Statistical Science
1997, Vol. 12, No. 3, 162–176

R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922

John Aldrich

Abstract. In 1922 R. A. Fisher introduced the method of maximum likelihood. He first presented the numerical procedure in 1912. This paper considers Fisher's changing justifications for the method, the concepts he developed around it (including likelihood, sufficiency, efficiency and information) and the approaches he discarded (including inverse probability).

Key words and phrases: Fisher, Pearson, Student, Bayes's postulate, inverse probability, maximum likelihood, sufficiency, efficiency, information.

INTRODUCTION

The making of maximum likelihood was one of the most important developments in 20th century statistics. It was the work of one man but it was no simple process: between 1912 and 1922 R. A. Fisher produced three justifications—and three names—for a procedure that existed in two forms. [Pratt (1976) shows that Edgeworth (1908) anticipated a good part of the 1922 version, but nobody noticed until a decade or so after Fisher had redone it.]

The "absolute criterion" of 1912 was derived from the "principle of inverse probability." The "optimum" of 1921 was associated with the notion of "likelihood" as a quantity for appraising hypothetical quantities on the basis of given data. The "maximum likelihood" method of 1922 gave estimates satisfying the criteria of "sufficiency" and "efficiency." There were two forms for sometimes Fisher based the likelihood on the distribution of the entire sample, sometimes on the distribution of a particular statistic.

The making of maximum likelihood was the making of the world of ideas and nomenclature including "parameter," "statistic," "likelihood," "sufficiency," "consistency," "efficiency," "information" —even "estimation." It was the unmaking of the method's inverse probability connection and (for Fisher) the unmaking of inverse probability as well as of the related "Bayes postulate."

John Aldrich is a member of the Department of Economics, University of Southampton, Southampton, SO17 1BJ, United Kingdom (e-mail: jca1@soton.ac.uk).

Fisher did not finish with maximum likelihood in 1922 (more was to come in 1925 and 1934), but the method of Cramér and Wald, of countless applications and textbook expositions, had arrived. The 1921 version also has a continuing place in Fisher's writings and elsewhere—as likelihood analysis.

This paper follows the development of the ideas of Fisher alone but the syllabus extends beyond his writings. His problems came out of the work of Karl Pearson and "Student" (W. S. Gosset). Their writings belong in the syllabus. [E. S. Pearson put them there and wrote very well about them; see, e.g., E. S. Pearson (1968, 1990). Box, Edwards, MacKenzie and Zabell have also discussed them.] So does the least squares teaching to which Fisher was exposed. This material presents a complicated and confusing scene. Pearson promoted at least four estimation methods, and Gosset found sampling distributions in the cause of Bayesian analysis while the least squares texts reproduced disconnected and mutilated bits of Gauss. The classics discussed by Stigler (1986) are not part of the syllabus, for Fisher did not know them.

1. THE ABSOLUTE CRITERION

Fisher published "An absolute criterion for fitting frequency curves" (Fisher, 1912) as a third-year undergraduate. [For biographical information on Fisher see Box (1978).] It is a very implicit piece of writing and, to make any of it explicit, we have to read outside the paper *and guess*. We start with the usual questions: where is the author coming from; to whom is the paper addressed; what does it say?

The author comes out of the theory of errors, a speciality of astronomers and surveyors. Fisher mentions Chauvenet's *Spherical Astronomy* and

T. L. Bennett's tract on the theory of errors. The only other reference is to his tutor, "to whose criticism and encouragement the present form of this note is due." F. J. M. Stratton, an astronomer, lectured on the theory of errors. He wrote on agricultural statistics and was in touch with Gosset. [See E. S. Pearson (1990, p. 47) for information on Stratton.] Brunt's textbook (Brunt, 1917), based on Stratton's lectures, shows what Fisher had to work with. [See Edwards (1974, 1994); Edwards brought this book into the Fisher Syllabus.] It lacks theoretical finesse, and important constructions are mutilated or left out. Least squares is justified following Gauss's *Theoria Motus* (Gauss, 1809), but omitting the prior! By 1930 Fisher (1930, p. 531) knew the original and by 1934 he knew (Fisher, 1934a p. 616) the "Gauss–Markov" justification of 1821; in 1912–1922 he seems to have known only bowdlerized Gauss. Karl Pearson's knowledge was no better [see, e.g., Pearson (1920) and Section 17 below]. However, Edgeworth (1908) knew more.

Fisher's title refers to "frequency curves" but he begins by criticizing least squares and the method of moments as general methods of curve-fitting. There are no references but the text reads like a critique of Pearson's "On the systematic fitting of curves to observations and measurements" (Pearson, 1902) which treats these methods. [Brunt refers to this work. Edwards (1994, Appendix 2) reproduces Fisher's notes on it. These were written much later and do not shed any light on my conjecture that the piece was the target of Fisher's criticism.] Fisher already knew of Pearson and his biometric program. In his 1911 essay "Mendelism and biometry" (Bennett, 1983, p. 56) he described Pearson as the one "on whose mathematical work the whole science of biometrics has been based."

Pearson thought that the method of moments was more easily applied than least squares; Fisher (Fisher, 1912, pp. 155–156) concurred. [Fisher's notes, in Edwards (1994, Appendix 2), are an interesting commentary on these claims and Fisher's earlier acceptance of them.] Pearson (1902, p. 267) judged that it gave "sensibly as good results" as least squares though admittedly the definition of best fit is "more or less arbitrary". He made a virtue out of the agreement between least squares and the method of moments where both could be used. Fisher (Fisher, 1912, 1983, p. 155) appealed to another standard:

> This mutual approximation [of results], though convenient in practice . . . is harmful from a theoretical standpoint as tending to obscure the practical discrepancies, and the theoretical indefiniteness which actually exist.

Fisher (Fisher, 1912, p. 156) objects to least squares that "an arbitrariness arises in the scaling of the abscissa line." The problem is to choose a function f, from a family indexed by θ, to approximate a curve y. The method of least squares is to choose the value of θ to minimize

$$\int (y(x) - f(x:\theta))^2 \, dx.$$

Now suppose the abscissa line is rescaled with

$$x = x(z).$$

Then, as Fisher observes, the value of θ that minimizes

$$\int (y(x(z)) - f(x(z):\theta))^2 \, dz$$

is *not* the same as that which minimizes the original objective. His reasoning seems to be that in rescaling account needs to be taken of the change of variable. Noninvariance is important in this paper and in Fisher 1912–1922 generally but it was not important in the statistical literature he knew.

This argument applies to approximating one curve by another. There was a further objection when fitting a curve to data: "if a finite number of observations are grouped about a series of ordinates, there is an additional arbitrariness in choosing the position of the ordinates and the distances between them in the case of grouping observations." These objections do not apply to the method of moments but there is the question of *which* equations to use: a choice is made "without theoretical justification."

Dismissing these methods, Fisher asserts "we may solve the real problem directly." His (Fisher, 1912, p. 156) complete argument is as follows. Let

$$p = f\,\delta x$$

be the chance of an observation falling within the range δx. Then P' with

$$P' = \prod p = \prod f\,\delta x$$

"is proportional to the chance of a given set of observations occurring." As the factors δx are independent of the theoretical curve, "the probability of any particular set of θ's is proportional to P, where

$$\log P = \sum \log f.$$

The most probable set of values for the θ's will make P a maximum."

This final sentence stating the absolute criterion appears at the head of the derivation of least squares in Chauvenet (1891, p. 481), Bennett (1908, p. 15) and Brunt (1917, p. 77): for example, Chauvenet writes "The most probable system of values

of the unknown quantities . . . will be that which makes the probability P a maximum." Fisher's criterion is their method applied to *all* estimation problems.

