The Annals of Statistics
1980, Vol. 8, No. 3, 457–487

MINIMUM CHI-SQUARE, NOT MAXIMUM LIKELIHOOD!

BY JOSEPH BERKSON

Mayo Clinic, Rochester, Minnesota

The sovereignty of MLE is questioned. Minimum χ^2_λ yields the same estimating equations as MLE. For many cases, as illustrated in presented examples, and further algorithmic exploration in progress may show that for all cases, minimum χ^2_λ estimates are available. In this sense minimum χ^2 is the basic principle of estimation. The criterion of asymptotic sufficiency which has been called "second order efficiency" is rejected as a criterion of goodness of estimate as against some loss function such as the mean squared error. The relation between MLE and sufficiency is not assured, as illustrated in an example in which MLE yields ∞ as estimate with samples that have different values of the sufficient statistic. Other examples are cited in which minimal sufficient statistics exist but where the MLE is not sufficient. The view is advanced that statistics is a science, not mathematics or philosophy (inference) and as such requires that any claimed attributes of the MLE must be testable by a Monte Carlo experiment.

Some times some reflections which are presented in a tone which is simple and modest enclose the truth and are as sure a guide to it as an accumulation of formulas which are in part a trompe-l'oeil–Émil Borel.

Some months ago I submitted a communication for publication in this journal that took issue with part of a paper by Efron (1975). After deliberate consideration the editor advised me to elaborate my views somewhat, and to comment on some related other papers which he designated, and to provide illustrative numerical material. This present notation is in response to that invitation.

In an unpublished paper, "An extended view of chi-square testing and estimation," Berkson (1976), I presented a view that the basic principle of estimation is minimum chi-square, not maximum likelihood. A chi-square function is defined as any function of the observed frequencies and their expectations (or estimates of their expectations) that is asymptotically distributed in the tabular chi-square distribution. Some of these when minimized yield RBAN estimates. Five specific chi-square functions which have been used were presented; Pearson $\chi^2_P = \Sigma((o - e)^2/e)$, Neyman's reduced $\chi^2_1 = \Sigma((o - e)^2/o)$, likelihood $\chi^2_\lambda = 2\Sigma o \ln(o/e)$, Kullback's discrimination information $\chi^2_I = 2 \times e \ln(e/o)$, logit $\chi^2_l = \Sigma npq(\text{lgt } p - \text{lgt } P)^2$, where o is an observed frequency out of n trials, e is an estimate of the expectation of the corresponding frequency, $p = 1 - q = o/n$, $P = 1 - Q = e/n$, $\text{lgt } p = \ln(p/q)$, $\text{lgt } P = \ln(P/Q)$.

Received July 1977; revised November 1978.

AMS 1970 *subject classifications.* Primary 62F10; secondary 62F20.

Key words and phrases. Estimation, criteria of estimate, maximum likelihood, minimum chi-square, efficiency, second order efficiency.

An outline of this view with the observational basis of its discernment was presented in Berkson (1972).

Minimizing χ^2_λ yields the same estimate as the MLE. Minimizing any of the other chi-squares yields asymptotically equivalent estimates, in the sense that they are asymptotically consistent, normal, with the same minimum variance. Inasmuch as minimizing χ^2_λ yields the MLE, it is argued that since the MLE can be derived as a minimum chi-square estimate, minimum chi-square is the primary principle of estimation. It is to be noted that Cramér, (1946), page 426, derived the MLE as a large sample approximation of the minimum χ^2_p estimate and refers to the derived MLE as obtained by the "modified χ^2 minimum method."

Statistics is an applied science and deals with finite samples. Asymptotic theorems refer to limits which, by definition, are never attained. There is no mathematical demonstration that there is any method of estimation that assures that the estimate attained is best in an operational sense in all circumstances, particularly not the method of maximum likelihood which was abandoned by Gauss. In this writer's experience there are situations in which sometimes the minimum χ^2_λ (maximum likelihood), sometimes one of the other minimum χ^2 estimates is better. The position here advanced is that every problem should be studied and if it is not known which is best, the minimum chi-square estimate which is simplest to compute for that case should be used.

In the course of time I have published opinions questioning the sovereignty of maximum likelihood estimation. In commenting on one of these Rao, (1961a), page 440, said: "There has been a tendency to consider estimation as part of decision theory, which requires as a datum of the problem the specification of the loss for a given difference between estimate and the true value of the unknown parameter . . . . This may be appropriate in certain situations but I am not sure whether one can support Berkson (ref.) when he wants to estimate the · · · regression line in a bio-assay using the criterion of 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."

Professor Rao's expressed opinion, reflecting dissociation from a view of statistics that may have operationally meaningful import, i.e., decision theory, as well as his dismissal of the mean squared error as an appropriate criterion of goodness of estimate, contrasts singularly with the view expressed by Gauss as reported by Plackett (1972). Gauss, Plackett noted, "anticipated ideas of decision theory". Gauss himself in a quoted letter to Bessel said: " . . . I must consider it less important in every way to determine the value of an unknown parameter for which the probability is largest, although still infinitely small, rather than that value, by relying on which one is playing the least disadvantageous game . . . ". The principle for estimation emphatically rejected by Gauss—"less important in every way"—is that of maximizing a probability (or likelihood); he favored instead minimizing a loss function, namely the expected squared error.

With all due regard for latter day developments in statistics, it is not necessarily a lack of insight that may prompt one to prefer the guidance of Gauss in contrast with sponsors of subjective theories of "inference" in vogue at the moment in some current statistical literature.

Mention of Gauss prompts the presentation of an example which contrasts maximum likelihood with least squares and points up some paradoxical consequences involved with application of maximum likelihood.

We consider a linear functional relation

$$(\mu|x_i) = \mu_i = \alpha + \beta x_i \tag{1}$$

where μ_i is the value of μ at x_i and α, β are parameters whose values are unknown. The μ_i are measured with error ϵ_i as Y_i, ϵ_i being distributed $N(0, \sigma_i)$.

$$Y_i = \alpha + \beta x_i + \epsilon_i. \tag{2}$$

We have an observation y_i at each x_i ($i = 1, 2, \cdots, n$) and wish to estimate μ_i, which means to estimate α, β.

If σ_i is the same at all i, independent of μ_i, it is well known that the ML and least squares estimate of α, β, are identical. However, if σ_i is not constant but is proportional to μ_i with $\sigma_i = C\mu_i$ where C, the coefficient of variation, is the same at all x_i, as is often reasonable, then the least squares and ML estimates are not the same.

If all the observations y_i fall on (1), then the least squares estimates, which are determined by minimizing $\Sigma(y_i - Y_i)^2/\sigma_i^2$ will yield the true values $\tilde{\alpha} = \alpha$, $\tilde{\beta} = \beta$. The ML estimates will in these circumstances yield

$$\hat{\alpha} = \alpha F$$
$$\hat{\beta} = \beta F$$

where

$$F = \frac{(1 + 4C^2)^{1/2} - 1}{2C^2}.$$

As an example

$$\alpha = 1, \quad \beta = 1, \quad C = 0.2$$
$$F = 0.962912.$$

We suppose observations at $x = 0, 2, 4, 6, 8$.

x	y	$\tilde{Y}_i$	$\ln \tilde{L}_i$	$\hat{Y}_i$	$\ln \hat{L}_i$
0	1	1	0.6905	0.9629	0.7283
2	3	3	−0.4081	2.8887	−0.3703
4	5	5	−0.9189	4.8146	−0.8812
6	7	7	−1.2554	6.7404	−1.2176
8	9	9	−1.5067	8.6662	−1.4689
	Total		−3.3986		−3.2097

Likelihood with $\tilde{Y} = 3.342 \times 10^{-2}$. Likelihood with $\hat{Y} = 4.037 \times 10^{-2}$.

Thus it is seen that the ML estimates, though they do not recover the true μ's do have larger likelihood.

The question suggests itself as to whether this example reflects an inconsistency of the ML estimate. Many years ago in conversation with the late "Jimmie" Savage, discussing consistency, I mentioned examples of some empirical estimates used in bioassay that had the awkward characteristic that they did not recover the true values of the pertinent parameters even when the observation followed the assumed function exactly. He commented, "That's not inconsistent, that's wrong." Barnard, in discussion of a paper by Rao, (1962), page 67, referred to the idea of consistency used by Gauss which requires that if the observations are free from error the estimate should give the true value of the parameter. Fisher (1922), (1925), (1938), (1956), to whom the statistical term is due, gave several definitions of "consistency". In the last of these Fisher, (1956), page 144 , says: "A CONSISTENT STATISTIC may then be defined as: a function of the observed frequencies which takes the exact parametric value when for these frequencies their expectations are substituted." By this definition the MLE in the present example is not consistent.

Fisher, (1925), page 714, gives a definition of efficiency which may be summarized as follows: if x represents an observation on a variable X whose probability density is $\phi(X) = f(x, \theta)$ which depends on a parameter θ, then the efficiency of a statistic T_n that is based on n observations is the ratio of the information per unit observation contained in T_n to the information contained in x.

$$\text{Efficiency} = \frac{1}{n}\frac{I_{T_n}}{I_x} \tag{3}$$

where

$$\text{Information in } T_n = I_{T_n} = E\left(\frac{\partial \ln \phi(T_n)}{\partial \theta}\right)^2$$

$$\text{Information in } x = I_x = E\left(\frac{\partial \ln \phi(x)}{\partial \theta}\right)^2.$$

A statistic T is sufficient if, and only if, the value of (3) is unity. It cannot be greater but if it is less than unity, (3) measures proportionately the departure from sufficiency—"the loss of information." This definition, says Fisher, " . . . *has the advantage of applying to finite samples and to other cases where the distribution is not normal.*" (italics added).

Rao (1961a, 1961b, 1962) has used various versions of (3) in an effort to show the superiority of the MLE. But he has considered it in its asymptotic aspects, not as it applies to finite samples. That is, he has identified efficiency of an estimator with asymptotic sufficiency of a statistic and called it "second order efficiency". He believes, or wishes us to believe, that an estimator is not measured by some index of its closeness to the parameter estimated, such as the mean squared error, the minimax criterion, or Pitman's index of proportion of closer cases, but by some

index of Fisher's information. This view has not been embraced by mathematical statisticians generally.

Bartlett, in discussion of the paper by Rao, (1962), page 64, remarked, "Professor Rao has . . . discussed properties of large-sample estimates and maximum likelihood ones in particular . . . . Now to do this comprehensively does . . . require correct knowledge of their sampling properties as well as how asymptotically 'sufficient' they are."

