Statistical Science
2005, Vol. 20, No. 1, 32–49
DOI 10.1214/088342305000000025
© Institute of Mathematical Statistics, 2005

Fisher in 1921

Stephen Stigler

Abstract. Ronald A. Fisher's 1921 article on mathematical statistics (submitted and read in 1921; published in 1922) was arguably the most influential article on that subject in the twentieth century, yet up to that time Fisher was primarily occupied with other pursuits. A number of previously published documents are examined in a new light to argue that the origin of that work owes a considerable (and unacknowledged) debt to a challenge issued in 1916 by Karl Pearson.

Key words and phrases: R. A. Fisher, Karl Pearson, Kirstine Smith, maximum likelihood, minimum chi square, sufficiency, history of statistics.

1. INTRODUCTION

On November 17, 1921, Ronald A. Fisher read a paper to the Royal Society of London titled "On the mathematical foundations of theoretical statistics." It was published in the Royal Society's *Transactions* the following year (see Figures 1 and 2). The paper is an astonishing work: It announces and sketches out a new science of statistics, with new definitions, a new conceptual framework and enough hard mathematical analysis to confirm the potential and richness of this new structure.

The paper opened with a list of definitions that were in 1921 entirely novel to statistical theory, but they startle us now only by their familiarity; they include consistency, efficiency, estimation, likelihood, optimum and sufficiency. Not in the list but hardly out of sight (it even appears seven times in the definitions of the new terms) is another, even more basic statistical novelty: It is in this paper of Fisher's that the word "parameter" is first used in the modern statistical sense. "Parameter" signals the key to Fisher's framework, namely a limitation to parametric families. This was a crucial limitation, one that gave him the structure where the other concepts became meaningful and he could explore questions that could not previously have been addressed.[1]

Fisher's parametric inference built on a century of work by others, but in ways that none of the others had foreseen. Fisher's paper was to become a watershed for twentieth century mathematical statistics. For most of the last three-quarters of the twentieth century, hardly any work in any statistical school was immune from its influence. Neyman–Pearson, Wald and Bayesian statistics—all of these as they were developed in the twentieth century bore the hallmark of the structures first presented by Fisher in 1921. The full story of Fisher's paper and its consequences is the story of twentieth century statistics, but that is not the story I will tell here. Rather, I seek to address a different question: Where did this epochal work come from?

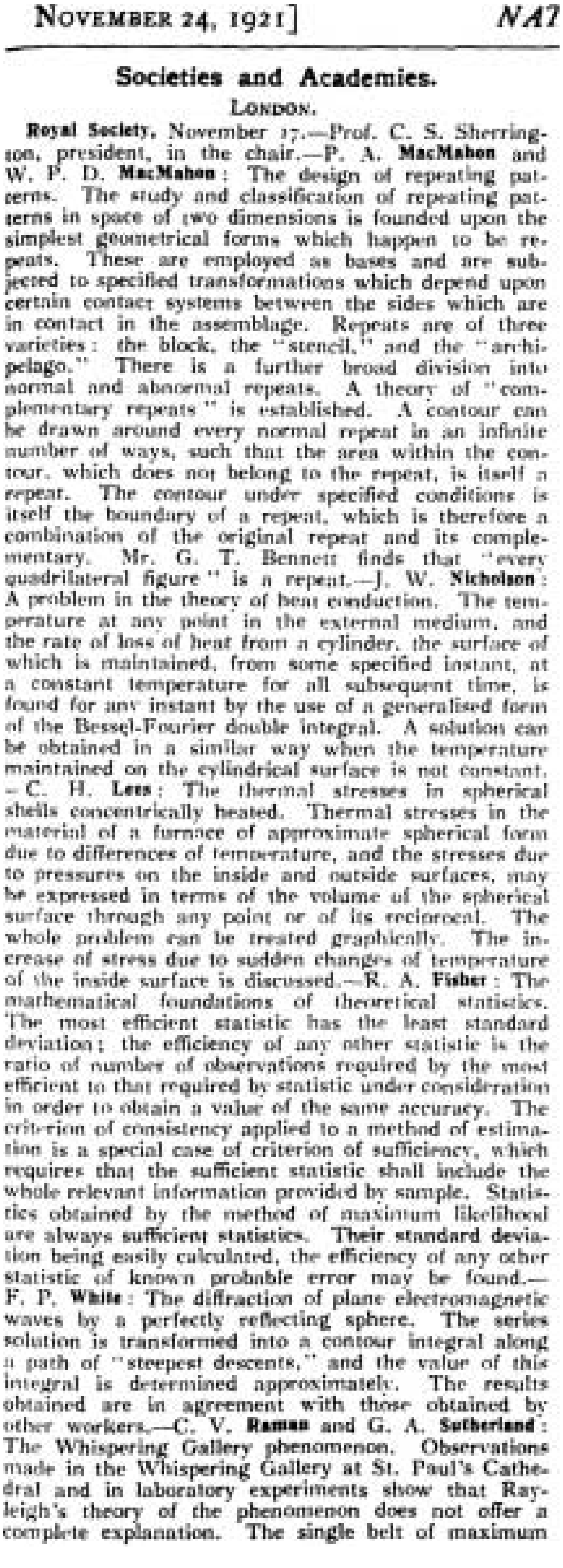
2. FISHER'S EARLY LIFE AND WORK

The source of the 1921 paper is a genuine puzzle. Neither Fisher's superficial biography before 1921 nor his record of publications for that period gives any clear indication of a tendency toward such a grand theory, and there is much in this record that points in other, contrary directions. I will make a stronger assertion: To the outside world the author of the 1921 masterpiece was not in evidence, certainly not before 1920.

Fisher's life history before 1921 was that of a very bright individual who was not fulfilling his early

Stephen Stigler is the Ernest Dewitt Burton Distinguished Service Professor, Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA (e-mail: stigler@galton.uchicago.edu).

[1] By my count the words "parameter" and "parameters" appear a total of 57 times in Fisher (1922). The word "parametric" also appears, once. There are a trivial number of appearances of "parameter" in earlier statistical literature, but only in senses that were ubiquitous in mathematics at that time, not in Fisher's statistical sense, as a quantity that was the object of estimation and determined the distribution from within a parameterized family. See Stigler (1976); the appearances listed there can be expanded through a modern search using JSTOR, but not greatly.



NOVEMBER 24, 1921] NAT

Societies and Academies.

LONDON.

