APPARENT ANOMALIES AND IRREGULARITIES IN MAXIMUM LIKELIHOOD ESTIMATION*

(*with discussion*)

By C. RADHAKRISHNA RAO
Indian Statistical Institute, Calcutta.

1. Introduction

Maximum likelihood (m.l.) estimation is criticised mainly on the following grounds:

(i) It does not always provide consistent estimates.

(ii) There exist estimates with lower asymptotic variance than that of the m.l. estimate, and therefore the m.l. method does not lead to most efficient estimates as claimed in the literature on this subject.

(iii) The computations involved in determining m.l. estimates are in most cases unduly heavy. On the other hand, there exist simpler methods of estimation which provide estimates which are asymptotically as efficient as m.l. estimates.

(iv) No adequate justification has been put forward for m.l. estimation in finite samples. Judged by the criterion of mean squared error in finite samples, there are examples where certain other procedures are better than m.l.

The criticism (iii) on grounds of computational difficulties will be relatively unimportant when high speed electronic computers become easily available for use by research workers. The computations, involving an iterative procedure for solving m.l. equations and inversion of matrices for obtaining standard errors, can be easily programmed on any modern electronic computer. Recently, routine programmes have been constructed at the Indian Statistical Institute for obtaining m.l. estimates of gene frequencies, standard errors, expected frequencies and goodness of fit χ^2, from observed phenotypic frequencies of various blood group systems such as OAB, MN, CDE, etc. The time taken for these computations is of the order of a minute for each blood group system, even on a comparatively slow machine like the HEC (Hollerith Electronic Computer).

I shall, therefore, confine my comments to the other points of criticism relating to consistency, efficiency, and properties of estimates in small samples.

2. Purpose of estimation

It will help in our discussion if we agree on the purpose of estimation, on which will depend the criteria for the choice of a suitable method of estimation. Much of the controversy in the literature on estimation could be dismissed once this problem is properly answered. There has been a tendency to consider estimation as a part of decision theory, which requires as a datum of the problem the specification of the loss for a given difference between the estimate and the true value of the unknown parameter. The criterion in such a case is naturally the minimisation of expected loss. This may be appropriate in certain situations but I

* This paper originally presented at the 32nd Session of the International Statistical Institute is being reprinted here with the kind permission of the Editor, *Bulletine of the International Statistical Institute.*

am not sure whether one can support Berkson (1955) when he wants to estimate the slope of the probit regression line in a bio-assay using the criterion of minimum expected squared error, unless of course he believes or makes us believe that the loss to society is proportional to the square of the error in his estimate. I suppose a bio-assayer, when he obtains an estimate of the standard deviation of a tolerance distribution, or the LD 50, uses it in a variety of ways besides playing a game with nature or with society.[1] He would like to compare it with an estimate of LD 50 for another insecticide, combine it with a previous estimate for the same insecticide to obtain a better estimate, preserve it for comparison or combination with future estimates, or indulge in some assertions (with some confidence) that LD 50 is less than a specified value or lies between two specified values and so on, or use the estimate itself more conveniently in the place of basic data in reaching optimum decisions for a specified loss function.

It may be argued that all these problems could be answered directly, and in theory more satisfactorily, from given data without considering the intermediate methodological problem of estimation. If then we insist on estimating the unknown parameters and use the estimates for purposes of inference such as those indicated above it can only be due to some convenience in handling the estimates rather than the original data, in addition to the resulting economy in recording only the estimates for future use, instead of preserving the entire mass of observed data, much of which may be irrelevant. If, therefore, we define the purpose of estimation as condensation of data, what criteria can we lay down for choosing a method of estimation?

Most statisticians would probably agree that statistical inference consists, in general, in discriminating between alternative possible situations on the basis of given data, and as such it should be based on the likelihood $P(S, \theta)$ of the parameter θ given the sample S, which is same as the probability (or probability density) of S given θ. More precisely, we need the ratio of the likelihoods for two given values θ_1 and θ_2 of the parameter. There may be, however, some controversy about the form in which the uncertainty in the choice of θ_1 or θ_2 given S, is to be expressed.

If there exists a statistic T such that

$$P(S, \theta_1)/P(S, \theta_2) = P(T, \theta_1)/P(T, \theta_2) \qquad \ldots \quad (2.1)$$

for all admissible θ_1 and θ_2, nothing is lost by replacing the sample S by the statistic T, which is for all relevant purposes equivalent to S. Such a statistic T is said to be sufficient in the sense of Fisher (1922). There will be a multiplicity of statistics T satisfying (2.1), one of them (in the extended sense of the term statistic) being the sample itself. In general we can choose one among them, say T_0, which is minimal in the sense that T_0 is essentially a function of every sufficient statistic T (Lehmann and Scheffé, 1950). A minimal sufficient statistic thus provides an exhaustive summary of the sample for purposes of statistical inference. If $x_1, \ldots, x_n$ is a sample of observations, the observed mean $\bar{x}$ and variance s^2 are jointly sufficient when the population distribution is normal with unknown mean and variance.

[1] A simple example given by Silverstone (1957) illustrates the point. In the case of the ordinary binomial distribution with probability θ, for number of trials $n = 3$, the estimate $T_1 \equiv 1/2$ of θ has smaller mean square error than the observed proportion r/n for all true values of θ between 1/4 and 3/4. This cannot be advanced as a cogent reason for using T_1 instead of r/n when nothing is known about the true value of θ.

On the other hand, the minimal sufficient statistic may be the ordered values in the whole sample as in the case of a Cauchy population with an unknown location parameter, i.e. no reduction of the data is possible without disturbing the relation (2.1). In such a case we may look for a statistic which belongs to a specified simple class and provides the maximum possible discrimination.

For a statistic T belonging to a specified class let us denote the likelihood ratio by $P(T, \theta_1)/P(T, \theta_2)$. The larger the deviation of this ratio from 1, the greater will be the discrimination between the parameters θ_1 and θ_2. For purposes of comparison, it is convenient to have a measure of the amount of discrimination provided by a statistic T. There is, indeed, some amount of arbitrariness in the choice of such a measure. Some natural measures are, 'the amount of overlap' between distributions corresponding to θ_1 and θ_2 as defined by the author (Rao, 1948), or the quantity

$$J_T(\theta_1, \theta_2) = \underset{\theta_1}{E} \log \left\{ \frac{P(T, \theta_1)}{P(T, \theta_2)} \right\} + \underset{\theta_2}{E} \log \left\{ \frac{P(T, \theta_2)}{P(T, \theta_1)} \right\} \quad \ldots \quad (2.2)$$

considered by Kullback and Liebler (1951) following the concepts of information theory. One may prefer even a pure distance measure like

$$\rho_T(\theta_1, \theta_2) = \int \sqrt{P(T, \theta_1)\, P(T, \theta_2)}\, dT \quad \ldots \quad (2.3)$$

introduced by Hellinger (1909) (see also Bhattacharya, 1946). Each of these measures is not more than the corresponding expression when T is replaced by the whole sample. The ratio of the amount of discrimination provided by T to that contained in the whole sample may be considered as an index of the effectiveness of T. When T is sufficient this ratio is unity for all these measures. For simplicity, let us consider $J_T\ (\theta_1, \theta_2)$ in the further discussion, observing that the same or similar results will be valid for the other measures mentioned.

When the sample S consists of n independent observations on a variate X we have

$$J_S\, (\theta_1, \theta_2) = n \left\{ \underset{\theta_1}{E} \log\, [P(X, \theta_1)/P(X, \theta_2)] + \underset{\theta_2}{E} \log [P(X, \theta_2)/P(X, \theta_1)] \right\} \quad \ldots \quad (2.4)$$

where the expression within the brackets is the value of $J(\theta_1, \theta_2)$ for a single observation and is therefore independent of n. As $n \to \infty$, $J_S(\theta_1, \theta_2) \to \infty$, and we have perfect discrimination between θ_1 and θ_2, as was shown by Basu (1954). A rigorous demonstration of this result was given earlier by Kakutani (1948) using the fact that $\rho_S(\theta_1, \theta_2 \to 0$ as $n \to \infty$. He showed that the distributions of the sample sequences in the infinite dimensional space for two different values of θ are 'orthogonal'. A statistic T_n which replaces a sample would not be of much use if it did not provide complete discrimination between any two values of θ as $n \to \infty$, i.e., if $P(T_n, \theta_1)$ and $P(T_n, \theta_2)$ are not orthogonal in the limit. This is possible if[2] $T_n \to \phi(\theta)$ with probability 1 as $n \to \infty$, where $\phi(\theta)$ is a function of θ, having one-to-one correspondence with the possible values of θ. This is exactly what the criterion of consistency[3] laid down by Fisher (1922) demands.

[2] Generally, orthogonality is possible only if T_n tends to a particular value but examples may be found where for each θ, T_n has a non-degenerate limiting distribution, with distributions corresponding to different values of θ being non-overlapping.

[3] We are not demanding that $T_n \to \theta$. It is enough, if for any two different values of θ, T_n tends to two different constants. In such a case T_n is defined to be consistent for θ in the wide sense (see section 4 of this paper).

For given n, $J_S(\theta_1, \theta_2)$ and $J_T(\theta_1, \theta_2)$ depend on how widely separated are the distributions corresponding to θ_1 and θ_2. Therefore, the ratio of $J_T(\theta_1, \theta_2)$ to $J_S(\theta_1, \theta_2)$ may not represent the true effect of replacing S by T if the distributions corresponding to θ_1 and θ_2 are widely different. We may, therefore, consider the ratio of these quantities as $\theta_2 \to \theta_1$ assuming that this implies closeness of distributions. It is easy to see that

$$J_S(\theta_1, \theta_1+\delta\theta_1) \sim \frac{n}{2}\, i(\theta_1)(\delta\theta_1)^2$$

$$J_T(\theta_1, \theta_2+\delta\theta_1) \sim \frac{n}{2}\, i_T(\theta_1)(\delta\theta_1)^2 \qquad \dots \quad (2.5)$$

where $i(\theta_1) = E_{\theta_1}[P'(x, \theta_1)/P(x, \theta_1)]^2$ is the information per observation as defined by Fisher (1922, 1925) and $i_T(\theta_1)$ the corresponding information per observation in the statistic T. It is shown by Fisher (1925) that $i_T(\theta_1) \leqslant i(\theta_1)$, which suggests the criterion of maximising the information per observation in the choice of a statistic.

Reference may also be made to earlier work by the author (Rao, 1945) where the distance between two distributions differing by small quantities in the parameters is [illegible] by an argument similar to that used here, as a quadratic differential metric of which (2.5) is a special case. In the general case $i(\theta_1)$ and $i_T(\theta_1)$ are matrices, and it is known that $\{i(\theta_1) - i_T(\theta_1)\}$ is a positive semi-definite matrix. The efficiency of a statistic may be measured by some expression reflecting the deviations from zero of the elements in the matrix $\{i(\theta_1) - i_T(\theta_1)\}$. We shall consider only the single parameter case in further discussions.

The information function seldom provides a complete ordering of the statistics for all values of θ in the admissible range. It is, of course, possible to obtain a complete ordering with respect to the average amount of information based on an *a priori* distribution of θ, if this last distribution can be specified. In other situations no satisfactory solution seems to exist, although information can be used to eliminate some statitstics which are worse than others in a range of the parameters in which we are interested. Fortunately under favourable circumstances, there exist methods of estimation for which $i_T(\theta) \to i(\theta)$ as $n \to \infty$, so that we have an assurance that at least in large samples the relative information lost is small.

In small samples, we could examine the performance of any statistic by computing the ratio i_T/i. If this quantity is small, we need not insist on replacing the observations S by the statistic T, but strengthen T by considering other statistics in addition to T, so that all taken together provide information per observation comparable to i. In the worst case, when the sample size is small, it may be necessary to retain the entire sample or the likelihood function either in the form of a graph or tabulated for some values of the parameters, which would enable us to reconstruct the function without much error, if needed in future.

3. Efficiency

3.1. *A new formulation of the concept of efficiency.*

Having discussed certain broad principles for summarising data we may examine some easily recognisable properties of statistics by which we can judge their effectiveness and discuss methods by which such statistics are obtained.

