The Annals of Statistics
1976, Vol. 4, No. 3, 501–514

F. Y. EDGEWORTH AND R. A. FISHER ON THE EFFICIENCY OF MAXIMUM LIKELIHOOD ESTIMATION[1]

BY JOHN W. PRATT

Harvard University

F. Y. Edgeworth's 1908–9 investigation is examined for its contribution to knowledge of the sampling properties of maximum likelihood and related estimates, especially asymptotic efficiency. The nature and extent of his progress and anticipation of R. A. Fisher are described. Fisher's relevant work is briefly examined in relation to Edgeworth's and to the Cramér–Rao inequality.

1. Introduction. Francis Ysidro Edgeworth (1845–1926), the notable statistician (of the Edgeworth series) and economist (of the Edgeworth box), has been more noted by economists than statisticians. His work in mathematical statistics has been surveyed extensively by Bowley (1928) and, more briefly but more cogently for modern readers, by Pearson (1967). For broader sketches, see Hildreth (1968), who gives further references, or Kendall (1968).

In formal public discussions, Bowley (1935, with reference to 1928) and Neyman (1961; see also 1951) have said that R. A. Fisher's remarkable results on maximum likelihood estimation were considerably anticipated by Edgeworth (1908–9). On both occasions Fisher denied Edgeworth all credit without coming to grips with the central issue. Others grant Edgeworth a modest claim (Le Cam, 1953; Pearson, 1967) or almost none (Rao, 1961; Norden, 1972, citing Rao and Le Cam). L. J. Savage's (1976) interest stimulated me to look into the matter.

The questions at issue are primarily:

1. To what extent did Edgeworth derive the method of estimation he advocates (which coincides with maximum likelihood) via sampling theory (direct probability) as well as via inverse probability?

2. However he derived it, did he advance the idea that it was a general method with desirable sampling-theory properties, especially asymptotic efficiency?

3. How far did he go toward proving such sampling-theory properties?

Of the commentators mentioned, only Bowley gives explicitly a positive answer to question 1. It is of interest in part because the current expert view (Edwards, 1974) is that Fisher (1912) introduced the method of maximum likelihood as a new concept in the modern era, the few 18th century uses of it having been virtually ignored during more than 100 years of disuse.

Received November 1975.

[1] Research supported by the Associates of the Harvard Business School.

AMS 1970 *subject classifications.* Primary 62-03, Secondary 62F20.

Key words and phrases. F. Y. Edgeworth, R. A. Fisher, asymptotic efficiency, maximum likelihood, M-estimators, Cramér–Rao inequality, history.

Furthermore, some seem to feel that a positive answer to question 1 greatly enhances any contribution Edgeworth made in the direction of questions 2 and 3, or even is prerequisite to recognition of it. The questions are not easy to answer because Edgeworth starts from and emphasizes inverse probability and because his own obscurity is compounded by the absence at the time of standard notation and terminology systematically distinguishing sample and population values and direct and inverse probabilities.

Section 2 describes the relevant aspects of Edgeworth's 1908–9 work, as objectively as possible and including some details, because there is general agreement that he writes obscurely and because his contribution has never been fully presented in current terms.

As the related work by Fisher is far better known and far more accessible in every way, Section 3 merely summarizes its relationship to Edgeworth's, with comments and judgments interposed where convenient.

Section 4 gives my conclusions on the questions at issue, along with various remarks, including some on the relation of Fisher's work to the Cramér–Rao inequality.

2. Edgeworth's 1908–9 paper. This is an ambitious work, in four installments. As well as new material, it presents Edgeworth's view of estimation theory as it then stood, heavily through examples, and with various philosphical and technical discursions. I will not attempt to outline the paper as a whole, but will discuss the parts relevant to the question: How much did Edgeworth conjecture about the sampling properties of maximum likelihood and maximum probability estimates, and how much did he succeed in proving? (When quoting, I will silently make insignificant changes of notation and omit Edgeworth's plentiful footnotes.)