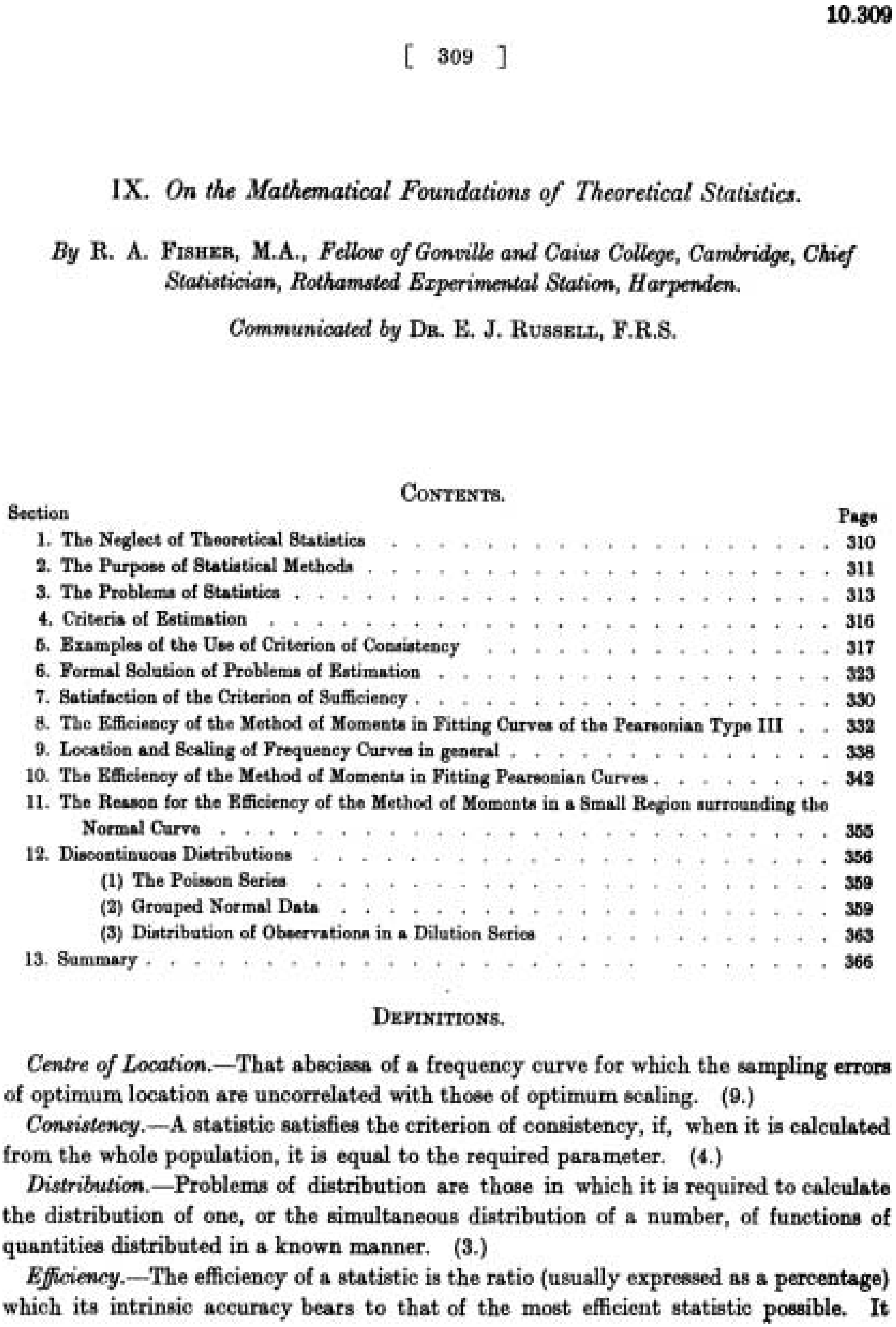
Royal Society, November 17.—Prof. C. S. Sherrington, president, in the chair.—P. A. **MacMahon** and W. P. D. **MacMahon**: The design of repeating patterns. The study and classification of repeating patterns in space of two dimensions is founded upon the simplest geometrical forms which happen to be repeats. These are employed as bases and are subjected to specified transformations which depend upon certain contact systems between the sides which are in contact in the assemblage. Repeats are of three varieties: the block, the "stencil," and the "archipelago." There is a further broad division into normal and abnormal repeats. A theory of "complementary repeats" is established. A contour can be drawn around every normal repeat in an infinite number of ways, such that the area within the contour, which does not belong to the repeat, is itself a repeat. The contour under specified conditions is itself the boundary of a repeat, which is therefore a combination of the original repeat and its complementary. Mr. G. T. Bennett finds that "every quadrilateral figure" is a repeat.—J. W. **Nicholson**: A problem in the theory of heat conduction. The temperature at any point in the external medium, and the rate of loss of heat from a cylinder, the surface of which is maintained, from some specified instant, at a constant temperature for all subsequent time, is found for any instant by the use of a generalised form of the Bessel-Fourier double integral. A solution can be obtained in a similar way when the temperature maintained on the cylindrical surface is not constant.—C. H. **Lees**: The thermal stresses in spherical shells concentrically heated. Thermal stresses in the material of a furnace of approximate spherical form due to differences of temperature, and the stresses due to pressures on the inside and outside surfaces, may be expressed in terms of the volume of the spherical surface through any point or of its reciprocal. The whole problem can be treated graphically. The increase of stress due to sudden changes of temperature of the inside surface is discussed.—R. A. **Fisher**: The mathematical foundations of theoretical statistics. The most efficient statistic has the least standard deviation; the efficiency of any other statistic is the ratio of number of observations required by the most efficient to that required by statistic under consideration in order to obtain a value of the same accuracy. The criterion of consistency applied to a method of estimation is a special case of criterion of sufficiency, which requires that the sufficient statistic shall include the whole relevant information provided by sample. Statistics obtained by the method of maximum likelihood are always sufficient statistics. Their standard deviation being easily calculated, the efficiency of any other statistic of known probable error may be found.—F. P. **White**: The diffraction of plane electromagnetic waves by a perfectly reflecting sphere. The series solution is transformed into a contour integral along a path of "steepest descents," and the value of this integral is determined approximately. The results obtained are in agreement with those obtained by other workers.—C. V. **Raman** and G. A. **Sutherland**: The Whispering Gallery phenomenon. Observations made in the Whispering Gallery at St. Paul's Cathedral and in laboratory experiments show that Rayleigh's theory of the phenomenon does not offer a complete explanation. The single belt of maximum

FIG. 1. *The summary of Fisher's 1921 paper as read to the Royal Society, from Nature, November 24, 1921. Note that mathematician P. A. MacMahon and physicists J. W. Nicholson and C. V. Raman also presented papers at the same meeting.*

promise (Box, 1978; Kruskal, 1980). He was born February 17, 1890, and his mother died when he was 14. His father went broke when he was $15\frac{1}{2}$, but not before Fisher was enrolled in Harrow School, where his mathematical talent earned him a medal in 1906 and sufficient scholarship support to see him through Harrow and Cambridge University. At Cambridge, he chose mathematics as his field and was a Wrangler in the 1912 Tripos, a distinguished achievement that helped earn him a further, postgraduate year of study. However, after he left Cambridge following the spring term in 1913, the promise receded from view.

Fisher spent the summer of 1913 working on a farm in Winnipeg, Canada (ostensibly to rest his congenitally weak eyes), and then he returned to London where he took a post with an investment company. He was ill-suited for and soon left that employment. He volunteered for military service in August 1914, but he was rejected by virtue of his weak eyes, and he then supported himself (and, after he married in 1917, a growing family) with a succession of teaching positions. He taught mathematics and physical science in secondary schools in Rugby, in Haileybury, on the training ship *Worchester* and then, from 1917–1919, at Bradfield College in Kent. By all accounts he was a poor teacher; he did not like his duties and the students did not understand him. In 1919 he moved to a research position at the agricultural experimental station at Rothamsted, a situation that proved to be much more congenial (but still appears a far cry from the theory of statistics).

If you seek a hint of the author of the 1921 paper, you must look for it in his published work from 1912–1920. There indeed hints are to be found, but the preponderance of evidence points in a different direction. That record is one of a mathematically able student who was turning his mind and energy to eugenics. Table 1 summarizes Fisher's 97 publications from 1912 to 1920; of these, 91 were in the *Eugenics Review*, two others were on questions in genetics related to eugenics (including an important work published in 1918), two more were papers written at Cambridge and published in the general mathematics magazine *The Messenger of Mathematics* and the other two (in 1915 and 1920) were on mathematical statistics. The publications in the *Eugenics Review* were predominantly short reviews of books involving eugenics and filled few pages, but they unfailingly showed he had, notwithstanding his weak eyes, read the book and understood its message, often better than the book's author. This is a record of great



10.309

[309]

IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. FISHER, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by DR. E. J. RUSSELL, F.R.S.

CONTENTS.

DEFINITIONS.

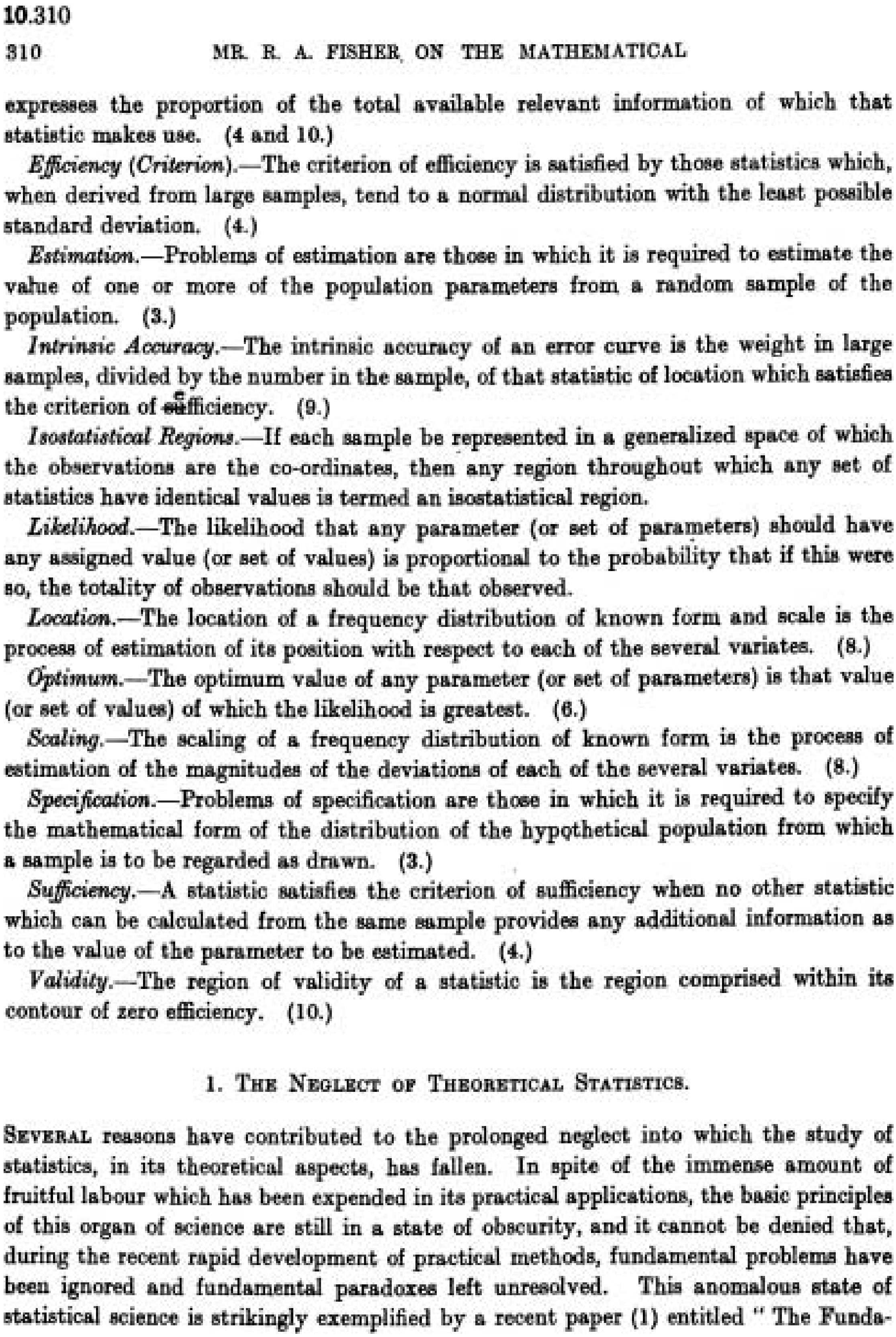
Centre of Location.—That abscissa of a frequency curve for which the sampling errors of optimum location are uncorrelated with those of optimum scaling. (9.)