Let us consider the consequences of replacing the observations by a statistic in discriminating an alternative value of a parameter close to a specified value θ. It was shown by Rao and Poti (1946) by an application of the important lemma of Neyman and Pearson that a test which discriminates best small departures from a given value of θ, *for any given* n, is of the form : "Reject if and only if $Z_n \geqslant \lambda$" where

$$Z_n = (\sqrt{n})^{-1} \sum_1^n z_i, \quad z_i = P'(x_i, \theta)/P(x_i, \theta) \qquad \dots \quad (3.11)$$

λ is a constant, $P(x, \theta)$ is the probability density of x, and $x_1, \dots, x_n$ are n independent observations. Or, if we denote by L the likelihood of the parameter given the sample, the statistic Z_n is simply $(d \log L/d\theta)/\sqrt{n}$.

Can we construct a statistic independent of θ and with a performance[4] as good as that of Z_n? This is possible when there exists a function T_n of the observations such that

$$Z_n = \lambda(\theta)f(T_n)+\mu(\theta) \qquad \dots \quad (3.12)$$

or, more generally, when the variance of Z_n given T_n is zero, a situation in which T_n is sufficient for θ. On the other hand, it may be possible to construct a statistic such that its asymptotic correlation with Z_n is unity as $n \to \infty$. Such a statistic, if it exists[5] is as good as Z_n in sufficiently large samples, i.e., is best for discrimination between two neighbouring values of the parameter in sufficiently large samples. Based on these considerations we give a new formation of the concept of efficiency.

Definitions. A statistic is said to be efficient if its asymptotic correlation with the derivative of log likelihood is unity. The efficiency of any statistic may be measured by ρ^2, where ρ is its asymptotic correlation with Z_n.

In the case of more than one unknown parameter, a statistic consistent for a parameter is said to be efficient if its multiple correlation with the derivatives of the log likelihood with respect to the unknown parameters is unity. The efficiency of any statistic is measured by the square of the multiple correlation.

3.2. *'Super efficient' estimates and their efficiency.*

An efficient statistic is defined by Fisher (1922) as one whose asymptotic variance is $[n\, i(\theta)]^{-1}$, or alternatively as one whose asymptotic variance is the least. Although Fisher formally stated the criterion of efficiency in terms of least asymptotic variance it is clear from his writings that by an efficient estimate he meant a statistic for which the loss of information per observation tends to zero. Fisher gives the following extended definition of efficiency on page 714 of his 1925 paper : 'The efficiency of a statistic is the ratio of the intrinsic accuracy of its random sampling distribution to the amount of information in the data from which it has been derived.' He argued that since the reciprocal of information for the mean of the

[4] The power of the test based on Z_n, when n is large and the alternative to θ is $\theta+\delta\theta$, is nearly $\phi[n\, i(\theta)]d\theta$ where ϕ is an increasing function of the argument. The quantity $i(\theta)$ which appears in the expression (2.5) for the distance between distributions close to one another is also explicitly involved in the power function.

[5] For instance the unique consistent root T_n^* of the m.l. equation (ref. Huzurbazar, 1948) under the conditions given by Doob (1934) or Cramer (1946) satisfies that property, for $|\sqrt{n}(T_n-\theta)-Z_n| \to 0$ with probability 1. An m.l. estimate when referred to in the sequel is assumed to have this property.

distribution is variance when the distribution is normal and the information in a statistic is bounded above by $n\,i(\theta)$, an efficient statistic is recognised when it has a limiting normal distribution with variance $[n\,i(\theta)]^{-1}$ which is the least possible. In 1951, J. L. Hodges[6] and later Le Cam (1953) constructed examples of consistent estimates with an asymptotic variance $\leqslant [n\,i(\theta)]^{-1}$, with strict inequality for certain values of θ. In fact, at these exceptional points the asymptotic variance can be made arbitrarily small. These examples of what are called 'super efficient' estimates show that there is no non-zero bound to the asymptotic variance of a consistent estimate, contrary to what is stated by Fisher. One might think that super efficient estimates with asymptotic variance $\leqslant [n\,i(\theta)]^{-1}$ should be preferred to efficient estimates with asymptotic variance $[n\,i(\theta)]^{-1}$. We shall examine these notions in the light of the new definition of efficiency given here.

First it may be noted that super efficiency arises, when the statistic is not an explicit function of the sample distribution function and therefore not satisfying the consistency condition as originally defined by Fisher.[7] Assuming Fisher consistency (FC) and certain regularity conditions (mainly Frechet differentiability) on the statistic, Kallianpur and Rao (1955) demonstrated that $[n\,i(\theta)]^{-1}$ is, indeed, a lower bound to the asymptotic variance, thus justifying Fisher's argument. It is also deducible from the results of Kallianpur and Rao that a FC statistic with asymptotic variance $[n\,i(\theta)]^{-1}$, under the regularity conditions assumed, has asymptotic correlation unity with Z_n. This demonstrates the equivalence of Fisher's definition of efficiency with that proposed here under the regularity conditions imposed on the estimate. Earlier work by Neyman (1949) and Barankin and Gurland (1950) also tend to confirm Fisher's results.

Now let us see how the new definition of efficiency enables us to judge the effectiveness of any statistic, whether it satisfies regularity conditions or not. What happens when FC and other regularity conditions imposed on the statistic are not satisfied ? In this case 'super efficient' estimates do exist as shown by Hodges and Le Cam. We shall show that when a super efficient estimate (i.e. with a possibly smaller asymptotic variance than that of the m.l. estimate) exists, one of the following two possibilities holds.

[6] The example by Hodges is quoted in a paper by Le Cam (1953). Consider the mean X_n of n independent observations on X from a normal distribution with mean θ and standard deviation unity. As is well known X_n is the m.l. estimate of the mean with variance $\sigma_n^2 = 1/n$. Let T_n be the function defined by

$$\begin{aligned} T_n(X_n) &= X_n \quad \text{if } |X_n| > \frac{1}{n^{1/4}} \\ &= \alpha X_n \text{ if } |X_n| < \frac{1}{n^{1/4}} \end{aligned}$$

It is easy to see that T_n is also asymptotically normally distributed about θ, with variance $= 1/n$ for $\theta \neq 0$. and α^2/n for $\theta = 0$. Since α is arbitrary, the asymptotic variance is less than that of the m.l. estimate when $\theta = 0$. This example of Hodges was generalized by Le Cam to improve the asymptotic variance at a countable number of values of the parameter θ.

[7] If S_n is the sample distribution function and $F(\theta)$ the true distribution function a functional $f(S_n)$ is said to be Fisher consistent (FC) for θ if $f[F(\theta)] \equiv 0$. For a discussion on this subject see Kallianpur and Rao (1955). The estimates of Hodges and Le Cam are not FC.

(1) In large samples, it is equivalent to the m.l. estimate (which is efficient in the new sense), i.e., has asymptotic correlation unity with the m.l. estimate, and therefore it is efficient in the new sense.

(2) It is not efficient in the new sense, in which case it is definitely worse than the m.l. estimate, for purposes of inference such as testing of hypothesis, interval estimation, etc.

Let us assume that a statistic T_n consistent for θ is such that $\sqrt{n}(T_n-\theta)$ and Z_n have a joint asymptotic distribution. Denote the asymptotic variance of $\sqrt{n}(T_n-\theta)$ by $v(T)$ and its asymptotic covariance with Z_n by $\alpha(\theta)$. Since the asymptotic variance of Z_n is $i(\theta)$ we have the obvious inequality

$$v(T) \geqslant \alpha^2(\theta)/i(\theta), \qquad \dots \quad (3.21)$$

From this relation, it follows that T_n has asymptotic correlation unity with Z_n, or it is fully efficient (in the new sense) if and only if $v(T) = \alpha^2(\theta)/i(\theta)$. If T_n^* is an m.l. estimate for which the observation made in footnote (5) is true, then $\sqrt{n}(T_n^*-\theta)$ has asymptotic variance equal to $1/i(\theta)$, and asymptotic correlation unity with Z_n. Therefore, when the equality in (3.21) is attained, T and T^* have asymptotic correlation unity *whatever* may be the inequality satisfied between their asymptotic variances[8], $\alpha^2(\theta)/i(\theta)$ and $1/i(\theta)$. If $\alpha^2(\theta) \leqslant 1$ for all θ, with strict inequality for some θ, we have an example of super efficiency as in the case of Hodges' example.[9] In fact, we can use the device of Hodges to construct examples of 'sub efficiency' i.e. where $\alpha^2(\theta) \geqslant 1$. In either case, when the equality in (3.21) is attained, T_n is equivalent to the m.l. estimate T_n^* in the sense that essentially the same type of inference is possible by using T_n or T_n^* in large samples, whether T_n is super or sub efficient in the earlier sense.

We may now ask what happens when the equality in (3.21) is not attained. Such a statistic T_n has asymptotic correlation $-1 < \rho < 1$ with Z_n, and therefore is not as good as Z_n (and therefore not as good as m.l.) for local discrimination, although T_n may be super efficient, i.e.

$$\frac{1}{i(\theta)} \geqslant v(T) > \frac{\alpha^2(\theta)}{i(\theta)}. \qquad \dots \quad (3.22)$$

Consider for example a sample $x_1, \dots, x_n$ from a normal distribution with an unknown mean μ and variance unity and denote by $\bar{x}$ and x_m, the sample mean and median respectively. Define the statistic.

$$\begin{aligned} T &= \alpha x_m \qquad \text{if } \bar{x} < n^{-1/4} \\ &= \bar{x} \qquad\quad \text{if } \bar{x} \geqslant n^{-1/4} \end{aligned} \qquad \dots \quad (3.23)$$

It is easy to see that the asymptotic distribution of T is normal with variance $\alpha^2\pi/2$ when $\mu = 0$, and 1 when $\mu \neq 0$. By choosing α arbitrarily small, $\alpha^2\pi/2$ can be made less than 1. The statistic T is therefore super efficient. But for testing the hypothesis $\mu = 0$, it is obvious that the test criterion is essentially the median when the null hypothesis is true and consequently the power of the test is smaller than that of $\bar{x}$. In this connection we may also refer

[8] We may compare this result with that of Fisher (1925), that the asymptotic correlation between two efficient estimates having the same least asymptotic variance is unity.

[9] It may be observed from the example given in footnote (6) that T_n and X_n have asymptotic covariance α for $\theta = 0$ and 1 for $\theta \neq 0$. The value of α, can be chosen to be >1 or <1 arbitrarily. The technique of Hodges and Le Cam provides a statistic which is essentially equivalent to the statistic with which they start.

to an interesting but a different type of example due to Basu (1956), where the ratio of a limiting variance of one statistic to that of another → ∞ but the corresponding ratio of the probabilities of concentration wthin any given limits of the true value → 0. So it would appear that the criterion of minimum asymptotic variance is misleading.

We may now raise the problem whether given a super efficient estimate, it is possible to find a function of the m.l. estimate which is consistent for the parameter and which has a smaller asymptotic variance than the given super efficient estimate. This means that a super efficient estimate can be uniformly improved from the point of view of asymptotic variance by using a function of the m.l. estimate. This is true of the known examples of super efficiency. Further, given a super efficient estimate, i.e., when $v(T)$ satisfies (3.22), we can construct a function of m.l. estimate, by using Le Cam's technique, such that its asymptotic variance is smaller than $v(T)$ or even $\alpha^2(\theta)/i$ at a countable set of values of θ. To examine whether improvement is possible for all values of θ, we have to study the function $\alpha(\theta)$. Under some assumptions Le Cam (1953) proved that $|\alpha(\theta)|$ can be less than unity only for a set of points of Lebesgue measure zero. This is encouraging but does not solve the problem posed here. We may have to explore the asymptotic sufficiency of the m.l. estimate (Wald, 1943; Le Cam, 1953) to prove this property.

3.3. *Information in the limit.*

We shall examine the limiting information contained in an efficient estimate, i.e., one which has asymptotic correlation unity with the first derivative of the log likelihood. Let T_n be such an estimate whether it is super or sub efficient with respect to the asymptotic variance. Suppose further that

$$(\sqrt{n}(T_n-\theta),\, Z_n) \to (T,\, Z) \text{ in distribution} \qquad \ldots \quad (3.31)$$

where (T, Z) is bivariate normal with mean zero and covariance matrix

$$\begin{bmatrix} \sigma_T^2 & \sigma_T\sqrt{i} \\ \sigma_T\sqrt{i} & i \end{bmatrix} \qquad \ldots \quad (3.32)$$

Then the variable $\left[T-\frac{\sigma_T}{\sqrt{i}}\,Z\right]$ has zero variance. Therefore

$$\left\{\sqrt{n}(T_n-\theta)-\frac{\sigma_T}{\sqrt{i}}\,Z_n\right\} \to 0 \text{ in probability.} \qquad \ldots \quad (3.33)$$

For a statistic T_n which satisfies the condition (3.33), under some regularity conditions on $P(x, \theta)$, the probability (or density) of a single observation, Doob (1936) has demonstrated[10] that

$$\lim_{n\to\infty} \{i_{T_n}(\theta)\} = i(\theta) \qquad \ldots \quad (3.34)$$

where $i_{T_n}(\theta)$ is the information per observation contained in the statistic T_n computed in the usual way. A simple proof of Doob's proposition is given in a recent paper by the author (Rao, 1960).