Fisher works through an example. It is a simple version of the standard least squares problem based on a random sample of size n from the normal error curve

$$f = \frac{h}{\sqrt{\pi}} \exp(-h^2(x-m)^2)$$

[h corresponds to $1/\sigma\sqrt{2}$ in modern notation]. Differentiating log P, Fisher obtains the most probable values

$$m = \bar{x} \quad \text{and} \quad h^2 = \frac{n}{2\sum v^2},$$

where v is $(x - \bar{x})$. The notation does not distinguish true from most probable values.

Fisher was following the textbooks when he maximized P with respect to m, but they used $(n - 1)$ in the estimate of dispersion. [Gauss never gave a Bayesian analysis of mean and dispersion *together*. In 1809 he treated the mean (regression coefficients) and in 1816, dispersion.] Fisher dissects two arguments for $(n - 1)$. The more significant dissection (of Bennett's argument) is treated in the next section. Chauvenet's (1891, p. 493) argument is a mangled version of Gauss's (1823) treatment of unbiased estimation of σ^2. Although the arguments are radically different, Brunt (1917, pp. 32–34) distinguishes the Gauss–Chauvenet argument from Bennett's as being from a "purely algebraic standpoint."

The paper's message to estimators seems to be: adopt the one valid method used by astronomers; shun the others as well as both the methods used by Pearson.

2. JUSTIFYING THE CRITERION

How did Fisher justify the absolute criterion? Two views are current: the argument is a rough draft of the likelihood position clearly articulated in 1921 [see E. S. Pearson (1968, p. 412), Edwards (1974, p. 14) or Box (1978, p. 79)]; the argument involves maximizing a posterior density obtained from a uniform prior [see MacKenzie (1981, p. 244), Geisser (1980, p. 62) or Conniffe (1992, p. 485)].

As Edwards (1994) indicates, there are difficulties with both accounts. I find them less plausible than Fisher's (1922, p. 326) own explanation:

> I must indeed plead guilty in my original statement of the Method of Maximum Likelihood [i.e., in the 1912 paper] to having based my argument upon the principle of inverse probability . . .

He (Fisher 1922, p. 326) explains the principle as follows: "if the same observed result A might be the consequence of one or other of two hypothetical conditions X and Y, it is assumed that the probabilities of X and Y are in the same ratio as the probabilities of A occurring on the two assumptions, X is true, Y is true." In the 1912 argument (given above) the principle appears as: "the probability of any particular set of θ's is proportional to P."

The 1922 account is plausible (Fisher's retrospective statements are not always so) because it fits the record. In 1912 the proposition just quoted *was* a step in the derivation of the absolute criterion. Fisher did *not* signal that a significant principle is involved and name it, but its role in the argument is clear. In 1917 he said that the absolute criterion is based on the principle of inverse probability. In 1921 he denied angrily that he had assumed a uniform prior in his correlation work, and in 1922 he treated the principle separately from the postulate of a uniform prior. These episodes are examined in detail below.

If we accept this account, then we have to recognize that Fisher's beliefs and usage were idiosyncratic—and a source of confusion. Contemporaries using "inverse probability" [such as Edgeworth (1908) or Keynes (1911) used it in the context of Bayes's theorem. [In his sketch of the history of inverse probability Edwards (1994) gives a formulation from De Morgan in 1838 that is the same as Fisher's.] Fisher's *criterion* was widely used for estimating regression coefficients but I cannot find anyone who based it upon his *principle*. Sophisticated authors like Keynes (1911, p. 324) follow early Gauss and use Bayes's theorem with a uniform prior. Conniffe (1992, p. 485) suggests Fisher was following Keynes but, like the textbooks, Fisher does not mention prior probability. Chauvenet, Bennett and Brunt go directly from maximizing the probability of observations to maximizing the probability of parameter values. For Keynes such proceedings assume a uniform prior; for Fisher they assume the "principle of inverse probability."

In the 1912 paper the phrase "inverse probability" appears twice: "the inverse probability system" meaning P as a function of the parameters and in a second context we consider below. In 1922 Fisher

(1922a, p. 326) insisted, "I emphasised the fact that such inverse probabilities were relative only." Fisher (1912, p. 160) did indeed emphasize that "fact" which proved a fixed point around which justifications would come and go:

> We have now obtained an absolute criterion for finding the relative probabilities of different sets of values for the elements of a probability system of known form. ... P is a relative probability only, suitable to compare point with point, but incapable of being interpreted as a probability distribution, or giving any estimate of absolute probability.

The principle of inverse probability involves ratios and generates "relative probabilities." Fisher gave an explicit argument against integrating P to obtain an "expression for the probability that the true values of the elements should lie within any given range." *If we could*, we would obtain incompatible probability distributions depending on the parametrization chosen: "the probability that the true values lie within a region must be the same whether it is expressed in terms of θ or ϕ." While "the relative values of P would be unchanged by such a transformation," the probability that the true values will lie in a particular region will only be unchanged if the Jacobian is unity, "a condition that is manifestly not satisfied by the general transformation." His objection to least squares in curve fitting was based on the same piece of calculus. Presumably Fisher's is an "absolute" criterion because it does not depend upon the parametrization chosen or the scaling of the variables.

Fisher used this insight to object to Bennett's argument for using $(n-1)$ when estimating h^2. Bennett's procedure was to integrate m out of P,

$$\begin{aligned}\int P\,dm &= \left(h^n \exp\left[-h^2 \sum (x-\bar{x})^2\right]\right) \\ &\quad \cdot \left(\int \exp[-h^2 n (m-\bar{x})^2]\,dm\right) \\ &\propto \left(h^n \exp\left[-h^2 \sum (x-\bar{x})^2\right]\right) h^{-1} \\ &= h^{n-1} \exp\left[-h^2 \sum (x-\bar{x})^2\right]\end{aligned}$$

and maximize the resulting marginal density with respect to h to obtain

$$h^2 = \frac{n-1}{2\Sigma(x-\bar{x})^2} = \frac{n-1}{2\Sigma v^2}.$$

Fisher (1912, p. 159) objects, "the integration with respect to m is illegitimate and has no definite meaning with respect to inverse probability."

Bennett is a well-chosen target (for a paper on point estimation) but Fisher's argument could do much more damage. The most basic use for the "inverse probability system" as *probability density function* was for finding probable errors of parameters (see Sections 10 and 17). Chauvenet (1891, p. 492) obtains the probable error of m in this way. When he (Chauvenet, 1891, p. 500) finds the probable error of $\bar{x}$, it has the same numerical value. Chauvenet treats them as the same thing.

Many of the ideas and attitudes of the 1912 paper were to persist in FIsher's work: attention to theoretical indefiniteness associated with scaling both observations and parameters; insistence on a single principle of estimation free from indefiniteness; belief that such a principle was available and had found restricted use in the theory of errors; identification of a link between this principle and inverse probability; recognition that inverse probability was subject to an important limitation. Some of these ideas were old and some of the new were not substantiated. However, if not contributions to knowledge, they were contributions to Fisher's intellectual make-up. The ideas varied in durability. Around 1920 Fisher realized his principle of inverse probability was unsatisfactory: it had the same weakness as the postulate of a uniform prior. When "Bayes's postulate" fell, the inverse principle fell with it.

3. THE SECOND CRITERION

Fisher published only one other application of the "absolute criterion" before 1922—to the correlation coefficient. However, this was *not* an application of the criterion expounded in Section 1: the formula Pearson gave in 1896 was that. It was an application of another criterion, hinted at in the 1912 paper.

Fisher closes his account of estimating h by stating: "we should expect the equation $\partial b/\partial h = 0$ to give the most probable value of h." Here b is the frequency distribution, not of the sample but of the statistic μ^2. Fisher did not know b in 1912. As he showed later (Fisher, 1923), the density of the observations is proportional to the product of the densities of $\mu^2\ (= s^2)$ and $\bar{x}$,

$$\left(\frac{s^{n-2}}{\sigma^{n-1}} \exp\left(-\frac{ns^2}{2\sigma^2}\right)\right) \cdot \left(\frac{\sqrt{n}}{\sigma} \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-m)^2\right)\right)$$

(omitting constants). Now the "most probable" value of $h\ (= 1/\sigma\sqrt{2})$ found by maximizing the first factor, the density of s^2, with respect to h is not the

same as that found by maximizing the whole expression, the sample density, with respect to h and m.