Consistency.—A statistic satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter. (4.)

Distribution.—Problems of distribution are those in which it is required to calculate the distribution of one, or the simultaneous distribution of a number, of functions of quantities distributed in a known manner. (3.)

Efficiency.—The efficiency of a statistic is the ratio (usually expressed as a percentage) which its intrinsic accuracy bears to that of the most efficient statistic possible. It

FIG. 2. *The first two pages of Fisher's 1921 paper, as reprinted in 1950 with a small correction in Fisher's hand.*



10.310

310 MR. R. A. FISHER ON THE MATHEMATICAL

expresses the proportion of the total available relevant information of which that statistic makes use. (4 and 10.)

Efficiency (*Criterion*).—The criterion of efficiency is satisfied by those statistics which, when derived from large samples, tend to a normal distribution with the least possible standard deviation. (4.)

Estimation.—Problems of estimation are those in which it is required to estimate the value of one or more of the population parameters from a random sample of the population. (3.)

Intrinsic Accuracy.—The intrinsic accuracy of an error curve is the weight in large samples, divided by the number in the sample, of that statistic of location which satisfies the criterion of ~~su~~efficiency. (9.)

Isostatistical Regions.—If each sample be represented in a generalized space of which the observations are the co-ordinates, then any region throughout which any set of statistics have identical values is termed an isostatistical region.

Likelihood.—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.

Location.—The location of a frequency distribution of known form and scale is the process of estimation of its position with respect to each of the several variates. (8.)

Optimum.—The optimum value of any parameter (or set of parameters) is that value (or set of values) of which the likelihood is greatest. (6.)

Scaling.—The scaling of a frequency distribution of known form is the process of estimation of the magnitudes of the deviations of each of the several variates. (8.)

Specification.—Problems of specification are those in which it is required to specify the mathematical form of the distribution of the hypothetical population from which a sample is to be regarded as drawn. (3.)

Sufficiency.—A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated. (4.)

Validity.—The region of validity of a statistic is the region comprised within its contour of zero efficiency. (10.)

1. The Neglect of Theoretical Statistics.

Several reasons have contributed to the prolonged neglect into which the study of statistics, in its theoretical aspects, has fallen. In spite of the immense amount of fruitful labour which has been expended in its practical applications, the basic principles of this organ of science are still in a state of obscurity, and it cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved. This anomalous state of statistical science is strikingly exemplified by a recent paper (1) entitled " The Funda-

FIG. 2. *Continued.*

TABLE 1
Titles of Fisher's early published works, 1912–1920

1912	On an absolute criterion for fitting frequency curves (Mess. Math.)
1913	Applications of vector analysis to geometry (Mess. Math.)
1914	Some hopes of a eugenist (ER)
	Review of *Mechanism, Life, and Personality* by J. S. Haldane (ER)
	Review of *The Family in its Sociological Aspects* by J. Q. Dealey (ER)
1915	The eugenic aspect of the employment of married women (with C. S. Stock; ER)
	Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population (*Biometrika*)
	Cuénot on preadaption (with C. S. Stock; ER)
	The evolution of sexual preference (ER)
	Review of *The Shadow on the Universe* by I. M. Clayton (ER)
	Review of *Kinship and Social Organization* by W. H. R. Rivers (ER)
	Review of *The Progress of Eugenics* by C. W. Saleeby (ER)
1916	Racial repair (ER)
	Ethnology and the war (ER)
	Note on bibliography of eugenic literature (ER)
	After the war problems (ER)
	Notice about *Biometrika* (ER)
	Plus 23 more book reviews related to eugenics (ER)
1917	Disabled soldiers and marriage (ER)
	Positive eugenics (ER)
	Plus 13 more book reviews related to eugenics (ER)
1918	The correlation between relatives on the supposition of Mendelian inheritance (Roy. Soc. Edin.)
	The causes of human variability (ER)
	Plus 12 more book reviews related to eugenics (ER)
1919	The genesis of twins (*Genetics*)
	Plus 8 more book reviews related to eugenics (ER)
1920	A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error (Roy. Astron. Soc.)
	Review of *Inbreeding and Outbreeding* by East and Jones (ER)
	Plus 17 more book reviews related to eugenics (ER)

NOTES. ER denotes *Eugenics Review*; Mess. Math. denotes *The Messenger of Mathematics*; Roy. Soc. Edin. denotes *Transactions of the Royal Society of Edinburgh*; Roy. Astron. Soc. denotes *Monthly Notices of the Royal Astronomical Society*.

intellectual energy, nearly all of it directed to issues involving eugenics, a subject he had become passionate about while at Cambridge (Box, 1978). To most who knew him, Fisher must have seemed to have become a eugenicist. But the great 1921 work has no apparent connection to eugenics: What did inspire it?

3. FISHER AT CAMBRIDGE

Fisher had excelled in mathematics at Cambridge, and some signs of this are evident in the first two publications. The two articles he published in 1912 and 1913 appear to be based on university term papers, and the first of these is frequently cited as his earliest statement on maximum likelihood (Fisher, 1912). Box (1978, page 33) tells us that Fisher's Cambridge postgraduate year was to be devoted to the study of statistical mechanics and quantum theory under James Jeans, and the theory of errors under the astronomer Frederick John Marrian Stratton (1881–1960), a recently appointed lecturer in astrophysics. The first published paper, a short (6 page) article, "On an absolute criterion for fitting frequency curves," must have been dashed off early that year or even completed while he was an undergraduate, for it was already published in 1912. It has been the subject of extensive and perceptive recent commentary, notably by Edwards (1997b) and in more detail by Aldrich (1997). Other commentaries that address the place of this paper include Hald (1998) and Zabell (1989, 1992). The article was clearly the product of a very bright undergraduate, but the article itself would not be notable were it not for the author's subsequent career.

In that first article, Fisher acknowledges Stratton's assistance, and the text indicated he was engaged in some reading in error theory, reading that included Chauvenet (1891), a standard source on that topic, and a published 1908 lecture on "Errors of observations" by a British surveyor in Egypt, T. L. Bennett.[2] In a 1937 letter (J. H. Bennett, 1990, page 84) Fisher indicates his paper was motivated by a question T. L. Bennett (1908) had treated, namely in determining the scale parameter of a normal population, should the sum $\sum(x_i - \bar{x})^2$ be divided by $n - 1$ (the usual choice at that time) or by n?