[10] Doob (1936) states the required condition in terms of strong convergence. I believe this is not necessary.

We have no such assurance about the limiting information in the case of estimates not efficient in the new sense. In fact, if the asymptotic correlation between T_n and Z_n is ρ and that between T_n and $P'(T_n, \theta)/P(T_n, \theta)$ is nearly unity (as may be expected) we have the relation

$$\lim_{n\to\infty} (i_{T'_n}) = \rho^2 i \qquad \dots \quad (3.35)$$

emphasizing the importance of ρ^2 as a measure of efficiency mentioned in the definition of Section 3.1. The use of T_n entails a loss of information equal to that contained in a fraction $(1-\rho^2)$ of the observations.

3.4. *Efficiency in non-regular cases.*

In non-regular cases such as the rectangular distribution over the range $(0, \theta)$, i.e., where the probability measures corresponding to different values of the parameters are not equivalent, the quantity $i(\theta)$ is not properly defined so that the foregoing theory is not applicable. We shall not discuss such situations in full generality but only consider a special example given by Basu (1952), where the maximum likelihood estimate has a uniformly larger variance than an alternative estimate proposed by him.

Let $x_1, \dots, x_n$ be n observations from a rectangular distribution in the range $(\theta, 2\theta)$, where $0 < \theta < \infty$. The maximum y and the minimum z of the observations are jointly sufficient for θ and the m.l. estimate of θ is $T_1 = y/2$. The asymptotic variance of T_1 is $1/4n^2$ while that of $T_2 = (2y+z)/5$, which is also consistent for θ, is $1/5n^2$. Judged by the criterion of ratio of asymptotic variances the m.l. estimate has only 80% efficiency compared to the alternative estimate. One might be tempted to infer that discrimination based on T_2 is therefore better than that based on T_1, the m.l. estimate, for small differences in the parameter. A computation of the power functions of the tests based on T_1 and T_2 for any sample size shows however that for alternatives close to a given value of θ the power of T_1 is much higher than that of T_2, although T_2 has smaller asymptotic variance than T'_1. On the other hand, for alternatives not close to a given value of θ, T_2 is better than T_1.

3.5. *Concentration.*

As is stated above, in the absence of regularity conditions on an estimate T_n, the asymptotic or actual variance of T_n does not necessarily give a good indication of the concentraton of T_n about the true value of θ. An approach to estimation which is concerned explicitly with comparing concentrations has been given recently by Bahadur (1960). This approach may be outlined as follows. Let T_n be a consistent and asymptotically normal estimate of θ based on n independent and identically distributed observations. For any n and any $\epsilon > 0$, let $\tau = \tau(T_n, \epsilon, \theta)$ be defined by the equation

$$P(|T_n - \theta| \geqslant \epsilon | \theta) = 2 \int_{\epsilon/\tau}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt. \qquad \dots \quad (3.51)$$

τ is called the 'effective standard deviation' of T_n, when θ obtains. It is shown by Bahadur, under mild regularity conditions on the sample space of single observation that

$$\underline{\lim}_{\epsilon\to 0} \; \underline{\lim}_{n\to\infty} \{n\,\tau^2\} \geqslant \frac{1}{i(\theta)}. \qquad \dots \quad (3.52)$$

provided only T_n is consistent. He also shows, under stronger regularity conditions, that the equality holds in (3.52) when T_n is the m.l. estimate of θ. The appearance of Fisher's measure of information in this analysis provides further evidence that this measure is of central importance to estimation.

3.6. *Concluding remarks on efficiency.*

We observe that E_n, for each n, has the maximum local power of discrimination between two neighbouring values (Rao and Poti, 1946) and demand the existence of a statistic T_n independent of θ and having asymptotic correlation unity with Z_n. This ensures that T_n has the same local properties as Z_n. Further it is shown by Wald (1942) that asymptotically shortest confidence intervals can be obtained by inverting regions of the type $Z_n(\theta) > A_n(\theta)$, $Z_n(\theta) < B_n(\theta)$ (one sided regions) and $| Z_n(\theta) | > C_n(\theta)$ (two sided regions). It is clear that any statistic having asymptotic correlation unity with Z_n has the same property in large samples.

It is immaterial what the asymptotic variance of the statistic is provided its asymptotic correlation with Z_n is unity. It may be super or sub efficient in the sense of having smaller or higher asymptotic variance than $[n \; i(\theta)]^{-1}$. If we are placing emphasis on the asymptotic correlation with Z_n being unity we can achieve this by restricting the class of statistics to well-behaved functions of observations. This is for convenience in drawing inferences on θ given the statistic. The m.l. estimate, under some conditions, satisfies our requirements.

Le Cam (1953) suggests asymptotic variance as a measure of concentration of the statistic round the true value in large samples. It may be argued that our interest does not lie in such a measure of concentration. But it is observed that even with respect to such a measure, so far as the existing illustrations suggest, a function of the m.l. estimate serves the purpose.

4. Consistency

A number of quite different examples of inconsistencey of m.l. estimates are now available (Neyman and Scott, 1948; Basu, 1955; Kraft and Le Cam, 1956; Kiefer and Wolfowitz, 1956; Bahadur, 1958). The examples have been useful in leading to a proper understanding of the concept of consistency.

Let us consider the concept of consistency as originally introduced by Fisher (1922). We have already referred to it as Fisher consistency (FC) to distinguish it from probability consistency (PC) which figures prominently in statistical literature (ref. Kallianpur and Rao, 1955). A statistic is said to be FC for a parameter θ if

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

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

There is some virtue in such a definition since it is reasonable to demand that the procedure we adopt should give us the true value of the parameter when applied to the entire distribution. Further the definition places some restriction on functions of observations to be considered, whereas in the case of PC there is no restriction on the statistic which can be arbitrary for any finite sample size, however large it may be. So pinning our faith in PC may be somewhat dangerous in many practical situations, where we have to deal with samples of a finite size.

It is easy to show that an m.l. estimate is FC without any restrictions whatsoever. For if we consider the multinomial situation the log likelihood

$$p_1 \log \pi_1(\theta)+\ldots+p_k \log \pi_k(\theta)$$

when $p_i = \pi_i(\phi)$, is a maximum for $\pi_i(\theta) = \pi_i(\phi)$ which implies, when there is one-to-one correspondence between $\pi_i(\theta)$ and θ, that $\theta = \phi$. In the continuous case the log likelihood

$$\int \log p(x, \theta)\, dS_n$$

has a maximum for $p(x, \phi) = p(x, \theta)$ or $\theta = \phi$ when S_n the sample distribution function is replaced by $F(\phi)$.

What can we say about the m.l. estimate when S_n or $(p_1, \ldots, p_k)$ is close to the true distribution function $F(\phi)$ or $[\pi_1(\phi), \ldots, \pi_n(\phi)]$? We may demand that the distribution in the admissible set maximising the likelihood, $F(\hat{\theta})$ or $[\pi_1(\hat{\theta}), \ldots, \pi_k(\hat{\theta})]$, if it exists, should be close to the true distribution. This is true under no condition whatsoever on the admissible class of distributions in the case of a finite multinomial (Hotelling, 1930; Rao, 1957), under the sole condition $\Sigma\pi_i \log \pi_i$ is covergent in the case of the infinite multinomial (Kiefer and Wolfowitz, 1956; Rao, 1958), and under slightly more restrictive conditions in the case of continuous distributions (Wald; 1949; Kraft, 1955). Examples of inconsistency of the estimated distribution functions due to Bahadur (1958), in the cases of an infinite multinomial distribution and a continuous distribution function, show that they are of a very special character, and it appears that it should be possible to prove convergence of the m.l. estimate of the distribution function under fairly weak conditions.

The situation thus appears to be extremely satisfactory so far as the estimated distribution function is concerned. The corresponding convergence in the estimated parameter then takes place when a continuity condition is satisfied, i.e., $F(\theta)\to F(\phi)$ (or $\pi_i(\theta)\to\pi(\phi)$) implies that $\theta\to\phi$. It may be noted that a parameter is, after all, a code number used to identify a distribution and as such it can be arbitrary and need not satisfy any condition. In the examples of inconsistency of m.l. estimates given by Basu (1955)[11] and Kraft and Le Cam (1956) the continuity condition is not satisfied and the examples depend, in a sense, on an unnatural choice of the parameter.

The anomaly regarding inconsistency of the m.l. estimate of a parameter can be resolved to some extent if we consider consistency in a broader sense as mentioned in Section 2 of this paper. It was observed that if for any two given values of the parameter the distributions of the observations tend to be orthogonal as the sample size→∞, it is reasonable to demand that the distributions of the estimate also behave in the same way. When this is so we may say the estimate is consistent for the parameter in the wide sense. Such wider consistency is ensured when the estimate tends to two different constants for two different

[11] Basu (1955) gave the example of a binomial distribution where the probability of success $p(\theta)$ is defined as follows:

$$\begin{aligned} p(\theta) &= \theta && \text{if } \theta \text{ is rational} \\ &= 1-\theta && \text{if } \theta \text{ is algebraic irrational} \end{aligned}$$

The m.l. estimate of θ, which is the observed proportion of success, tends to θ when θ is rational and to $(1-\theta)$ when θ is algebraic irrational, and is thus not consistent. Basu also shows that there exists another estimate which is consistent for θ. The example of Kraft and Le Cam is more complicated.

values of the parameter. We need not insist that the constant to which the estimate tends should be equal to the true value of the parameter. In Basu's example, (footnote 11) the m.l. estimate tends to θ when θ is rational and to $(1-\theta)$ when θ is algebraic irrational; thus the m.l. estimate is consistent in the wide sense. The same is true of the example[12] considered by Neyman and Scott (1948) where the m.l. estimate of σ^2, the structural parameter, tends to $(n-1)\ \sigma^2/n$ and not to exactly σ^2; clearly, the m.l. estimate is consistent in our sense. The m.l. estimate is, however, not consistent even in the wide sense in Bahadur's examples.

We may also consider a slightly different kind of example due to H.E. Daniels (quoted in a paper by Kendall and Babington Smith, 1950). Observations $(x_1, y_1), \ldots, (x_n, y_n)$ are such that

$$x_i = \alpha_i+\epsilon_i, \qquad y_i = \alpha_i+\mu+\eta_i$$

where ϵ_i is $N(0, \sigma^2)$, η_i is $N(0, \zeta^2)$, and ϵ_i and η_i are independently distributed. Simultaneous estimation of α_i, μ, σ^2 and ζ^2 by the m.l. method leads to the same value for the estimates of σ^2 and ζ^2, so that the estimates of these two parameters are clearly inconsistent in any sense. This result is perhaps not surprising, for the data *themselves* do not seem to provide satisfactory discrimination between σ and ζ (or between one pair of values of σ, ζ and another pair) however large may be number of pairs of observations, when nothing is known about the behaviour of the incidental parameters α, as $i \to \infty$.

5. Conclusion

Since the main aim of this paper is to consider apparent anomalies in the m.l. method, no reference has been made to the superiority of the m.l. method over others. It may be claimed that certain other methods also provide estimates which have the same properties as the m.l. estimates in large samples, although they may be subject to similar criticism in other respects. This may be true, but we cannot use asymptotic properties as sole criteria for the selection of a technique which has to be applied in finite samples in practice. So we have to look for other properties, which hold good for all sample sizes. We may list here some properties of this type which support the claims of m.l. estimates.

The m.l. method has wide applicability. The m.l. estimate is a function of a minimal sufficient statistic, and in special cases is itself a minimal sufficient statistic, a property which may be considered desirable (ref. Rao, 1945, 1946, 1948) and which is not shared, in general, by other general methods of estimation. Finally, consideration of the likelihood function enables us to recognise the minimal sufficient statistic, and if necessary, to supplement the m.l. estimate with other statistics to recover part of the information lost in using the m.l. estimate alone. A more accurate measure of loss of information, based on the variance of Z_n given an estimate T_n (Fisher, 1925), the asymptotic value of which is more appropriate for comparing statistics when the sample size is not very large, shows that the loss associated with the m.l. estimate is smaller when compared to many other procedures. A detailed study of this aspect is undertaken in Rao (1960).