In the course of discussion on a paper by Stein (1962) which takes the mean squared error as the central criterion in an estimation problem, and which has been hailed as the most important statistical paper of the decade, Stein remarked, page 295, "I feel that the aim in a statistical analysis is to get as close to the true value, or make as nearly correct a decision as can reasonably be expected, and I reject any principle that conflicts with this."

Pfanzagl (1973), page 1006, said forthrightly, " . . . any definition of second order efficiency should be based on covering probabilities . . . . An unmotivated concept of second order efficiency like that of C. R. Rao (1962) . . . should be abandoned."

Aside from the questionable character of Rao's "second order efficiency" as a criterion of goodness of estimation, there appears to be some question as to the technical validity of his claim for the necessary superiority of the MLE. Daniels, in discussion on the paper of Rao (1962), page 65, challenged the superiority of the MLE, in respect of second order efficiency, over estimation based on order statistics. Also Rao's putative proof, page 50, following Table 1, of the inferiority of minimum chi-square estimates vis-a-vis MLE cannot be correct, since, for the elementary binomial or multinomial parameter, the minimum chi-square estimates are identical with the MLE. Then again, the second order efficiency cannot be the critical criterion of estimation, as may be exemplified with the bioassay experiment discussed by Berkson (1955). There are three "dosages" x, 10 animals exposed at each, among which the number of deaths is observed. There are 1331 possible sets of results. Suppose they are coded T in order of the deaths at the three doses. For 0, 0, 0, $T = 1$; 0, 0, 1, $T = 2$; $\cdots$; 10, 10, 10, $T = 1331$. Then T is in one-to-one correspondence with the observations, therefore sufficient, and the second order efficiency is unity. But T could not be used as an estimator.

Ghosh and Subramanyam (1974) in a detailed and mathematically rigorous article deal with relevant questions respecting the MLE. They give serious consideration specifically to the papers Berkson (1955) and Berkson and Hodges (1961). I deeply appreciate the attention they pay to my work. It is unusual. This is the first time, to my knowledge, that these papers have been referred to in any statistical journal. Similarly I am gratified with their remark: "Our second order expansions seem to agree quite well with the Monte Carlo values in a few examples of Berkson that we studied." I may point out in this connection that I obtained similar results in respect of the smaller mean squared error of the minimum normit chi-square

estimate in comparison with the MLE (Berkson, 1957b) and a similar comparison between the MLE and minimum logit chi-square estimation has been reported by Little (1974). Corroboratory evidence is to be found also in Wetherill (1963), page 22. Recently Amemiya (1978) has investigated the mean squared error of the minimum logit chi-square estimator and the MLE to the order of n^{-2} in many examples and found that in an overwhelming majority of cases the minimum chi-square estimator has the smaller mean squared error.

Ghosh and Subramanyam defended Rao's second order efficiency but acknowledged, page 326, " . . . its decision theoretic implications are far from clear." They go on to develop extensions and draw conclusions in favor of the MLE. My reaction will be presented by some quotations from their article and comments upon them.

They say, page 327: "It is shown that if a correction is made to the maximum likelihood estimator so that the bias is the same as Berkson's minimum logit chi-square estimator up to terms $O(1/n)$, then the maximum likelihood estimator has lower variance up to terms of $O(1/n^2)$."

I recognize the mathematical interest in a finding, if it can be validly established, that the MLE can be modified in a way that other minimum chi-square estimators cannot be, and which results in the MLE being better. But I hold that statistics is an applied science, Berkson (1977), not mathematics or philosophy. A statistical proposition must, in principle, be testable by a Monte Carlo experiment. Let us, therefore, examine the matter from a perspective of application. In the first place, it is unclear as to whether the correction is or is not a function of the parameter θ to be estimated, and hence there is a question whether it could be applied at all. Be this as it may, if it is to be applied it must be calculable. For instance there are given in column 1 of Table 1 the MLE (minimum χ^2_λ) estimates for the data shown. I wish to compare the Ghosh-Subramanyam modified MLE with them. This is not possible with the information presently available. It would be helpful if the authors issued an expository article explaining how the modified MLE is to be computed, desirably with a numerical example as for the present data. It would then be possible to carry out a Monte-Carlo investigation with the new modified MLE. We might then discover hitherto unforeseen characteristics, as we did when we found that the MLE itself yielded infinite estimates.

My reported experiment which Ghosh and Subramanyam commented on is a bioassay experiment in which the probability of death at dose x is given by the logistic function with which the scale parameter is β and the location parameter is α. The experiments dealt with the case in which β known α to be estimated, α and β both to be estimated. In each case the minimum logit chi-square estimate was found superior to the MLE by the criterion of mean squared error.

Ghosh and Subramanyam consider the case with β known α to be estimated. There is a minimal sufficient statistic for α, namely the total number of observed deaths, Berkson (1955), page 142, Cox (1959), page 238, symbolized by them Σp_i^n.

They symbolize the minimum logit chi-square estimate as T_n^*, the maximum likelihood estimate as $\hat{\alpha}$, the Rao-Blackwellized estimator as T_n'. They say, page 351: "Here one has a complete sufficient statistic, namely Σp_i^n but T_n^* is not a function of it. If one considers the so-called Rao-Blackwellized $T_n' = E(T_n^*|\Sigma p_i^n)$ then it is indistinguishable from $\hat{\alpha}$ up to $O_E(1/n)$." These general remarks are in tenuous relation with immediate reality. The MLE $\hat{\alpha}$ is a function of the sufficient statistic, but it is not a one-to-one function. There are groups of samples (called "sufficiency groups," each sample in the group having the same value of the sufficient statistic) corresponding to different values of the sufficient statistic Σp_i^n, but with the same value of $\hat{\alpha} = \infty$, and $\hat{\alpha}$ is, therefore, not sufficient. On the other hand, the minimum logit chi-square estimate, T_n^*, is also not a one-to-one function of the sufficient statistic, but survey of the sufficiency groups disclosed that while there were some groups with more than one value of T_n^*, there was no instance of the same value of T_n^* corresponding to different values of the sufficient statistic Σp_i^n. Therefore, T_n^* is a function of the sufficient statistic and it is sufficient, though not minimal.

The comparison between the estimates is briefly this: the minimum logit chi-square estimate T_n^* has lower mean squared error than the MLE $\hat{\alpha}$, indeed, even lower than the computed value of the Cramér-Rao lower bound of $\hat{\alpha}$. The Rao-Blackwellized minimum logit chi-square estimate T_n' has the same bias as T_n^* but smaller variance and it attains its lower bound, which distinguishes it from T_n^* and from $\hat{\alpha}$, and the variance *is lower than the lower bound of an unbiased estimator*, which was widely thought to be impossible. The contrast is attributable to the fact that the bias function of the minimum logit chi-square estimate is negative, while that of the MLE is positive, as explained in the article referred to. The example also serves to refute the generalization enunciated by Fisher and others to the effect that if an estimator exists which is sufficient, or if it attains the lower bound, the MLE will be such. Fisher (1937–1938), page 151, said, "For example . . . when estimation without loss of information is possible, maximizing the likelihood will always furnish such an estimate." *The Rao-Blackwellized minimum logit chi-square estimate is sufficient and attains the lower bound for variance and mean squared error* (*which are less than* $1/(I)$), *but it is not the* MLE, i.e., *the estimate which maximizes the likelihood*, Berkson (1955), page 143, Table 7.

The authors mentioned that Silverstone and also Rao, in indicated papers, have defended the use of the maximum likelihood estimator from certain other points of view. I must note that as regards Silverstone, a reply was published by Berkson (1960) in which some of the counter-considerations outlined here were presented. Similarly as regards Rao, I commented at the meeting at which his paper was presented, and the discussion was published, Berkson (1961).

I now turn to the paper by Efron (1975) which is projected on the premise of the sovereignty of maximum likelihood estimation. In the course of the discussion following the presentation, Lucien Le Cam referred to a paper of mine, Berkson

(1951), which presented contravening evidence. In reply Efron said, "A function of the MLE may be better than the MLE itself for any specific estimation problem. This is the case in the Berkson example quoted. Berkson finds a 'better' estimator than the MLE which eventually is improved by Rao-Blackwellizing it on the sufficient statistic. This gives a function of the MLE!" Efron did not cite any reference to support his general assertion, which is surprizing. His remark about Berkson's finding does not refer to the article mentioned, but presumably it refers to Berkson (1955), the substance of which has been briefly discussed above. The Rao-Blackwellized estimate is the expected value of the estimate conditioned on the sufficient statistic. It is possible to Rao-Blackwellize the minimum logit chi-square estimate because, as was mentioned earlier, this estimate is not always the same in each sufficiency group. The MLE on the other hand is the same in each sample of the sufficiency group, or it is ∞. Hence the Rao-Blackwellized MLE is the same as the MLE, or it is infinite. The Rao-Blackwellized estimate which was computed is therefore a modified minimum logit chi-square estimate, not a MLE.

Efron stated that "Le Cam's criticism of the MLE as a point estimator should not be confused with Fisher's preference for it as an information gatherer." As a matter of fact the MLE, in the experiment referred to, loses information, whereas the minimum logit chi-square estimator does not because the same value of the MLE, namely ∞, corresponds to different values of the sufficient statistic and the loss of information by a formula of Fisher is directly related to the variance of the sufficient statistic, conditioned on a particular value of the estimate. The information lost is small when the experimental dosages x are disposed symmetrically around x_{50}, but increases toward 100 percent as the dosages are asymmetrically located. With central dosage at x_{90}, in the experiment referred to, the loss with the MLE was estimated to be about 90 percent. In these circumstances, to say that the Rao-Blackwellized minimum logit chi-square estimate, which itself, of course, is sufficient is a function of the MLE is patently incorrect.

Pervading these arguments on behalf of maximum likelihood estimation are two assumptions with respect to sufficiency that are untenable. The first is the idea that a sufficient statistic contains all the information in the observations required for statistical inference. We find this strongly stated in the posthumous article by Savage (1976), page 453: "I know of no disagreement that when an experiment admits a given statistic as sufficient then observation of that statistic is tantamount for all purposes to observation of all the data of the experiment." The invalidity of this notion is exemplified in the investigation of the present writer into whether radioactive disintegration events follow the Poisson exponential function as they are said to, Berkson (1966) (1975). In this case there is only a single parameter and the observed mean is minimal sufficient. The mean was not actually sufficient for the investigation. A variety of chi-square tests were carried out, and the entire body of observations was necessary.

That the sufficient statistic is not enough for a problem of this kind is elementary and is usually expressed by saying that it is only when the model can be assumed to

be true, that a sufficient statistic is sufficient. De Finetti in the discussion, page 487, corrects his friend and disciple: " . . . *sufficiency* for a statistic, is a property which can hold (only) on the hypothesis of a given model . . . ". Pearson (1974), page 7, said: "Let me go a little more into detail about my . . . disagreement with Fisher's views. The discovery of the property of sufficiency was a brilliant one, yet it seemed to me that it could be dangerous . . . . One or more than one statistic will contain all the information to be extracted from the data if and only if the probability density law in the population has the assumed mathematical form."