2.1. *The* 1908 *installments.* Edgeworth deals throughout with a sample (usually large) of independently, identically distributed observations. The method of estimation he uses in the three 1908 installments, which he calls "the genuine inverse method," and says is "familiar" from "hosts of references" (page 384), is to find the posterior mode for a uniform prior distribution (the joint mode if there are several parameters, page 393). He says (page 387), "· · · I submit that very generally we are justified in assuming an equal distribution of *a priori* probabilities over that tract of the measurable with which we are concerned. And even when a correction is required by what is known about the *a priori* probability, this correction is in general, as I have elsewhere shown, of an order which becomes negligible as the number of the observation is increased." He notes (page 392) that uniform prior distributions for a parameter and for functions of a parameter are inconsistent but says, presumably referring to irregularity at 0 (or ∞), "· · · the objection is only serious when the [parameters] are extraordinarily small (or large), and the functions not those occurring in ordinary practice." I find no whiff of "likelihood" or any other novelty in Edgeworth's *definition* of the estimates in the 1908 installments, but they are identical to maximum likelihood estimates.

Early on, while discussing some univariate and bivariate normal examples, Edgeworth shows that posterior distributions are approximately normal (pages 389–391, citing Gauss and Laplace), and since, for a normal distribution, the mode is "most advantageous" for any "detriment" (loss function) of a rather general type (page 386), it follows that the posterior mode approximates the a posteriori "most advantageous" estimate generally (pages 385–387, 391, 504).

Edgeworth sets up (page 499) a general multivariate model of given form with unknown "*primary* constants, the averages (mode, arithmetic mean, etc.) of the [variables]" (location parameters) and other unknown "*secondary* frequency-constants." Their most probable values and probable errors are to be determined. He refers to Karl Pearson [and L. N. G. Filon, 1898] for discussion and illustration of this general problem. He begins some murky "following reflections· · ·largely suggested by [their] 'general theorem' on the probable errors of a system of frequency-constants," and also by Laplace, by restricting consideration to the location-parameter model $f = f(x - \theta)$. Along the way, he notes, for a purpose which wouldn't seem to require it, that, in large samples, the derivative of the log likelihood at the true θ is approximately

$$n \int f \frac{d}{dx} \log f \, dx = n \int \frac{df}{dx} \, dx = 0$$

and the log likelihood decreases from its value at the true θ approximately quadratically with coefficient $\frac{1}{2}n \int (1/f)(df/dx)^2 \, dx$, provided $df/dx = 0$ at the extremes of the range. (Of course he doesn't use the term likelihood, and he doesn't mention the factor $\frac{1}{2}$.)

Soon he comes (page 505) to "· · ·the question whether the probable error of a frequency-constant determined by the proper inverse method is necessarily smaller than the probable error incident to some other methods." He says, "The compared method must, of course, be 'other,'" and gives examples where it coincides ("hitting on one aspect of sufficiency," S. M. Stigler commented in correspondence). Then he says (page 506):

> Certainly it is very natural to associate increased accuracy in the ordinary sense of the term with increased *precision* in a technical sense. But the following objection occurs. Grant that the increased knowledge obtained by taking account of all the data affords a more accurate determination of the probabilities that the observed event, the given set of observations, should have resulted from each of the possible causes, the different values of the *quaesitum*. But what if, in this corrected distribution of probabilities, the outlying causes as distinguished from the central become relatively

> more probable; and accordingly the 'spread' of the curve of frequency for the values of the *quaesitum* is increased!

This paragraph is quoted because it could be read as narrowing the question. It says that the posterior distribution based on an insufficient statistic might be more peaked than that based on all the data (which can happen, though not for all samples) and hence, conceivably, more precise, but for the sense of "precise" we must rely on the context. The next paragraph is central. It is quoted with my interpretations in brackets:

> The objection is specious only while there is ignored what is known about the applicability of the normal curve. The following answer may suffice. Consider any particular set of observations, $x_1, x_2, \cdots, x_n$, forming one of a series of sets, such as are encountered in practice. The probability that any particular point should be the true one is given by inversion proper as above; the most probable value being$\cdots$say $\phi(x_1, x_2, \cdots, x_n)$; and the probabilities of other points being disposed about that maximum in conformity with a normal curve. [That is, the posterior distribution is (approximately) normal with center ϕ.] Now consider some other formula, some other function of the observations known to coincide with the true value of the *quaesitum*, in the long run formed by a series of sets; e.g., $(x_1 + x_2 + \cdots + x_n)/n$, the Arithmetic Mean. [The condition sounds more like sampling-theory consistency than anything else.] The point designated by this formula being generally different (for any particular set) from $\phi(x_1, x_2, \cdots, x_n)$, the centre of the normal curve assigning the frequency with which each point is the true value of the *quaesitum*; it follows that in the long run formed by the different originations of the particular set [a posteriori], the mean square of deviation from the true point for $(x_1 + x_2 + \cdots + x_n)/n$ is greater than the mean square of deviation for $\phi(x_1, x_2, \cdots, x_n)$. The like is true of any other particular set. [That is, for any sample, ϕ has smaller (or equal) posterior mean square error.] It is therefore true for the whole series [marginally over prior and sampling distributions] that the Mean Square of deviation from the true point, and accordingly the probable error, is less for the formula given by inversion proper than it is for the Arithmetic Mean, and, by parity of reasoning, for any other rival method, say, $\chi(x_1, x_2, \cdots, x_n)$. If then* we take numerous sets of observations, each set numbering n, and form for each set the value ϕ and also

> χ, while both series—that of the ϕ's, and that of the χ's—will fluctuate according to a normal law of frequency, the probable error for the ϕ's will be less than what it is for the χ's.
>
> * If any hesitation is felt as to the connection between this and the preceding statement, it may be removed by the following illustration. Imagine a long line of soldiers shooting bullets at a wall-shaped target parallel to the long line; each man aiming at a point on the target straight in front of him. The deviation, measured horizontally, of the bullets fired by each man, from the point he aimed at, obeys the same law of frequency, namely a normal error-curve with one and the same probable error. Considering any particular shot-mark on the target, let us determine by inverse probability the most probable position of the man that fired that shot. The probable error affecting this determination is the same as the probable error shown by the dispersion of the bullets fired by any particular man (supposing that the distance between two adjacent men is small, compared with the probable error in question). In this parable a single shot stands for a combination of n observations, such as $\phi(x_1, x_2, \cdots, x_n)$, or $\chi(x_1, x_2, \cdots, x_n)$; each of which is known, by the Law of Error, to fluctuate according to a normal law of frequency.

Edgeworth doesn't say explicitly that the last statement quoted before the footnote refers to a fixed true value, and hence is a sampling-theory statement, but that is the only possibility remaining. Furthermore, the beginning of the footnote definitely suggests that the statement is new, not a restatement, and the rest of the footnote confirms this impression. Thus, the statement appears to be that ϕ has smaller mean square error, in a sampling-theory sense, than any rival χ (in large samples), i.e., is asymptotically efficient. Unfortunately, the statement does not follow, even at the level of rigor appropriate here. It does follow in the translation case, which Edgeworth is considering, if χ is translation invariant, which he may be tacitly assuming. (The argument generalizes easily to invariant procedures in problems with suitable group structure—perhaps too easily in view of difficulties with unbounded Haar measures.) Edgeworth buries, if he gives at all, the conditions needed for his argument, but at least he is sufficiently unsure of it at this point to welcome proofs of special cases (see below).

Edgeworth restates the proposition in an Appendix (page 662):

> Let $x_1, x_2, \cdots, x_n$ be a set of n observations forming a random selection from the indefinitely large group of the

> observations ranging under the given frequency curve.... Then, if we take (at random) a series of sets, such as—
>
> $$\begin{matrix} {}_1x_1 & {}_1x_2 & \cdots & {}_1x_n\,, \\ {}_2x_1 & {}_2x_2 & \cdots & {}_2x_n\,, \\ \vdots & \vdots & & \vdots \\ {}_mx_1 & {}_mx_2 & \cdots & {}_mx_n\,; \end{matrix}$$
>
> and form for each set the corresponding value of ϕ, the series of mean values thus formed—say, ${}_1\phi, {}_2\phi, \cdots, {}_m\phi$— will be such that ($m$ and n being large numbers) the mean square of their deviation from the true point, say θ, viz.,
>
> $$\frac{({}_1\phi - \theta)^2 + ({}_2\phi - \theta)^2 + \cdots + ({}_m\phi - \theta)^2}{m},$$
>
> will be less than the mean square of deviation presented by any other set of mean values ${}_1\chi, {}_2\chi, \cdots, {}_m\chi$, each formed from a set of n observations, where χ (like ϕ) is a symmetrical function of observations, having the properties of an average.