Fisher did not pursue the point in the paper but he did in a letter to Gosset. Gosset reports how "with two foolscap pages covered with mathematics of the deepest dye...[Fisher] proved, by using n dimensions" that the divisor should be $n - 1$ after all. [Gosset's letter is reprinted in E. S. Pearson (1968, p. 446).] Fisher (see Fisher, 1915) obtained the density for s^2 "using n dimensions." Maximizing with respect to σ gives the estimate with divisor $n - 1$. Fisher never published this argument but he did (Fisher, 1922b, p. 599) for the analogous regression case. Gosset did the same maximization, in Bayesian mode assuming all values of σ are equally likely (see Section 5). In his (Student, 1908b) work on correlation the likelihood component was based on the distribution of the correlation coefficient not on that of the sample. However, Fisher does not seem to have known Gosset's work (Student, 1908a, or 1908b) at this time.

Fisher never discussed the relationship between the original and the second criterion, why he preferred the latter in 1912–1921 and changed his mind in 1922. He (Fisher, 1922a, p. 313) recognizes that there has been "some confusion" but does not clear it up beyond stating that the relevance of other parameters is relevant. He never even formulates the second criterion. It seems to be: (i) use the original criterion to find the relevant statistic; (ii) use the distribution of that statistic to find the best version of the statistic. For Savage (1976, p. 455, fn 20) to use the modified criterion is to "cheat a little bit." Such cheating makes honesty seem easy for first Fisher had to get the distribution of μ^2 and of the correlation. He did. Perhaps this is why.

4. THE CORRELATION COEFFICIENT

Between 1912 and 1916 Fisher used ideas from his paper in estimating the correlation coefficient and criticizing minimum χ^2. Both applications led to disagreement with Pearson. The more consequential dispute concerned the basis to Fisher's calculation of the "most probable value" of the correlation coefficient. The ensuing precipitation of prior probabilities made Fisher differentiate his position from the Bayesian one.

Fisher sent copies of his paper to Pearson and Gosset. The only letters to survive are the recipients's comments to each other, giving their overall impressions. [Gosset's letter is reprinted with comments in E. S. Pearson (1968, p. 446) and an extract from Pearson's is reproduced in E. S. Pearson (1990, pp. 47–48).] Gosset described the criterion to Pearson: "A neat, but as far as I could understand it, quite unpractical and unserviceable way of looking at things. (I understood it when I read it but it's gone out of my heat...)" Pearson wrote some time later, "I did not think it of any importance at the time & had some communication with [Fisher] on the subject."

Why did the paper fail to register with Gosset or impress Pearson? The method was the oldest in the book, illustrated by the easiest of cases; Pearson had even used it in his first work on correlation. Fisher paid no attention to the practicality of the method, the matter that most worried Pearson. The arguments against the method of moments, least squares and treating P as a probability distribution may be telling but so what? Pearson was interested in the first two methods but placed them in the realm of the practical.

Fisher's next project was the correlation coefficient. Correlation was central to the study of heredity and so central to Fisher's interests. There was also an outstanding problem: despite the efforts of Student (1908b) and Soper (1913) the exact distribution of the correlation coefficient was still open.

Pearson produced some large sample theory for the correlation coefficient (see Section 17). Gosset worked with small samples. In his formulation (Student, 1908b) a prior for the population correlation is combined with the distribution of the sample correlation. He conjectured the form of this distribution for the case of zero correlation. When Fisher (1915) found the form for the nonnull case, Gosset told him "there still remains the determination of the probability curve giving the probability of the real value (for infinite population) when a sample x has given r" (i.e., the posterior distribution). [Reprinted in E. S. Pearson (1990, p. 25). Pearson and Welch (1958) discuss Student's Bayesianism.]

In a way parallel to Fisher's second criterion, the biometricians reacted to the sampling distribution of r (and s^2) by reconsidering whether these statistics were reasonable estimate. Soper (1913, p. 91) remark on the implications of the "markedly skew character" of the distribution of r:

> the value of r found from a single sample will most probably be neither the true r of the material nor the mean value of r as deduced from a large number of samples of the same size, but the modal value of r in the given frequency distribution of r for samples of this size.

In "Appendix I to Papers by 'Student' and R. A. Fisher" Pearson applies this reasoning to the stan-

dard deviation: $\tilde{s}$, the modal value of s, is given by

$$\tilde{s} = \frac{\sigma\sqrt{(n-2)}}{n}.$$

Pearson (1915, p. 528) argues "If we take the most probable value, $\tilde{s}$, as that which has most likely been observed, then the result [s] should be divided by [$\sqrt{(n-2)}/n$] to obtain the most reasonable value for σ." So he advocates the use of the divisor $(n-2)$. Against this Gosset proposed the maximum posterior estimate based on a uniform prior; this uses the divisor $(n-1)$.

We now turn to Fisher's paper, not to the distribution theory but to the implications Fisher saw for estimation. In the following passage (Fisher, 1915, p. 520) he criticized Soper's reasoning; r is the sample correlation, $\bar{r}$ the mean of the sampling distribution of r, ρ the population correlation and $t = r/\sqrt{(1-r^2)}$:

> The fact that the mean value $\bar{r}$ of the observed correlation is numerically less than ρ might have been interpreted as meaning that given a single observed value r, the true value of the correlation coefficient from which the sample is drawn is likely to be greater than r. This reasoning is altogether fallacious. The mean $\bar{r}$ is not an intrinsic feature of the frequency distribution. It depends upon the choice of the particular variable r in terms of which the frequency distribution is represented. When we use t the situation is reversed.

This criticism recalls the noninvariance objection to least squares. The reasoning is "fallacious" because it depends on the scaling of the variable *not* because there is no sense in making probability statements about ρ.

Fisher goes on to use the "absolute criterion," which is "independent of scaling," to obtain the "relation between an observed correlation of a sample and the most probable value of the correlation of the whole population." The most probable value, $\hat{\rho}$, is obtained by the modified criterion, that is, by maximizing the sampling distribution of r. This is not the maximum likelihood estimator; r is that. Fisher (1915, p. 521) gives the approximate relationship

$$r = \hat{\rho}(1 + (1-r^2)/2n).$$

He writes the formula not with $\hat{\rho}$ but with ρ, a symbol he uses both for the parameter and its most probable value. He concludes that "It is now apparent that the most likely value [$\hat{\rho}$] of the correlation will in general be less than that observed [r]." "Most likely" has arrived. It means "most probable."

5. MINIMUM CHI-SQUARED VERSUS THE GAUSSIAN METHOD

There are further signs of immobility in Fisher's thinking on fundamentals in an abortive publication criticizing Smith (1916). She proposed minimum χ^2 as a criterion of "bestness" for the values of constants in frequency distributions. The methods she reviewed included the method of moments.

The method of moments estimates of the normal mean and variance are the sample mean and variance. She comments (Smith, 1916, p. 262) that "if we deal with individual observations then the method of moments gives, with a somewhat arbitrary definition of what is to be a maximum, the 'best' values for σ and $\bar{x}$ [the population mean]." She criticized (Smith, 1916, p. 263n) the "Gaussian" choice of maximand,

> [P]robability means the frequency of recurrence in a repeated series of trials and this probability is in the [Gaussian] case supposed indefinitely small.

The "Gaussian" maximand was textbook (priorless) Gauss. In correspondence Gauss actually gave Smith's objection to maximizing a quantity that was still infinitely small as a reason for changing from maximizing posterior probability to minimizing sampling variances (see Plackett, 1972, p. 247).