But even assuming the model, and even in consideration of estimation alone, it is still an abuse of language to say that a sufficient statistic contains all the information in the data, unless the only permissible estimate is the MLE.

The other mistaken idea is that if a minimal sufficient statistic exists the MLE will be sufficient. We have noted that in the bioassay experiment referred to with the logistic function for which the parameters have sufficient statistics, the MLE is not sufficient. But this is not an isolated case. With the elementary exponential distribution itself, it was pointed out by Berkson and Elveback (1960), page 420, that while there are minimal sufficient statistics for the parameter, the MLE is not sufficient. Simplified, the case is as follows.

$$\text{Prob}\{T \geqq t_i\} = e^{-\beta t_i} \tag{4}$$

where T is time of death, β the parameter. N individuals are followed for various periods from some defined origin and d have been observed to die at times t_i $(i = 1, 2, \cdots, d)$ and $s = N - d$ last observed living at T_j $(j = 1, 2, \cdots, s)$. The density is

$$\phi = \beta^d e^{-\beta\Sigma(t_i + T_j)}. \tag{5}$$

From (5), d and $\Sigma(t_i + T_j)$ are jointly sufficient for β. The MLE of β is

$$\hat{\beta} = \frac{d}{\Sigma(t_i + T_j)}. \tag{6}$$

The MLE is not sufficient because there may be more than one sample with the same value of $\hat{\beta}$, but these may not have identical values of the sufficient statistics.

Indeed a similar situation exists, even in situations where it is frequently stated that the MLE is sufficient, as with the normal function $N(\mu, 1)$ where the density is

$$\phi = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\Sigma^n(x - \mu)^2}{2}\right) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\Sigma x^2 - 2\mu\Sigma x + n\mu^2}{2}\right). \tag{7}$$

From (7) Σx, n are jointly sufficient for μ. If n is constant in all samples the MLE estimate $\hat{\mu} = \Sigma x/n$ is sufficient. But in a sequential experiment n may vary and the mean is not in general sufficient.

For similar reasons the observed relative frequency p which is the MLE of the binomial parameter is not in general sufficient, and the question is raised whether the MLE is ever completely sufficient for all experiments.

How the viewpoint advanced here works in practice can be conveyed by presentation of some illustrative examples with the computed values of the various minimum chi-square estimates provided. Methods of fitting used to obtain the estimates are presented in Berkson (1949) (1957a) (1957b) (1972) and in Berkson and Nagnur (1974). For a bioassay experiment with the logistic model Table 1 shows the five minimum chi-square estimates, minimum χ^2_λ, (maximum likelihood) followed in the successive columns by the minimum χ^2_I, χ^2_I, χ^2_1, χ^2_p. To be noted is that the respective χ^2's are smallest for the corresponding estimates, which serves as a check on the theory and computations. The computations for all the estimators except the minimum χ^2_I require iterative procedures; the number of cycles needed are shown in parentheses at the tops of the columns. The minimum logit chi-square estimates here require no iterative procedures but instead are given directly in definitive form. It is therefore the estimate preferred.

Table 2 shows a contingency table testing the hypothesis of "no interaction." The data are those used by Bartlett (1935) in the article which launched the statistical problem of "interaction," but the model assumes linear additivity as defining "no interaction" instead of logit additivity which is the Fisher-Bartlett model. The five minimum chi-square estimates are shown together with the five χ^2's corresponding to each estimate. Again we note that the respective χ^2's are smallest for the corresponding estimates. All the estimates except the minimum χ^2_1 (Neyman) require iterative procedures. The minimum χ^2_1 estimate here is obtained explicitly without iteration and is therefore preferred for this case.

I will terminate this brief exposition by referring to an episode that bears on the questions considered. In the winter of 1967 I was invited to present a paper at a conference on the future of statistics to be held in Madison, Wisconsin, and noting

TABLE 1
Bioassay, logistic model.

Dose			Dead Minimum chi-square estimates				
x	Total	Observed	(5) χ^2_λ	(0) χ^2_I	(8) χ^2_I	(10) χ^2_1	(9) χ^2_p
0	10	1	1.901432	2.132060	1.692677	1.569061	2.165845
1	10	6	3.445099	3.650354	3.249082	3.111830	3.645872
2	10	3	5.405506	5.494770	5.320097	5.230441	5.435554
3	10	8	7.247963	7.212541	7.286402	7.269252	7.119426
Total	40	18	18.000000	18.489725	17.548258	17.180584	18.366697
		χ^2_λ	5.985436	6.040200	6.036494	6.128020	6.053979
		χ^2_I	5.564378	5.523352	5.684914	5.814846	5.528284
2 D.F.		χ^2_I	6.209156	6.389490	6.153299	6.177290	6.421971
		χ^2_1	6.731597	7.075613	6.567784	6.538177	7.128706
		χ^2_p	6.031680	5.968458	6.210858	6.399564	5.962296

TABLE 2
Data of Bartlett
Estimates, no interaction, linear model.

Time of planting	Type	Total	Observed	Dead				
				Minimim chi-square estimates				
				(4) χ^2_Λ	(0) χ^2_1	(4) χ^2_I	(7) χ^2_I	(4) χ^2_p
At once	Long	240	84	82.888682	82.882860	82.885142	82.890240	82.889880
	Short	240	133	134.212435	134.213297	134.213137	134.213064	134.212704
In spring	Long	240	156	157.112743	157.117138	157.114858	157.110216	157.110120
	Short	240	209	208.436496	208.447574	208.442843	208.433040	208.432944
	1 D.F.			.08195635	.08196044	.08195726	.08195735	.08195740
				.08185052	.08184496	.08184597	.08185471	.08185475
				.08190230	.08190150	.08190054	.08190490	.08190489
				.08200855	.08201615	.08201363	.08200630	.08201033
				.08201216	.08202095	.08201583	.08201159	.08201158

that L. J. Savage and G. Barnard were scheduled for the program, I suggested that they be invited to discuss my paper, and the suggestion was accepted. As it happened, I was unable to attend the meeting, but unknown to me, Dr. Marvin Kastenbaum read my submitted paper, and both Savage and Barnard discussed it. The following are some excerpts from their remarks, Berkson (1968), page 197, page 200. Said Savage; "It was hotly contested whether, in this particular case, Berkson's estimate was better, but there is no categorical reason why the maximum likelihood estimate should be absolutely the best. Indeed, it is hardly to be expected . . . . The truly amazing thing about this dispute is that maximum likelihood should ever have been so entrenched in some minds as to make Berkson's contention surprizing . . . ". Barnard said: "I agree that Berkson's early work was badly treated and it is, I think, quite a useful idea to have something even in computing days, which one can work out on paper. It's not the only method like this."

REFERENCES

[1] AMEMIYA, TAKESHI (1978). The n^{-2}-order mean squared errors of the maximum likelihood and minimum logit chi-square estimator. Publication of the Institute for Mathematical Studies in the Social Sciences, Stanford Univ.

[2] BARTLETT, M. S. (1935). Contingency table interactions. *J. Roy. Statist. Soc. Suppl.* **2** 248–252.

[3] BERKSON, J. (1949). Minimum χ^2 and maximum likelihood solution in terms of a linear transform, with particular reference to bio-assay. *J. Amer. Statist. Assoc.* **44** 273–278.

[4] BERKSON, J. (1951). Relative precision of minimum chi-square and maximum likelihood estimates of regression coefficients. *Proc. Second Berkeley Symp. Math. Statist. Probability* 471–479. Univ. California Press.

[5] BERKSON, J. (1955). Maximum likelihood and minimum χ^2 estimates of the logistic function. *J. Amer. Statist. Assoc.* **50** 130–162.

[6] BERKSON, J. (1957a). Tables for the maximum likelihood estimate of the logistic function. *Biometrics* **13** 28–34.

[7] BERKSON, J. (1957b). Tables for the use in estimating the normal distribution function by normit analysis. Part I. Description and use of tables. Part II. Comparison between minimum normit χ^2 estimate and the maximum likelihood estimate. *Biometrika* **44** 411–420.

[8] BERKSON, J. (1960). Problems recently discussed regarding estimating the logistic curve. *Bull. Internat. Statist. Inst.* 37 Part 3, 207–211.

[9] BERKSON, J. (1961). The purpose of estimation. Some remarks on Mr. C. R. Rao's paper. *Bull. Internat. Statist. Inst.* **38** Part 1, 200–205.

[10] BERKSON, J. (1966). Examination of randomness of α-particle emissions. In *Festschrift for J. Neyman Research Papers in Statistics*. (F. N. David, ed.) 37–53. Wiley, London, New York, Sydney.

[11] BERKSON, J. (1968). Application of minimum logit χ^2 estimate to a problem of Grizzle with a notation on the problem of "no interaction," with discussion. In *The Future of Statistics*. 175–200. Academic Press, New York and London.

[12] BERKSON, J. (1972). Miminum discrimination information, the 'no interaction' problem, and the logistic function. *Biometrics* **28** 443–468.

[13] BERKSON, J. (1975). Do radioactive decay events follow a random Poisson-exponential? *Internat. J. Appl. Radiation and Isotopes* **26** 543–549.

[14] BERKSON, J. (1976). An extended view of chi-square testing and estimation, Part I and II. Presented at a seminar of Prof. Jerzy Neyman, Berkeley, California.

[15] BERKSON, J. (1977). My encounter with neo-Bayesianism. *Internat. Statist. Rev.* **45** 1–8.

[16] BERKSON, J. and ELVEBACK, L. (1960). Competing exponential risks, with particular reference to the study of smoking and lung cancer. *J. Amer. Statist. Assoc.* **55** 415–428.

[17] BERKSON, J. AND HODGES, J. L., JR. (1961). A minimax estimator for the logistic function. *Proc. Fourth Berkeley Symp. Math. Statist. Probability* **4** 77–86.

[18] BERKSON, J. and NAGNUR, B. N. (1974). A note on the minimum χ^2_1 estimate and a L.A.M.S.T. χ^2 in the "no interaction" problem. *J. Amer. Statist. Assoc.* **69** 1038–1040.

[19] BOREL, E. (1964). Apropos of a treatise on probability, a review of John Maynard Keynes. Reprinted in *Studies in Subjective Probability*. (Henry E. Kyburg, Jr., Howard E. Smokler, eds.) 47–60. Wiley, New York.

[20] COX, D. R. (1959). Acknowledgement of priority of Berkson in noting sufficient statistics for the logistic function. *J. Roy. Statist. Soc. Ser. B* **21** 238.