This again seems clearly to refer to a sampling property (as pointed out by Neyman, 1961). Indeed, it introduces explicit notation for repeated samples yet nowhere mentions varying θ.

Edgeworth notes (page 507, page 667) that application of the general proposition to examples yields specific inequalities which he thinks might be hard to prove directly. He gives direct proofs of several (pages 664, 667, 81–82), all based on Schwarz's inequality, the first being due to Love[2] and inspiring the others. I omit them because the translation cases are covered by a more general case discussed below. Of course we now know Schwarz can do them all in the same way (by the usual proof of the Cramér–Rao inequality).

On pages 667–668, Edgeworth says, "The fundamental principle, it will be remembered, is even wider than the theorem [just cited] to which the examples have been referred. The general principle applies not only to the direct probability that the result of observations which are about to be made will be such or such, but also to the inverse probability that the origin of observations which have been made was such or such. Indeed, it is the inverse side which seems to form the easiest approach to the general proof." He then restates more fully the argument given in the footnote quoted earlier. All this further confirms that he identified and believed in the asymptotic efficiency of the "genuine inverse method" from the sampling-theory viewpoint as well as a posteriori.

[2] Augustus Edward Hough Love, 1863–1940, Sedleian Professor of Natural Philosophy at Oxford, wrote articles on Calculus of Variations, etc. in Encyclo. Brit. 11th ed. See *Nature* (1929) **123** 850; *J. London Math. Soc.* (1941) **16** 69–80 (art. by E. A. Milne); *Obituary Notices of Fellows of the Roy. Soc.* (1941) **3** 467–482. (Information supplied by S. M. Stigler.)

In passing (pages 677–678) Edgeworth derives the asymptotic efficiency of the method of moments for a Pearson Type III location parameter when the other parameters are known.

2.2. *The* 1909 *Addendum.* The foregoing would, I believe, justify more credit to Edgeworth than some have granted him, but there is more. Edgeworth's 1909 Addendum (discussed by Bowley, 1928) contains an important new approach and proof. After giving one of the Schwarz-inequality examples, Edgeworth says (page 82):

> The general theorem which comprehends the preceding proposition as a particular species may itself likewise be proved by a direct method free from the speculative character which attaches to inverse probability. For this purpose we no longer take our stand on a particular set of observations in order thence to remount a posteriori to the originating cause; rather we watch in prior experience the distribution of observations and determine that function of which the several values, each formed from a large set of observations, hover with minimum dispersion about the true value of some constant represented by a symmetrical function of the observations.

Thus, Edgeworth plans to choose among estimators by directly considering their sampling distributions. He thereupon (pages 82–83) introduces the class of estimators $\hat{\theta}$ satisfying an equation of the form

$$(1) \qquad \textstyle\sum_t k(x_t - \hat{\theta}) = 0$$

for some function k, nowadays called M-estimators, and gives quite a clear proof, of at least Fisherian rigor, that if the density is $f(x - \theta)$ and if $\hat{\theta}$ has asymptotic mean θ, then $\int kf = 0$ and $\hat{\theta}$ is asymptotically normal with variance V/n where

$$(2) \qquad V = (\textstyle\int k^2 f)/(\int k' f)^2 .$$

(He loses the n at this point.)

He then proceeds (pages 83–84) to show, equally clearly and rigorously but more cumbersomely, that V can be minimized with respect to k only for $k = A\, f'/f$, and he verifies by what amounts to a second proof (page 84) that this gives a minimum, not some other kind of stationary value. (He doesn't notice that one step in the verification itself, namely the last inequality on page 84, contains the result originally desired.)