Fisher's treatment was based on a classical approach of error theory, specifically in his recommending choosing as the "most probable set of values for the θ's" those that made the probability density of the data a maximum. This was known then by some as the Gaussian method; we (after Fisher, 1922) would see

[2] T. L. Bennett was with the Finance Ministry in Cairo, Egypt, and had at least one other publication, a later article in the *Journal of the Royal Statistical Society* on economics (Bennett, 1920).

this as maximizing the likelihood function. Such maximization had even then a long history, going back before Gauss even. Daniel Bernoulli, Lagrange, Lambert, Gauss and many others had made much the same statement many years before, with various degrees of clarity (see Edwards, 1974; Stigler, 1986, 1999a, Chapter 16). Fisher's language and notation echoed that of T. L. Bennett and error theory texts, as he focused essentially on one narrow point to decide the issue: The criterion for fitting functions or curves should avoid "theoretical indefiniteness." As Aldrich (1997) acutely observed, Fisher used the word "absolute" in his title in describing the criterion to denote the fact that it did not give inconsistent answers when the quantity sought was transformed by a nonlinear transformation. In the scale parameter case in question, Fisher would for a normally distributed sample maximize the probability density of the data x_i,

$$P = \frac{h^n}{\pi^{n/2}} \exp\Bigl(-h^2 \sum (x-m)^2\Bigr),$$

with respect to both mean m and scale h simultaneously, and he referred to his solution as maximizing "the inverse probability system." This gave him $m = \bar{x}$ and $2h^2 = n / \sum (x_i - \bar{x})^2$. That is, he would divide $\sum (x_i - \bar{x})^2$ by n, not by $n - 1$.

Fisher also stated that inverse probability should not be used to make statements about the probabilities that the unknowns would fall in intervals. His evident concern was not any objection to inverse probability *per se*, since his preferred solution was described in such terms. Rather, since the problem was not changed in essentials by arbitrary changes in scale for expressing the unknown, he worried about the inconsistency that would be introduced by adopting uniform priors for different choices of scale. Probability statements, when based upon uniform priors for an arbitrarily selected scale, were not invariant to nonlinear transformations of the unknown; an interval found for σ^2 would not simply be the square of an interval found for σ. He specifically criticized T. L. Bennett, who argued for the divisor $n - 1$ by integrating out m and maximizing $\int P\,dm$ with respect to h, to get $2h^2 = (n-1)/\sum (x_i - \bar{x})^2$. Fisher stated that such integration "is illegitimate and has no definite meaning with respect to inverse probability." That is, it was illegitimate because it gave no *definite* answer; nonlinear changes in scale before integration would yield different results.

Fisher here showed an interesting but statistically undeveloped focus on mathematical issues, and no considered view of the theoretical statistical basis for what he did. Indeed, as Aldrich notes, Fisher was himself inconsistent. Just after rejecting the usual solution (T. L. Bennett's), he endorsed another approach: Let b be the probability density of $\sum (x_i - \bar{x})^2$; then "we should expect the equation $\frac{\partial b}{\partial h} = 0$ to give the most probable value of h." This procedure, had Fisher then known the density b, would have given Bennett's solution, as he evidently later discovered in correspondence with Gosset (Pearson, 1968; Aldrich, 1997). As Aldrich (1997) said of the 1912 paper, "to make any of [its theoretical basis] explicit, we have to read outside the paper *and guess*." But we do not have to make it explicit. It is not that Fisher in 1912 had some considered view of statistical theory that either agreed with or differed from his later views. In 1912 Fisher had no such considered view at all.

Fisher's (1913) article, "Applications of vector analysis to geometry," which cites the mathematical physicist John William Nicholson (1881–1955) for assistance, was a more substantial effort. Nicholson was a lecturer at Trinity College, Cambridge before moving in 1912 to a Chair in Mathematics at the University of London, and presumably James Jeans had put Fisher and Nicholson in contact.[3] Fisher's article was much more than the title suggests: It was a *tour de force* review of the state of differential geometry at that time, developed in terms of J. Willard Gibbs's vector analysis. It may not have been an original work, but it showed an astonishing mastery of all the tools of a modern mathematician, and he displayed a visualization of high-dimensional spaces that shows more about his subsequent statistical work than does the 1912 article. However, there was no sign of any follow-up: The two university papers remain as the echo from an immensely promising student career that at least his mathematics teachers must have felt had run upon the shoal of eugenics.

4. READING *BIOMETRIKA*

Fisher's initial short pieces on eugenics—really talks he gave to society meetings—were primarily concerned with social questions, but they did show some scientific promise. Already in 1911 he had given a talk at Cambridge that showed he had encountered Karl Pearson's work on biometry and understood its use

[3] Nicholson's work in 1911–1912 played a crucial role in Niels Bohr's pioneering work on the spectra of atoms. In 1923 Nicholson married Dorothy Wrinch, Harold Jeffreys' coauthor in a frequently cited 1921 article on Bayesian inference that noted the insensitivity of the posterior to choice of prior when the sample size is large.

in the study of heredity (Norton and Pearson, 1976). Much of Pearson's work was presented in *Biometrika*, the journal Pearson had founded in 1901. *Biometrika* was already the preeminent scientific periodical in mathematical biology, and once Fisher opened it, he did not limit his view to the biological articles. One early sign of this was two letters he wrote to William Sealy Gosset ("Student") in September 1912 after he had first encountered Gosset's 1908 article on the t test. The letters do not survive, but it is clear from Gosset's correspondence with Pearson that month (Pearson, 1968) that Fisher had sent Gosset his 1912 article and had at least sketched out a high-dimensional geometric argument showing that what Gosset had conjectured in 1908 was indeed correct: The actual distribution of the t statistic was in fact what we now call the t distribution. Another sign, three years later, had a published outcome.

In 1914 Fisher had come upon a laborious study by H. E. Soper in the previous year's *Biometrika*, a study of approximations to the moments and distribution of the sample correlation coefficient. Fisher, with little apparent effort, used his powerful understanding of high-dimensional spaces to write down almost directly the exact sampling distribution of the correlation coefficient r, and he sent the result to Pearson for consideration. After responding to Pearson's prodding for ways to approximate the distribution by a normal distribution, prodding that among other things led to Fisher producing his hyperbolic tangent transformation of r, the article was published in *Biometrika* in May 1915 (Fisher, 1915; Pearson, 1968).

Mathematically the derivation was breathtaking. As brilliant as that short article was, however, it was still more a technical *tour de force* than a conceptual advance. This was Fisher the problem-solver at his best, but little more. He found the distribution, he found expressions for moments, he found ingenious transformations and distributional relationships (including describing his earlier solution to Gosset's t-distribution problem), he found expressions for the bias of r as an estimate of ρ, and he even applied the 1912 "absolute criterion" to find the "most likely" value for ρ. Since he based this maximization calculation on the marginal density of r, his result here was in conflict with Pearson and Filon (1898), where the maximization was for the five-parameter bivariate normal. If Fisher was aware of this conflict, he prudently did not point it out. Still, if this is all Fisher had accomplished, Neyman's later faint-praise description of Fisher as "a very able 'manipulative' mathematician" would have been defensible (Neyman, 1951). For all its brilliance and extensive calculation, it was for Fisher a quick effort, seemingly "tossed off" as a side dish from his main course of eugenics. Even so, it was one more bite from the statistical mix in *Biometrika*, and one bite did lead to another.

5. KIRSTINE SMITH

In the year following the publication of his article on the correlation coefficient, Fisher was teaching uncomfortably at Haileybury, reviewing books on eugenics and working on what would be his first major effort in genetics. That paper, published eventually in 1918 as "The correlation between relatives on the supposition of Mendelian inheritance," was an important work that carried the seeds of some of Fisher's later work on the analysis of variance and multivariate analysis. At the same time it addressed a question Karl Pearson had raised earlier: Was variation in human populations consistent with the Mendelian model for inheritance? Even while Fisher worked at all of that, however, he evidently also continued to develop the formulas for the distribution of the correlation coefficient, with the intention of producing another paper for *Biometrika*. Pearson was encouraging and made some suggestions, but the correspondence lapsed as both Pearson and Fisher worked on other things and Pearson pushed a group in his laboratory forward on a "cooperative study," without including Fisher as one of the cooperators. The publication of *Biometrika* was sporadic during the war, and in May of 1916 the first (and only) issue of that year appeared, inspiring another letter from Fisher to Pearson.

> Dear Professor Pearson,
>
> There is an article by Miss Kirstine Smith in the current number of *Biometrika* which, I think, ought not to pass without comment. I enclose a short note on it.
>
> I have recently completed an article on Mendelism and Biometry which will probably be of interest to you. I find on analysis that the human data is as far as it goes, not inconsistent with Mendelism. But the argument is rather complex.
>
> Yours v. truly,
> R. A. Fisher

This brief letter left much unstated. First, the article in question (Smith's "On the 'best' values of the constants in frequency distributions") stated clearly that it was written in Pearson's laboratory, and she thanked him specifically "for his aid throughout the work."