[12] Neyman and Scott (1948) consider an increasing sequence of s series of measurements x_{ij} $(i = 1, 2, \ldots, s \to \infty,\ j = 1, \ldots, n)$, all independently distributed. The probability law of x_{ij} is normal with mean α_i and variance σ^2. The parameter σ^2 is called structural and is the same for all observations, while the α_i, which vary from series to series, are called incidental parameters. The maximum likelihood estimate of σ^2 is $\Sigma\ \Sigma.\ (x_{ij}-\bar{x}_i)^2/sn$ which is consistent for $(n-1)\ \sigma^2/n$ and not for σ^2.

References

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

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

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

Bhattacharya, A. (1946): On a measure of divergence of two multinomial populations. *Sankhyā*, **7**, 401.

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

——— (1954): Choosing between two simple hypotheses and the criterion of consistency. A chapter in the D. Phil thesis submitted to the Calcutta University.

——— (1955): An inconsistency of the method of maximum likelihood. *Ann. Math. Stat.*, **26**, 144.

——— (1956): The concept of aymptotic efficiency. *Sankhyā*, **17**, 193.

Berkson, J. (1955): Estimation by least squares and by maximum likelihood. *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**.

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

Doob, J. (1934): Probability and Statistics. *Trans. Am. Math. Soc.*, **36**, 766.

——— (1936): Statistical estimation. *Trans. Am. Math. Soc.*, **39**, 410.

Fisher, R. A. (1922): On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London*, Series A, 222, 309.

——— (1925): Theory of statistical estimation. *Proc. Cambridge Phil. Soc.*, **22**, 700.

Hellinger, H. (1909): Neue Begründung der Theorie quadratischer Formen von unendlich-vielen veränderlichen. *J. für reine und angew, Mathematik*, 136, 210.

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

Huzurbazar, V. S. (1948): The likelihood equation, consistency and the maximum of the likelihood function. *Ann. Eugenics*, **14**, 185.

Kallianpur, G. and Rao, C. R. (1955): On Fisher's lower bound to asymptotic variance. *Sankhyā*, **15**, 331.

Kendall, M. G. and Smith, B. Babington (1950): Factor analysis. *J. Roy. Stat. Soc.*, Series B (Methodological) **12**, 60.

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

Kraft, C (1955): Same conditions for consistency and uniform consistency of statistical procedures. *Univ. California Publications in Statistics*, **2**, 125.

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

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

Le Cam, L. (1953): On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. of California Publ. in Stat.*, **1**, 277.

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

Neyman, J. (1949): Contributions to the theory of the χ^2 test. *First Berkeley Symposium on Mathematical Statistics and Probability*, **239**.

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

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

——— (1946): Minimum variance and the estimation of several parameters. *Proc. Cambridge Phil. Soc.*, **43**, 280.

——— (1948): Sufficient statistics and minimum variance estimates. *Proc. Cambridge Phil. Soc.*, **45**, 215.

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

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

——— (1960): Asymptotic efficiency and limiting information (to appear in the *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics*).

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

SILVERSTONE, H. (1957) : Estimating the logistic curve. *J.A.S.A.*, 52, 567.

WALD, A. (1942) : Asymptotically shortest confidence intervals. *Ann. Math. Stat.*, 13, 127.

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

——— (1949) : A note on the consistency of the maximum likelihood estimation. *Ann. Math. Stat.*, 20, 595.

RESUME

La méthode de maximum de vraisemblance (m.l.) pour l'estimation des paramètres inconnus a été censuré pour les raisons suivantes : (i) elle ne donne pas un estimateur convergent et (ii) il y a des estimateurs plus efficients que ceux de maximum de vraisemblance et (iii) l'usage du m.l. entrainet de la computation difficile. Cette dernière critique ne serait pas importante si les machines electroniques capable d'accepter les instructions compliqueés concernant des opérations numériques, se font disponibles chez travailleurs.

L'on indique ici, d'abord, que le but de l'estimation est la condensation des données sans perte de l'information essentielle et puis l'on fournit une certaine justification pour la mesure de l'information de Fisher et le critère de maximisation de l'information dans une statistique. Les conceptions de l'efficience et de la consistence ont été reformulées afin que l'on puisse fournir un critère pour une telle choix de l'estimateur que la perte de l'information donnée soit negligible dans les grands échantillons. Un estimateur efficient a été définé comme une estimateur qui a une corrélation asymptotique de mesure d'unité avec la derivé du logarithme de la vraisemblance. Un estimateur de maximum de vraisemblance sous quelques conditions, est efficient dans ce sens.

L'équivalence de cette définition avec celle de Fisher qui constate que la consistence est l'atteinte de moindre variance asymptotique, est établiée dans quelques conditions de régularité sur la statistique. Mais la définition nouvelle résout la difficulté qui a apparué grâce à l'existence des estimateurs super-efficients ayant, peutêtre, une variance asymptotique plus petite que la variance auprès des estimateurs de maximum de vraisemblance.

L'on montre ici, que les estimateurs super-efficients sont équivalents aux estimateurs de maximum de vraisemblance (quand ils sont efficients dans ce sens nouveau) ou sont inférieurs auprès des estimateurs de maximum de vraisemblance pour servir le but de l'inférence statistique, (quand ils ne sont pas efficients dans le sens nouveau).

On dit qu'un estimateur soit consistent dans le sens plus ample si ses distributions asymptotiques pour deux valeurs différentes du paramètre, soient orthogonales. Plusieurs exemples de l'inconsistence des estimateurs de maximum de vraisemblance dans le sens ordinaire paraissent remplir la condition de la consistence plus étendue.

L'on indique ici la consistence de la fonction estimée de distribution est plus fondamentale que celle de l'estimateur du paramètre particulier. Cette dernière consistence suive naturellement si le paramètre est définé comme un fonctionnel continu de la fonction de distribution. Mais un paramètre est rien qu'un nombre de code définé pour l'identification d'une distribution et par conséquence, le paramètre ne doive remplir la condition posée. Les irregularités à l'ègard de l'inconsistence de l'estimateur de maximum de vraisemblance laissent s'expliquer par le défaut de cette condition.

Cette étude conclut avec une note sur les propriétés essentielles des estimateurs de maximum de vraisemblance en cadre des échantillons petits.

APPARENT ANOMALIES AND IRREGULARITIES IN MAXIMUM LIKELIHOOD ESTIMATION

PRESIDENT : E. J. G. PITMAN

1. Apparent anomalies and irregularities in maximum likelihood estimation

L'auteur, M. Rao, présente sa communication[1]

MR. NEYMAN : 1. Mr. Rao's very interesting paper brings out certain philosophical questions regarding criticisms levelled at maximum likelihood estimation and, in addition, presents an extensive history of the problem going back to 1922 when the term Maximum Likelihood Estimate (MLE) was first used. The purpose of the present note is to contribute to both subjects: to express my views on the philosophy of theoretical statistical research and to push the historical sketch back to 1908 when the ideas or certain properties of the MLE seem to have been first expressed.

2. The philosophical aspect of the problem is connected with the two different points of view on statistics, one having to do with intensities of belief and the other behavioristic. To me personally[2] the intensity-of-belief theory of statistics appears dogmatic and, as reflected in the writings of the various authors, is reducible to proposals, occasionally quite insistent proposals, to adopt specified formulas as measures of intensities of belief which an individual should experience in specified circumstances. One such theory, or creed, advises special formulas as *a priori* probability distributions for unknown parameters to be used in cases where the circumstances of the problem do not imply specific *a priori* distributions or even do not imply that the parameter considered is a random variable. Further advice is to use the recommended *a priori* distribution for substitution in the familiar Bayes' formula.

Another modification of essentially the same dogmatic school of thought is based on the premise that the concept of probability is a measure of intensity of belief which is applicable in some cases but not in all. For these cases where the probability is not applicable to measure the uncertainty, the proponents of the relevant school of thought devise new measures of confidence or diffidence and one of them is the mathematical likelihood. In thinking of these and similar attempts at foundations of mathematical statistics, I recall the expressive title of two articles of our recently deceased colleague and friend, D. van Dantzig: "Statistical Priesthood" I and II.[3]

The alternative point of view on foundations of statistics, the behavioristic or operational point of view, stems from some ideas of Laplace and, expressed somewhat more clearly, of Gauss. Leaving aside the question of confidence and diffidence, the behavioristic point of view concentrates on those cases where the mathematical probability is an idealization of relative frequencies as experienced in the realm of natural phenomena. Here, as is most frequently the case, we are confronted with the necessity of a choice among a number of possible actions and the desirability of each action depends upon the value of a parameter intervening in the distribution of the observable random variables. If the value of this parameter is known or assumed known, there is no problem. Statistical problems arise when the relevant parameter is not known and the choice of action has to be based on the values of the observable random variables, that is, on the value of an estimator of the parameter. The problem of estimation is, then, to devise the estimator. This problem splits into a number of detailed problems. One is to establish the properties of all the different estimators, that are available to choose from in a given problem. Another detailed problem is to devise the method, just as easy a method as possible, of calculating the estimator having the properties that fit the situation best.

The properties of an estimator which may be considered desirable vary from one particular problem to the next. Also undoubtedly, they depend on subjective elements: it is quite conceivable that two different persons contemplating the same problem will have different preferences for the properties that an estimator should have. In some problems and to some individuals, unbiasedness of the estimator and the smallness of its variance appear of paramount importance. Here, Gauss' method of least squares is frequenlty the answer. In other cases, unbiasedness and small variance are secondary or irrelevant, and some other property appears important. Thus, for example, in the problem of estimating the degree of contamination of drinking water, it may appear most important not to underestimate the contamination and, from the point of view of public health, the most desirable estimator is certainty not unbiased.

[1] *Bull. Inst. Int. Stat.*, XXXVIII, 4, p. 439.

[2] J. Neyman, "Inductive behaviour as a basic concept of philosophy of science." *Rev. Int. Stat. Inst.*, Vol. 25 (1957), pp. 7-22.

[3] D. van Dantzig, "Statistical Priesthood" I and II, *Statistica* Neerlandica, Vol. II (1957) and Vol. 12 (1958).

With reference to Rao's discussion of criticisms of maximum likelihood estimation, I wish to make it clear that my own criticisms are directed not towards the use of MLE but to the insistence that these estimators, or indeed any other estimators, be used as a matter of principle. In my opinion, any user of statistical methods should have complete freedom of choice and the role of theory in the matter is to elucidate the properties of the methods that are available. Thus, for example, the famous inequality giving the greatest lower bound of the variance of an unbiased estimator, first found by Fréchet and then, independently, by Harald Cramer and Rao, is a very important result. It is purely behavioristic or operational and tells us that, under certain conditions, if we insist on using unbiased estimators then, no matter what we do, the variance of the estimator cannot be less than a calculable limit. As indicated by a slightly more general version of the same inequality, there may be a possibility of finding an estimator which has mean square error less than the bound of Rao; then this estimator must be biased. With this in mind, the consumer of statistical theory may perhaps decide to drop the requirement of unbiasedness.

For quite some time the possibility of biased estimators with mean square errors less than the bound for the variance of an unbiased estimator remained just a theoretical possibility with no live example to show that they really exist. Then Mr. Joseph Berkson appeared on the scene and produced a real case of an estimator, obtained by minimizing the classical Karl Pearson χ^2, which not only has its mean square error less than that of MLE but also less than the indicated bound! I submit that this particular result of Berkson is of considerable interest and importance. Quite apart from the possibility of estimating the parameter with precision, in the sense of mean square error, better than other known estimators, this result raises a host of novel theoretical problems: what are the situations in which biased estimators exist with their mean square errors less than the Fréchet-Cramér-Rao lower bound for variances of unbiased estimators? Can one invent a method, a machinery such as the maximalization of the likelihood or the minimalization of the χ^2, which, at least in some cases, would lead to such estimators if they exist ?

I note that Rao does not particularly like Berkson's result, apparently for the reason that in the particular problem considered, Rao's own interest centers on a parameter different from the one estimated by Berkson. Evidently, we consider the question from different points of view.