Now (1) with $k = Af'/f$ is the equation for the maximum likelihood estimate, which Edgeworth has thus derived entirely within the sampling-theory framework as having minimum asymptotic variance among consistent estimates of the form (1), i.e., among consistent M-estimates, in the location case. He considers (1) "a sufficiently general type of those symmetrical functions of the observations with which we are concerned," except percentiles and the like, arguing that such functions must be at least approximately of the form (1) to be asymptotically normal (pages 85–87, with references to other papers of his). In other words, he purports to show something like Best Asymptotic Normality for the maximum likelihood estimate in the location case, but I can't vouch for his proof beyond M-estimators.

Edgeworth then discusses scale parameters along similar lines (pages 87–90). In place of (1), however, he uses $\sum k(x_t) = \int k(x)f(x/\theta)\,dx/\theta$, which is not exactly analogous and does not yield the maximum likelihood estimate as a special case, though is apparently serves the same purpose in his argument for Best Asymptotic Normality.

Edgeworth concludes (page 90):

> Analogous reasoning is applicable to other cases: the case which is compounded of the two preceding, the case in which a combination of observations in two dimensions stands for a coefficient of correlation, and more complicated cases relating both to primary and secondary constants.
>
> Universally the combination of (n) observations that is subject to least fluctuation proves to be the same function whether sought *a posteriori* or by way of prior experience. It may be said that this identity is evident to those at least who are conversant with the first principles of Probabilities. Yet, as far as I know, the theorem has not been clearly stated by the writers who have dealt with the Method of Least Squares. Nor, I think, is it so self-evident—considering how often appearances in Probabilities prove fallacious—as to render demonstration superfluous; especially as the general theorem comprehends particular propositions which are far from self-evident.

3. Fisher's related work.

This section aims to describe briefly the parts and aspects of Fisher's work on maximum likelihood estimation which are relevant to the question of how much Edgeworth anticipated him. In no sense does it attempt to summarize all of Fisher's work, even on MLE, and the focus on comparison with Edgeworth naturally shifts emphasis even in the domains touched on. Since it must be just a selective sketch, and since rereading Fisher is encouraged anyway (Savage, 1976), my personal judgments and reactions have not been kept separate.

3.1. *Fisher's papers of* 1912, 1922, 1925 *and* 1935.

The relevant work is contained in four of Fisher's papers. His other papers and books add essentially nothing for the purpose at hand.

In a curious paper of 1912, Fisher notes that other methods have an unconvincing basis and anomalous properties, and proposes maximum likelihood with breathtaking assurance but almost no justification. He mentions no sampling properties, asymptotic or otherwise. (As a first publication by a 22-year-old, the paper is remarkable in maturity and for some hints of later work, but its actual content I am at a loss to assess. Perhaps its main contribution is to introduce the concept of likelihood as relative probability, of interest in its own right, not to be integrated like inverse probability. The terms "likelihood" and "maximum likelihood" are not used until later. See Edwards, 1974.)

In a monumental paper of 1922, which lays out the problem of estimation and related concepts, Fisher clearly thinks the maximum likelihood estimate is asymptotically efficient. His reasoning is that he thinks it is always sufficient (pages 323, 330, 331, 367, ranging from "I believe" to "A proof is given"), and sufficient estimates are asymptotically efficient, as he shows (page 317). His "proof" of the sufficiency of the maximum likelihood estimate (pages 330–331) is of more than Edgeworthian obscurity, as it would have to be. He derives the asymptotic distribution of the maximum likelihood estimate (pages 328–329). He gives extensive studies of a variety of examples, including the efficiency of the method of moments for all the Pearson type III parameters, with nuisance parameters known and unknown.

An important paper of 1925 contains Fisher's first real proof of the efficiency of maximum likelihood. He inserts the assumption (page 707) that the maximum likelihood estimate is asymptotically normal with variance $\propto 1/n$, or (page 710) that an asymptotically efficient estimate exists (i.e., one minimizing n times the asymptotic variance among asymptotically normal and unbiased estimates with asymptotic variance $\propto 1/n$). Apparently he had recognized that the maximum likelihood estimate is not always sufficient (pages 714, 718) and was unsure what assumptions he needed. His proof is valid, at this level of rigor, but seems rather unilluminating to me. A better proof is easy to obtain (see Section 4.1 below) from the information-reduction result he proves later in the paper, namely that no statistic has greater Fisher-information than the sample. Indeed the efficiency of maximum likelihood is implicit in his use of this result, but he doesn't point out that it provides another proof, although the properties of information are a major concern and contribution of the paper.