Second, Fisher would have known of Pearson's disputes with Mendelian biologists (including a violent quarrel with Bateson a few years earlier), with Pearson expressing skepticism that the Mendelians could account for observed variation in human populations (Porter, 2004, page 266ff). However, at that point Fisher's relations with Pearson by all accounts were relatively cordial, and on neither issue was he throwing down a gauntlet. As Porter argues, Pearson was not hostile to Mendelian genetics, only to biologists like Bateson who claimed that was the only legitimate approach and denied validity to his biometry. Fisher would have viewed his long paper as showing how biometry could be reconciled with Mendel, if one admitted the existence of a large number of unspecified traits. Pearson himself had found Mendel's rules congenial with some parts of biometry (Galton's "law of ancestral inheritance"), and Fisher was showing how with an intricate calculus of correlations the same ideas could be pushed further. The other issue Fisher raised, the observation on Smith's article, was critical, but Fisher still must have assumed Pearson would be interested in the point he made, inasmuch as it was based upon principles Pearson had previously endorsed and it defended Pearson's method of moments.

Kirstine Smith (1878–1939) was a bright young Dane who had arrived at Pearson's laboratory for doctoral study a year or so before. She is not well known to the history of statistics, although a long paper she published in *Biometrika* in 1918 is sometimes cited as a pioneering work in the design of regression experiments (Kiefer, 1959, page 295). The 1916 paper that animated Fisher's pen was short and had a limited goal: She suggested that when fitting a frequency curve with grouped data, the constants should be selected to minimize the chi-squared statistic, and she illustrated the use of this criterion through a series of examples. She granted that the labor involved was great and that the improvement in her examples over fits found by the method of moments was small, and so the practical advantages were slight at best. However, she did clearly state that compared to the use of the chi-squared measure of fit, other approaches were arbitrary in this setting, including what she termed "the Gaussian 'best' value," the approach that Fisher had adopted from error theory and embraced in 1912.

Fisher's handwritten submission survives at University College and occupies only a single page. It was printed in full in Pearson (1968). It begins:

> In your issue of May 1916 Miss Kirstine Smith proposes to use the minimum value of χ^2 as a criterion to determine the best form of a frequency curve; and proceeds to compare in a number of cases the distributions obtained by ordinary methods with those 'improved' by the use of χ^2. It should be observed that χ^2 can only be determined when material is grouped into arrays, and that its value depends upon the manner in which it is grouped.
>
> [Fisher gave a worked example showing this effect for fitting a normal curve to $n = 53$ data values, with five different groupings.]
>
> The Gaussian [method] would have to be 'improved' by shifting the mean not only by different amounts, but in opposite directions, in several cases.
>
> 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.

6. EDITOR PEARSON

Given Pearson's sponsorship of Smith, Fisher probably knew he was not sending Pearson entirely welcome news. On the other hand, he was limiting the point to the "arbitrary" effect of grouping and defending the method of moments, so there was some chance Pearson would agree to publish the note, and to Fisher in 1916 there was no other important point: The "Gaussian" method he used was widely accepted and traditional—even Pearson himself had employed it. Fisher's only personal stake in the method was that he had noted it to be "absolute"; it avoided indefiniteness due to choice of scale transformations for the quantities of interest or, the point here, to choice of grouping. He may or may not have expected a favorable reaction, but I doubt he expected the response he did get.

> June 26, 1916
>
> Department of Applied Statistics
> University College
>
> Dear Mr. Fisher,
>
> I am afraid that I don't agree with your

> criticism of Frøken K. Smith (she is a pupil of Thiele's and one of the most brilliant of the younger Danish statisticians). In the first place you have to demonstrate the logic of the Gaussian rule. I have the more right to ask for a proof as I followed it in 1897, but very much doubt its logic now. In the next place your argument that χ^2 varies with the grouping is of course well known and is one of the modes of finding the best grouping. What we have to determine, however, is with *given* grouping which method gives the lowest χ^2. Frøken shows there is extremely little difference, but she can get better fits by making χ^2 a minimum, always on the hypothesis that χ^2 a minimum is a more reasonable thing to start from than P a [maximum*]. I think the keynote to this is the footnote on page 263. Now if you look at Frøken's illustrations you will see that no choice of grouping is possible in Illustrations III, IV and V. In II she takes the actual facts as given by Bessel and asks whether values can be chosen which give a better fit than Bessel's moment values. I can see nothing whatever valid with the argument that if another grouping were taken χ^2 would change. Data must be grouped in all series of astronomical and anthropometric observations, even if only owing to the limitation in reading accuracy.
>
> It is clear to me that your true position for criticism must arise, not from saying that χ^2 a min. does not give a 'better value' for m and σ—it obviously must if you accept the χ^2 test—than the method of moments, but that you must demonstrate that the Gaussian method of making the ordinate of a certain contour of the multiple frequency-surface a maximum is more legitimate than making a minimum the chance of a series of observations as bad or worse than the observed series. I think the latter is the true test, not the Gaussian method. I frankly confess that I approved the Gaussian method in 1897 (see *Phil. Trans.* Vol. 191, A. p. 232), but I think it logically at fault now.
>
> If you will write me a defence of the Gaussian method, I will certainly consider its publication, but if I were to publish your note, it would have to be followed by another note saying that it missed the point, and that would be a quarrel among contributors.
>
> Yours very sincerely,
> Karl Pearson
>
> P.S. Of course the reason I published Frøken Smith's paper was to show that by *another* test than the Gaussian, the method of moments gave excellent results, i.e. her second conclusion.
>
> [*Pearson mistakenly wrote "minimum" here.]

Pearson's letter is remarkable. It is the letter of an editor of the first rank telling a novice author exactly what is wrong with his submission and exactly what would be needed to fix it. It is direct, it is honest and it is dead-on correct in its assessment. From his first undergraduate effort in 1912, Fisher had confidently and uncritically taken as self-evidently true the classical justification for maximization of what he would in 1921 call a likelihood function. This justification had been phrased by the error theorists in terms of a naïve appeal to inverse probability, but was used by them and Fisher in a way that amounts to saying that among all explanations being entertained for the data observed, choose the one that maximizes the chance that the data in hand would have been observed or, with continuous distributions, maximizes the probability density. How could such a "most probable" choice be criticized? Pearson himself had followed this route to find the sample correlation coefficient in 1897 (Pearson and Filon, 1898), but he rejected its logic now, he stated. Kirstine Smith, in the footnote Pearson called attention to in his letter, put the matter quite clearly, almost as if she were directly addressing Fisher's 1912 assertions:

> There is a point of some philosophical interest here which deserves further consideration. As is well known the Gaussian demonstration depends on making the product $P = [\prod\{\phi((x_s - \bar{x})/\sigma)\}$ where ϕ is the standard normal density], s being taken so as to include each individual observation, a maximum by varying σ and $\bar{x}$, the result being that the 'best' values are found from the first two moments. Now it will be observed that this is not the same idea as lies in the χ^2 test of goodness of fit. The conception of 'goodness' in that case

> is that we should measure the probability of a drawing from a certain population giving as divergent or *a more divergent result than that observed*. In other words while the Gaussian test makes a *single ordinate* of a generalized frequency surface a maximum, the χ^2 test makes a real probability, namely the *whole volume* lying outside a certain contour surface defined by χ^2, a maximum. Logically this seems the more reasonable, for the above product used in the Gaussian proof is not a probability at all. To make it a probability it must be multiplied by the product $\{\delta x_s\}$, and then the probability of the actually observed result, namely $x_1, x_2, \ldots, x_s, \ldots, x_q$, will be of course infinitely small, and what is made a maximum is an infinitely small probability. The exact meaning of $P\{\delta x_s\}$ when x_s is an actual observation is obscure, but it appears that the probability for constant indefinitely small ranges in the variates *in the neighborhood* of the observed values is made a maximum. But probability means the frequency of the recurrence in a repeated series of trials and this probability is in the case supposed *infinitely small*. It seems far more reasonable to make a finite probability, i.e. the probability of a divergence as great *or greater* than the observed a maximum, i.e. to use the χ^2 test and not the Gaussian principle. (Smith, 1916, footnote on page 263, her italics.)

Put simply, she stated that greatest probability is not impressive unless the probability is large. In continuous cases the maximum probability is infinitesimal, and even in discrete cases it tends to be very, very small. The best explanation for the data in this Gaussian sense may be a very poor explanation; a better case for the method is needed.