[21] CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

[22] EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242.

[23] FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London, Ser. A* **222** 309–368.

[24] FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725.

[25] FISHER, R. A. (1937–1938). Note of comment following article by Harold Jeffereys, "Maximum likelihood, inverse probability and the method of moments". *Ann. Eugenics* **8** 146–151.

[26] FISHER, R. A. (1938). *Statistical Methods for Research Workers*. Oliver and Boyd, London.

[27] FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Hafner, New York.

[28] GHOSH, J. K. and Subramanyam, K. (1974). Second order efficiency of maximum likelihood estimators. *Sankyā, Ser. A* **36** 325–358.

[29] LITTLE R. E. (1974). The mean square error comparison of certain median response estimates for the up-and-down method with small samples. *J. Amer. Statist. Assoc.* **69** 202–206.

[30] PEARSON, E. (1974). Memories of the impact of Fisher's work in the 1920's. *Internat. Statist. Rev.* **42** 5–8.

[31] PFANZAGL, J. (1973). Asymptotic expansions related to minimum contrast estimators. *Ann. Statist.* **1** 993–1026.

[32] PLACKETT, R. L. (1972). The discovery of the method of least squares. *Biometrika* **59** 239–251.

[33] RAO, C. R. (1961a). Apparent anomalies and irregularities in maximum likelihood estimation. *Bull. Internat. Statist. Inst.* Part 4 **38** 439–453.

[34] RAO, C. R. (1961b). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Probability* **1** 531–545.

[35] RAO, C. R. (1962). Efficient estimates and optimum inference procedures in large samples, with discussion. *J. Roy. Statist. Soc., Ser. B* **24** 46–72.
[36] SAVAGE, L. J. (1976). On rereading R. A. Fisher, with discussion. *Ann. Statist.* **4** 441–500.
[37] STEIN, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution, with discussion. *J. Roy. Statist. Soc., Ser. B* **24** 265–296.
[38] WETHERILL, G. B. (1963). Sequential estimation of quantal response curves. *J. Roy. Statist. Soc., Ser. B* **25** 1–48.

MEDICAL RESEARCH STATISTICS
MAYO CLINIC
ROCHESTER, MINNESOTA 55901

DISCUSSION

BRADLEY EFRON
Stanford University

Before tearing into the paper, let me first applaud Professor Berkson's skeptical attitude toward asymptotics and fancy theory in general. Throughout his productive career he has always been primarily concerned with the practical, the computable and the verifiable—the right attitude for a good scientist doing good science. His mistake is not crediting Fisher (and Rao, Savage, Ghosh, Subramanyam, and me) with some of the same good sense.

Why is maximum likelihood estimation so popular? Certainly not because of any mystique connected with maximizing likelihoods. This process is actually rather unintuitive compared with, say, the method of moments, as teachers of elementary courses soon discover, and it is computationally more difficult to boot. The appeal of maximum likelihood stems from its universal applicability, good mathematical properties, by which I refer to the standard asymptotic and exponential family results, and generally good track record as a tool in applied statistics, a record accumulated over fifty years of heavy usage. If this last point were not true, the first two would be irrelevant, and we would not be having this discussion.

Consider Berkson's censored exponential example (4). None of the chi-square methods Berkson lists even give an answer in this situation, unless some artificial grouping is imposed on the data. The MLE (6) is intuitively reasonable ($\hat{\beta}$ equals the observed intensity of deaths), asymptotically wonderful, and *automatic*; the applied statistician in the field faced with model (4) does not have to know any fancy theory at all to get a quite good estimate. Maximum likelihood is the original "jackknife", a dependable tool for almost any estimation purpose. These comments in no way forbid criticism of the method in specific cases, or in general for that matter, but they do underline the temerity of this paper's title. [Incidentally, the density (5) represents a "curved exponential family", Efron 1975 (Berkson's references), in which the MLE is well known not to be sufficient.]

Now for some specific points:

1. It is misleading to say, as Berkson does in regard to his logistic regression example, that the MLE is not minimal sufficient in an exponential family. The difficulties Berkson alludes to are semantic in nature, and can be avoided by not calling the MLE the same name (such as "∞") for values of the sufficient statistic vector outside the range of the expectation space, see Efron (1978).

2. Berkson's 1955 example to which I referred is a logistic regression with three independently observed biniomial proportions $p_i \sim Bi(10, \pi_i)/10$,

$$\pi_i = 1/\left[1 + \exp\{-(\alpha + \beta x_i)\}\right] \qquad i = 1, 2, 3. \tag{1}$$

The regression variable x_i takes on values $-1, 0, 1$ for $i = 1, 2, 3$, and β is fixed and known to equal .84730. It is desired to estimate α on the basis of the observations p_1, p_2, p_3. This situation is a one-parameter exponential family with sufficient statistic $\Sigma_{i=1}^{3} p_i$. The MLE $\hat{\alpha}$ is indeed a minimal sufficient statistic here, while the minimum logit chi-square estimator $\tilde{\alpha}$ is not.

Berkson considers a sampling experiment with the true value $\alpha = 0$, and notes that $\tilde{\alpha}$ has smaller mean square error of estimation than $\hat{\alpha}$ in this case (and in several others). Efron and Holland (1968) show that this is due to the different biases of $\hat{\alpha}$ and $\tilde{\alpha}$; both estimators are unbiased at $\alpha = 0$, but, to a good approximation,

$$E_\alpha \hat{\alpha} = 1.027\alpha, \qquad E_\alpha \tilde{\alpha} = .936\alpha \tag{2}$$

for α near 0. (No extra observational data is required to calculate (2), of course.) If $\hat{\alpha}$ is adjusted to have the same bias structure as $\tilde{\alpha}$, by defining $\hat{\alpha}' = (.936/1.027)\hat{\alpha}$, then $\hat{\alpha}'$ has smaller M.S.E. than $\tilde{\alpha}$ for values of α near 0, as it must by the Rao-Blackwell theorem. The important point here is that it is the bias structure, and not the particular method of estimation, that is determining the comparison; $\hat{\alpha}/1.1$ has smaller M.S.E. than either $\hat{\alpha}$ or $\tilde{\alpha}$ over a large range of α values near 0, but that does not mean we should always use $\hat{\alpha}/1.1$ to estimate α.

3. The James-Stein estimator of a multivariate normal mean vector is an example where deliberately induced biases give impressive global improvements over the MLE. A systematic theory of biased estimation would be of the greatest importance, but does not yet exist. See Efron (1975).

4. Fisher and Rao hardly need my protection, but second order efficiency was certainly not an "unmotivated concept". It was based on a deep and subtle understanding of the estimation process. Decision theory is a powerful weapon, and in the hands of masters like Neyman, Wald and Stein has subdued some formidable problems, but the inference school of statistics has many valuable trophies hung upon its wall too. Hard line decision theorists like Berkson and Pfanzagl are useful policemen in the often chaotic world of statistical theory, but the rest of us must be vigilant for signs of police brutality.

REFERENCES

[1] EFRON (1975). Biased versus unbiased estimation. *Advances in Math.* **16** 259–275. Reprinted in *Surveys in Applied Mathematics*. (1976) Academic Press.
[2] EFRON (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376.
[3] EFRON and HOLLAND (1968). On the misuse of the mean square error to compare biased estimators. Memorandum NS-110, Depart. Statist., Harvard Univ.

J. K. GHOSH
Indian Statistical Institute

It appears that Professor Berkson has revived an old debate about the MLE at least partly because of the recent interest in the results of Fisher and Rao on second order efficiency. In view of this it may be worth recording here what second order efficiency does and does not mean for Berkson's famous example from bioassay.

Adopting the notations of Ghosh and Subramanyam (1974 Section 3) (but writing E for the rather pedantic E^U) one may write

$$E(T_n) = \alpha + b(\alpha)/n + o(n^{-1})$$

$$E(\hat{\alpha}_n) = \alpha + b_o(\alpha)/n + o(n^{-1})$$

where T_n is the minimum logit chi-square estimate, $\hat{\alpha}_n$ is the MLE,

$$b(\alpha) = \Sigma\pi_i(1 - \pi_i)(2\pi_i - 1)/I^2 - \Sigma k(2\pi_i - 1)/2I$$

$$b_o(\alpha) = \Sigma\pi_i(1 - \pi_i)(2\pi_i - 1)/2I^2.$$

The corrected MLE may be taken to be a truncated version of

$$\hat{\hat{\alpha}}_n = \hat{\alpha}_n + \{b(\hat{\alpha}_n) - b_o(\hat{\alpha}_n)\}/n.$$

To truncate $\hat{\hat{\alpha}}_n$ one must choose some $d > 0$, such that the true α may be assumed to lie in $(-d, d)$ and then replace $\hat{\hat{\alpha}}_n$ by d or $-d$ according as it exceeds d or falls below $-d$. (The asymptotic theory is insensitive to the choice of d). Let the estimate T_n be truncated in a similar way. *Then the mean squared error of the truncated* $\hat{\hat{\alpha}}_n$ *is strictly smaller than that of the truncated* T_n *if terms of* $o(n^{-2})$ *are neglected.* This result remains true for quite general loss functions, see Ghosh, Sinha and Wieand (1977). As explained in Ghosh and Subramanyam (1974, Section 4) the reason for this is that the MLE approximates Bayes estimates better than its common rivals. When Subramanyam and I started studying second order properties of the MLE we were looking for a Bayes estimate which would be better than the MLE. It came as a surprise to us that this is impossible (up to $o(n^{-2})$ in the mean squared error).

A general result of this sort should make one prefer the MLE to the other BAN estimates commonly used as alternatives provided two rather strong conditions hold. First, the assumption about the form of the likelihood function is correct. Secondly, the terms of $o(n^{-2})$ are negligible for actual samples. It seems to me the first assumption is the more serious one and consequently the main criticism of the MLE should be based on its lack of robustness.

If Berkson's object is to provoke us into a critical reappraisal of the MLE, then I am in complete agreement with him. However, if he means all that he says in his provocative title then we must part ways at some point. Minimum chi-square estimates may be all right in certain forms of data analysis when very little is known about the data. In all other cases the likelihood is too useful a part of the data to be ignored. We should be looking for an estimate which makes use of the likelihood but in a more robust way than the MLE.

I will end by making a few comments on Professor Berkson's examples.

Consider first the bioassay example treated above. The following three statements are easily verified. If all the subjects are killed, i.e., $\Sigma p_i^n = k$, the MLE is ∞. If all the subjects survive, i.e., $\Sigma p_i^n = 0$, the MLE is ∞. In all other cases, i.e., if $0 < \Sigma p_i^n < k$, the MLE is the unique solution of the likelihood equation. Thus, the MLE is a one-one function of the minimal sufficient statistic Σp_i^n, contrary to a claim of Berkson. On the other hand, the definition of the minimum logit chi-square estimate becomes ambiguous if $p_i^n = 0$ or 1 for any i.