Fisher's 1935 Royal Statistical Society invited survey has a different proof of the efficiency of maximum likelihood, this one incorporating (page 43) a proof of the information-reduction result without explicit mention of it, though it is stated later without proof (page 47). At the end of the proof, in order to show that maximum likelihood achieves the lower bound (page 44, last sentence; page 46, paragraph 1, last sentence) he considers solutions of

$$\sum_t k(x_t, \hat{\theta}) = 0 \tag{1'}$$

and shows, provided $\int kf = \int k(x, \theta) f(x, \theta)\, dx = 0$ to make $\hat{\theta}$ consistent, that $\hat{\theta}$

is asymptotically normal with variance V/n where

$$V = \left(\int k^2 f\right) \Big/ \left(\int \frac{\partial k}{\partial \theta} f\right)^2 . \tag{2'}$$

He then shows that (2′) is minimized by the maximum likelihood estimate. His proof of (2′) and its minimization by maximum likelihood are essentially streamlinings of Edgeworth's at (2) above, improving parts but leaving others less complete. (Actually Fisher considers discrete distributions, Edgeworth continuous. Fisher's generalization from (1) to (1′) is nice, but turns out to make no real difference to the mathematics.) These results seem superfluous, since Fisher already knew the asymptotic variance of the maximum likelihood estimate and had shown that it was a lower bound. Strangely enough, however, he attached considerable importance to them, not only at this time (passages cited, and inclusion in this summary paper), but also earlier (1928, pages 97–98, where the proof is even more abridged and the results serve another purpose) and later (1956, pages 148, 157; see also 1950, page 11.699b).

All Fisher's proofs of efficiency are for one-parameter models. To my knowledge, he never indicated what would be involved in extending any proof to several parameters, though in 1922 he was already clear and confident about the result.

3.2. *Fisher's references.* Fisher relates the estimators obtained by maximum likelihood to earlier estimators, especially Gauss's (see Savage, 1976), and he provides a long footnote (1922, page 329) on formulas for asymptotic variances emphasizing Karl Pearson's confusion about the method of moments. Except for the latter, however, Fisher gives essentially no references to anyone else for the idea or properties of maximum likelihood, either in the papers cited here or elsewhere.

When others have cited Edgeworth's contribution, Fisher's replies have not clarified the issue. The substance of his reply to Bowley (1935) is a rehearsal of the history of formulas for asymptotic variance "presumably known to Edgeworth, writing in 1908 and 1909" (see above). He also does not recognize so much as a contribution by Edgeworth, let alone a reasonable claim to considerable priority, then or even much later (1956, loc. cit.; 1961, see below).

When Neyman (1961) credited Edgeworth with proposing the method of maximum likelihood, conjecturing its asymptotic efficiency in a sampling-theory sense, and realizing that some sort of restriction on the alternative estimators is necessary, Fisher's complete reply was:

> Mr. Neyman surprised many of us by his claim in his recent memorandum that Edgeworth introduced the Method of Maximum Likelihood. Edgeworth in fact bound his method on the theory of inverse probability and ascribed his notion to K. Pearson and Filon in 1898; the Method

> of Maximum Likelihood may equally be found in this paper, only Pearson and Filon were under the misapprehension that the errors of random sampling were the same as those of the Method of Moments regarded as axiomatic by these authors.
>
> Edgeworth, however, ends his paper with the reservation that all he had said referred only to Measures of Central Tendency and not to the more complex problem of "The Fluctuation."

In fact, Edgeworth did not bind his method on the theory of inverse probability. If he was too generous (unlike Fisher) to Pearson and Filon, ascribing ideas to them which were really his own or doing correctly what they did wrong, this cannot strengthen Fisher's own claim to priority. If the method of maximum likelihood and its asymptotic distribution precede Edgeworth, why does Fisher so often seem to claim them for himself? And what about the efficiency of maximum likelihood, which is deeper and newer and hence a far more important question?