3. Turning to the other part of my contribution, concerned with a detail in the history of MLE I find it interesting that, at least on two occasions, the idea of MLE sprang up from the dogmatic intensity-of-belief approach to statistics. However, in both cases the original dogmatic approach was followed by studies of a distinctly behavioristic or operational character. The two approaches can be roughly summarized as follows :

(i) *First statement :* MLE should be used because this use is implied by such and such principle.

(ii) *Second statement :* The consistent use of MLE will guarantee such and such long range advantages.

As far as I am aware, the priority in the approach to MLE just described, involving both statements (i) and (ii), belongs to F.Y. Edgeworth.[4] The dogmatic intensity-of-belief ideas of Edgeworth, which are also noticeable in Laplace, were connected with the arbitrary *a priori* distributions and the use of Bayes' formula. The fact that this brought Edgeworth to the use of what we now call MLE is occasionally noted in the literature. For example, an appropriate reference is found in M.G. Kendall's book.[5] However, it is much less generally known and seems to have escaped the attention of Rao, that, after making statements roughly equivalent to (i), Edgeworth proceeded to formulate a conjecture in the spirit of the statement (ii) above. The passage I have particularly in mind, published in 1908, is printed in the Appendix of a very large and involved paper. It so happens that this conjecture of Edgeworth is now known to be broadly true but with some exceptions. Also, as reflected in the excellent historical summary given in the present paper by Rao, although 52 years have elapsed since the publication of Edgeworth's conjecture, the limits of its validity are still the subject of numerous studies all over the world. In these circumstances and because of the general lack of awareness of the identity of the author of the conjecture, it appears appropriate to reproduce here a brief quotation from the Appendix of Edgeworth's paper.

[4] F. Y. Edgeworth, "On the probable error of frequency constants" *J.R.S.S.*, Vol. 71 (1908), pp. 651- 678.

[5] M. G. Kendall, "*The Advanced Theory of Statistics*" Vol. II, Griffin, London, 1946.

Appendix

This Appendix is designed as a receptacle for some mathematical developments which might have interrupted the course of the preceding arguments.

I *Love's proof of some preceding propositions.*

A foremost place is due to Love's confirmation of certain propositions above stated by means of an independent proof.

The first proposition thus verified is a particular case of a general theorem which may thus be provisionally restated. Let $y = e^{\psi(x)}$, be a frequency-function apt to represent the distribution of statistical observations. Let $x_1, x_2, \ldots, x_n$ be a set of n observations forming a random selection from the indefinitely large group of the observations ranging under the frequency curve. Let $\phi\ (x_1, x_2, \ldots, x_n)$ be that function of the given observations, which affords the *most probable* value (as determined by inverse probability) of the sought point to which the observations relate; a symmetrical function when, as will be here supposed, the observations are all of equal weight or worth. Then, if we take (at random) a series of sets, such as

$$\begin{array}{cccc} {}_1x_1 & {}_1x_2 & \ldots, & {}_1x_n, \\ {}_2x_1 & {}_2x_2 & \ldots, & {}_2x_n, \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ {}_mx_1 & {}_mx_2 & \ldots. & {}_mx_n; \end{array}$$

and form for each set the corresponding value of ϕ, the series of mean values thus formed say, ${}_1\phi, {}_2\phi, \ldots, {}_m\phi$ will be such that (m and n being large numbers) the mean square of their deviation from the true point, say x, viz.,

$$\frac{({}_1\phi - x)^2 + ({}_2\phi - x)^2 + \cdots + ({}_m\phi - x)^2}{m}$$

will be less than the mean square of deviation presented by any other set of mean values ${}_1\chi, {}_2\chi, \ldots {}_m\chi$, each formed from a set of n observations, where χ (like ϕ) is a symmetrical function of observations, having the properties of an average.

In line with the Victorian style, the above passage is interspersed with footnotes. I take the liberty of omitting these.

In contemplating this passage, one is struck by the change in style, terminology and precision of expression which have occurred during the half-century that elapsed since the publication of Edgeworth's paper. However, the translation of the passage into modern terms presents little difficulty.

The function $y = \exp\{\psi(x)\}$ is the probability density of an observable random variable, say X. Further context suggests that, in addition to x, the particular value of X, the function ψ depends upon a parameter, say θ, and that the actual value of this parameter, say θ_0, is unknown. The value θ_0 is described by Edgeworth as "the sought point to which the observations relate" and, later on, as "the true point, say x, ... ". In order to avoid the use of the same letter x in several different meanings, I introduce the symbol θ_0.

The symbols ${}_ix_j$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ represent independent observations on the variable X arranged in m samples of n observations each. The function ϕ described as "that function of the given observations, which affords the *most probable* value (as determined by inverse probability)" is simply the maximum likelihood estimator of θ_0.

Edgeworth's assertion is that, if X is an alternative estimator of θ_0, based on the same observations as ϕ and subject to some not distinctly stated limitations: "where χ (like ϕ) is a symmetrical function having the properties of an average," then the asymptotic mean square error of ϕ will be less than (presumably, not greater than) that of χ.

Edgeworth was not able to prove his conjecture to his own satisfaction and tried to enlist the help of Love. Unfortunately, Love's success was limited. However, Edgeworth's assertion compares favourably with those found in a number of recent books on statistics which flatly assert that, as proved by somebody or other, the asymptotic variance of MLE is a minimum, without any limitations. In favour of Edgeworth is the realization that some sort of restriction on the alternative estimator χ is necessary.

The ideas of Edgeworth did not seem to have much influence on the thinking of the contemporary statisticians, and the above clear cut statement of the presumed optimal property of MLE went unnoticed.

The idea reappeared in the literature fourteen years later in the famous paper of R.A. Fisher to whom we owe a great number of other concepts and terms, consistency, efficiency, sufficiency, etc. Here, again, the origin of MLE was in the degree-of-belief approach to statistics but based on principles different from those of Edgeworth. Also in this case, the original approach, which appears to me dogmatic, was followed by increasingly accurate behavioristic studies. This part of history appears adequately covered by Rao, and I need not enter in any details.

4. Before concluding I would like to request Professor Rao to explain his philosophical standpoint a little more clearly than he does in his paper. Some passages in his paper suggest the possibility that, in Rao's opinion, MLE should always be used irrespective of the properties it may have. Does Rao really mean this? The passages I have in mind include Rao's *Introduction* and his *Conclusions*.

In his Introduction Rao lists four different reasons for which maximum likelihood estimation has been mainly criticised. In the Conclusions there are listed the various advantages of MLE. It looks as if the choice of a method of estimation is treated more or less like the choice of an automobile which a family will have to use for a number of years. All automobiles on the market are open to some criticisms and some of them have certain advantages. The standpoint of Rao seems to be that the advantages of the automobile MLE outweigh the disadvantages. This impression is fortified by Rao's dealing with what he considers as criticisms of MLE. One reason listed is that in certain cases MLE have been shown to be inconsistent in the usual sense of the term. This fact is not denied by Rao. Instead, he introduces a distinction between PC and FC, consistency in the sense of convergence in probability and consistency in the sense of Fisher. Also there are some other interesting connections in which the term consistency is used. It is then shown that in some cases where the MLE are inconsistent in one sense, they are consistent in another sense.

Another ground for criticisms discussed by Rao is that, in some specified cases, consistent estimators of a parameter θ are readily available with mean square errors that are less than those of MLE. In one such case, Rao's stand seems to be that it is pointless to try to estimate θ. In my opinion, the difference between selecting an automobile and selecting a method of estimation is that the car is, so to speak, indivisible. It is impossible for a purchaser to take some characteristics of a Volkswagen, very desirable for short trips in town, and combine them with certain other characteristics of a Rolls Royce, most desirable for extensive travel. If the family is limited to a single car, it must face the necessity of weighing the relative advantages of each make against the disadvantages. No such necessity exists in the choice of a method of estimation. Provided one knows the properties of the several methods available for the given problems, and provided one is clear as to what one wants to achieve, one is at liberty to use MLE in certain cases and some alternative estimators in others. However, in order to avoid disappointments, it is quite essential to know what the properties of the different estimators are.

From this point of view, the authors whom Rao considers as critics of MLE, are not really critics. They just provide us with valuable information.

In order to make Rao's philosophical stand quite clear, I suggest that he gives an unambiguous answer to a trivial question which, however, is both specific and illustrative. Suppose that an association of manufacturers of certain measuring instruments is anxious to have a formula for estimating the error variance σ^2 of each instrument. Suppose that with each instrument a moderate number n of independent measurements are made of a large number N of different objects. With the usual assumptions and with the usual notation, the two contemplated estimators are

$$S_1^2 = \frac{1}{nN} \sum_{i=1}^{N} \sum_{j=1}^{n} (x_{ij} - x_{i.})^2$$

and

$$S_2^2 = n S_1^2/(n-1).$$

The first estimator S_1^2 is the ML estimator. The second is not. However, the first estimator has an operational property which may seem undesirable: it is inconsistent. In fact, as $N \to \infty$, the first estimator tends in probability not to σ but to a smaller number $(n-1)/n\sigma^2$ On the other hand, the second estimator is consistent and, in fact, unbiased.

The question is: which of the two estimators would Rao recommend? I hope that Rao's advice will be behavioristic, in favour of S_2^2. If it is not and if he insists on MLE, there may be trouble. In fact, there may be a lawsuit for damages. For, if one of the manufacturers, say A, has his $n = 10$ and another manufacturer B has $n = 2$, the manufacturer A will have a legitimate reason to complain.

Here is another question. In his paper Rao writes that someone has suggested that the super efficient estimators constructed by Hodges and Le Cam be used in practice. Would Rao kindly indicate who made this suggestion. The point is that, with Rao's statement as it now stands, the reader is likely to think that the suggestion came from either Hodges or Le Cam or from both. At this I would be most surprised.

The Hodges-Le Cam estimators are the usual count examples showing that certain theorems, thought to have been proved rigorously, are in fact, false. In the present case the theorem in question is : out of all consistent and asymptotically normal estimators, the MLE has a minimum asymptotic variance for all values of the estimated parameter. Hodges' example showed that this theorem is false and that, for the assertion to be true, it is necessary to consider not all the estimators of the kind described but those of some limited class.

Résumé

Le papier intéressant de M. Rao me suggère quelques réflections d'ordre philosophique et quelques autres d'ordre historique. Premièrement, il me parait frappant que les estimateurs de "Maximum Likelihood" (M.L.E.) semblent être recommandés par certains auteurs, dont M. Rao, pour des raisons de deux espèces différentes : parce que dans certains cas ces estimateurs possèdent des propriétés désirables et parce que leur usage systématique est une affaire de principe, indépendamment des conséquences. La premiére raison est toute naturelle, mais la deuxiéme me parait étrange. Voici un exemple. Considérons quelques instruments à mesurer. Pour charactériser leur précision, on emploie chaque instrument pour faire m mesures indépendantes sur chacun des n objets différents. Soit $X_{ij} = N(\xi_i, \sigma^2)$ une de ces mesures. Une agence publique, dont le but est de charactériser la précision moyenne des instruments produits par différentes fabriques, a besoin d'un estimateur de la variance σ^2 de l'erreur de mesure. La formule

$$S^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} - X_{i.})^2 / mn \tag{1}$$

représente l'estimateur M. L. Supposons que deux usines, A et B, produisent des instruments identiques, avec $\sigma = 1$. Supposons que dans la fabrique A on a $m = 2$. Alors, comme on le sait bien, lorsque n augmente, lim $p\, S^2 = 0.5$, D'autre part, si dans la fabrique B on a $m = 10$, alors lim $p\, S^2 = 0.9$. Donc, dans ce cas, l'application de M.L.E. conduirait à une conclusion fausse que les instruments venant de A sont beaucoup plus précis que ceux venant de B. D'autre part, il est aisé de définir un estimateur de σ^2 n'ayant pas cet inconvenient. Ce qui m'intéresse c'est si M. Rao recommandrait l'usage de (1), meme dans les conditions indiquées, pour l'unique raison que cet usage est prescrit par le principe de M. L. de M. Fisher.—Un détail historique : à ma connaissance, la première tentative de formuler un théorème impliquant les propriétés désirables des M. L. E. se trouve dans un travail de F. Y. Edgeworth publié en 1908. J'en cite un passage dans mon texte anglais.

Mr. Kitagawa : I believe that Mr. Rao has been most successful in attaining his main purpose in this paper, namely, in resolving apparent anomalies and irregularities in maximum likelihood by reformulating the notions of efficiency and consistency in a very natural and elegant way and also in closer connection with the original ideas of Sir Ronald Fisher. It is the merit of the present paper that further discussions can and must be done from any more essential standpoints including philosophical ones than those connected merely with mathematical techniques.

