n, \theta)$. If it were assumed that $l(t, \theta)$ were quadratic for each θ, i.e.

$$l(t, \theta) = c(\theta)\,(t-\theta)^2,$$

with $c(\theta) > 0$, then the mean squared error would be a relevant measure of the goodness of the estimator. However, the quadratic loss is not always applicable. Worse, according to the theory of utility, utility must be bounded and therefore it is never really applicable when θ is unbounded. On the other hand, one may frequently assume that $l(t, \theta)$ is locally quadratic, i.e. one may be able to approximate $l(t, \theta)$ by $c(\theta)\,(t-\theta)^2$ for t sufficiently close to θ. What happens then to the expected loss as $n \to \infty$? Under mild conditions, the expected loss is somewhere between the asymptotic variance of the estimator T_n and the variance of T_n; the latter quantity may be larger. Furthermore, Commins proved that for the maximum likelihood estimator (subject to regularity conditions on the distribution function of the data) the expected loss is given by the asymptotic variance. This, together with Stein's version of the Cramer–Rao theorem, shows that no estimator can do substantially better in any open interval. In this sense we have the efficiency of the maximum likelihood estimator without assuming a particular form for the loss function.

It must be admitted that in spite of the general applicability of locally quadratic loss functions, there are cases where other loss functions are called for. It would be interesting if one could prove that, given an arbitrary loss function, there is an appropriate function of the maximum likelihood estimator $f_n(T_n) \approx T_n$ for which the expected loss is asymptotically minimized. The proof or disproof of such a generalization of Commins's result would undoubtedly relate to and depend on many of the ideas expressed in Professor Rao's paper.

Professor G. A. Barnard: I wish to add my congratulations to Professor Rao. I am sure we have all been most impressed with the range as well as with the depth of the contributions he has made during the course of his visit to London.

If I may follow Professor Daniels in not being too closely relevant, and perhaps follow also Professor Rao in that respect, I would like to comment on the first section of his paper where he discusses the question of consistency. This is not altogether relevant to efficiency, but it is a topic worth dwelling on. Professor Rao's definition of consistency is formulated with reference to the possibility of distinguishing asymptotically between two distributions and it has the property that the usual estimate s^2 of the variance is a consistent estimator of σ as well as of σ^2. This rather affronts our usual notions of what is meant by consistency and I want to point out that one can get over the difficulty in this particular case by referring to the group properties of the problem. The reason we consider usually that s^2 estimates σ^2 and does not estimate σ can be expressed by saying that if all the observations are multiplied by a constant α, s^2 is multiplied by α^2, and so is σ^2, but σ is multiplied by α.

The group covariance property is relevant also to the examples of super-efficient estimates due to LeCam. These fail to satisfy the reasonable group properties that one would normally require of them (they are not covariant under the translation group), and I am glad to learn this evening from Professor Chernoff that there is a theorem which implies that one cannot construct such examples which are group covariant.

I do not wish to suggest that the difficulty of this definition of consistency can always be repaired by group considerations, although the range of cases in which it can be so repaired can be considerably extended if we use the idea of local group. That is to say, if we apply these notions under the restrictions that α lies close to the unit element of the relevant group. But in fact I think that no wholly satisfactory single definition of consistency is ever likely to be given. The original idea of consistency used by Gauss in connexion with the theorem of least squares, which condition has so long in the textbooks been so confusingly and misleadingly replaced by the condition of unbiasedness, was based on the fact that in the problems that Gauss was considering one could distinguish between true values and errors. Gauss's consistency requirement was simply that when the observations were free from error the method of estimation used should give the true value. This is very close to the Fisher consistency definition which Professor Rao mentions, but that generalization, ingenious though it is, suffers from the disadvantage that it is restricted to functionals on the empirical sampling distribution function. Not all the usual statistics are definable as such, for it is a general property of all such functionals that if the same set of observations occurs twice over in a sample of double size the sample distribution function is unaltered. But, for example, s^2 will become $(2n-2)s^2/(2n-1)$ instead of remaining unaltered. Therefore the usual estimate of variance fails to be a statistic in this sense. One might try to repair this defect by introducing well-behaved factors, such as functions of n which tend monotonically to 1; but I really think that the tendency to look for a single definition of consistency, applicable to a wide variety of different circumstances, is an example of the tendency to over-simplification in a mathematical sense, which has been endemic in the field of statistics for many years past. I do not think the search is likely to succeed.