Fisher sent Pearson a draft of a note on Smith's paper. He did not respond to her criticism of the Gaussian method but he reacted to her use of "arbitrary":

> There is nothing at all "arbitrary" in the use of the method of moments for the normal curve; as I have shown elsewhere it flows directly from the absolute criterion ($\sum \log f$ a maximum) derived from the Principle of Inverse Probability. There is on the other hand, something exceedingly arbitrary in a criterion which depends entirely upon the manner in which the data happen to be grouped.

[Fisher's draft note and covering letter are in E. S. Pearson (1968, pp. 454–455).] Fisher's objection to grouping follows his 1912 objection to least squares.

Pearson agreed with Smith, "I frankly confess I approved the Gaussian method in 1897 . . . I think it logically at fault now." He invited Fisher to write a defence of the method. When Fisher submitted a further piece on Smith's work, Pearson rejected it. [Pearson's letters are reprinted in E. S. Pearson (1968, pp. 455–456).] For Pearson's earlier approval, see Section 17 below.

Fisher's work 1912–1916 contains applications of his ideas on the absolute criterion, on invariance to

rescaling of the data and the arbitrariness of grouping but no new basic ideas. "Appendix II to the Papers of 'Student' and R. A. Fisher" by Pearson and his team roused Fisher from his dogmatic slumber.

6. THE PRECIPITATION OF PRIOR PROBABILITIES

The Cooperative Study of Soper, Young, Cave, Lee and Pearson (Soper et al., 1917) has a section on the "'most likely' value of the correlation in the sampled population." The framework is Gosset's. Writing $\phi(\rho)\,d\rho$ for "the law of distribution of ρ's," that is, the prior distribution of the population correlation, they state (Soper et al., 1917, p. 352) "we ought to make" the product of $\phi(\rho)$ and the density of r, the sample correlation, "a maximum with ρ."

They state (Soper et al., 1917, p. 353) "Fisher's equation ... is deduced on the assumption that $\phi(\rho)$ is constant"—with good reason! Fisher told Pearson that the absolute criterion was derived from the principle of inverse probability. It was reasonable to presume that prior probabilities are not mentioned because they are taken to be equal. Fisher's work fitted into Gosset's program: having derived the distribution of r he was using a uniform prior to make probability statements about ρ. Gosset was proposing the same for σ. The cooperators (Soper et al., 1917, p. 353n) expected clusterings of values and rejected the uniform prior for σ. The views that Pearson attacked were Gosset's. It was reasonable to suppose that Fisher held them too—though it turned out that he did not.

Pearson (1892, p. 175) argued that the "equal distribution of ignorance" could have a justification in "experience" if there is prior experience to the effect that all values of the parameter have been met with equal frequency. "Bayes's theorem" in his paper "On the influence of past experience on future expectation" (1907) gives the probability density of the chance of an event happening given p occurrences in n Bernoulli trials *on the assumption* of the equal distribution of ignorance. "Bayes's theorem" in the cooperative study is similarly based on a uniform prior for the parameter (the correlation).

The cooperators (Soper et al., 1917, p. 358) work through an example of data on 25 parent–child pairs. As to "equal distribution of ignorance" they argue that "there is no such experience in this case." So they judged that (Soper et al., 1917, p. 354) the fuller solution of Fisher's equation "appears to have academic rather than practical value." The cooperators close their account of Bayes by preaching:

> Statistical workers cannot be too often reminded that there is no validity in a mathematical theory pure and simple. Bayes' theorem must be based on experience ...

The cooperative study made Fisher angry. A letter from L. Darwin to Fisher (Bennett, 1983, p. 73) on a draft of Fisher (1921) shows how much: "you speak of someone's interpretation of your remarks being 'so erroneous etc., etc.' ... Again, you imply that your opponents have criticized you without reading your paper. ..." Fisher's intellectual response was to state that he and his method had been misrepresented and to redraft his method.

7. THE ABSOLUTE CRITERION REVISITED

Fisher distinguished his method from that of Bayes by articulating the notion of "likelihood." This was not merely a search for the *mot juste* as Edwards (1974, p. 14) implies. Genuine conceptual development was involved, though at first Fisher did not admit this. It spreads over two papers: the combative 1921 paper was written by a very injured party; the 1922 paper is almost serene—it even admits that the author makes mistakes (cf. Section 2).

The 1921 paper picks up two issues from Fisher's earlier correlation work. We will not be concerned with the (extensive) material on distribution theory. The material on estimation bristles with criticisms: the cooperators are criticized for ascribing to Fisher a position he had not taken; that position is criticized; finally the position Fisher had taken is criticized!

Fisher (1921, p. 16) objected to the cooperators' portrayal of his correlation work as based on a uniform prior:

> a reader, who did not refer to my paper ... might imagine that I had used Boole's ironical phrase "equal distribution of ignorance," and that I had appeared to "Bayes' Theorem". I must therefore state that I did neither.

Against the possible charge that he had implicitly assumed a uniform prior for the correlation, Fisher begins by pointing out that he would have obtained the same value of the "optimum," his new name for the most probable or likely value, if he had used a transformation (z) of the correlation (r): "Does the value of the optimum therefore depend upon equal numbers of parental correlations having occurred in equal intervals dz? If so, it should be noted this is inconsistent with an equal distribution in the scale of r."

Fisher gives a somewhat edited account of his 1912 and 1915 papers. He recalls his use of "most likely value" but not of "most probable value." He mentions (Fisher, 1921, p. 4) the determination of "inverse probabilities" as "resting" on Bayes's work but does not mention his own use of the notion. The account is also somewhat enhanced: the simple facts—that there was no mention of prior probabilities in 1912 and that the "criterion" delivers the "optimum"—become (Fisher, 1921, p. 16):

> As a matter of fact, as I pointed out in 1912 . . . the optimum is obtained by a criterion which is absolutely independent of any assumption regarding the a priori probability of any particular value. It is therefore the correct value to use when we wish for the best value for the given data, unbiassed by any a priori presuppositions.

These were additions: he had *not* pointed out the criterion was "independent of any assumption. . . ."

8. THE REJECTION OF BAYES AND INVERSE PROBABILITY

Fisher appended a "Note on the confusion between Bayes' Rule and my method of the evaluation of the optimum" to his 1921 paper. He recalled Bayes: (Fisher, 1921, p. 24):

> Bayes (1763) attempted to find, by observing a sample, the actual probability that the population value [ρ] lay in any given range. . . . Such a problem is indeterminate without knowing the statistical mechanism under which different values of ρ come into existence; it cannot be solved from the data supplied by a sample, or any number of samples, of the population.

In 1922 he wrote (Fisher, 1922a, p. 326): "probability is a ratio of frequencies, and about the frequencies of such [parameter] values we can know nothing whatever."

In 1922 Fisher used the binomial distribution to illustrate maximum likelihood and to contrast it with Bayes's method. Against Bayes's "postulate," that is, that it is reasonable to assume that the a priori distribution of the parameter p is uniform, he argued (Fisher, 1922a, pp. 324–325) that "apart from evolving a vitally important piece of knowledge, that of the exact form of the distribution of p, out of an assumption of complete ignorance, it is not even a unique solution," This nonuniqueness charge was the final form of the 1912 noninvariance argument. He shows that a uniform prior for p and a uniform prior for θ, defined as

$$\sin \theta = 2p - 1,$$

leads to inconsistent posterior distributions. The point was not new. For instance, Edgeworth (1908, p. 392) made the point in the context of different parametrizations of dispersion for the normal distribution: a uniform prior for h is inconsistent with a uniform prior for c $(= 1/h)$. However, Edgeworth did not think the objection fatal.