Mr. Berkson : Professor Rao has presented a novel concept of estimation, and it is interesting to visualize its operation in the bio-assay case. Recall how the bio-assay problem arises in its medical application. The physician has ordered the administration of, say, 200 units of insulin in a unit volume and the pharmacist prepare a solution with that concentration. If his stock solution contains 400 units per unit volume, he will dilute it to half its strength. If it contains m units he will dilute it in a proportion of $1/m$. He makes a bio-assay to find out what is the value of m. In terms of the decision concept of estimation, the decision involved here is the number of cubic centimeters of diluent to add to the stock solution, in order to bring it to the strength required by the physician. The measure of the efficacy with which the bio-assay is accomplished is some average of the error made in estimating m, and the classic measure used, though not the only conceivable one, is the mean square error. Perhaps we will give it loftier statistical prestige if we call it a "loss function."

MR. Rao suggests that in statistics, estimation is only an incidental procedure—that the serious statistical objective is the condensation of the data to economical form so that, for instance, they may be added to similar data obtained in another bio-assay. We can imagine doing this. We perform a bio-assay and record the results in a statistically efficient way, perhaps as a minimal sufficient statistic, perhaps as the whole likelihood function. When we perform another bio-assay we will, in an efficient manner, add the efficiently summarized data of that bio-assay to the efficiently summarized data of the first bio-assay. When we make another bio-assay, we will add the efficiently summarized data to those already accumulated, and so forth. Now, if instead of making a point estimate in the first instance and diluting the solution to the required 200 units in accord with that point estimate, the statistician faithfully pursues the avowed purpose of estimation to condense the data and to prepare a whole series of fiducial limits in the sense of Fisher, or confidence limits in the sense of Neyman, what will happen ? Well, what will happen in the meantime is that the patient will die in diabetic coma ! This, of course, is irrelevant to the logical development of the fundamentals of statistics, but it is a point. Another point is that the law, in its benightedness, does not allow the killing of a patient with an overdose of inference theory and an underdose of insulin. If I am the statistical bio-assayist following Rao's theories—and this is conceivable since I am a great admirer of Rao—I will be committed to the hoosegow on the charge of malpractice. I hope that while I am in durance vile, my friend Rao will visit me. It will be a consolation to contemplate with him the ultimate nature of statistics and to realize that while I am suffering on bread and water it is in the noble cause of statistics considered as right thinking and correct rational inference, regardless of practical consequences.

Now a word about the summarization of data. Suppose we accept Fisher's measure of efficiency as the proportion of available information (in a certain sense), extracted by a statistic T[8].[6] It is obvious that this cannot be a measure of the efficacy of T as an estimator. Any "random" number or even a meaningless symbol T that is a one-to-one function of the possible samples will be completely efficient (sufficient) in this sense. However, it would hardly do as an estimator. But considering Fisher's efficiency only as a measure of effective condensation of data, what is the relation of it to maximum likelihood estimation ? It should be emphasized that Rao definitely did not say that this estimator necessarily extracts as much information as possible. But I have the impression that such a claim has been made, though this may be a misunderstanding. It seems to be widely believed for instance that where a sufficient statistic other than the sample exists, the maximum likelihood estimate will be sufficient and hence will extract the total amount of information available [8] [9] [10] [11]. But this is not strictly true. What seems to be true is that the maximum likelihood estimate will be a function of the minimal sufficient statistic, but it will not necessarily be a one-to-one function, and therefore it will not necessarily be sufficient. In such cases—and they seem to be of fairly common occurrence, e.g. [5]—the maximum likelihood estimate would not be sufficient even for storing the total "information." For this purpose, one should have to store at least the sufficient statistics themselves. In the instance of the logistic function with binomial variation, there are minimal sufficient statistics for its parameters α, β. For the case of a "bio-assay" experiment with three equally spaced "doses" x, $n = 10$ animals exposed at each dose, both parameters to be estimated, a minimum χ^2 estimate which I call the "minimum logit χ^2 estimate" is consistent FC as well as consistent PC, and is asymptotically efficient. For finite samples it is sufficient, and extracts the total amount of available information. The same is true for an infinite number of other controllable experimental arguments, though not in all such arrangements. The maximum likelihood estimate is also consistent FC and PC and is asymptotically efficient, but for finite samples it has larger mean square error than the minimum logit χ^2 estimate, and it is not sufficient. It loses a calculable amount of information, which is small for an experiment in which the probability P_c of response at the central dose is 50 per cent, but the proportion of information lost increases as the experiment is asymmetrically placed, and approaches unity as P_c approaches 1 or zero. I should like to ask Professor Rao whether, with an experiment such as described, he would still prefer the maximum likelihood estimator to the minimum logit χ^2 estimator.

[6] Rao reiterates this definition, but one should note that Fisher did not limit it to asymptotically normal estimators, and specifically applied it "to finite samples and to other cases where the distribution is not normal." The definition is pertinent as a measure of the sufficiency of a statistic, but not as the efficiency of an estimator. This distinction is widely recognized.

My position on estimation is crudely simple to the point of simple-mindedness. Statistics is used for many purposes. One of these, and in my own opinion it is the centrally important one, is to estimate a specified parameter. When this is the objective, the measure of the relative worth of an estimator is the value of some loss function such as the mean square error. In using the mean square error, I do not mean to make the world believe that its loss is always proportional to the square of its error, as Rao implies, but take it only as a representative loss function though, of course, it is the classic measure of error and the most widely used. It does seem to be a good working rule to suppose that the probability of making an error greater than a critical size is probably larger with increase of the error variance of the measure, though the conditions in which this is certainly true are limited indeed. I do not know any theorems in which the probability of significant errors is smaller, the larger the mean error, but I know some in the opposite sense, for instance the normal law as a precise case, and the Tschebyshoff rule for an approximate evaluation. I do not think that it is a matter of indifference whether the mean square error of an estimate is large or small, and I see no reason for preferring an estimate with large mean square error. Knowing nothing relevant to the contrary about an assay, I should want medicine that I prescribed to be assayed by a method with known small mean error. Indeed, I should feel duty bound to insist on it. And if there were a reason for my disregarding the small mean error, it could not be because there was another method that better condensed the data. I quite disagree with Rao when he defines the purpose of estimation as the condensation of data. The object of estimation is to evaluate the parameter with as little error as possible, in some acceptable definition of "error." To define the objective as condensation of data, irrespective of error, seems to me not to point up the essential purpose of estimation, but to divert us from it. Rao's apparent predilection for an assay with large average error seems to me unnatural. His present belittling of small mean square error is puzzling. I notice that elsewhere he characterizes an unbiased estimate as "best" if it has minimum attainable variance [12].[7]

Now this does not mean that the loss function of mean square error is the only conceivable one, or that it is necessarily definitive. If, in a particular application, some other loss function suggests itself, let it be investigated. Rao has questioned my use of the mean square error, which is the loss function of Gauss,[8] when comparing some minimum χ^2 estimates with the maximum likelihood estimate of the parameters of the logistic function and of the integrated normal function. In these investigations it was found

[7] Rao has informed me that he was only using accepted terminology here, without implying that such an estimator is best from a practical view. Even so, the use reflects a generally accepted attitude and does not support Rao's suggestion that his view is shared by most statisticians.

[8] Since Edgeworth and Gauss have been mentioned in this discussion, the following quotation from Edgeworth [7] referring to Gauss is interesting:

The reflections of the great mathematician on this branch of mathematical physics deserve to be transcribed here —"That the metaphysic employed in my Theoria Motus Corp. Coel...to justify the method of least squares has been subsequently allowed by me to drop (Dass ich....habe fallen lassen) has occurred chiefly for a reason that I have myself not mentioned publicly. The fact is, I cannot but think it in every way less important to ascertain that value of an unknown magnitude the probability of which is the greatest—which probability is nevertheless infinitely small—rather than that value by employing which we render the Expectation of detriment a minimum (an welchen sich haltend man das am wenigsten nachtheilige Spiel hat). Thus if $f(a)$ represents the probability of the value a being assumed by (für) the unknown quantity x, it is not so important (ist weniger daran gelegen) that $f(a)$ should be a maximum as that $\int f(x)\, F(x-a)dx$, the integral extending over all possible values of x, should be a minimum; when for F is selected a function that is continually positive and continually increases in a due degree (auf eine schichliche Art) with the increase of the variable. That the square is selected for this purpose is "purely arbitrary, and is in the nature of the subject that there should be this arbitrariness (Willkürlichkeit). Except for the well-known very great advantages....which the choice of the square secures, one might have chosen any other function satisfying the above conditions."

Rao has asked me why I thought it pertinent to consider the square error as a measure of the efficiency of the estimators which I studied. My reasons are the same as Gauss's.

that some minimum χ^2 estimators are more efficient [1] [2] [3] [4]. But he has presented no parallel analysis of these estimates, on the basis of some othêr loss function, disagreeing with this result. Until he does so, judgement of his criticism of my work should be held in abeyance.

Although at one point Rao criticises the use of the criterion of mean square error, at another point he seems to approve it, for he suggests that my "difficulties" are due to having applied it to the estimate of the wrong parameters. In my papers I was concerned with estimating parameters α and β (location and scale) of the logistic function. I shall explain that some 10 years ago, when I became concerned with the problem of statistical bio-assay, I found that not only were there innumerable articles but there were even several books concerned with this problem (or its equivalent in terms of other functions). Itwould seem that this itself is sufficient justification for examining the estimate of these parameters. But I may go further. Wishing to get data and examples in actual use, I communicated with several pharmaceutical firms, and from one very important house I received copious data of bio-assays that had been performed "by the method of Bliss" ("probits," with maximum likelihood). In all those assays a value of β was assumed as known from previous experience, and the problem was to estimate α (from which the E.D. 50 followed directly). This was the origin of my taking as the paradigmatic problem for mathematical statistical bio-assay the estimation of α with β known. Rao says I should instead have considered estimating the probabilities of death at various doses. But it is not for the mathematician to say what parameters should be estimated. It is his function only to say how parameters that are specified for him can best be estimated —if he can ! If Rao does not say that the P_i's should be estimated *instead of* α, β, but that it would be interesting and important to consider the estimate of the P_i's *also*, my *answer* is: "Yes, and I have thought of it, but with my limited and primitive means of computation it was important to do first things first, and besides, this particular programme is not so easy to define, much less to carry out, as it may appear."[9] If Rao desires it, I shall undertake some computations along these lines. I may say in advance, however, that (1) whatever the results may turn out to be, they will not mitigate the results already obtained in estimating α, β, which have their own primary importance and (2) I do not anticipate an essential reversal of my previous conclusion of the general relative inefficiency of the maximum likelihood estimates compared with some minimum χ^2 estimates, in these experiments.

We possess no principle of estimation the application of which ensures a best estimate in terms of the mean square error or any other objective operationally meaningful loss function. For the case of multinomial variation, I have defined an extended class of minimum χ^2 estimates which provides asymptotically efficient estimates [2], and this can frequently, but not always, be interpreted as estimates with approximately minimum variance in large samples. The maximum likelihood estimator can most simply be regarded as just one of the estimators in this class of minimum χ^2 estimators. For finite samples, really of any size but euphemistically referred to as small samples, we do not in general know which of these estimators has smallest mean square error. Certainly there is no reason to believe that the maximum likelihood is necessarily the best. The only way to find out is to investigate. Let us not stifle investigation by assuming that we already know. For some cases, I have found that what I have called the minimum transform χ^2 estimate has smaller mean square error than either the maximum likelihood estimator or the minimum Pearson estimator, and, incidentally, smaller than the lower bound for the variance of an unbiased regular estimator, which was widely thought to be impossible. For a special case with the logistic function I found another estimator—the Rao-Blackwellized estimator—which has even smaller mean square error. Mr. Joseph Hodges and I[6] will present, at the forthcoming Berkeley Symposium, another estimator, the H estimator, for the same case, which in a certain minimax sense is still better than the Rao-Blackwellized estimator. Different estimators can be developed for particular cases which have different operationally defined optimum properties. We do not have to have a monolithic statistics. Let investigation flower along different paths, and let a thousand estimators bloom!

I should like to take the opportunity to express my gratitude to Mr. Rao for his interest in my work and for the many invaluable suggestions that he has given me in the course of our correspondence. I have a very lively appreciation of the generosity reflected in so renowned a mathematician taking the trouble to help me. I realize that his presentation was not made to derogate my work, but to provoke a clarification of my views. I hope my attempt to reply in a forthright manner is understood in the same spirit.