Now I wish to comment on the main part of the paper. In spite of the emphasis which Professor Rao has placed on the importance of regarding statistics as discriminators, it still tends to be overlooked that most of the definitions of consistency, second-order efficiency and so on can be applied to "estimators" of the form (t_1, t_2) used for estimating a single quantity θ. There is no restriction to a single number. If one is reducing data one can reduce it to a pair of numbers, or to a triplet, as well as to a single number, and this possibility should be borne in mind. It would help to get us away from the misleading concept of point estimation, which I think does a great deal of harm.

Finally, I would like also to comment, and I do not doubt that Professor Rao will agree with this, that nothing in this paper, or indeed nothing that has been said tonight, detracts from the desirability, in the small finite samples which unfortunately we are given in practice, of looking at the whole of the likelihood function. Thanks to computers, we can now so often do this.

Professor D. V. LINDLEY: Tonight's paper cannot be considered apart from two others that Professor Rao has given at the Berkeley Symposium (1961) and at the Tokyo meeting of the International Statistical Institute (1960a). In the trio he has developed a highly interesting account of maximum likelihood theory approached from an original and illuminating angle, and our Society is much honoured by his presenting his paper. Professor Rao follows in the footsteps of Fisher in basing his thesis on intuitive considerations of estimation that people, like myself, who lack such penetrating intuition, cannot aspire to. We are forced to follow more pedestrian paths and it is such a pedestrian approach to estimation that I would like to consider this evening.

The problem of large-sample estimation was discussed in the paper that I gave to the Berkeley Symposium, but there it was treated as a decision problem with an explicit statement of both a utility function and a prior distribution. The best estimate was one that maximized the expected utility for a given observation. Only the product of utility and prior probability enters into the calculations, and this function can be written $w(d, \theta)$, for an estimate (or decision) d and a parameter θ, and called a weight function. It was shown that the expected utility may be expanded in an asymptotic series which is proportional to

$$w+w\left\{\frac{L_4}{8L_2^2}-\frac{5L_3^2}{24L_2^3}\right\}+\left\{\frac{L_3 w_1}{2L_2^2}-\frac{w_2}{2L_2}\right\} \qquad (*)$$

as far as the terms of order n^{-1}. The notation here is as follows: L is the log-likelihood and L_i is its ith derivative with respect to θ; w_i is the ith partial derivative of $w(d, \theta)$ with respect to θ; all the functions are evaluated at arguments d and $\hat{\theta}$, the maximum likelihood estimate. The omitted constant of proportionality is a function of $\hat{\theta}$.

The first term in (*) is $O(1)$, the other two are $O(n^{-1})$. If only the first is retained the expected utility is $w = w(d, \hat{\theta})$ and the optimum d is that which maximizes $w(d, \hat{\theta})$. In other words one should act as if θ was known to be the maximum likelihood value. Thus, to first order, the maximum likelihood estimate is the best, and this corresponds to its known first-order efficiency properties. It is important to notice that this result does not depend on the form of the weight function, provided only that it is sufficiently well behaved. In all this discussion the appropriate regularity conditions are assumed to obtain.