In 1921 Fisher quietly uncoupled the absolute criterion from the method of inverse probability. In 1922 he confronted the method and argued (Fisher, 1922a, p. 326) that it did not give a unique solution either: "irrelevant" distinctions between "conditions" alter relative probabilities. Although he begins by distinguishing the principle of inverse probability from the assumption of a uniform prior, he ends by concluding that inverse probability "amounts to assuming that . . . it was known that our universe had been selected at random from an infinite population in which X was true in one half, and y true in one half." This sounds very like an admission that the absolute criterion *is* based on Bayes's postulate. From "Inverse probability" (Fisher, 1930) onwards Fisher treated the principle of inverse probability and Bayes's postulate as one.

9. THE MAKING OF LIKELIHOOD

Fisher (1921, p. 4) tacitly criticizes himself when he states "two radically distinct concepts have been confused under the name of 'probability' and only by sharply distinguishing between these can we state accurately what information a sample does give us respecting the population from which it was drawn." The concepts are probability and likelihood (Fisher, 1921, p. 25):

> We may discuss the probability of occurrence of quantities which can be observed . . . in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of hypotheses . . . [We] may ascertain the likelihood of hypotheses . . . by calculation from observations: . . . to speak of the likehood . . . of an observable quantity has no meaning.

He develops the point with reference to correlation (Fisher, 1921, p. 24):

> What we can find from a sample is the likelihood of any particular value of ρ, if we define the likelihood as a quantity proportional to the probability that, from a population having the particular value of ρ, a sample having the observed value of r, should be obtained.

Fisher still worked with the distribution of r rather than with the distribution of the entire set of obser-

vations. In 1922 that changed (Fisher, 1922a, p. 310):

> The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.

Fisher (1921, p. 24; 1922a, p. 327) redrafted what he had written in 1912 about inverse probability, distinguishing between the mathematical operations that can be performed on probability densities and likelihoods: likelihood is not a "differential element," it cannot be integrated. Fisher had not been concerned with *non*optimal values of the likelihood since 1912. In 1921 he reemphasized the importance of the whole function. He points out (Fisher, 1921, p. 24) that "no transformation can alter the value of the optimum, or in any way affect the likelihood of any suggested value of ρ." In the note on Bayes he describes how probable errors encode likelihood information:

> "Probable errors" attached to hypothetical quantities should not be interpreted as giving any information as to the probability that the quantity lies within any particular limits. When the sampling curves are normal and equivariant the "quartiles" obtained by adding and subtracting the probable error, express in reality the limits within which the likelihood [relative to the maximum] exceeds 0.796542...

This view of "reality" leads directly to the exposition of likelihood as a "form of quantitative inference" in *Statistical Methods and Scientific Inference* (1971, p. 75) or to works like Edwards's *Likelihood*.

The 1921 paper was not consistent likelihood analysis. In Examples II (Fisher, 1921, p. 11) and VIII (p. 23) probable errors are used to calculate prob-values, tail areas, not relative likelihood values. In Example II he confuses the two, concluding from prob-values of 0.142 for one hypothesis and 0.00014 for another that the latter is roughly 1,000 times "less likely" than the former.

Fisher (1922, p. 327fn) wrote "in an important class of cases the likelihood may be held to measure the degree of our rational belief in a conclusion." He wrote (Fisher 1925b, pp. 10–11) of likelihood as "the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations." Despite such declarations there was no development of likelihood methods. Edwards (1972, p. 3) writes of Fisher's use of likelihood, "From 1921 until his death, Fisher... quietly and persistently espoused a... measure by which he claimed rival hypotheses could be weighed." The measure came in with a bang but the volume was not sustained. In 1922 the main case for maximum likelihood was based on its sampling properties and on its association with information and sufficiency.

10. GAUSS AND THE SAMPLING PROPERTIES OF ESTIMATORS

We have seen nothing to suggest that an estimator's sampling properties could matter for Fisher. However, while he was developing likelihood he was working on what seems to have been an unrelated project, comparing the sampling properties of estimators. In 1919–1920 he compared different estimators of the standard deviation and the correlation coefficient. He did not say that the estimator given by the absolute criterion was best and that was that! Absolute criterion and optimum method do not figure here, only size of probable error.

Again Gauss (or textbook Gauss) is a presence. Earlier developments can be counted remote consequences of his 1809 justification of least squares; those we about to treat can be linked to his 1816 paper on estimating the parameter h (see Section 1) of the normal distribution. Gauss gave a Bayesian analysis using a uniform prior for h (m is assumed known) and obtained a normal approximation to the posterior. He also compared the large-sample properties of estimators of h based on different powers of the errors and showed that the value 2 gives the estimator with least dispersion. He distinguished the two investigations but they are linked: the probable error of the posterior distribution of h is equal to the probable error of the sampling distribution of the squared deviation estimator of h.

Merriman's (1884, pp. 206–208) textbook reproduces the Bayesian argument (without the prior and the sampling theory results, linking them with the phrase "a similar investigation." He reports Gauss's finding that with absolute values 114 observations "give the same uncertainty... as 100 observations" with squared errors. Merriman's book is significant as it seems to have been Pearson's reference for least squares. [Pearson (1894, p. 21) uses data from Merriman, and Yule, Pearson's assistant, used it as his least squares reference in his (Yule 1897, p. 818).] In Section 17 we consider Pearson's use of Merriman–Gauss and its connection with Fisher's 1922 analysis.

11. FISHER ON COMPARING ESTIMATES

Fisher first compares sampling properties of estimators in his 1919 critique of Thorndike's work on

twins. Thorndike used a measure of resemblance, $2xy/(x^2 + y^2)$, where x and y are intraclass bivariate normal. Fisher obtained its distribution and mentioned (Fisher, 1919, p. 493) that data on resemblances might be used to estimate the population correlation but added at once that "The value of the correlation that best fits the data may be found with less probable error, from the product–moment correlations." This remark is not developed, nor its basis made clear, but in 1920 he devoted a paper to comparing estimates of the standard deviation of the normal distribution.

The 1920 paper was a reply to Eddington's (1914, p. 147) claim that, "contrary to the advice of most text-books" it is better to use σ_1, based on absolute values of residuals, than σ_2, based on squares of residuals. As distribution theory, Fisher's response was a continuation of his work on s^2. Correcting Eddington took much less than Fisher gave. Finding the large-sample probable errors of the two estimators was dispatched in the first third of the paper and demonstrating that 2 is the best choice for the power of the residuals took another page or so.

The rest of the paper is concerned with the joint distribution of the two estimators, σ_1 and σ_2. Fisher (1920, p. 763) writes that "Full knowledge of the effects of using one rather than another of two derivates can only be obtained from the frequency surface of pairs of values of the two derivates." It is not clear whether this principle guides the investigation or is a conclusion from it. Fisher's route to the marginal density of σ_1 is through the joint density of σ_1 and σ_2. This accident of technique may have led him to the principle.

Fisher (1920, p. 758) wrote "The case is of interest in itself, and it is illuminating in all similar cases, where the same quantity may be ascertained by more than one statistical formula." This suggests a program of comparing different estimators on a case-by-case basis. The absolute criterion has dropped out and so have the universal pretensions associated with it. In 1922 all are back, thanks to an idea in the 1920 paper.

12. SUFFICIENCY IN 1920

The important new idea is sufficiency, though it was only named in 1922. Fisher (1920, p. 768) ends by discussing a "qualitative distinction, which reveals the unique character of σ_2." This property had not been identified in earlier investigations of the merits of σ_2 versus σ_1. [Stigler (1973) discusses related work of Laplace.] "For a given value of σ_2, the distribution of σ_1 is independent of σ." Fisher added the gloss: "The whole of the information to be obtained from σ_1 is included in that supplied by σ_2." He went further and argued that σ_1 could be replaced by any other statistic: "The whole of the information respecting σ, which a sample provides is summed up in the value of σ_2."

Fisher does not state that the "qualitative distinction" underlies the superiority of σ_2. Perhaps it was obvious. He finishes (Fisher, 1920, pp. 769–770) by investigating the curve related to σ_1, "in the same way as the normal curve is related" to σ_2. His treatment is, he admits, not very satisfactory. However, there is an implicit generalization: the sufficient statistic provides an "ideal measure" of the parameter.