[9] Which P_i's I should consider estimating ? The particular P_i's corresponding to the three experimental doses have no special interest.

REFERENCES

[1] BERKSON, J. (1957): Tables for use in estimating the normal distribution function by normit analysis. *Biometrika*, **44**, 411–435.

[2] ——— (1956): Estimating by least squares and by maximum likelihood. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1 University of California Press, Berkeley and Los Angeles, 1–11.

[3] ——— (1955): Estimate of the integrated normal curve by minimum normit chi-square with particular reference to bio-assay. *J.A.S.A.*, **50**, 529–549.

[4] ——— (1955): Maximum likelihood and minimum χ^2 estimates of the logistic function. *J.A.S.A.*, **50**, 130–162.

[5] BERKSON, J. and ELVEBACK, LILA (1960): Competing exponential risks, with particular reference to the study of smoking and lung cancer. *J.A.S.A.*, **55**, 415–428.

[6] BERKSON, J. and HODGES, J. L. JR. (1960): A minimax estimator for the logistic function. *Proceedings of the Fourth Berkeley Symposium on Probability and Statistics*. In press.

[7] EDGEWORTH, F. Y. (1908): On the probable errors of frequency-constants. *Jour. Roy. Stat. Soc.*, **71**, 386–87.

[8] FISHER, R. A. (1950): Theory of statistical estimation, in *Contributions to Mathematical Statistics*, John Wiley and Sons, Inc., New York, Paper 11, 700-725, See p. 714.

[9] MATHER, K. (1947): *Statistical Analysis in Biology*, Interscience Publishers, Inc., New York, Offset Lithoprint Reproduction, 212–213.

[10] MOOD, A. M. (1950): *Introduction to the Theory of Statistics*, McGraw Hill Book Company, Inc., New York, 160.

[11] QUENOUILLE, M. H. (1956): Notes on bias in estimation, *Biometetrika*. **43**, 353–360, 363.

[12] RAO, C. R. (1952): *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, Inc., New York, 139–140.

SIR RONALD FISHER: Mr. Neyman surprised many of us by his claim in his recent memorandum that Edgeworth introduced the Method of Maximum Likelihood. Edgeworth in fact bound his method on the theory of inverse probability and ascribed his notion specifically to K. Pearson and Filon in 1898; the Method of Maximum Likelihood may equally be found in this paper, only Pearson and Filon were under the misapprehension that the errors of random sampling were the same as those of the Method of Moments regarded as axiomatic by these authors.

Edgeworth, however, ends his paper with the reservation that all that he had said referred only to Measures of Central Tendency and not to the more complex problem of "The Fluctuation".

MR. KATZ lit les observations suivantes soumises par M.G.A. BARNARD.

It seems to the writer that the so-called anomalies in maximum likelihood estimation arises from misunderstanding of the problem which the method sets out to solve. The idea has grown up that the object of an estimation procedure is to find a single value for a parameter which may in some sense be regarded as "best", given a set of data. Alternatively, an interval is required within which the true value of the parameter may be supposed to lie. Neither of these formulations corresponds with the requirements of scientific inference. These can, in the first place, be roughly specified as requiring both a single value, to be regarded as "estimate" and indissolubly associated with it, some means of specifying the "error" to which this estimate is liable.

The method of maximum likelihood, in its simplest form, answers this requirement by giving as the "estimate" the point at which the log likelihood function has its maximum value, together with the inverse of the second derivative of this function at the maximum, which is used as an indication of the error. This procedure may be "justified" in several ways, but perhaps the principal justification can now be seen to consist in the facts: (1) that the log likelihood function is always minimal sufficient, so that for problems of the type considered we need only aim to specify this function. (2) The log likelihood function is often well approximated in the neighbourhood of its maximum, by a quadratic expression; so that a specification of the location of the maximum, together with the second derivative there, gives us a good idea of the general course of the function.

From this point of view it is evident that we may expect "anomalies" to arise when the log likelihood function is far from being parabolic, and it is trivial that such instances can be constructed, starting from non-anomalous cases, by sufficiently pathological transformations of the parameters. More serious difficulties may arise when it is the form of the probability (density) function of the observations which makes the parabolic approximations poor—as may arise, for example, with certain configurations of small samples from the Cauchy distribution. In such cases we should bear in mind the principle of serendipity, according to which, if we are lucky enough to have obtained a sample which happens to give a parabolic log likelihood function, we need not concern ourselves with the problem of what we should have done had we been less lucky. In other cases, where serendipity does not come to our aid, we may either follow the suggestion made many years ago by Fisher, of specifying higher derivatives of the log likelihood, or we may use the sequence of moments of the likelihood function, rather than the sequence of its Taylor coefficients as the basis for our specification. In the case of the Cauchy distribution this would lead us to the Pitman estimator, though with an interpretation different from his, since we would think of it as associated with an "error" given by the second moment of the likelihood function, rather than as a "point estimate."

The problem of approximating to the specification of the log likelihood function, by way of the form indicated, is thus seen to have the same limited degree of arbitrariness associated with it as do other problems of approximation of functions.

In certain particular contexts a practical decision problem may be represented as leading to what has been called the problem of point estimation, and in such cases the loss function and a Bayesian prior distribution require to be specified before a unique solution can be arrived at. The fact that the data enter the solution of this problem through the likelihood function which they generate can be seen as another mode of justification of the likelihood approach. Evidently, under suitable regularity conditions, the solutions to wide classes of problems of this type could be seen to be estimators which are functions of the maximum likelihood estimator, together with the second and perhaps a few higher derivatives of the log likelihood function.

The need for a simplified description of the likelihood function by means of parabolic approximations, or otherwise, can be thought of as considerably reduced by the possibility, now existent, of drawing contours of constant likelihood, for up to 3 unknown parameters with the help of automatic computers. A specimen of such a contour map (for the programme for with I am indebted to Mr. H. Whitfield of Imperial College) for the unknown parameters p_1, p_2 arising from the 2×2 table is attached. The effect of the skewness of the likelihood function for p_2 can be seen quite clearly and the limitations of the paraboloidal approximation are apparent. It is also evident that these limitations would be reduced considerably if the logistic transformation

$$\alpha_1 = \log p_1/(1-p_1), \quad \alpha_2 = \log p_2/(1-p_2)$$

were applied to the parameters.

A	not-A	Total	
3	7	10	$P_r(A) = p_1$
1	11	12	$P_r(A) = p_2$

All the essential ideas mentioned above seem to the present writer to have been implicit in Fisher's classical papers, and the only excuse for restating them here is that subsequent developments have shown that these classical papers have not always been studied with the attention they deserve.

APPARENT ANOMALIES AND IRREGULARITIES IN M. L. ESTIMATION

MR. BIRNBAUM: I would like to congratulate Mr. Rao on his presentation of very interesting contributions to the mathematical theory of estimation, and also to thank him for his clear statement of his general standpoint concerning the nature and purpose of the estimation problem. His view of estimation is a broad one, in which a single point estimate may be used in a variety of specific ways, some of them having the character of decision-making or specific inference problems, and some of them serving the purpose of efficient recording and interpretation of basic scientific or technical information of more general interest. Thus Mr. Rao's view of estimation is a kind of combination of the two standpoints presented by Messrs. Neyman and Barnard, and his general problem is that of showing how well a single point estimator can serve these broad and varied functions.

The standpoint which Mr. Neyman stated concisely is one upon which I based the first part of my own contribution here last week: in the problem of point-estimation, which formally includes confidence limit estimation, in general all possible estimators should be considered, and for a specified situation of application of choice of one estimator should (at least in principle) be used on comparisons of the probability distributions of all estimators. Such comparisons and choices may be informal or may utilize formal criteria, but they should reflect appropriately the situation of application and the statistician's purposes and judgements in the given situation—but the subject-matter of such comparisons and choices is basically those properties of estimators, represented by probabilities of errors of many kinds, which admit direct frequency interpretations. This standpoint leads in typical problems to large classes of admissible estimators, often including maximum likelihood estimators among many others. None of these admissible estimators can be eliminated from consideration as a matter of principle on the grounds mentioned; choices can be based only on grounds of specific judgements in specific problems and situations.

Mr. Barnard considers the point-estimation problem itself to be an incomplete and inadequate formulation of another inference problem. He states that the solution to this other problem is in principle the likelihood function itself, and that the role of the maximum likelihood point-estimate is simply to give a partial description of the likelihood function. What is this other problem whose solution is the likelihood function? I would call it the problem of informative inference, and define it as the problem of reporting efficiently, in meaningful objective terms, the statistical evidence, provided by an observed experimental outcome, which is relevant to the statistical hypotheses (possible parameter values) under consideration. Although the term "statistical evidence" is not in common use in mathematical statistics, I believe that it should be, because it represents accurately an essential feature of many important applications of statistical techniques. What is the nature of statistical evidence, and what are its objective qualitative and quantitative properties? As a familiar example, when an outcome of a scientific experiment indicates rejection of one statistical hypothesis in favour of another, on the basis of a test having very small probabilities of both types of errors, what seems most relevant and useful for typical purposes is the character of the outcome as strong evidence against the first hypothesis. It is a familiar fact that results of statistical tests are customarily interpreted in this way; one may wonder how often any standard statistical techniques would be used in scientific research if they did not admit such interpretations, which we may call evidential interpretations, of outcomes. The objective basis for interpreting the test outcome "reject" as strong evidence against a hypothesis is the small magnitude of its error-probabilities. For the familiar purpose of evidential interpretation of one given outcome of a test, it is enough that the latter probabilities admit an objective frequency interpretation in the conceptual sense. Certain relative frequencies which correspond to error-probabilities could in principle be realized physically, but will not be so realized in connection with the given experimental investigation; although this objective interpretation of these probabilities is purely conceptual, it suffices to support the interpretation of a single given outcome as statistical evidence.

The full analysis of the nature and structure of statistical evidence, in such objective probabilistic terms turns out to be a well-defined mathematical problem, as illustrated in the second part of my contribution here last week. Such analysis may be said to constitute a mathematical theory of informative statistical inference, and its subject-matter is quite distinct from intensities of belief or subjective probabilities. Such analysis leads to a certain central position for the likelihood function; and this analysis unfolds systematically, in objective probabilistic terms, the evidential significance inherent in the likelihood function itself. Such analysis gives support to the claim that informative inference should in principle be based just on the likelihood function itself, and in my opinion eliminates the need for some of the other kinds of justification of this claim, mentioned by Mr. Barnard, which seem somewhat less direct; on the other hand, it seems essential to develop the general theory of informative inference, including complete explicit probabilistic interpretations of the statistical evidence provided by experiments of various mathematical forms. The simplest type of such interpretations is illustrated by the following example: when two simple hypotheses are considered, an outcome which gives the likelihood ratio statistic the value 99, regardless of the structure of the experiment in which it was obtained, has the same qualitative and quantitative properties, as evidence, as the outcome "reject" obtained by a statistical test having probabilities of errors of both kinds equal to .01. Such examples and interpretations illustrate the nature of the objective probabilistic bridge which can be constructed to connect systematically the two standpoints presented here by Mr. Neyman and Mr. Barnard.

I believe that such analysis clarifies certain essential unities and certain essential differences between the two standpoints mentioned, and that it can throw further light on the possibilities and possible limitations of programmes, such as that of Mr. Rao, which aim to go as far as possible in developing a single type of inference method which will prove satisfactory from both of those standpoints.

Mr. Rao : It is my first duty to express thanks to all those who contributed to the discussion. I have intentionally made some provocative statements in my paper to invite criticism necessary for a proper understanding of the issues involved. I think my plan has borne fruit. I would like to consider the various points raised in the discussion under a number of headings. The first one is historical.

1. Historical aspects

I must admit I have not touched on the historical aspects of the m.l. method adequately, as that would be outside the scope of the subject assigned to me. But since Mr. Neyman raised some historical issues in his discussion I have to answer them.