If the terms $O(n^{-1})$ in (*) are included we might expect to obtain a better estimate by maximizing this over d, instead of just $w(d, \hat{\theta})$. This will lead us to an estimate having some sort of second-order efficiency and therefore perhaps comparable with Professor Rao's. It was shown in my Berkeley paper that, in general, the new estimate will not be the maximum likelihood estimate but will differ from it by a small amount. The exact form of this difference will depend upon the derivatives, w_1 and w_2, of the weight function and upon the derivatives of L up to L_4. An explicit form for it was obtained in a special case. Now this is quite contrary to the results given by Professor Rao, for example, in Table 1 of tonight's paper. He claims that the above difference is zero and that the maximum likelihood estimate is best, even to second order. There would appear to be a discrepancy somewhere.

It therefore occurred to me to see whether there was some special form of weight function for which even (*) would yield the maximum likelihood estimate. This would happen if the third term in (*) were to vanish, for then (*) would be proportional to $w(d, \hat{\theta})$ and the same argument as with the first-order case would show $\hat{\theta}$ to be best. For the third term to vanish it is necessary that

$$w_1 L_2 = w_2 L_1,$$

the arguments still being d and $\hat{\theta}$. This will certainly be true if the same equation holds for all d and θ. (For this to happen the weight function will also have to depend on the observations but the results persist even if this is so.) Again this equation is satisfied if w_1 is proportional to $L_1(\theta)$ which implies $w = AL_1(\theta)+B$, where A and B are functions of d, and possibly the observations. It never has any effect on the determination of the best estimate if an arbitrary function of θ is added to w: and by this device it is possible to work in terms of losses. Then we may take

$$w(d, \theta) = \{L_1(d)-L_1(\theta)\}^2.$$

Our result then says that with this weight function the best estimate (to second order) is the maximum likelihood estimate. I do not know whether there are any other weight functions with this property, but I suspect not.

The discrepancy is thus resolved. The weight function just derived is essentially equivalent to that used by Professor Rao (Definition (2.4)) and the argument given here provides a little support to his claim of the superiority of the maximum likelihood estimate. Nevertheless, an extremely important proviso must be inserted. Unlike the first-order efficiency, which is true for most weight functions, the second-order efficiency depends critically on the weight function. Professor Rao has hidden this dependence and thus made the method appear to be better than it is. For example, if the Hellinger distance had been used for the weight function it is quite likely that Hellinger's estimation method would have emerged victorious in Table 1. An estimation procedure can only have its second-order efficiency judged in connexion with some form of weight function.

Mr A. STUART: The preceding discussion shows that Professor Rao was too modest in apologizing for the subject-matter of his paper. He should have been warned by the experience of Professor Lindley, who once (*J. R. statist. Soc.* B, 15 (1953), 30) went so far as to apologize for the mathematics of his paper, was taken to task for doing so in the discussion by Professor Neyman, and in his reply to the discussion very nearly apologizes for his original apology. I do not wish to press Professor Rao this far, but only to say that he could hardly have chosen a more relevant or important topic.

I should like to confine my remarks to section 3 of the paper. My starting point is the definition (3.8) of the efficiency of a test. In the circumstances of the paper (variances of order n^{-1}, principally), this is exactly the square root of the Asymptotic Relative Efficiency (A.R.E.) of a consistent test as defined by Pitman in 1948; this definition was later generalized by Noether (*Ann. math. Statist.*, 26 (1955), 64) to statistics with variances of orders $n^{-2\delta}$. It turns out that for one-tailed tests, the A.R.E. raised to the power δ is the ratio of first derivatives at θ_0 of the power functions of the tests being compared, and for two-tailed tests (not discussed by Professor Rao) it is, when raised to the power 2δ, the ratio of second derivatives at θ_0 of the power function under fairly general conditions, results given in *Skand. Akt.* by myself in 1954, p. 163, by Noether, and more accessibly in Kendall and Stuart's *Advanced Theory of Statistics*, 2, Chapter 25. Moreover, as originally stated by me for $\delta = \frac{1}{2}$ (*J. Amer. statist. Ass.*, 49 (1954), 147), and for general δ by Kendall and Stuart, the A.R.E. is equivalent to the ratio of estimating efficiencies of those transformations of the test statistics which are consistent estimators of the parameter θ. It follows at once that the full efficiency of an estimator guarantees its local test efficiency and maximum power derivatives, which is essentially the content of Professor Rao's Theorems 1 and 2. These relationships between A.R.E., derivatives of power functions and estimation efficiency, together with Fisher's theorem that estimating efficiency is the square of the correlation ρ between an estimator and an efficient estimator, imply at once that $\rho^{2\delta}$ is the δth power of the ratio of power function derivatives. For $\delta = \frac{1}{2}$, this is Professor Rao's Lemma 3.3.