13. MAXIMUM LIKELIHOOD IN 1922

The name "maximum likelihood" finally appears in the "Mathematical foundations of theoretical statistics" (Fisher, 1922a, p. 323). The method is the link between two separate investigations. The first relates maximum likelihood to sufficiency and efficiency. The second uses the large-sample standard error of maximum likelihood to compare error curves and assess the method of moments.

The new theoretical scheme is this: maximum likelihood produces sufficient estimates; sufficient estimates are efficient, besides being important in their own right. The theory is thin and underdeveloped. The only sufficient statistic presented is the standard deviation from 1920, though Fisher (1922a, p. 357) alludes to the case of the Poisson parameter and gives a demonstration in Fisher, Thornton and MacKenzie (1922, p. 334).

In other ways the paper is a grand finale. Besides inverse probability and Bayes (see above) the method of moments, minimum χ^2 and the effects of grouping are all reexamined. Before 1922 Fisher worked on only three estimation problems: the mean and variance of the univariate normal and the correlation coefficient of the bivariate normal. The new paper has a miniature treatise on the Pearson curves and the method of moments. The efficiency of the method of moments is assessed by comparing the large-sample variance of such estimators with that of maximum likelihood. Sufficiency plays no role here, nor is the computability of the superior estimator considered.

14. EFFICIENCY AND SUFFICIENCY

Fisher (1922a, p. 316) presents three "Criteria of Estimation" two of them linked to maximum likelihood. He links the third, "consistency," to the method of moments, taking it for granted that maximum likelihood satisfies it.

The "Criterion of Efficiency" refers to large-sample behavior: "when the distribution of the statistics tend to normality, that statistic is to be chosen which has the least probable error." This criterion underlies Eddington's comparison and the 1920 paper. The efficiency of an estimator is given by the square of the ratio of its probable error to that of the most efficient statistic.

The "Criterion of Sufficiency" is that "the statistic chosen should summarize the whole of the relevant information supplied by the sample." The "mathematical interpretation" of sufficiency is the conditional distribution notion from the 1920 paper. The example is the same.

"Efficiency" is efficiency in summarizing information so, in the large-sample case, Fisher (1922a, p. 317) identifies sufficiency and full efficiency. Let θ be the parameter to be estimated and let θ_1 be the sufficient statistic and θ_2, any other statistic. In accordance with the scope of the criterion of efficiency θ_1 and θ_2 are taken to be normal in large samples, indeed jointly normal with common mean θ. Using the condition that the distribution of θ_2 given θ_1 does not involve θ, Fisher obtains the relation

$$r\sigma_2 = \sigma_1,$$

where r is the correlation between θ_1 and θ_2 and σ_1 and σ_2 the large-sample standard errors of θ_1 and θ_2, respectively. This shows that σ_1 is necessarily less than σ_2 and that the efficiency of θ_2 is measured by r^2. Although this analysis covers the 1920 efficiency comparison, the technique is different.

15. SUFFICIENCY AND INFORMATION

The move from efficiency to sufficiency was not simply from "large" to "all finite" samples. The objective seems to have changed: efficiency relates to estimation accuracy, sufficiency to "information."

"Information" has arrived. The 1920 notion of information was a comparative one: information in one statistic is "included" in that supplied by another. There is no explanation of what information *is*. Perhaps a statistic contains relevant information if its distribution depends on the parameter of interest. In 1922 (Fisher, 1922a, p. 311) information is written into the statistician's task.

The statistician constructs a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The phrase "hypothetical infinite population" was new in 1922. Fisher later told Jeffreys (Bennett, 1990, pp. 172–173) that it may have been used by Venn, "who ... was wrestling with the same idea." The idea, if not the phrase, was already commonplace in 1922. Galton (1889, p. 125) imagined an "exceedingly large Fraternity, far more numerous than is physiologically possible" from which random samples are taken, and Fisher (1918, p. 400n) describes an "infinite fraternity, ... all the sons which a pair of parents might conceivably have produced." "Indefinitely large population" appears in the title of Fisher (1915). In his correlation paper Student (1908b, p. 35) wrote that starting from the actual sample of 21 years "one can image the population indefinitely increased and the 21 years to be a sample from this."

The parameters associated with the law of distribution of this population are "sufficient to describe it exhaustively in respect of all qualities under discussion." Fisher continues (Fisher, 1922a, p. 311):

> Any information given by the sample, which is of use in estimating the values of these parameters is relevant information .... It is the object of the statistical processes ... to isolate the whole of the relevant information contained in the data.

The only "statistical process" discussed in 1922 paper is "estimation." Isolating relevant information is the business of sufficient estimates. Efron (1982) suggests that Fisher was not primarily interested in estimation but in "summarization." In 1922 Fisher could see no need to choose between them.

16. SUFFICIENCY AND MAXIMUM LIKELIHOOD

The normal standard deviation seemed to exemplify a connection beyond efficiency and sufficiency, namely, between sufficiency and maximum likelihood. Fisher (1922a, p. 323) conjectured that maximum likelihood "will lead us automatically" to a sufficient statistic. He had much more confidence in this proposition than in his argument for it. He invited "pure mathematicians" to improve upon his proof. The proposition is a cornerstone of the paper but of course it is not true, as Geisser (1980, p. 64) notes in his brief commentary.

Fisher's (1922a, pp. 330–331) argument is obscure in structure and in detail; it is not even clear whether he is showing that "optimality" implies sufficiency or vice versa. He considers the joint density $f(\theta, \hat{\theta}, \theta_1)$ of the maximum likelihood estimator $\hat{\theta}$ and any other statistic θ_1 and argues that the equation

$$\frac{\partial}{\partial\theta}\log f\left(\theta, \hat{\theta}, \theta_1\right) = 0$$

is satisfied irrespective of θ_1 by the value $\theta = \hat{\theta}$. He then shows correctly that this equation would be satisfied if the maximum likelihood estimator were sufficient. Fisher continued to work on the maximum likelihood–sufficiency connection, producing the factorization theorem in 1934. However, we now turn to maximum likelihood's other role in the 1922 paper: its use as an efficiency yardstick for other estimators, particularly for the method of moments.

17. PROBABLE ERRORS OF OPTIMUM STATISTICS

Fisher seemed to consider his work on the large-sample distribution of maximum likelihood as refining existing work. He noted (Fisher, 1922a, p. 329fn) how a "similar method" for calculating large-sample standard errors had been developed by Pearson and Filon (1898). In 1940 he saw their work differently: "If I remember right, they arrive at the formula by a very obscure argument involving inverse probability" (Bennett, 1990, p. 125). There is no evidence that Fisher began by developing their argument but it is not implausible that he did. In any case a comparison is instructive. Fisher's treatment is unambiguously frequentist while their method developed out of work that was as unambiguously Bayesian.

The Pearson–Filon treatment can be seen evolving in Pearson's lectures of 1894–1896. Pearson (1894–1896, vol. 1, p. 90) begins by treating the standard deviation along the lines of Merriman (see Section 10 above). His first new case is the correlation coefficient; he published this in 1896. Although he later referred to his use of the "Gaussian method" (see Section 5 above), Pearson's way of finding the "best" value and its probable error was Gauss without a prior.

Pearson (1896, pp. 264–266) considers a bivariate normal distribution with parameters, σ_1^2, σ_2^2 and r; Gauss–Merriman had one parameter. Pearson expresses the joint density of the observations in terms of r, not by integrating out the nuisance parameters or maximizing them out, but by treating σ_1^2 and σ_2^2 as *identical* to $(1/n)\Sigma x^2$ and $(1/n)\Sigma y^2$, respectively. He states (Pearson, 1896, p. 265) that the "best" value of r is found by choosing the value for which "the observed result is the most probable." A quadratic Taylor approximation to the log of the function around its maximum value (given by the product–moment formula) yields a normal density. The probable errors are obtained from this.