How did Fisher react to Pearson's remarkable letter? No reply survives. A comment in a October 21, 1918, letter from Pearson to Fisher ["also I fear I do not agree with your criticism of Dr. Kirstine Smith's paper and under present pressure of circumstances must keep the little space I have in *Biometrika* free from controversy..." (Pearson, 1968)] has been read by some as suggesting a revision was later submitted, but I doubt that. Fisher may simply have reminded Pearson of the earlier note while sending Pearson a triumphant offprint of Fisher (1918), or Fisher may have sent a note about Smith (1918), which had just appeared and which treated questions on experimental design.

Indeed, there is one good reason to suppose Fisher did not send a revised version: What could he have said in rebuttal? As I will argue below, he had no suitable reply until at least 1920. Rather, I suspect he grumbled to himself, entertained dark thoughts about Pearson and maybe even considered what he might do to show Pearson what real statistical science was. In short, he probably reacted in much the same way that rejected authors generally do.

In the event, Fisher's dark thoughts about Pearson were reinforced several times in the next two years. His major article reconciling Mendelism and biometry was discouraged by the Royal Society, Pearson being one referee. Pearson was not actually hostile to the paper: He did find the hypothesis of a large number of unspecified traits unconvincing without some empirical support and he probably did not fully understand the mathematics (which remain difficult today, even with the exegesis of the text by Moran and Smith, 1966). The article was published in 1918 in Edinburgh only with difficulty and the financial assistance of Leonard Darwin (Norton and Pearson, 1976; Bennett, 1983, pages 68–69). By 1917 Pearson's "Cooperative Study" of the correlation coefficient, inspired by Fisher's work on the distribution but done without Fisher's cooperation, was published. To make matters worse it included a section that made comments about part of Fisher's 1915 paper that were seen (with some justice) by Fisher as misrepresenting him. Fisher prepared a strongly worded rejoinder which only was printed in 1921 by Corrado Gini's new journal *Metron*, and then only after some of the language was softened from the first draft (Bennett, 1983, page 73). Although Fisher had been misunderstood in one passage, for the most part he had not been. As Pearson's letter and Kirstine Smith's footnote show, he had, like Daniel Bernoulli and others before him, simply failed to make the case for the "Gaussian method," maximum likelihood.

7. FISHER'S REPLY

Fisher put Pearson's letter to the side—it was one of a small amount of early correspondence he saved to the end of his life. The question Pearson posed was a difficult one and initially Fisher avoided it. His aforementioned reply to the "Cooperative Study" expressed indignation that he had been in one passage construed unfairly to be taking a Bayesian stance, but he would surely have known from Pearson's letter and Smith's carefully worded footnote that he was faced with a deeper and more difficult challenge. He was being asked to provide a mathematical and logical basis for his earlier approach, the Gaussian method, not merely to differentiate it from a naïve Bayesianism.

The germ of his reply came apparently by accident in the late spring of 1919. Fisher was again in problem-solving mode, considering the relative merits of two alternative estimates of the standard deviation of a normal distribution: one was based on the mean absolute deviation

$$\sigma_1 = \frac{1}{n}\sqrt{\frac{\pi}{2}}\sum |x_i - \bar{x}|$$

and the other was the value he found by his "absolute criterion" of 1912,

$$\sigma_2 = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}.$$

Fisher later stated that he was led to consider the question by a passage in Eddington's 1914 book *Stellar Movements*, and a surviving July 1919 letter from Eddington and a footnote Eddington contributed to the published article (Fisher, 1920) would support this. Indeed, Eddington (1914, page 147) had written, "... in calculating the mean error of a series of observations it is preferable to use the simple mean residual irrespective of sign rather than the mean square residual," adding as a note, "This is contrary to the advice of most text-books; but it can be shown to be true."

Fisher, with his newly developed facility for distribution theory, could not resist looking into this claim, which was so contrary to prevailing opinion as well as to what he had maintained in his 1912 work. Without much difficulty he was able to show that Eddington's recommendation, σ_1, had a standard deviation 14% greater than his own choice, σ_2, for large n. He went on to show that among all estimates based on the pth power of the residuals, that for $p = 2$ was best for large n. Thus far he was, without knowing it, echoing an 1816 investigation by Gauss. But to see what the situation was when the sample size was small and the standard deviations did not fully describe the distributions, he went on to consider the exact distributions for the case $n = 4$, and there he made an astounding discovery. His choice, σ_2, he found had a "unique character":

> From the manner in which the frequency surface has been derived, ... it is evident that:—
>
> *For a given value of σ_2, the distribution of σ_1 is independent of σ.*
>
> On the other hand, it is clear ... that for a given value of σ_1 the distribution of σ_2 does involve σ. In other words, if, in seeking information as to the value of σ, we first determine σ_1, then we can still further improve our estimate by determining σ_2; but if we had first determined σ_2, the frequency curve for σ_1 being entirely independent of σ, the actual value of σ_1 can give us no further information as to the value of σ. The whole of the information to be obtained from σ_1 is included in that supplied by a knowledge of σ_2.
>
> This remarkable property of σ_2, as the methods which we have used to determine the frequency surface demonstrate, follows from the distribution of the frequency density in concentric spheres over each of which σ_2 is constant. It therefore holds equally if σ_3 or any other derivate be substituted for σ_1. If this is so, then it must be admitted that:—
>
> *The whole of the information respecting σ, which a sample provides, is summed up in the value of σ_2.* (Fisher, 1920; Fisher's italics.)

In the era before the word processor, italics were marked by the author with a firm underscore, often a way of showing excitement as if by raising the voice. Fisher's excited italics shouted out: this was a remarkable phenomenon that must have been entirely unexpected. Not only did σ_2 have smaller standard deviation, it captured all the information in the strongest possible sense. Once σ_2 was reported there was nothing more to learn from the data about σ, and any attempt to improve was doomed to failure.

In this example he had an answer to Pearson that was absolutely compelling. This type of dominance by an estimate, the estimate found by the "Gaussian method" in fact, was a quantum leap beyond anything known before in statistics. But it was only one example, and a rather special one at that. Could it be generalized? Fisher must have set to work nearly immediately to explore this altogether new phenomenon, which he would later term "sufficiency.[4]" Already in that same 1920 article he could report that the dominant role of σ_2 for normal distributions could be nearly achieved by σ_1 when the curve instead of being normal was of the form

$$\frac{1}{\sigma\sqrt{2}}\exp\left(-\frac{|x-m|}{\sigma}\sqrt{2}\right)dx,$$

but how to go further?

Fisher never described how he went from the discovery of sufficiency to the full-blown theory he presented in 1921 and published in 1922, but since the results of what was clearly a major investigation were already sent to the Royal Society in late June 1921, only a year after the 1920 article appeared, there could have been little time for revision. So we may look to the published version for indications, and there, almost at the beginning, we find a particular telling clue.

The first three sections of the published article Fisher (1922) were rhetorical in nature: comments on the nature of statistics, gratuitous digs at Karl Pearson, and discussion of the logic of statistical method. With Section 4, "Criteria of estimation," he comes finally to the first technical material and, to no great surprise, he opens with the example of σ_1 and σ_2 from the 1920 article. After verbally introducing the criterion of efficiency ("That in large samples, when the distributions of the statistics tend to normality, that statistic is to be chosen which has the least probable error"), he moved to sufficiency and to a telling calculation that I suspect dates from the early weeks after his exciting discovery.

The discovery of sufficiency in 1919 had been precipitated by Fisher asking the following question: Grant that σ_2 is the superior estimate, could it perhaps be improved upon by exploiting the bivariate distribution of it and an alternative estimate? It is natural that he would then ask that same question more generally. Of course the answer must depend on the particular bivariate distribution involved, but he knew that frequently the distribution would be at least approximately bivariate normal, so that is where he started. He set up the problem with his definition of a sufficient statistic θ_1:

> [I]f θ be the parameter to be estimated, θ_1 a statistic which contains the whole of the information as to the value of θ, which the sample supplies, and θ_2 any other statistic, then the surface of the pairs of values of θ_1 and θ_2, for a given value of θ, is such that for a given value of θ_1, the distribution of θ_2 does not involve θ. In other words, when θ_1 is known, knowledge of the value of θ_2 throws no further light upon the value of θ.

Fisher then gave this far-reaching consequence of sufficiency:

> It may be shown that a statistic which satisfies the criterion of sufficiency will also fulfil the criterion of efficiency, when the latter is applicable. For if this be so, the distributions of the statistics will in large samples be normal, the standard deviations being proportional to $n^{-1/2}$.