I am sure that the reason for Professor Rao's not having connected his work with that on A.R.E. and the derivatives of power functions is that the latter has been developed in the specialized field of distribution-free methods for non-parametric problems. Non-parametric statisticians are the beat generation of statistics, and we cannot expect respectable maximizers of likelihoods to have too much to do with them. I would only ask Professor Rao to square his measure of test efficiency to make it operational (and equivalent to A.R.E.) in the sense that it will then equal the limiting inverse ratio of sample sizes required by two tests to attain equal local power, as does Pitman's A.R.E. As it stands, his definition (3.8) measures local test efficiency in a manner flattering to inefficient tests in the same way as the comparison of standard errors rather than variances flatters inefficient estimators.

Finally, I want to make a brief point about Sundrum's work, referred to at the end of section 3. Sundrum was not concerned with local test efficiency; indeed, his paradox (that greater estimating efficiency does not automatically carry greater power with it) arises specifically because he considers *fixed* alternatives not in the immediate neighbourhood of θ_0, and the paradox disappears as they approach θ_0, when we get left with the A.R.E. In Sundrum's treatment, large-sample size is necessary only to ensure normality of the test statistics, and is essentially irrelevant to his argument. There is thus no contradiction between his result and Professor Rao's, or, for that matter, the equivalent A.R.E. results I have mentioned.

Mr P. Wegner: One of the difficulties of maximum likelihood estimation is that of computation of the estimates. I should like to draw attention to a class of models where computation is feasible in both the univariate and multivariate case, although the likelihood function is not differentiable. Consider the case where the parameters are known to lie within a convex linearly bounded subspace of the parameter space (i.e. a zero prior distribution is imposed on all points outside the convex subspace but nothing is assumed about the prior distribution within the subspace). In the general n dimensional case assume that the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ is subject to restrictions of the form

$$\sum_{j=1}^{n} a_{ij}\,\theta_j \leqslant b_i \quad (i = 1, 2, \ldots, m).$$

These constraints make it impossible to maximize the likelihood function by differentiation, since $\partial L/\partial\theta$ is not necessarily zero at the maximum. However, the resulting problem is computationally tractable for a linear model with quadratic or linear loss functions. The inequality restrictions lead naturally to a mathematical programming formulation of the likelihood maximization problem. In particular, a quadratic loss function leads to a quadratic programming problem, while the linear loss function that has been mentioned by Professor Chernoff (minimization of the sum of absolute deviations) leads to a linear programming problem.

The optimality properties of the point estimates obtained in the quadratic case have been investigated by H. O. Hartley both for large and small samples. He has derived formulae for the variance of estimates in constrained convex parameter spaces. He has shown that the resulting estimates are more efficient than the corresponding unconstrained estimates, and that the estimates are in general biased.

The class of problems mentioned here is both of practical and of theoretical interest, and I feel that further investigation along these lines will prove fruitful. More generally, I feel that mathematical programming techniques will be found useful in the analysis on non-differentiable multivariate statistical problems.

Professor RAO replied briefly at the meeting and subsequently more fully in writing as follows:

I wish to thank all who contributed to the discussion and gave me material for further research. I am also grateful to the various speakers for the kind references to my visit to the U.K. and to my lectures.