In his first example Professor Berkson wants an estimate to be equal to the estimated parameter if all the observations are equal to their expectations. I shall call it Berkson consistency (with respect to the Y's) to distinguish it from Fisher consistency which requires equality of estimate and parameter when the sample distribution function coincides with the true distribution function. (Thus Fisher consistency is Berkson consistency with respect to the sample distribution function.) I will now give an example where no estimate which is admissible with respect to the squared error loss can be Berkson consistent with respect to Y.

Let Y be a single observation from $N(\theta, 1)$ and assume $a \leqslant \theta \leqslant b$ where $a < b$ are known constants. To be Berkson consistent an estimate $T(Y)$ must equal Y if $a \leqslant Y \leqslant b$. A standard argument involving analyticity of Bayes estimates then shows T cannot be proper Bayes and hence T is inadmissible. Surely in this example Berkson consistency (with respect to Y) should be repugnant to Bayesians as well an Neyman-Pearsonians.

Here is another example which is instructive in a different way. Consider a single observation Y from $N(\mu, \sigma^2)$ and assume that $\mu = \sigma^2$. In this case $\hat{\mu}$ is not Berkson consistent with respect to Y but it is Berkson consistent with respect to Y^2 which is minimal sufficient. In the example given by Berkson something of this sort happens. If at each dose ΣY as well as ΣY^2 equals its expected value, then the MLE would recover the true values. Of course it is impossible to get such data if one has only one observation for each dose. However, even for Berkson's example the match between y^2 and $\hat{Y}^2$ is no worse than that between y^2 and $\tilde{Y}^2$.

REFERENCES

GHOSH, J. K. and SUBRAMANYAM, K. (1974). Second order efficiency of maximum likelihood estimators. *Sankhyā Ser. A* **36** 325—358.

GHOSH, J. K., SINHA, B. K. and WIEAND, H. S. (1977). Second order efficiency of the MLE for any bounded bowl shaped loss function. Unpublished manuscript.

L. LE CAM[1]

University of California, Berkeley

1. Introduction. It is a pleasure and a priviledge to have been asked to comment on the paper by our distinguished colleague, J. Berkson. One may readily disagree, as the present writer does, with certain details of Dr. Berkson's article. However, in view of the accumulated knowledge and experience in these matters one is forced to conclude that Dr. Berkson is right. In fact one wonders why standard texts continue to peddle the m.l.e. instead of more sensible procedures.

Since the facts seem to be so easily forgotten, we recall some of them below, but first present some general comments.

One comment concerns Dr. Berkson's assertion that: "There is no mathematical demonstration that there is any method of estimation that assures that the estimate attained is best in an operational sense in all circumstances."

When the "circumstances" are sufficiently circumscribed one may have recourse to Wald's theory of decision functions, but that theory does not offer any criteria for selection between several admissible estimates. Various "principles" such as minimax, unbiasedness, have long ago been shown to lead to unacceptable or even ludicrous results in suitable frameworks.

In fact one could argue that it is a virtue of Wald's theory that it does not tell us exactly what to do. In each particular practical case the "circumstances" are various and difficult to define in straight mathematical terms. Thus it seems that unless somebody comes along with a precise definition of what is a "method of estimation" and what are "circumstances", one will have to resign oneself to the idea that, in Dr. Berkson's words, "every problem should be studied". We have no mathematical demonstration that this is the best course to follow, but this seems to be what scientists do when they can and where they care.

In practical settings the working statistician should certainly beware of any and all so-called "principles", even the principle of sufficiency.

For instance, in fitting a two-parameters gamma distribution, "sufficiency" would tell us to use the sum of the observations and the sum of their logarithms. This may be highly irresponsible if the small observations are not measured with extreme precision. In such cases one may be led to accept procedures which are less than optimal under the ideal model, but remain reasonable under the particular circumstances at hand (see Huber [8]). Such a remark may not merit a Nobel prize in

[1]This research was supported by the National Science Foundation Grant MCS75-10376-A02.

economics, but it certainly should be kept in mind in general and when reading the statements which follow.

2. Some facts about maximum likelihood. It has been observed long ago that maximum likelihood does not always lead to estimates which are "best" in some desirable sense.

For instance if one takes n independent observations from a uniform distribution on $[0, \theta]$ the m.l.e. is the maximum, say Z_n, of the observations. One has

$$E\left[(Z_n - \theta)^2|\theta\right] = 2\theta^2\left[(n+2)(n+1)\right]^{-1}$$

but

$$E\left[(\theta_n^* - \theta)^2|\theta\right] = \theta^2(n+1)^{-2}$$

for $\theta_n^* = (n+2)(n+1)^{-1}Z_n$.

Some authors have observed that the m.l.e. technique is unproductive in estimating parameters of mixtures of Gaussian distributions [9] or even in the three parameter log normal distribution (that is, $X = a + b\exp\{cY\}$, with $\mathcal{L}(Y) = \mathcal{N}(0, 1)$), even though these families fall in the category to which some of the asymptotic results given below can be made applicable.

Neyman and Scott [12] gave a very smooth example in which m.l.e. tends to be about half of what any reasonable estimate should be.

Dr. Berkson provided us with the first example of a "straight" exponential family in which a very easily obtainable estimate is definitely better than the m.l.e. (see [4], [5] and [6]).

We are indebted to R. R. Bahadur [2] for the first example of a family in which the m.l.e. always exists, but tends to infinity almost surely, no matter what is the true value of the parameter. Bahadur's example is perhaps hard to grasp intuitively. Here is a variation of it where the reason for the misbehavior of m.l.e. is visible. Take $g(x) = \exp(1/x^2)$ for $x \in (0, 1)$ and let c be fixed number $0 < c < 1$. Define recursively values a_k in $(0, 1]$ by $a_0 = 1$ and by $\int_{a_k}^{a_{k-1}}[g(x) - c]dx = 1 - c$. For $\theta = 1, 2, \cdots$ let f_θ be that density with respect to Lebesgue measure which is equal to c for $x \in (0, 1)$ but $x \notin [a_\theta, a_{\theta-1})$ and to the average $[a_{\theta-1} - a_\theta]^{-1}\int_{a_\theta}^{a_{\theta-1}} g(x)dx$ for $x \in [a_\theta, a_{\theta-1})$.

This family has properties which show that reliance on the m.l.e. principle is rationally untenable. Assume that the observations X_j taken are i.i.d. from some f_θ. Then the m.l.e. of θ will almost surely go to infinity no matter what θ is. However one could imbed this family in a larger one, take the m.l.e. there *ignoring the fact that we assume that the observations do come from one of the* f_θ. A suitable imbedding can make the new m.l.e. consistent! If one flattens out a good part of the information contained in the X_j by recording instead of X_j a sum $Z_j = X_j + Y_j$ where the Y_j are i.i.d. independent of the X_j and $\mathcal{N}(0, \sigma^2)$, $\sigma = 10^{26}$, an m.l.e. computed from the distribution of the Z_j becomes consistent!

In the family f_θ described above one can transform θ into a continuous parameter, interpolating smoothly between the successive integer values of θ. One can do

so in such a way as to get a family which satisfies conditions of the Cramér type, but where m.l.e. always tends to infinity almost surely.

In the preceding example we have mentioned "inconsistency" properties of the m.l.e. This is of an asymptotic nature, but the reader can easily verify that even for moderate sample sizes m.l.e. is very badly behaved indeed and does not capture much of the "information" contained in the observations, contrary to what seems to be a prevailing belief.

One should perhaps also mention "Fisher consistency". This is another one of these principles which cannot be justified on any reasonable grounds. For instance, in the uniform $[0, \theta]$ example given above Z_n is "Fisher consistent" but θ_n^* is not. Furthermore, θ_n^* depends on the number n of observations, while it seems to be a tenet of faith in some quarters that estimates should have a form independent of n, whatever that may mean. See [15].

In view of examples of the kind listed above, one wonders why some authors still claim that m.l.e. is "best" in some sense. It seems that the reason is that occasionally and under severe restrictions m.l.e. happens to be close to estimates which have desirable properties. In the recent past there has been a considerable body of literature devoted to second or third order efficiency properties. We shall now discuss this subject briefly to point out that it really does not imply much about the m.l.e. itself.

3. Higher order efficiency. A most careful, delicate and highly creditable description of the situation can be found in the papers of Pfanzagl and his collaborators. (See, for instance, Pfanzagl and Wefelmeyer [14] and the references given there.)

Pfanzagl considers estimates $\theta^{(n)}$ which are called asymptotically maximum likelihood of order $n^{-1}l_n(2)$ (as m.l. of order $n^{-1}l_n(2)$), and shows that they, or functions of them, are asymptotically better than any one of the estimates of a fairly wide class.

One first remark, for the theoretically minded, is that the m.l.e. itself satisfies Pfanzagl's conditions only under fairly restrictive assumptions. A more important remark, for the practically minded, is that, as shown by Pfanzagl, the as m.l. estimates are often easily obtainable, starting with a good auxiliary estimate and using a Newton-Ralphson method *just twice*. (The cruder method proposed by the present author [10], using differences instead of derivatives, will also work if iterated just the same way.)

At first sight the complete class result of Pfanzagl, et al is at variance with previous results of Pfanzagl which shows that the as m.l.e. or similar objects are not asymptotically sufficient at the desired rate of approximation. For instance, Pfanzagl [13] has shown, in an unmistakable manner, that for testing purposes and for the desired rate of approximation one needs to involve more than one logarithmic derivative.

Pfanzagl's arguments in [13] and those of Pfanzagl and Wefelmeyer in [14] are very delicate. Thus it may not be out of bounds to suggest here a heuristic argument, similar to that proposed by Ghosh and Subramanyam in [7]. This does not replace proofs, but will lead us back to Berkson's article and a better view of the problem.

Suppose that we have i.i.d. observations from a density $f(x, \theta)$ which depends in a smooth way on the one dimensional parameter θ. (The k-dimensional situation introduces mostly notational problems but also some other phenomena, such as the one described by C. Stein [16].)

Take functions W such that $0 \leqslant W \leqslant 1$ and such that the sets $\{u;\ W(u) \leqslant \alpha\}$ are convex, symmetric around zero and bounded when $\alpha < 1$. If T_n is an estimate based on n observations, let $W[n^{\frac{1}{2}}(T_n - \theta)]$ be the loss and let $R(T_n, \theta)$ be the corresponding risk.