The demonstration was simple and elegant. By hypothesis, the large sample bivariate density of θ_1 and θ_2 is

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times \exp\left(-\frac{1}{1-r^2} \times\left\{\frac{(\theta_1-\theta)^2}{2\sigma_1^2}-\frac{2r(\theta_1-\theta)(\theta_2-\theta)}{2\sigma_1\sigma_2} +\frac{(\theta_2-\theta)^2}{2\sigma_2^2}\right\}\right)$$

and the univariate density of θ_1 is

$$\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{(\theta_1-\theta)^2}{2\sigma_1^2}\right),$$

so the conditional density of θ_2 knowing the value of θ_1 must be

$$\frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-r^2}} \times\exp\left(-\frac{1}{2(1-r^2)}\left\{\frac{r(\theta_1-\theta)}{\sigma_1}-\frac{(\theta_2-\theta)}{\sigma_2}\right\}^2\right).$$

However, by the definition of sufficiency this cannot depend on θ, and so the θ's in the exponent must cancel and $r\sigma_2=\sigma_1$. Since the correlation r is never larger than 1.0, $\sigma_2 \geq \sigma_1$ and the claim is proved! Subject to the important proviso that there was an efficient estimate (which implied that the assumed normal approximations would hold), no estimate could have a smaller

[4] Actually, there had been a few isolated hints in earlier literature by people who did not discover sufficiency. Among these were Köbel in 1535 (Stigler, 1999a, page 361), Laplace in 1818 (Stigler, 1973), Newcomb in 1860 (Stigler, 1978, page 252) and Edgeworth in 1908 (Pratt, 1976, page 503).

standard deviation than the sufficient one, and so no estimate could improve upon it in any sense (in terms of the approximating normal distributions).

We can well imagine that once he saw this simple demonstration, Fisher would have seen immediately the direction his program—his answer to Pearson—could take. For a very general class of problems, sufficiency guaranteed the best possible accuracy of estimation, at least for large samples. How general was this class? He did know that for the examples he had considered, maximum likelihood estimates were sufficient statistics, so that should be a compelling answer to Pearson indeed—a proof that maximum likelihood was always best! The core of the theory of 1921 was that estimates found from the "Gaussian method," or as Fisher and others from then on would call them, maximum likelihood estimates, were sufficient and therefore efficient, even though the conditions under which this was true were not spelled out fully and precisely, then or later.

This enthusiasm for a while convinced him that sufficiency was a more general phenomenon than it really is. In fact, Fisher's November 1921 summary of the paper stated baldly that "Statistics obtained by the method of maximum likelihood are always sufficient statistics" (see Figure 1). But even by the time the paper was published in 1922, Fisher had backed off of this view somewhat, writing:

> For the solution of problems of estimation we require a method which for each particular problem will lead us automatically to the statistic by which the criterion of sufficiency is satisfied. Such a method is, I believe, provided by the Method of Maximum Likelihood, although I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect. Readers of the ensuing pages are invited to form their own opinion as to the possibility of the method of maximum likelihood leading in any case to an insufficient statistic. For my own part I should gladly have withheld publication until a rigourously complete proof could have been formulated; but the number and variety of the new results which the method discloses press for publication... (Fisher, 1922, page 323).

By 1925 Fisher had learned that maximum likelihood did not always imply sufficiency, and when the 1922 paper was reprinted in 1950 he made one minor alteration: in the definition of "Intrinsic Accuracy" where he had at first had "sufficiency" he minimally altered that word to "efficiency," as has been pointed out (Hinkley, 1980; Geisser, 1992) (see Figure 2).

The 1922 article was filled with examples, but pride of place was given to a long and intricate discussion of the efficiency of Pearson's favored method of moments for fitting the "Pearson family" of frequency curves. The labor in producing this investigation must have been immense, but it had a satisfying payoff: the method of moments was shown to *only* have high efficiency when the curve was near the normal curve (where it was fully efficient, in fact maximum likelihood). For other cases Pearson's method could perform abysmally. For Type III curves (Gamma and chi-squared densities), Fisher quoted efficiencies for low degrees of freedom dropping off from 0.2727 to 0.

Fisher had not forgotten Kirstine Smith, and in Section 12 he returned to her and the method of minimum χ^2. By now he had realized that minimum χ^2 actually would be efficient—indeed agree approximately with maximum likelihood—for large samples if the grouping of the data was not too fine, and he granted that point. However, he emphasized the differences, that for general problems with fine grouping the efficiency would be lost and that the justification for minimum χ^2 hinged upon its agreement with maximum likelihood: when the two agreed, all was fine, but when they did not, maximum likelihood was superior. He reworked one of Smith's examples involving grouped normal data to show that maximum likelihood essentially agreed with the method of moments with Sheppard's correction, while Smith's minimum χ^2 "corrected" the method of moments in the wrong direction. When he rewrote his theory in 1925 (Fisher, 1925), he again returned to minimum χ^2 and showed that while that method was efficient in large samples, it suffered a second order deficiency for finite samples. This idea, that employing the method of minimum χ^2 was equivalent to discarding the information in a fixed finite number of observations (and so the fraction of information lost would become insignificant in very large samples), was later refined and generalized by C. R. Rao under the name "second order efficiency" (Rao, 1961, 1962).

8. AFTER 1922

Fisher's published paper was largely ignored at the time. It was long and it was difficult in every sense:

hard mathematics and a mode of expressing results that could leave the reader in doubt as to exactly what had been demonstrated. Fisher himself cited it and built on the structure in several important papers over the next few years. One immediate payoff was Fisher's discovery of Pearson's error in specifying the degrees of freedom for the chi-squared statistic when parameters were estimated, a discovery that would scarcely have been possible before Fisher started thinking in terms of parametric families (Stigler, 1999a, Chapter 19). In 1925 Fisher rewrote the theory, and he corrected some problems noticed in his first formulation and expanded upon it in various other ways. In 1925 he also inaugurated the study of information in finite samples and explored this there and in later work, particularly in the 1930s (e.g., developing the idea of ancillary statistic; see Savage, 1976; Fienberg and Hinkley, 1980; Stigler, 2001). However, most others (such as Karl Pearson) who read the 1921 memoir either missed its significance or found it too hard or uncongenial to their own approaches and passed it by without audible notice. In December 1925 the American A. R. Crathorne quoted Fisher's criticism of Bayesian inference in an address on how a course in mathematical statistics should be designed, but the address gave its full attention to the Pearsonian structures that would soon be swept away by Fisher's work (Crathorne, 1926).

The first serious published engagement with Fisher's ideas by an author other than Fisher came only in 1928 when Jerzy Neyman and Karl's son, Egon S. Pearson, took Fisher's conceptual approach and ran with it in the first of their papers that created the modern theory of hypothesis testing (Neyman and Pearson, 1928). Harold Hotelling also came early to a deep appreciation of the work, and in 1928–1929 he almost wrote a book with Fisher to flesh out the mathematical arguments (Stigler, 1999b). Indeed, the vast influence of Fisher's work came either indirectly, from the adoption of his framework by others such as Neyman and their students (with as grudging credit as Fisher gave to Karl Pearson), or from Fisher's own promulgation of the methods derived from the theory through his widely used book, *Statistical Methods for Research Workers*, first published in 1925. Still, not all readers were receptive, even to indirect influence.

9. PEARSON, AGAIN

Fisher's argument that maximum likelihood estimates were generally functions of sufficient statistics, and therefore were generally asymptotically normal and efficient, with the smallest possible asymptotic standard deviation among consistent estimates, provided a program for much of twentieth century statistical theory. This program arose as a reply to Karl Pearson's challenge to the young Fisher in 1916. In a perfect world, perhaps Fisher would have been at least eventually grateful for the advice, and Pearson would have smiled and blessed the achievements he had encouraged, but the world was not that perfect. Fisher never showed any sense of appreciation to Pearson and he tended to be ungenerous in all later commentary. Fisher was commissioned to write a biographical article about Pearson for the *Dictionary of National Biography*, but several drafts were so grudging in acknowledging Pearson's accomplishments that it was actually rejected by the editors (Edwards, 1994). And Karl Pearson died without ever accepting Fisher's work as an answer to his challenge.

Karl Pearson's final comment on Fisher appeared in 1936, in the journal he had founded 35 years earlier, *Biometrika*. Pearson had been stung by a pointed comment about the inefficiency of the method of moments in a book review Fisher (1935) published in the *Annals of Eugenics*, another journal which Pearson had founded in 1925 but with Fisher succeeding him as editor in 1934. Pearson wrote to Fisher on August 28, 1935, giving the appearance of ignorance of Fisher's work:

> Dear Professor Fisher,
>
> I am ever ready to adopt new methods, if they are quicker and more exact than the old.
>
> Now I do not suppose you spend much, if any, time in fitting frequency curves; nevertheless I should like to have your method of fitting them to observations, which avoids the "traditional but inefficient method of fitting them by moments." (*Annals of Eugenics* Vol VI p. 252) It would aid me in many inquiries, if you would let me know the more efficient way.
>
> I am, yours sincerely,
> Karl Pearson

Fisher's reply on August 30, 1935, was civil but uncompromising.