I wish to make a few general observations on the approach to estimation adopted in the paper, before answering the specific points raised during the discussion. It is, indeed, futile to attempt any general discussion of estimation by considering a particular loss function, or a particular type of loss function, without reference to the practical situation to which it actually corresponds. However, if in any practical problem a particular loss function suggests itself, and what is of interest is the expected loss in the long run, this should certainly be investigated. However, in the majority of investigations an estimate will have a variety of uses for all of which the concept of a loss function may not be meaningful; besides, it may be necessary to preserve an estimate as a substitute for the whole sample for possible future use. What is then needed is a more comprehensive approach for examining the usefulness of an estimate, and one way of doing this is to assess how good a substitute an estimate is for the entire sample, for drawing inferences about unknown parameters.

With such a criterion, estimation need not be confined to what are called point estimates, but may be looked upon from a wider point of view as reduction of data in a form convenient for drawing inferences about unknown parameters. The scheme thus envisages consideration even of vector estimates of a single unknown parameter, as suggested by Professor Barnard. This point was also stressed in my previous paper presented at the 32nd session of the International Statistical Conference, Tokyo.

I have deliberately given a definition of consistency without reference to any parameter, which fits in with the general approach, although I agree with Professor Barnard that consistency in other forms and with reference to particular parametric functions may have to be considered in some specific problems. Although likelihood by itself is an important concept, it is necessary to explore whether there exist more satisfactory forms in which uncertainty regarding unknown parameters can be expressed. The fiducial inference is one such example and it appears to be the ideal form for this purpose.

In view of the remarks I have already made, I do not agree with the general stand taken by Professor Lindley of seeking for a justification of an estimation procedure through a utility function and a prior distribution, both of which introduce some amount of arbitrariness. It may be interesting to note that a m.l. estimate is best for a certain type of loss function, but the insistence on point estimation is not a defensible proposition.

For the same reason the concept of minimum asymptotic variance has to be given up, although Professor Chernoff shows that it can be cleared of an apparent anomaly pointed out by LeCam. As shown in my paper, smaller asymptotic variance does not necessarily imply better properties of an estimate from the point of view of inference.

I must admit that I am not aware of the vast literature on the "efficiency of a test" referred to by Mr Stuart. Much of that work does not seem to be rigorous. I, however, agree with Mr Stuart that there is some advantage in defining the efficiency of a test as the square of the expression I have given.

Maximum likelihood estimation involves a certain amount of heavy computation, and any contribution such as that indicated by Mr Wegner will, no doubt, be useful.

Professor Daniels referred to his recent work on maximum likelihood estimation when the usual regularity conditions are not satisfied. I agree with him that it is worth examining whether some of the propositions proved in my paper are valid under less restrictive or a different set of conditions. It may be possible to do so by studying particular examples as Professor Daniels does.

Professor Bartlett raised a number of questions to which I do not have ready solutions. My statement, in one of my University of London lectures, on the comparison of the power functions of the chi-square and likelihood ratio goodness of fit tests, is based on partly theoretical and partly empirical investigations carried out at the Indian Statistical Institute. Further investigation is in progress and I hope a more convincing proof will be available. I am not, however, considering the rates of convergence of the distributions of these statistics to the chi-square approximation, on the null hypothesis. This would be an associated problem for which Professor Bartlett has already given a partial solution.

In the present paper I was concerned with the development of criteria for judging the usefulness of an estimate. One of the criteria suggested was the closeness of fit to the likelihood of a quadratic function of the estimate. The residual variance from the fitted function is, under some conditions, the actual loss of information due to estimation. The quadratic fit, however, would provide an approximate reconstruction of the likelihood when only the estimate is available. Regarding the test for linkage, I was only pointing out that the argument given by Sundrum cannot possibly be correct in view of what has been established in Theorem 1 regarding the efficiency of the test based on a m.l. estimate. It is quite possible that, when second-order terms are considered, the alternative test considered by Sundrum, which is now shown to be equivalent up to first-order terms, may turn out to be better. This, however, needs careful examination.