Let $\mathcal{P}$ be a class of prior distributions for θ. Assume that $\mathcal{P}$ is convex, compact for the ordinary convergence of distributions and such that all members of $\mathcal{P}$ have, with respect to Lebesgue measure, nonvanishing densities which are bounded and have fourth derivatives bounded by some given constant b.

If a given estimate T_n is not a Bayes estimate with respect to some member of $\mathcal{P}$, then there is an $\varepsilon > 0$ and some other estimate T_n' (which is a Bayes estimate relative to $\mathcal{P}$) such that

$$\int [R(T_n, \theta) - R(T_n', \theta)]\, \mu(d\theta) > \varepsilon \qquad \text{for all} \quad \mu \in \mathcal{P}.$$

Thus if one is not too worried about behavior of risks for large values of θ and does not relish local rapid variations of risks, the Bayes estimates obtainable from $\mathcal{P}$ look like acceptable candidates.

Under the conditions given, and some other conditions which are not too terrible (see ([11]), the Bayes estimates, say β_n, will be such that $\mathcal{L}\{n^{\frac{1}{2}}(\beta_n - \theta)|\theta\}$ is relatively compact. By contrast, m.l.e. may not even be consistent.

Now assume that $\theta_0 = 0$ is the true value of the parameter. The posterior density of θ will have a local maximum, say θ_n^*, situated near zero. Taking as new variable $s = n^{\frac{1}{2}}(\theta - \theta_n^*)$, the posterior density takes a form of the type

$$f_n(s) = C_n \exp\left\{ -\tfrac{1}{2} A_n s^2 \left[1 + \frac{s B_n(s)}{3n^{\frac{1}{2}}} \right] \right\}$$

where A_n and B_n are random, but not overly large. To minimize the posterior risk, one would minimize the expression $\int W(s - t) f_n(s)\, ds$ with respect to t. We can already assume (from [11] for instance) that the minimizing value t_n will stay bounded in probability. Substituting for f_n approximations of the type $C_n \exp\{-\frac{1}{2} A_n s^2(1 + \varepsilon)\}$ with $|\varepsilon|$ small, independent of s and using T. W. Anderson's lemma [1], one can even argue that t_n will tend to zero in probability. Then one may proceed in the manner of Laplace, using Taylor expansions near zero to obtain asymptotic expansions. One sees then that t_n is approximately equal to

$(D_n/6n^{\frac{1}{2}})$ where D_n is the product of $(B_n(0)/A_n)$ by a term of the type

$$\left[E(3\xi^2 - \xi^4)W\left(\xi/(A_n)^{\frac{1}{2}}\right)\right]\left[E(\xi^2 - 1)W\left(\xi/(A_n)^{\frac{1}{2}}\right)\right]^{-1}$$

for a variable ξ which is $\mathcal{N}(0, 1)$.

In these expressions, A_n, B_n and D_n depend on the observations in a complex manner. However, one may follow the suggestion of Ghosh and Subramanyam [7], and replace D_n by an estimated value which is a function of θ_n^* alone. This modifies t_n by terms of higher order only.

Thus, in this sense, the Bayes estimate $\theta_n^* + (t_n/n^{\frac{1}{2}})$ can be approximated by functions of the type $\theta_n^* + n^{-1}F_n(\theta_n^*)$. Since θ_n^* itself differs from the $n^{-1}l_n(2)$ as m.l.e. $\theta^{(n)}$ of Pfanzagl only by terms of order n^{-1} one can substitute $\theta^{(n)}$ in the above conclusion.

Note, however, that at this order of approximation θ_n^* does depend on the prior distribution $\mu \in \mathcal{P}$ used for the computation. The correction term t_n depends in addition on the function W used to define the loss. (A similar remark could be made about Wolfowitz maximum probability estimates [18] obtainable here by replacing $1 - W$ by the indicator of an interval $[-r, +r]$. The maximizing value will depend on r.)

Thus to use this approach effectively one would need to specify both μ and W. Otherwise the correction terms remain arbitrary.

The same feature is also present in Theorem 1 of [14]. At the order of magnitude considered there, the as. m.l.e. of order $n^{-1}l_n(2)$ differ mostly by a term which is analogous to the "bias" terms one meets in other circumstances. To say that one can match or better any regularly behaved estimate by a function of $\theta^{(n)}$ does not resolve the problem of deciding what to match, or if one wants, of selecting the "bias" term.

This is precisely the type of endeavor which is carried out by Berkson and Hodges in [6].

Another approach, using linear combinations of several "minimum chi-square" estimates was shown to this author by K. Sutrick [17]. The scope of this latter method is more limited, but it appears useful.

It is interesting that the above described procedure of estimating terms like D_n fails for testing problems [13]. A heuristic explanation is as follows. For the estimation problems with loss function of the type considered above, the "best" estimates depend on second derivatives of the likelihood function only in a "minor" way, thanks to T. W. Anderson's lemma. However, for testing problems the Neyman-Pearson lemma cannot be ignored.

Take, then, likelihood ratios $\Lambda_{n,i}$ for densities at $\theta_0 + (s_i/n^{\frac{1}{2}})$, $i = 1, 2$, and at θ_0. Let R_n be the combination $n^{\frac{1}{2}}[(\Lambda_{n,1}/s_1) - (\Lambda_{n,2}/s_2)]$ centered at its median.

Except for "straight" exponential families, the joint limiting distribution of R_n and $n^{\frac{1}{2}}(\theta^{(n)} - \theta_0)$ is a nondegenerate bivariate Gaussian distribution. Thus, at the order of approximation desired Λ_n is *not* a function of $\theta^{(n)}$ for all values of s.

Finally all the arguments described in the present section depend very strongly on the validity of the parametric model used in the derivations.

Recent results of R. Beran [3] show that one can modify the Newton-Ralphson type techniques to yield a certain amount of "robustness" in small Hellinger tubes around the assumed parametric models. However, as far as this author knows, none of the techniques proposed have been shown to be able to cope with one of the most dispiriting features of large sample theories. It is not unusual in our days to have samples of size $n = 5105$ or 10^6 or even 10^8, for instance in automated biological measurements on blood cells. Typically, this should make asymptotic arguments very applicable, but instead crude checks of the validity of the hypothesized models, for instance by Kolmogorov-Smirnov tests, show that they are disastrously far from reality, some of differences being "highly significant" statistically but invisible on standard graph paper and sometimes due to minute, trivial or indescribable features of the measurement processes. Thus the asymptotics fail precisely when one would feel that they are applicable.

REFERENCES

[1] ANDERSON, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* **6** 170–176.

[2] BAHADUR, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhyā* **20** 207–210.

[3] BERAN, R. (1978). Robust and efficient window estimates in parametric models. Unpublished manuscript.

[4] BERKSON, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *J. Amer. Statist. Assoc.* **48** 565–599.

[5] BERKSON, J. (1955). Maximum likelihood and minimum χ^2 estimates of the logistic function. *J. Amer. Statist. Assoc.* **50** 130–162.

[6] BERKSON, J. and HODGES, S. L. JR. (1961). A minimax estimator for the logistic function. *Proc. Fourth Berkeley Symp. Math. Statist. Probability* **4** 77–86.

[7] GHOSH, J. K. and SUBRAMANYAM, K. (1974). Second order efficiency of maximum likelihood estimators. *Sankhyā, Ser. A* **36** 325–358.

[8] HUBER, P. J. (1972). Robust statistics: a review. *Ann. Math. Statist.* **43** 1041–1067.

[9] KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

[10] LE CAM, L. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* **3** 37–98.

[11] LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

[12] NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika* **16** 1–32.

[13] PFANZAGL, J. (1974). Non-existence of tests with deficiency zero. Preprints in statistics, Univ. Cologne, #8.

[14] PFANZAGL, J. and WEFELMEYER, W. (1978). A third order optimum property of the maximum likelihood estimator. *J. Multivariate Anal.* **8** 1–29.

[15] SAVAGE, L. J. (1976). On rereading R. A. Fisher. *Ann. Statist.* **4** 441–500.

[16] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probability* **1** 197–206.

[17] SUTRICK, K. (1978). Private communication.

[18] WEISS, L. and WOLFOWITZ, J. (1967). Maximum probability estimators. *Ann. Inst. Statist. Math.* **19** 193–206.

JOHANN PFANZAGL
University of Cologne

I am in full agreement with Mr. Berkson's basic attitude to judge the quality of an estimator by the concentration of its distribution about the true parameter value. In particular, any claim for the superiority of the maximum likelihood estimator has to be based on this criterion.

Since the exact distributions of estimators are not accessible in general, the last recourse is numerical comparison. But numerical comparisons hold out no prospects of obtaining general insights (like the superiority of the maximum likelihood estimator). A reasonable way out of this dilemma is, perhaps, to approximate the distributions of the estimators by probability measures of sufficiently simple structure, for instance, by Edgeworth-expansions, to draw general conclusions, and to check the validity of these conclusions by numerical examples.

Since the discussion is on differences between efficient estimators, the normal approximation is too crude. For estimators with stochastic expansion—the maximum likelihood estimator as well as the minimum chi-square estimators are of this type—a refined analysis is relatively easy. The first outcome of such an analysis is that efficient estimators differ mainly by a median-bias of order $n^{-\frac{1}{2}}$ (resp. a bias of order n^{-1}). Hence it is meaningless to compare different efficient estimators without eliminating their difference in bias prior to the comparison. This point was already raised by Ghosh and Subramanyam (1974, page 321), but not fully recognized by Berkson.

A refined investigation for the case of i.i.d. observations reveals that, after the elimination of the bias-difference, the distributions of the estimators differ at most by an amount of order n^{-1}.

For the particular case of one real parameter these results can be stated as follows:

If the estimator $\hat{\theta}^{(n)}$ is asymptotically efficient, then $n^{\frac{1}{2}}(\hat{\theta}^{(n)}(\mathbf{x}) - \theta) = I(\theta)^{-1}\tilde{l}'(\mathbf{x}, \theta) + R_n(\mathbf{x})$, where $I(\theta) = \int l'(x, \theta)^2 P_\theta(dx)$ and $\tilde{l}'(\mathbf{x}, \theta) = n^{-\frac{1}{2}}\sum_{\nu=1}^{n} l'(x_\nu, \theta)$ with $l'(x, \theta) = (d/d\theta)\log p(x, \theta)$. R_n is an error term which converges to zero stochastically.