The value of r chosen appears to be "best" simply because of the way it is obtained. Pearson mentions the analogy of estimates of σ: "the error of mean square gives the theoretically best results." For Merriman the error of mean square was distinguished because it was given by the Gaussian method *and* because it had smaller probable error than other estimators. Pearson does not mention the second point here or in his lectures (Pearson, 1894–1896, vol. 1, 90–92; vol. 3, 15–22).

Pearson and Filon (1898) floated off the method of calculating probable errors from that of finding best values and applied the technique to "nonbest" estimates, in particular to the method of moments. Welch (1958, p. 780), MacKenzie (1981, pp. 241–243) and Stigler (1986, pp. 342–345) describe the method as implicitly Bayesian. I would put things differently. The method had Bayesian origins but there was no longer a Bayesian argument to be made explicit. Person had taken a method and changed it in a way that contradicted the basis of the method.

The Taylor expansion is now around the true value and the discrepancy is referred to as an "error," a usage that signals a concern with the probable error as a sampling distribution property. The new multiparameter treatment leads to a correction of the probable error formula for the correlation, but the more fundamental change is not registered. I will simplify their argument and consider a scalar variable x and parameter θ. Let P_0 be the density of the observations $x_1, \ldots, x_n$, and let P_Δ be the density corresponding to the parameter value $\theta + \Delta\theta$. Consider P_0/P_Δ and its logarithm

$$\log\left(\frac{P_0}{P_\Delta}\right) = \sum \left(\log f(x_i;\theta) - \log f(x_i;\theta+\Delta\theta)\right),$$

expanding using Taylor's theorem

$$\log\left(\frac{P_0}{P_\Delta}\right) = \Delta\theta \sum \frac{d}{d\theta}\log f(x_i;\theta) + \frac{1}{2}(\Delta\theta)^2 \sum \frac{d^2}{d\theta^2}\log f(x_i;\theta) + \cdots.$$

"Replacing sums by integrals" they obtain

$$\log(P_0/P_\Delta) = A\,\Delta\theta + \tfrac{1}{2}B(\Delta\theta)^2 + \cdots,$$

where A and B are defined by

$$A = \int \frac{d}{d\theta}\log f(x_i;\theta) f(x_i;\theta)\,dx,$$

$$B = \int \frac{d^2}{d\theta^2}\log f(x_i;\theta) f(x_i;\theta)\,dx.$$

As $A = 0$, they write

$$P_\Delta = P_0 \exp\left(-\tfrac{1}{2}B(\Delta\theta)^2\right) \cdot \exp(\text{cubic and higher-order terms}).$$

They (Pearson and Filon, 1898, pp. 233–234) interpret P_Δ as the density of $\Delta\theta$ an "error," that is, $\Delta\theta$ is the difference between the estimate and the true value of the frequency constant, and conclude that $\Delta\theta$ is approximately normally distributed with zero mean and standard deviation $B^{-1/2}$.

Fisher's 1922 procedure resembles Gauss–Merriman–Pearson but the objective is not to approximate the posterior distribution but a frequency distribution as in Pearson and Filon. Fisher (1922a, p. 328) expands the log likelihood around the value of a statistic θ_1:

$$\log \phi = C + (\theta - \theta_1)\sum a + \tfrac{1}{2}(\theta - \theta_1)^2 \sum b + \cdots.$$

For "optimum" statistics, that is, when θ_1 is maximum likelihood, $\Sigma a = 0$. For sufficiently large samples Σb differs from $n\bar{b}$ by a quantity of order $\sqrt{n}\sigma_b$. Eliminating all terms which converge to zero as n increases gives

$$\log \phi = C + n\frac{\bar{b}}{2}(\theta - \theta_1)^2.$$

So,

$$\phi \propto \exp\left(n\frac{\bar{b}}{2}(\theta - \theta_1)^2\right).$$

Fisher argues that the density of θ_1 is proportional to this function and that the large-sample standard error of θ_1 is given by $(-n\bar{b})^{-1/2}$.

In 1922 Fisher made no fundamental objection to Pearson and Filon's development. He criticized the application of their formulae to methods other than maximum likelihood. The novel application in Pearson and Filon was the construction of probable error formulae for the method of moments applied to the Pearson curves. The large part of the 1922 paper on the (in)efficiency of this method can be taken as an ironical commentary on their results. The formulae had already been quietly withdrawn by Pearson; Fisher was reinstating them as appropriate to the efficient method of maximum likelihood.

18. GOODBYE TO ALL THAT

Fisher's 1922 view of the place of maximum likelihood in statistics did not last. In particular, he discovered that there may be *no* sufficient statistic for maximum likelihood to find. However, he did not revise his refutation of Bayes's postulate and inverse probability nor his general diagnosis of what had been wrong with statistics. We close with an examination of that diagnosis.

Fisher (1922a, p. 310) points to the "anomalous state of statistical science." There has been an "immense amount of fruitful labour" on applications, including such topics as Pearson curves, but the basic principles remain in "a state of obscurity." The state of affairs is due to the "survival to the present day of the fundamental paradox of inverse probability which like an impenetrable jungle arrests progress towards precision of scientific concepts" (Fisher, 1922a, p. 311).

Fisher (1922a, p. 326) refers to the "baseless character of the assumptions" made under the names inverse probability and Bayes's theorem and the "decisive criticism to which they have been exposed." The critics—Boole, Venn and Chrystal—came to appear regularly in Fisher's historical accounts. Yet when he detailed their criticisms—in 1956 (Fisher, 1971, p. 31)—he reflected, "[Chrystal's] case as well as Venn's illustrates the truth that the best causes tend to attract to their support the worst arguments." Zabell (1989) was unimpressed when he examined their work. Fisher may have drawn comfort from the critics's existence; it is hard to believe he was influenced by them.

A major factor in the "survival" was a "purely verbal confusion" (Fisher, 1922a, p. 311): the use of the same name for true value and estimate. So Fisher introduced the terms "parameter" and "statistic." He wrote later (Bennett, 1990, p. 81) "I was quite deliberate in choosing unlike words for these ideas which it was important to distinguish as clearly as possible." The outraged and confused commonsense of the older generation was voiced by Pearson (1936, p. 49n): he objected "very strongly" to the "use of the word 'statistic' for a statistical parameter" and asked "are we also to introduce the words, a mathematic, a physic, ... for parameters ... of other branches of science?"

ACKNOWLEDGMENTS

I am grateful to librarians at University College London and Adelaide for their help. My thanks to Janne Rayner, Denis Conniffe, Raymond O'Brien and Grant Hillier for discussion. On submitting an earlier draft of this paper I learned that A. W. F. Edwards already had a paper reexamining Fisher's use of "inverse probability"; a revision of his 1994 preprint in printed here. However, though we use the same "data," our interpretations are somewhat different. I am grateful to Anthony Edwards for helpful discussions as well as to the Editor and an Associate Editor for their suggestions.

REFERENCES

BENNETT, J. H., ed. (1971). *Collected Papers of R. A. Fisher* **1–5**. Adelaide Univ. Press.

BENNETT, J. H., ed. (1983). *Natural Selection, Heredity, and Eugenics*. Oxford Univ. Press.

BENNETT, J. H., ed. (1990). *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Oxford Univ. Press.

BENNETT, T. L. (1908). Errors of observation. Technical Lecture 4, Cairo, Ministry of Finance, Survey Department Egypt.

BOX, J. F. (1978). *R. A. Fisher: The Life of a Scientist*. Wiley, New York.

BRUNT, D. (1917). *The Combination of Observations*. Cambridge Univ. Press.

CHAUVENET, W. (1891). *A Manual of Spherical and Practical Astronomy* **2**, 5th ed. [Reprinted (1960) by Dover, New York.]

CONNIFFE, D. (1992). Keynes on probability and statistical inference and the links to Fisher. *Cambridge Journal of Economics* **16** 475–489.