The nearest I have found to the reservation Fisher mentions is at the end of only an intermediate section (page 674). There Edgeworth says

> The proofs which I have offered for the superiority of the inverse method in determining a parameter relate only to the general principle and particular examples, to the *summum genus* and *infimae species*. I still desiderate a *specific* proof, of the kind which Professor Love has supplied in the case of a primary constant, for theorems like the following: that (under the assigned conditions) [a formula follows]· · ·

The italics are Edgeworth's and emphasize how far Edgeworth was from feeling that nothing he had said referred to the problem of "The Fluctuation."

4. Conclusions and remarks. Unquestionably parts of Edgeworth's 1908–9 work are highly related to all Fisher's work cited above. The most specific connections, perhaps, are the asymptotic efficiency of the method of moments for a Pearson Type III location parameter; the asymptotic distribution of M-estimators of location parameters, maximum likelihood included, if older references are considered inadequate; and most strikingly, the minimization of the asymptotic variance of M-estimators. More important however, is the conjecture of asymptotic efficiency and progress toward proving it.

To answer the questions raised in Section 1, in the translation case, Edgeworth actually derived the method of maximum likelihood (without the name or its connotation) via direct as well as inverse probability (by the minimization just mentioned). He was convinced of its sampling-theory asymptotic efficiency in general (like Fisher in 1922). He adduced enough evidence for it and made enough progress toward proving it to deserve very considerable credit.

Thus I believe that Edgeworth anticipated Fisher more than most commentators suggest (and the neglected 1909 Addendum is an important part of the evidence). If Edgeworth's contribution is minor, what of Fisher in 1922, thirteen years later? Fisher's view of the problem was clearer and grander, but his conjecture was the same and his 1922 "proof" was only an invalid gesture.

Fisher's great advance was to give a general proof of efficiency (in 1925, for one parameter, and in the senses of "general" and "proof" relevant at the time), and of course to introduce and explore very fruitful related concepts. Fisher's "derivation" by maximizing the likelihood is also valuable to whatever extent "likelihood" is a free-standing concept, meaningful in itself. But once good sampling properties are proved, the mode of derivation should be unimportant anyway (from a "classical" point of view). If Edgeworth or Fisher had proved the sampling-theory asymptotic efficiency of all posterior modes for smooth, nonvanishing prior densities, not merely maximum likelihood estimates, would not that have been still better? And of course Edgeworth's treatment of inverse probability does not diminish his contribution on the direct side, but only its visibility.

This is not at all to deny that Fisher's contributions to this subject, as to many others, were tremendous. Nor can one blame a man busy making giant progress on several fronts for not taking the time to trace the forerunners or sources of his ideas or even for being unaware of them. One can, however, note, with regret if not blame, that from his writing one will not learn and may be misled about even the existence of his intellectual debts (Neyman, 1951). This makes it difficult to attain proper perspective and realize continuity of development of the subject (Pearson, 1967).

How much of Edgeworth's paper did Fisher read? He quotes from the first part in another connection (1936, page 248). It is common practice to read others' work just enough to absorb what seems useful without concern for remembering sources, and not unusual later to mistake an absorbed idea for one's own. My conjecture (and no more) is that Fisher so read Edgeworth's paper at some time, including the Addendum. If Fisher was less than averagely concerned with seeking references when writing, one must expect the same when others pointed out connections.

4.1. *Fisher and information.* Fisher's proofs of the asymptotic efficiency of maximum likelihood do not clearly exploit his remarkable vision of the problem. In 1925, Fisher gave simple demonstrations that information is additive and that statistical reduction cannot increase it. His 1935 paper incorporates the latter in a proof of asymptotic efficiency and later states it. Yet if he had stated it first, he could then have argued as follows.

Any statistic T has information $\leqq ni$ where i is the information per observation. Furthermore, the variance of a normally distributed statistic equals the reciprocal of its information about the mean. Therefore:

A. If T is asymptotically normal with mean θ, then asymptotically its variance is at least $1/ni$, which maximum likelihood achieves. This is the efficiency of maximum likelihood.