Assume now that the estimator admits a stochastic expansion, i.e.,

$$(*) \qquad n^{\frac{1}{2}}(\hat{\theta}^{(n)}(\mathbf{x}) - \theta) = I(\theta)^{-1}\tilde{l}'(\mathbf{x}, \theta) + n^{-\frac{1}{2}}Q_1(\tilde{l}'(\mathbf{x}, \theta), \tilde{f}_1(\mathbf{x}, \theta); \theta) + n^{-1}Q_2(\tilde{l}'(\mathbf{x}, \theta), \tilde{f}_1(\mathbf{x}, \theta), \tilde{f}_2(\mathbf{x}, \theta); \theta) + n^{-1}R_n(\mathbf{x})$$

where Q_i are polynomials, $f_i(x, \theta)$ are vector-valued functions with $\int f_i(x, \theta)P_\theta(dx) = 0$, and $\tilde{f}_i(\mathbf{x}, \theta) = n^{-\frac{1}{2}}\sum_{\nu=1}^{n} f_i(x_\nu, \theta)$.

To simplify our diction, we shall say: an estimator-sequence with stochastic expansion (*) has property S if $Q_1(\tilde{l}'(\mathbf{x}, \theta), \tilde{f}_1(\mathbf{x}, \theta); \theta)$ with $\tilde{l}'(\mathbf{x}, \theta)$ replaced by zero becomes independent of $\mathbf{x}$.

Among the estimators sharing property S are the maximum likelihood estimator and Bayes-estimators with respect to symmetric loss functions.

If an estimator-sequence $\hat{\theta}^{(n)}$ has property S, and only in this case, it is asymptotically optimal of order $o(n^{-1})$ in the following sense: if $\theta^{(n)}$ is any other estimator-sequence with a stochastic expansion (*) (in general with different polynomials Q_i and different functions f_i), then there exists a function q of the parameter such that—simultaneously for all sufficiently smooth symmetric loss functions L which are bounded by polynomials—

$$\int L\left(n^{\frac{1}{2}}\left(\hat{\theta}^{(n)}(\mathbf{x}) + n^{-1}q\left(\hat{\theta}^{(n)}(\mathbf{x})\right) - \theta\right)\right)P_\theta^n(d\mathbf{x})$$

$$\leqslant \int L\left(n^{\frac{1}{2}}\left(\theta^{(n)}(\mathbf{x}) - \theta\right)\right)P_\theta^n(d\mathbf{x}) + o(n^{-1}). \qquad (**)$$

(The function q is determined such that the median-bias of $\hat{\theta}^{(n)}(\mathbf{x}) + n^{-1}q(\hat{\theta}^{(n)}(\mathbf{x}))$ agrees up to $o(n^{-\frac{1}{2}})$ with the median-bias of $\theta^{(n)}$).

If both estimators, $\theta^{(n)}$ and $\hat{\theta}^{(n)}$, are regular in the sense that Q_1 is even and Q_2 odd, then (**) even holds for nonsymmetric loss functions. (See Pfanzagl and Wefelmeyer 1978, Theorem 1′ and Remarks 3.15, 3.24; see 1979 for regular estimators.)

Such a result enables us to compare risks of estimators up to error terms of order $o(n^{-1})$ without computing their distributions. Regrettably, they are, so far, available only for the i.i.d. case and are, therefore, not applicable to Berkson's bioassay experiments, presuming the probability of death under dosage d_i equal to

$$\pi_i(\alpha, \beta) = 1/\left(1 + \exp\left[-\alpha - \beta d_i\right]\right).$$

They become applicable in the special case $\beta = 0$.

Specializing formulas (3.4) and (3.5) of Ghosh and Subramanyam (1974) (which presume an equal number of animals for each dosage) to this case, we obtain that the m.l. estimator for α, as well as the minimum logit chi-square estimator, have a regular stochastic expansion (*) with

$$Q_1 = \frac{2\pi - 1}{I^2}\frac{k}{2}n(p - \pi)^2 \qquad \text{for the m.l. estimator}$$

and

$$Q_1 = \frac{2\pi - 1}{I^2}k\left(n(p - \pi)^2 - \tfrac{1}{2}\Sigma_1^k\frac{n}{k}(p_i - \pi)^2\right) \qquad \text{for the minimum logit chi-square estimator}$$

where k is the number of dosages, p_i the fraction of deaths under dosage d_i, and p the average of p_i, $i = 1, \cdots, k$. Alone, the fact that the minimum logit chi-square estimator depends on $p_1, \cdots, p_k$ (and not on p only, even though p is sufficient) generates bad feelings. Since $\tilde{l}'(\mathbf{x}, \theta)$ becomes $n^{\frac{1}{2}}(p - \pi)$ in this case, the m.l.

estimator has property S (which is true in general), whereas the minimum logit chi-square estimator does not. Replacing $n^{\frac{1}{2}}(p - \pi)$ by zero leaves the second term which depends on the observations.

Hence, the result obtained by Ghosh and Subramanyam (1974, page 351) for the truncated quadratic loss function is true for arbitrary (i.e., not necessarily symmetric) loss functions which are sufficiently smooth and bounded by a polynomial. If the maximum likelihood estimator is corrected so that its median-bias matches the median-bias of the minimum logit chi-square estimator up to $o(n^{-\frac{1}{2}})$, its risk (defined as in (**)) falls short of the risk of the minimum logit chi-square estimator by an amount of order n^{-1}. (On the other hand, the minimum logit chi-square estimator, lacking property S, cannot be adjusted so as to match, let alone underbid the m.l. estimator.)

Are these theoretical results really conclusive? Of course not. First of all, they are limited to the case $\beta = 0$. This is, however, not essential. The general results cited above can certainly be extended to also cover regression models. A more serious objection is the asymptotic nature of these results. Strictly speaking, asymptotic results, however refined they might be, do not tell anything about samples of fixed size. Numerical computations in connection with other applications show, however, that asymptotic expansions of order $o(n^{-1})$ render in many cases excellent approximations to the actual performance. Hence, one can be confident that our final recourse, the numerical computations, will also in this case confirm the conclusion obtained by asymptotic expansions.

It is essential in our conception that this conclusion is not contingent on a particular loss function (such as the quadratic), but holds for a rather wide class of loss functions.

If the reference to a historical authority is a valid argument at all, it is certainly unwarranted to take Gauss as an advocate of the quadratic loss function, favored by Mr. Berkson. Gauss introduced the quadratic loss function for *technical* reasons[1] which are not essential any more for recent results.

REFERENCES

GHOSH, J. K. and SUBRAMANYAM, K. (1974). Second order efficiency of maximum likelihood estimators. *Sankhyā Ser. A* **36** 325–358.

PFANZAGL, J. and WEFELMEYER, W. (1978). A third-order optimum property of the maximum likelihood estimator. *J. Multivariate Anal.* **8** 1–29.

PFANZAGL, J. and WEFELMEYER, W. (1979). Addendum to "A third-order optimum property of the maximum likelihood estimator." *J. Multivariate Anal.* **9** 179–182.

[1]Says Gauss in his letter to Bessel, dated February 28, 1839: " . . . indem für L eine Function gewählt wird, die immer positiv und für grössere Argumente auf eine schickliche Art immer grösser wird. Dass man dafür das Quadrat wählt, ist rein willkürlich und diese Willkürlichkeit liegt in der Natur der Sache. Ohne die bekannten ausserordentlich grossen Vortheile, die die Wahl des Quadrats gewährt, könnte man jede andere jenen Bedingungen entsprechende wählen . . . ".

C. RADHAKRISHNA RAO
Indian Statistical Institute

Berkson raises several important issues in the theory of estimation, presumably based on his experience in using different methods of estimation in a bioassay problem. Such feedback from practical applications is extremely important for evaluating existing statistical techniques and making amendments and improvements, if necessary. Statistical inference has been a controversial subject and, depending as it does on inductive logic, will remain so. Any valid criticism of statistical methods will, no doubt, serve a useful purpose.

In advocating the method of minimum chi-square (MC) in preference to maximum likelihood (ML), Berkson makes some misleading statements and emphasizes some principles which may not be acceptable in all situations. One of the principles is *consistency* and another is *minimization of mean square error.* We shall examine how good these principles are through some examples.

Berkson gives an example where "ML did not recover the true value of the pertinent parameters even when the observations followed the assumed function exactly." He quotes Savage and Barnard in defense. Let $X = \theta + \varepsilon$ where θ is an unknown parameter and the error ε is such that $E(\varepsilon) = 0$ and $V(\varepsilon) = \theta^2$. According to Berkson X is a consistent estimator of θ since $X = \theta$ when $\varepsilon = 0$. Obviously $X/2$ is not consistent, but

$$E\left(\frac{X}{2} - \theta\right)^2 < E(X - \theta)^2 \tag{1}$$

uniformly for all θ. Then $X/2$ is better than X by the principle of minimum mean square error. The two principles seem to give contradictory results.

Consider independent random variables $X_i \sim N(\theta_i, 1)$, $i = 1, \cdots, p$. The unbiased estimator X_i of θ_i is consistent and the James-Stein estimator $t_i = X_i[1 - (p - 2)(\Sigma X_i^2)^{-1}]$ is not in Berkson's sense. But

$$E\Sigma(t_i - \theta_i)^2 < E\Sigma(X_i - \theta_i)^2 \tag{2}$$

uniformly for all $\theta_1, \cdots, \theta_p$. Which alternatives would Berkson recommend in the situations (1) and (2)?

Berkson defines Fisher consistency properly but applies it wrongly. Fisher's claim regarding the consistency of an ML estimator is the following. Let $\pi_1(\theta), \cdots, \pi_k(\theta)$ define a k-cell multinomial distribution and $0_1, \cdots, 0_k$ be the observed frequencies. If θ_0 is the true value, then

$$\Sigma\pi_i(\theta_0)\log \pi_i(\theta_0) \geqslant \Sigma\pi_i(\theta_0)\log \pi_i(\theta) \tag{3}$$

so that when $0_i \propto \pi_i(\theta_0)$, the ML estimate of θ is θ_0 and is, therefore, Fisher consistent.

Is the quadratic loss function appropriate in all situations? Let $X \sim N(0, \sigma^2)$. Then

$$E(X^2 - \sigma^2)^2 > E\left(\frac{X^2}{3} - \sigma^2\right)^2 \tag{4}$$

for all $\sigma^2 > 0$ so that $X^2/3$ is better than X^2 as an estimator of σ^2. Does it imply that $X^2/3$ is closer to the true value of σ^2 than X^2? For instance

$$\Pr(|X^2 - \sigma^2| < |3^{-1}X^2 - \sigma^2|) > 0.5 \tag{5}$$

for all σ^2. Should one prefer $X^2/3$ to X^2 as an estimator of σ^2?

How good is the principle of minimum asymptotic variance? Let $\bar{X}$ and X_m denote the average and the median of a sample of size n from $N(\theta, \sigma^2)$. Consider the Hodge type estimator

$$\begin{aligned} t &= \alpha X_m \quad \text{if} \quad |\bar{X}| \leqslant n^{-\frac{1}{4}} \\ &= \bar{X} \quad \text{if} \quad |\bar{X}| > n^{-\frac{1}{4}}. \end{aligned} \tag{6}$$

If α is sufficiently small $\bar{X}$ is inadmissible compared with t on the criterion of asymptotic variance. Should one prefer t to $\bar{X}$ as an estimator of θ in large samples?