EDDINGTON, A. S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, New York.

EDGEWORTH, F. Y. (1908). On the probable errors of frequency-constants. *J. Roy. Statist. Soc.* **71** 381–397, 499–512, 651–678; **72** 81–90.

EDWARDS, A. W. F. (1972). *Likelihood*. Cambridge Univ. Press.

EDWARDS, A. W. F. (1974). The history of likelihood. *Internat. Statist. Rev.* **42** 9–15.

EDWARDS, A. W. F. (1994). What did Fisher mean by 'inverse probability' in 1912–22? In Three contributions to the history of statistics, Preprint 3, Inst. Mathematical Statistics, Univ. Copenhagen.

EFRON, B. (1982). Maximum likelihood and decision theory. *Ann. Statist.* **10** 340–356.

FIENBERG, S. E. and HINKLEY, D. V., eds. (1980). *R. A. Fisher: An Appreciation*. Springer, New York.

FISHER, R. A. (1911). Mendelism and biometry. Unpublished manuscript. [Reproduced in Bennett (1983), pp. 51–58.]

FISHER, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41** 155–160. [CP1 in Bennet (1971), vol. 1.]

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521. [CP4 in Bennett (1971), vol. 1.]

FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52** 399–433. [CP9 in Bennett (1971), vol. 1.]

FISHER, R. A. (1919). The genesis of twins. *Genetics* **4** 489–499. [CP11 in Bennett (1971), vol. 1.]

FISHER, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society* **80** 758–770. [CP12 in Bennett (1971), vol. 1.]

FISHER, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32. [CP14 in Bennett (1971), vol. 1.]

FISHER, R. A. (1922a). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368. [CP18 in Bennett (1971), vol. 1.]

FISHER, R. A. (1922b). The goodness of fit of regression formulae, and the distribution of regression coefficients. *J. Roy. Statist. Soc.* **85** 597–612. [CP20 in Bennett (1971), vol. 1.]

FISHER, R. A. (1923). Note on Dr. Burnside's recent paper on errors of observation. *Proc. Cambridge Philos. Soc.* **21** 655–658. [CP30 in Bennett (1971), vol. 1.]

FISHER, R. A. (1925a). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725. [CP42 in Bennett (1971), vol. 2.]

FISHER, R. A. (1925b). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

FISHER, R. A. (1930). Inverse probability. *Proc. Cambridge Philos. Soc.* **26** 528–535. [CP84 in Bennett (1971), vol. 2.]

FISHER, R. A. (1934a). Two new properties of mathematical likelihood. *Proc. Roy. Soc. Ser. A* **144** 285–307. [CP108 in Bennett (1971), vol. 3.]

FISHER, R. A. (1934b). Discussion on "On the two different aspects of the representative method," by J. Neyman. *J. Roy. Statist. Soc.* **97** 614–619.

FISHER, R. A. (1971). *Statistical Methods and Scientific Inference*, 3rd ed. Hafner, New York. (First edition published in 1956.)

FISHER, R. A., THORNTON, H. G. and MACKENZIE, W. A. (1922). The accuracy of the plating method of estimating the density of bacterial populations. *Annals of Applied Biology* **9** 325–359. [CP22 in Bennett (1971), vol. 1.]

GALTON, F. (1889). *Natural Inheritance*. Macmillan, London.

GAUSS, C. F. (1809). *Theoria Motus Corporum Coelestium*. [Extract in Trotter (1957), pp. 127–147.]

GAUSS, C. F. (1816). Bestimmung der Genauigkeit der Beobachtungen. [See Trotter (1957), pp. 109–117.]

GAUSS, C. F. (1821). Theoria combinationis observationum erroribus minimum obnoxiae, first part. [See Trotter (1957), pp. 1–49.]

GEISSER, S. (1980). Basic theory of the 1922 mathematical statistics paper. In *R. A. Fisher: An Appreciation* (S. E. Fienberg and D. V. Hinkley, eds.) 59–66. Springer, New York.

KEYNES, J. M. (1911). The principal averages and the laws of error which lead to them. *J. Roy. Statist. Soc.* **74** 322–331.

MACKENZIE, D. A. (1981). *Statistics in Britain 1865–1930*. Edinburgh Univ. Press.

MERRIMAN, M. (1884). *A Textbook on the Method of Least Squares*. Wiley, New York. (Eighth edition, 1911.)

NEYMAN, J. (1934). On the two different aspects of the representative method (with discussion). *J. Roy. Statist. Soc.* **97** 587–651.

PEARSON, E. S. (1968). Some early correspondence between W. S. Gosset, R. A. Fisher and Karl Pearson, with notes and comments. *Biometrika* **55** 445–457.

PEARSON, E. S. (1990). "*Student*," *A Statistical Biography of William Sealy Gosset* (edited and augmented by R. L. Plackett with the assistance of G. A. Barnard). Oxford Univ. Press.

PEARSON, K. (1892). *The Grammar of Science*. White, London.

PEARSON, K. (1894). Contribution to the mathematical theory of evolution. *Philos. Trans. Roy. Soc. London Ser. A* **185** 71–110.

PEARSON, K. (1894–1896). *Lectures on the Theory of Statistics* **1–5**. (Five volumes of notes taken by Yule. Item 84/2 in Pearson collection, University College, London.)

PEARSON, K. (1896). Mathematical contributions to the theory of evolution, III. Regression, heredity and panmixia. *Philos. Trans. Roy. Soc. London Ser. A* **187** 253–318.

PEARSON, K. (1902). On the systematic fitting of curves to observations and measurements, parts I and II. *Biometrika* **1** 265–303; **2** 1–23.

PEARSON, K. (1907). On the influence of past experience on future expectation. *Philosophical Magazine* **13** 365–378.

PEARSON, K. (1915). On the distribution of the standard deviations of small samples: appendix I to papers by "Student" and R. A. Fisher. *Biometrika* **10** 522–529.

PEARSON, K. (1920). Notes on the history of correlation. *Biometrika* **13** 25–45.

PEARSON, K. (1936). Method of moments and method of maximum likelihood. *Biometrika* **28** 34–59.

PEARSON, K. 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. Ser. A* **191** 229–311.

PLACKETT, R. L. (1972). The discovery of the method of least squares. *Biometrika* **67** 239–251.

PLACKETT, R. L. (1989). Discussion of "R. A. Fisher on the history of inverse probability" by S. Zabell. *Statist. Sci.* **4** 256–258.

PRATT, J. W. (1976). F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation. *Ann. Statist.* **4** 501–514.

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

SMITH, K. (1916). On the "Best" values of the constants in frequency distributions. *Biometrika* **11** 262–276.

SOPER, H. E. (1913). On the probable error of the correlation coeffcient to a second approximation. *Biometrika* **9** 91–115.

SOPER, H. E., YOUNG, A. W., CAVE, B. M., LEE, A. and PEARSON, K. (1917). On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A cooperative study. *Biometrika* **11** 328–413.

STIGLER, S. M. (1973). Laplace, Fisher and the discovery of the concept of sufficiency. *Biometrika* **60** 439–445.

STIGLER, S. M. (1986). *A History of Statistics*. Belnap, Cambridge.

STUDENT (1908a). The probable error of a mean. *Biometrika* **6** 1–25.

STUDENT (1908b). Probable error of a correlation coefficient. *Biometrika* **6** 302–310.

TROTTER, H. F. (1957). Gauss's work (1803–1826) on the theory of least squares. Technical Report 5, Statistical Techniques Research Group, Princeton Univ.

WELCH, B. L. (1958). "Student" and small sample theory. *J. Amer. Statist. Assoc.* **53** 777–788.

YULE, G. U. (1897). On the theory of correlation. *J. Roy. Statist. Soc.* **60** 812–854.

ZABELL, S. (1989). R. A. Fisher on the history of inverse probability (with discussion). *Statist. Sci.* **4** 247–263.