B. If T has mean θ, regardless of normality, then the mean of m independent T_i is asymptotically normal with mean θ, whence $\operatorname{Var}\{\sum_1^m T_i/m\} \geqq 1/mni$ asymptotically as $m \to \infty$ for n fixed, whence $\operatorname{Var}\{T\} \geqq 1/ni$. This is the Cramér–Rao inequality.

Fisher seems not to have come out with either A or B. A is now common understanding. I haven't registered seeing the Cramér–Rao inequality deduced from the information-reduction theorem as in B, nor the information-reduction theorem from the asymptotic efficiency theorem. It would be good to have a way of viewing all three which makes their connection more apparent and less dependent on a trick.

Acknowledgments. For helpful comments, I am very grateful to M. G. Kendall and D. L. Wallace, and especially to S. M. Stigler.

REFERENCES

BOWLEY, A. L. (1928). *F. Y. Edgeworth's Contributions to Mathematical Statistics.* Roy. Statist. Soc., London.

BOWLEY, A. L. (1935). Discussion of Fisher (1935). *J. Roy. Statist. Soc.* **98** 55–57.

EDGEWORTH, F. Y. (1908–9). On the probable errors of frequency-constants. *J. Roy. Statist. Soc.* **71** 381–97, 499–512, 651–78. Addendum *ibid.* **72** 81–90.

EDWARDS, A. W. F. (1974). The history of likelihood. *Internat. Statist. Rev.* **49** 9–15.

FISHER, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messeng. Math.* **41** 155–160.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368. (Reprinted in Fisher (1950)).

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725. (Reprinted in Fisher (1950)).

FISHER, R. A. (1928). On a property connecting the χ^2 measure of discrepancy with the method of maximum likelihood. *Atti Congr. Internaz. Mat., Bologna* **6** 95–100. (Reprinted in Fisher (1950)).

FISHER, R. A. (1935). The logic of inductive inference (with discussion). *J. Roy. Statist. Soc.* **98** 39–82. (Reprinted in Fisher (1950)).

FISHER, R. A. (1936). Uncertain inference. *Proc. Amer. Acad. Arts Sci.* **71** 245–258. (Reprinted in Fisher (1950)).

FISHER, R. A. (1950). *Contributions to Mathematical Statistics.* Wiley, New York.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh; 3rd ed. (1973). Hafner, New York.

FISHER, R. A. (1961). Discussion of Rao (1961). *Bull. Inst. Internat. Statist.* **38** No. 1 200–201.

HILDRETH, CLIFFORD (1968). Edgeworth, F. Y. *Internat. Encycl. Social Sci.* **4** 506–509.

KENDALL, M. G. (1968). Studies in the history of probability and statistics. XIX. Francis Ysidro Edgeworth (1845–1926). *Biometrika* **55** 269–275.

LE CAM, LUCIEN (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. California Publ. Statist.* **1** 277–329.

NEYMAN, JERZY (1951). Review of *Contributions to Mathematical Statistics* by R. A. Fisher. *Scientific Monthly* **62** 406–408.

NEYMAN, JERZY (1961). Discussion of Rao (1961). *Bull. Inst. Internat. Statist.* **38** No. 1 193–200.

NORDEN, R. H. (1972-3). A survey of maximum likelihood estimation. *Internat. Statist. Rev.* **40** 329-354, **41** 39-58.

PEARSON, E. S. (1967). Studies in the history of probability and statistics. XVII. Some reflections on continuity in the development of mathematical statistics, 1885-1920. *Biometrika* **54** 341-355.

PEARSON, KARL and FILON, L. N. G. (1898). Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philos. Trans. Roy. Soc. London Ser. A* **191** 229-311.

RAO, C. RADHAKRISHNA (1961). Apparent anomalies and irregularities in maximum likelihood estimation. *Bull. Inst. Internat. Statist.* **38** No. 4 439-453. Discussion *ibid.* No. 1 193-214. (Reprinted (1962) in *Sankhyā Ser. A* **24** 73-101.)

SAVAGE, LEONARD J. (1976). On rereading R. A. Fisher. *Ann. Statist.* **4** 441-500.

GRADUATE SCHOOL OF BUSINESS ADMINISTRATION
HARVARD UNIVERSITY
SOLDIERS FIELD
BOSTON, MASSACHUSETTS 02163