How robust is the minimum mean square criterion? Consider the James-Stein estimator for $p = 3$. Moran pointed out (see Rao and Shinozaki, 1978) that although

$$E\Sigma_1^3(t_i - \theta_i)^2 < E\Sigma_1^3(X_i - \theta_i)^2 \tag{7}$$

uniformly for all $\theta_1, \theta_2, \theta_3$,

$$E\Sigma_1^3(t_i - \theta_i)^4 = \infty, \qquad E\Sigma_1^3(X_i - \theta_i)^4 = 9 \tag{8}$$

so that t_i stand in no comparison with X_i on the basis of a quadratic loss function. This is not an artificial example and shows the danger of pinning one's faith to a quadratic loss function. (The James-Stein estimator can be, however, modified to make the expectation of the fourth moment finite, but this is a different issue.)

Berkson has obviously misunderstood the criterion of second order efficiency. He seems to have relied on hasty opinions expressed by mathematical statisticians like Pfanzagl that the concept of second order efficiency (SOE) was not properly motivated. On the contrary, SOE is more basic than any other proposed criterion as it refers to the performance of an estimator when used as a substitute for the sample in drawing inferences on unknown parameters. One should not be led to believe that a problem is highly motivated simply because it is formulated in terms of quadratic loss function. No research worker engaged in a practical investigation or a business executive who makes decisions based on forecasts has ever specified his loss as a quadratic function of the error in an estimate. It is chosen by the mathematical statistician for convenience of mathematical investigation, and with the hope that an estimator with a smaller mean square error is closer to the true

value in some sense and is also robust with respect to a wider class of convex loss functions. As the examples considered show, such hopes are not always realized.

Fisher considered his information measure as a more intrinsic property of an estimator than its bias and variance, and suggested the choice of an estimator which preserves the maximum information. Fisher did not claim that the ML estimator has the maximum information in small samples, although in many cases he found that ML estimator fared better than others. He, however, suggested a method of examining the information in a statistic as the sample size $n \to \infty$, but did not develop the theory fully. Following Fisher's ideas, I introduced SOE to measure the amount of information up to terms of order $(1/n^2)$ and showed that in the case of the multinomial distribution the ML estimator contains the maximum information in large samples (Rao, 1961). In two later papers (Rao, 1962, 1963), I showed that SOE also provides a comparison of variances (up to terms of $O(1/n^2)$) of alternative estimators. The ML estimator scored better than the MC even with respect to asymptotic variance (up to terms of $O(1/n^2)$. The recent work of Efron, Ghosh and Subramanyam quoted in Berkson's paper has thrown further light on SOE and stressed its importance in statistical inference.

Berkson says that "Rao's putative proof, page 51, following Table 1, of the inferiority of MCE vis-a-vis MLE cannot be correct, since for the elementary binomial . . . the MCE's are identical with MLE's." I wish Berkson had not questioned the correctness of my expressions for SOE for various methods of estimation. In the case of the elementary binomial, the expression Δ in my table has the value zero so that ML and MC have the same SOE and there is no contradiction. ML is better than MC when $\Delta \neq 0$. For instance, if we have a multinomial distribution in $(k + 1)$ classes with probabilities specified by a binomial distribution with index k and unknown probability π, the ML and MC estimators of π differ in SOE. I hope Berkson would not prefer MC to ML estimator in this case.

Large sample theory cannot be brushed aside as irrelevant in practice. Generally a point estimate provided for use in a variety of situations in the place of an unknown parameter has to be very precise and a large sample may be necessary to achieve the desired precision. In such cases techniques which are efficient in large samples, and generally robust with respect to the choice of a loss function and prior information, have to be used. There may be situations where a point estimator has to be made from a small sample. In such cases one has to be careful in the choice of a loss function, use of prior information and selection of a particular technique.

Although Fisher expressed strong views against certain techniques, he advocated the use of different methods in different situations and mentioned the need to develop other methods of inference. Berkson might have chosen the appropriate technique in his problem, although his arguments based on consistency and minimum mean square error are not very convincing, as the examples considered in this note show. But in view of the general applicability of ML, its large sample properties and its superiority in small samples in a variety of situations, a better title to Berkson's paper might be "ML, sometimes MC" and *not* "MC, not ML."

REFERENCES

RAO, C. RADHAKRISHNA (1961). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Probability* **1** 531–546.

RAO, C. RADHAKRISHNA (1962). Efficient estimates and optimum inference procedures. *J. Roy. Statist. Soc. Ser. B* **24** 46–72.

RAO, C. RADHAKRISHNA (1963). Criteria of estimation in large samples. *Sankhyā Ser. A* **25** 189–206.

RAO, C. RADHAKRISHNA and SHINOZAKI, N. (1978). Precision of individual estimators in simultaneous estimation of parameters. *Biometrika* **65** 23–30.

REPLY TO DISCUSSANTS

BY JOSEPH BERKSON

Professor Efron is right in emphasizing the relatively broad applicability of maximum likelihood estimation. Minimum chi-square is, in general, limited to discrete or grouped data, though not always so. He also is right in saying that the MLE is sometimes difficult to compute. An example of a practical problem in which no algorithm for its computation has been put forward is dealt with by Ireland and Kullback (1968). These authors provide two minimum chi-square estimates, the minimum χ_I^2 and minimum χ_1^2.

However, the chief reason that the MLE is so widely advocated is that the academic and editorial establishments hold it to be virtually sacrosanct on principle. An example from my own experience is that when I submitted an article, later published (Berkson (1972)) advancing essentially the present viewpoint, it was found unacceptable unless I limited it to the MLE. Bradley Efron will recall this incident, since he was the editor of *JASA* at the time. Other statisticians have reported similar experiences.

He taxes me with being insufficiently appreciative of the inference school of statistics. I have to admit that I do not comprehend what inference is. In this I am at one with LeCam (1977), page 134. A recent formulation of its theory, Wilkinson (1977), discloses that it is irreconcilable with Kolmogorov, which will make it unacceptable to most mathematicians, and it is not subjectable to testing by a Monte Carlo experiment, which obviously excludes it from science.

With regard to Professor Ghosh's remarks, I respectfully assure him that I did not mean to provoke him with my title. As I explained in the text, I had observed that minimizing χ_λ^2 yields the estimating equations for the MLE, a fact that I do not know has been noted before. I argued that therefore the MLE can be thought of as a particular minimum chi-square estimator. This does not mean that every MLE can now be derived as a minimum χ_λ^2 estimate, as my title may have appeared to imply, and it is to be conceded that the title may have been too enthusiastic.

As for the MLE's failure to attain full sufficiency, Ghosh cites two examples with different values of the sufficient statistic that have the same value, infinity, for the

MLE. The examples cited in Berkson (1955), page 156, footnote, may be recalled. We deal with the logistic function, parameters $\alpha = 0$, $\beta = \text{logit } 0.99 \approx 4.595$, $y = \text{L.D.}_{50} = -\alpha/\beta$. The statistics $\Sigma n_i p_i$, $\Sigma n_i p_i x_i$ are jointly sufficient for α β, and for y, where p_i is the observed relative frequency among n_i exposed at dose x_i. In the 3 dose experiment with $n_i = 10$ at each, the following are the MLE's of β y for the class of samples 0, p, 1.

p_i at x_i			Sufficient statistic		MLE	
-1	0	$+1$	$\Sigma n_i p_i$	$\Sigma n_i p_i x_i$	$\hat{\beta}$	$\hat{y}$
0	0.1	1.0	11	10	∞	0
0	0.2	1.0	12	10	∞	0
0	0.3	1.0	13	10	∞	0
0	0.4	1.0	14	10	∞	0
0	0.5	1.0	15	10	∞	0
0	0.6	1.0	16	10	∞	0
0	0.7	1.0	17	10	∞	0
0	0.8	1.0	18	10	∞	0
0	0.9	1.0	19	10	∞	0
0	1.0	1.0	20	10	∞	0

The MLE's are not in one-to-one relation with the sufficient statistics. Considering that the probability of this set of samples is about 65 percent and a similar situation exists with other sample sets such as 0 0 p, p 1 1, p 0 0, 1 p 0, 1 1 p, the MLE loses a large fraction of the available information.

Professor Rao poses a number of conundrums that call into question use of the quadratic loss function as a definitive criterion of goodness of an estimator. I have not advocated what he ridicules. I have advanced the use of RBAN estimators which include the MLE, and among these the simplest to compute. If among these one has smaller mean squared error than the MLE, I favor that one and so do other statisticians as reflected in the applied literature.

I questioned Rao's proof that the minimum χ^2 estimate is inferior to the MLE, since for the elementary binomial and also for the unconstrained multinomial, these are identical. This was based on the text and Table 1 of Rao (1962). The table lists an index of inefficiency of the MLE as μ and that of the minimum χ^2 estimate as $\mu + \Delta$. Rao explains that for the elementary binomial $\Delta = 0$, so that the efficiency of the minimum χ^2 is also μ. Perhaps $\Delta = 0$ also for the elementary multinomial. Rao did not indicate anywhere in his text that Δ could equal 0 and I failed to explore this possibility.

In regard to consistency of the MLE, Fisher gave several definitions of consistency. By the definition quoted the MLE in the example cited is to be interpreted as inconsistent. Whether this interpretation is reasonable will have to be judged by the reader.

To say, as Mr. Pfanzagl does, that it is meaningless to compare the MLE with the minimum logit chi-square estimate without correcting for the bias, is incomprehensible to me. The mean squared error is, for biased estimators, a natural

extension of the variance as a measure of closeness for unbiased estimators. In any case, in the reported experiments Berkson (1955), the variances as well as the mean squared errors are smaller for the minimum logit chi-square estimator than for the MLE.

The bias corrected MLE is apparently not computable. However elegant the mathematics of Ghosh and Subramanyan is, until computed values of the modified MLE are provided, enabling their Monte Carlo testing, it remains pure mathematics, outside of the science of statistics. In science a proposition is considered meaningless if it is not testable experimentally, Bridgman (1928).

REFERENCES

BRIDGMAN, P. W. (1928). *The Logic of Modern Physics*. Macmillan, New York.

IRELAND, C. T. and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188.

LE CAM, L. (1977). A note on metastatistics or an essay toward stating a problem in the doctrine of chances. *Synthese* **36** 133–160.

WILKINSON, G. N. (1977). On resolving the controversy in statistical inference, with discussion. *J. Roy. Statist. Soc., Ser. B* **39** 119–171.