Mr. Neyman states, "as far as I am aware, the priority in the approach of m.l.e..... belongs to F. Y. Edgeworth (1908a)"[10], a statement which Edgeworth himself would have contradicted as he attributed the method to Gauss, Laplace, and Pearson (footnotes on pages 384 and 395 of Edgeworth, 1908a). It also appears from Edgeworth's understanding of the earlier writers that the justification of m.l.e. consists in the inverse probability argument. Edgeworth (1908b, p. 500)[11] himself supported this view and was also aware of the contradictions involved in assigning the same *a priori* probability distributions to different functions of parameters, but contended that the matter was not serious in large samples and for functions not out of the ordinary (p. 392, Edgeworth, 1908a). It is, indeed, surprising that Mr. Neyman, paraphrasing Edgeworth's work, asserts that the estimate obtained by the method of inverse probability (i.e., by maximising the *a posteriori* distribution) is in fact the m.l. estimate. If $\hat{\theta}$ is an m.l. estimate of θ, then $\phi(\hat{\theta})$ is an m.l. estimate of any one-to-one function $\phi(\theta)$, while such a property is not true of estimates obtained by the inverse probability argument.

As for Laplace's work, it is clear from the interpretation by Todhunter (1865, p. 578, 585)[12] that Laplace never stressed the choice of "most probable result" nor did he justify its use in preference to any other method. I had no access to contributions by Gauss on this subject, but I take the liberty of quoting Mr. Barnard who thought that Gauss's justification of maximising the probability for estimation of parameters is not free from inverse probability.

[10] Edgeworth, F. Y. (1908a) : On the probable errors of frequency constants. *JRSS*, LXXI, 381.

[11] Edgeworth, F. Y. (1908b) : On the probable errors of frequency constants. *JRSS*, LXXI, 499.

[12] Todhunter, I. (1865) : *A History of the Mathematical Theory of Probability*. Chelsea Publishing Company, New York. (1949 edition).

Karl Pearson (1896)[13] used the m.l. method in estimating the correlation coefficient without offering justification (p. 125), but his subsequent work with Filon (1898)[14] on standard errors of 'frequency constants' (estimates of parameters) did not show that he was considering m.l. estimates. It appears that they were attempting to compute the variance-covariance matrix of the asymptotic *a posteriori* distribution of the parameters presumably derived from a uniform *a priori* distribution and a sufficiently large sample. The analysis was not, however, rigorous.[15] We may thus infer that the authors were attempting to derive the asymptotic standard deviation of estimates which are the mean values of the *a posteriori* distribution, or those obtained by maximising the *a posteriori* probability density of the parameters. It is, however, somewhat puzzling to note that Pearson used the same expressions to determine the asymptotic standard errors of estimates obtained by the method of moments as well. It is also well known that Pearson did not advocate the use of m.l. in his subsequent writings.

A reference has also been made by Neyman to the result of Edgeworth, proved with the help of Love, that the most probable value (as determined by inverse probability) has the "smallest mean square deviation from the true point." This, indeed, is a remarkable attempt although the class of alternative estimates was very much restricted and the estimation was confined to location and scale parameters. A different argument is necessary to establish this result for the estimate of any general parameter and under less restrictive conditions on the class of estimates. The result, however, is not true, as observed by Hodges and mentioned by Neyman in the present discussion, without any restriction on the class of estimates.

We, therefore, do not have any literature supporting prior claims to the method of m.l.e., as a principle capable of wide application and justifying its use on reasonable criteria (such as efficiency in a sense wider than that used by Edgeworth and consistency) and not on inverse probability argument, before the fundamental contributions by Fisher in 1922 and 1925.

2. Philosophical standpoint

Mr. Neyman wants me to explain my philosophical standpoint on estimation. "Should m.l.e. always be used irrespective of the properties it may have ?" If I understand correctly the spirit of this question and the emphasis on point estimation by Neyman (in the case of contamination of water) and by Berkson (in determining the concentration of a solution), I must differ from their philosophy quite sharply. I think for the estimation of contamination of water, instead of giving a point estimate, *sufficiently overestimated and considered safe* (in some sense), a statistician should ideally provide the customer with a whole series of inferences about the unknown value and the associated risks or consequences. For instance, in large samples under fairly general conditions, an estimate such as that obtained by m.l. together with its standard error estimable from the data themselves provides the complete answer. In small samples, mechanisms exist, under favourable circumstances, for providing fiducial probability statements or a whole series of fiducial limits in the sense of Fisher or confidence limits (interval, upper and lower) in the sense of Neyman.

I do not see how considerations of bias, under or over estimation arise. Again in the example of estimation of variance of an instrument, Neyman suggests that I should preferably give an unbiased estimate, if I have to escape "the lawsuit for damages by the manufacturer of the instrument." Assuming that a lawsuit is filed whenever there is an error in the estimate, an unbiased estimate can only give a mental consolation that errors made, however large they are, even out in the long run, although heavy damages may have to be paid every time ! It must be noted that if one adopts the "minimum mean square error" criterion for the choice of an estimate, the unbiased estimate may not even be admissible in the sense of decision theory. If the damage to be paid is proportional to the square of the error, I should not give an unbiased estimate.

[13] Pearson, K. (1896) : Mathematical contributions to the theory of evolution IV. Regression, Heredity and Panmixia. *Phil. Trans. Roy. Soc.* London Series A, 187, 253.

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

[15] This problem is now under investigation and it is hoped to publish some of the results elsewhere.

I am aware that in some methodological problems such as obtaining a pooled estimate by averaging parallel estimates, one need to consider unbiased or nearly unbiased estimates. This can be achieved, in many cases, by a suitable adjustment of an available estimate.

What I maintain is that m.l.e. provides a convenient summary of data, demonstrably better than other methods in large samples, for answering questions of interest concerning an unknown parameter, and not that a point estimate obtained by maximising the likelihood is *the answer* to any specified question. Neyman and Berkson were repeatedly asking me during the discussion whether I would suggest the m.l. estimate in all situations. I do not know how the misunderstanding has arisen.

I am glad to note that Neyman looks upon "super efficient" estimates only as examples to show that the definition of efficiency as the attainment of minimum asymptotic variance is void without some restriction on the estimate. But I do not see why, *in large samples*, Hodges-LeCam "super efficient" estimates or the "super efficient" estimate given in the present paper for the mean of a normal population should not be preferred to $\bar{x}$, the sample mean (from behavioristic viewpoint). On the basis of decision theory, there is, perhaps, justification in doing so, or at least $\bar{x}$ has no definite claims over the other. My objection is, however, for other reasons. The super efficient estimate of the present paper is a function of the median and the mean of a sample of observations and is, therefore, less useful than $\bar{x}$, for purposes of statistical inference. The "super efficient" estimate of Hodges-LeCam is, however, equivalent to m.l.e. in large samples, i.e., efficient in the sense defined in the present paper and one may expect no substantial difference in the inferences associated with the two estimates in large samples. But it has certain defects. For instance, its asymptotic standard deviation being a discontinuous function of the unknown parameter does not admit reasonable estimation. Consequently, the inversion of a "super efficient" estimate for inference on the unknown parameter becomes a little complicated.

3. Minimum mean square error

It was not my intention to be unfair to Berkson in pointing out certain defects in the criterion of minimum mean square error. The example due to Silverstone of estimating the probability of success by the constant 1/2 may be of a special nature. But we have a number of examples to illustrate that smaller variance does not necessarily mean higher concentration round the true value. It does not also imply that an estimate with a smaller variance provides a better discrimination between alternative values of the parameters. Recently, at my suggestion, Sethuraman (1960)[16] examined the relative powers of two statistics ξ and $2\xi+\eta$, (where ξ and η are the maximum and minimum respectively in a sample of size n from a rectangular population in the range $(\theta, 2\theta)$), for testing the hypothesis that $\theta = \theta_0$. Although as an estimate of θ, the m.l. estimate $\xi/2$ has uniformly larger variance than $(2\xi+\eta)/5$, an alternative estimate, it has better power as a test criterion for values of θ close to the assigned one. Since estimates with minimum mean square may not have *other* desirable properties, I was, naturally inclined to ask Berkson about the significance of, or the motivation for the choice of this criterion.

Berkson observes that if he follows my philosophy on theory of estimation, it will be disastrous in routine practice as in the use of a bio-assay for medical purposes, because one has to wait indefinitely collecting more and more observations before a decision can be reached. I have not said that decisions should not and cannot be made on the basis of available data, however meagre they are. But I am not convinced that an estimate which has minimum mean square error will be of help in minimising the mortality among his patients. I will only be too glad to accept Berkson's procedure if the latter were to be true. I am sure that for a statistical procedure to be made available for routine practice the approach should be somewhat different. Past data, as they accumulate, must be effectively used to improve the existing procedure. The theory of estimation as developed by Fisher is most suitable for such situations. It is not claimed anywhere that the m.l. estimates are minimal sufficient statistics, although they are explicit functions of the latter. It may be seen that in the problem of fitting a logistic function (specified by two parameters

[16] Sethuraman, J. (1960): Conflicting criteria of "Goodness" of statistics. *Sankhyā*, series A (in press).

α, β) to bio-assay data, the m.l. method does provide minimal sufficient statistics in samples where the m.l. estimates are properly defined. If a suitable convention of specifying an estimate, like the one suggested by Silverstone, is adopted in the case of samples for which the m.l. estimates of one or both of α and β are not finite, sufficiency of m.l. estimates can be claimed for all samples.[17] The situation is not so simple in the case of minimum logit χ^2, advocated by Berkson. Generally, the estimates obtained by this method are not sufficient. But Berkson insists on quoting one example with 3 doses and 10 animals at each dose, in which case, with the application of a dubious rule such as $2n$-th, the minimum logit χ^2 estimate happens to be sufficient.

Berkson gives no other argument in favour of mean square error, except that "it is a representative loss function, and if, in a particular application, some other loss function, suggests itself, let it be investigated: suppose $\hat{\alpha}$ and α^* are two alternative estimates of a parameter α such that,

$$E(\hat{\alpha}-\alpha)^2 > E(\alpha^*-\alpha)^2$$

and there exists a function ϕ such that

$$E_n[\phi(\hat{\alpha})-\phi(\alpha)]^2 < E\,[\phi(\alpha^*)-\phi(\alpha)]^2$$

then the estimate of α obtained by using one loss function is not good with respect to another loss function. Such situations are not rare and any number of examples with a reasonable choice of the function ϕ can be given. In the problem of fitting the logistic function, I venture to suggest that some increasing function of the differences between hypothetical and estimated probabilities of success or failure at each dose, may be a better indicator of the goodness of estimation than the deviations in the estimates of parameters α and β themselves. I do not know whether a minimum logit χ^2 estimate would have smaller expected loss than other types of estimates when loss functions of the type indicated are considered.

4. Other aspects

The views expressed by Barnard on point estimation, the role it plays in specifying the likelihood, and its relation to a practical decision problem do not seem to be in conflict with those in my paper. Both of us have tried to interpret Fisher's work on estimation, though not completely and not in exactly the same way. I hope they will serve to remove some wrong notions about m.l. found in recent literature.

I am particularly interested in Birnbaum's contribution to the theory of estimation as it provides a small sample justification to certain estimation procedures including the m.l. This is a far more difficult task than what I have attempted to do confining my remarks mainly to the case of large samples. I cannot think of situations where serious decisions are taken on meagre evidence supplied by small samples, while in routine practice such as the application of control charts in industry one may think of specifying rules of action based even on very small samples to minimise certain risks in the long run. Further discussion on the theory of estimation in small samples as attempted by Birnbaum would, no doubt, be of great value.

I would also wish to take the opportunity of mentioning a few results in connection with the investigation mentioned in the last paragraph of my paper. It was thought that no distinction could be made in large samples among estimation procedures such as m.l., minimum chi-square, modified minimum chi-square, etc. since they all provide asymptotically efficient estimates in a wider sense of $i_T \to i$ (i_T and i are informations, *per observation* contained in the statistic T and the sample respectively). But as mentioned by Fisher in the 1925 paper, differences in the *actual* amounts of information contained in different estimates are more relevant. It has been possible to compute a quantity, analogous to, if not same as, the limiting difference in the total information contained in the statistic and in the sample and establish that the m.l. method has the least limiting loss. The minimum chi-square, modified minimum chi-square, and other related methods have a greater loss. The actual values are given by the author in a paper under print in the Proceedings of the 4th Berkeley Symposium on Statistics and Probability.

[17] The emphasis should be not on estimating the parameters α and β but on probabilities of death at various doses. The parameter space has then to be properly defined in terms of these probabilities. Once this is done, many difficulties mentioned by Dr. Berkson would disappear.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Dr. D. Basu

Shri S. John

Dr. R. G. Laha

Dr. P. K. Menon

Dr. C. R. Rao

Shri T. V, Hanumantha Rao

Dr. J. Roy

Dr. D. K. Roychowdhury

Editor,
Sankhyā : The Indian Journal of Statistics.