> Dear Professor Pearson,
>
> The fullest examination of the method of moments in fitting the Pearsonian curves

> is in a paper "On the mathematical foundations of theoretical statistics," *Phil. Trans.* A, ccxxii. 309–368. High efficiencies are only obtained in the neighborhood of the normal curve. Efficient equations of estimation may always be obtained by the method of maximum likelihood. These equations are often transcendental, or, if algebraic, of an inconveniently high degree, and their solution, therefore, usually requires the devices ordinarily employed in solving transcendental equations.
>
> A method very generally applicable of obtaining an efficient solution approximate to the maximum likelihood solution is given in "Theory of statistical estimation," *Proc. Camb. Phil. Soc.*, xxii. 700–725 [i.e. Fisher's method of scoring], and an example in which even the theoretical cell-frequencies are functions difficult to manipulate (a heavily grouped Type I distribution) was worked by Koshal about two years ago in the Statistical Society's Journal, using the absolute values of the likelihood instead of its differential coefficients.
>
> Yours sincerely,
> R. A. Fisher

(Letters from the Fisher papers, University of Adelaide.)

As Pearson worked on what was to be his very last article (Pearson, 1936), he evidently also wrote in December 1935 to his son, Egon, asking if there was indeed any principle underlying Fisher's work. Egon's straightforward reply gave a good contemporary view of Fisher's work by a knowledgeable but skeptical reader.

> 16 xii 35
> My Dear Father,
>
> Thanks for your's of 14th. I quite agree as to the clearer meaning of fitting based on χ^2; and on the question of practicality, it is clear that to obtain the true maximum likelihood solution, immense labour would be required.
>
> The "principle" that Fisher would probably give to support his method is that (on certain assumptions) it obtains from the sample data estimates of the unknown population parameters that (in repeated sampling & fitting) have the *smallest standard errors*. The "proof" of this (of "Fisher"-type) has been given, I think, both in the Phil. Trans paper & the Cambridge Phil. Soc. Papers to which Fisher referred you to [Fisher 1922, 1925]. It depends however on the use of very *large samples*, and on the sampling distributions of the sample estimates of parameters such as a_1, a_2, m_1, m_2 in
>
> $$y = y_0\left(1 - \frac{x}{a_1}\right)^{m_1}\left(1 + \frac{x}{a_2}\right)^{m_2}$$
>
> being *Normally* distributed *about population values*.
>
> My impression is that under these limiting circumstances the solution obtained by the method of maximum likelihood will be precisely that of min χ^2.
>
> In writing, all that I think is necessary is to allow that Fisher *has* some principle behind his advocacy of Max. Likelihood. It is that of minimum standard errors for parameters (or constants) of fitted curve. There may be a lot of gaps in the proof that in a given case, his method in application will lead to such a result, but the principle is there in his mind.
>
> Yrs,
> E. S. P.

[University College London, Pearson Papers 563.]

Egon Pearson seems to have been unwilling to point out to his father that in the specific example given, maximum likelihood was far superior to the method of moments, the main topic of his father's article, even in terms his father would have at one time accepted (smaller standard errors); otherwise his comments were fair and accurate. His measured tone had little effect on his father, who died on April 26, 1936, but not before finishing his paper, which appeared in *Biometrika* in June 1936 and begins with the italicized sentence, "Wasting your time fitting curves by moments, eh?" It runs to 25 dense pages of quotations, criticism of numerical work and rehashing the old arguments as if the past 20 years had never occurred. Kirstine Smith's paper was again discussed and Fisher's 1922 use of it was criticized. Fisher's (1937) reply to this matched Pearson in spirit, in myopic attention to detail and nearly in length.

This last exchange did not show either of these statistical giants to advantage. In his biography of his father, Egon Pearson gently described Karl Pearson's position in this last paper as emphasizing "the difference between the world of concepts and the world of perceptual experience," philosophical views that Pearson had held since writing *The Grammar of Science* in 1892 (E. S. Pearson, 1938, page 123). However, it is hard not to believe that a younger or less defensive Karl Pearson might have seen a more generous Fisher as addressing the challenge in terms of accurate representation (smaller standard errors) that would have been even philosophically congenial. Porter (2004, page 245) called attention to Pearson's inconsistency in such matters in his recent biography.

After this last gasp from the old school, the general professional discussion in statistical circles accepted maximum likelihood and moved to other issues, although not without occasional returns to minimum χ^2. One such publication was a 1980 discussion paper by Joseph Berkson, where much of the discussion echoes that of 45 years earlier, although Kirstine Smith was now lost from view and uncited (Berkson, 1980). Smith had received her D.Sc. from the University of London in January 1918 and then returned to Denmark, where she worked as an applied researcher and teacher until her death in 1939. She published a third paper in *Biometrika* in 1922 on correlation coefficients in statistical genetics (Smith, 1922). It did not refer to Fisher.

10. CONCLUSION

Fisher's theory of estimation appeared almost full grown in 1921. It is a rare event in the history of science when a single work launches a new era in a science: Darwin on evolutionary biology, Gauss on number theory, Kolmogorov on probability, and Adam Smith on economics. It is rarer still when the work is a single, unanticipated article. Fisher was one such case. The intellectual child he produced after short gestation in 1921 was not without flaws. The purposefully vague statements Fisher made about the extent of the theory's range frustrated many readers; his unwillingness to acknowledge that others could advance his work and his grudging acknowledgment of earlier work infuriated others.[5] However, Fisher's was an epochal work, and the waves it made have outlasted that century.

Prior to 1921 Fisher had been immersed in eugenics, and when he had raised his head to investigate mathematical problems, he had only had the narrow vision of a problem-solver. In 1912 the problem had been to choose between two scale estimates for normal populations, and his solution had not looked beyond the avoidance of what he saw as mathematical inconsistency. In 1915 the problem was the distribution of the correlation coefficient, and his solution had not gone much beyond the direct answer produced by his brilliant visualization of the problem in high-dimensional space. In 1916 the problem was the analysis of Mendelian inheritance in human populations, and his solution was a *tour de force* in the calculation of correlations for high-dimensional data that would bear later fruit in statistical genetics, but introduced no new basic conceptual statistical structure. But in 1919 when the problem was to investigate Eddington's claim, again on the merits of two scale estimates for normal populations, and his (again brilliant) solution led to unexpected results: Fisher saw the problem with eyes that had been opened by Pearson's 1916 challenge. He was no longer simply a brilliant problem-solver, he was now working on a new plateau. He now had a much broader vision to go along with the talents he had already shown, and he would go on to use that vision to great advantage often in his remarkable subsequent career.

Fisher did not underestimate his own 1921 achievement. In a 1937 letter to the economist Henry Schultz at the University of Chicago, Fisher wrote a candid self-assessment of his work on statistical theory.

> Perhaps I ought to say that I do not personally agree with your remark that clearing up the χ^2 problem is the most useful thing I have done in statistical theory. The series of exact distributions on which the tests of significance are based was certainly more immediately fruitful, and I think the theory of estimation is certainly of more permanent value.

It would be hard to disagree with this summary.

[5] The one who came the closest was Francis Edgeworth at the end of a fine but obscure series of papers in 1908–1909 that remained unread and unappreciated for several decades; see the measured discussion by Pratt (1976). Arthur Bowley had called attention to Edgeworth's work in a 1934 Royal Statistical Society discussion, and later Neyman (1951) used it to undermine Fisher's priority for the theory. Edgeworth's work gives a nice illustration of Whitehead's statement, "To come very near to a true theory, and to grasp its precise application, are two very different things, as the history of science teaches us. Everything of importance has been said before by somebody who did not discover it" (Alfred North Whitehead, *The Organization of Thought*, 1917, as quoted by Robert K. Merton, 1967, page 1; Sills and Merton, 1991).

ACKNOWLEDGMENTS

This paper is based on material presented at the Morris H. DeGroot Memorial Lecture at Carnegie Mellon University, September 12, 2003. The present version has benefited from comments made on that occasion and since by many friends and colleagues, including Rob Kass, Steve Fienberg, Teddy Seidenfeld, Joan Fisher Box, Ted Porter and Brad Efron. I am grateful to the University of Adelaide for permission to consult the Fisher Papers in the University Library.

REFERENCES

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912–1922. *Statist. Sci.* **12** 162–176.

Bennett, J. H., ed. (1983). *Natural Selection, Heredity, and Eugenics. Including Selected Correspondence of R. A. Fisher with Leonard Darwin and Others*. Oxford Univ. Press, New York.

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

Bennett, T. L. (1908). *Errors of Observation* (Technical Lecture No. 4, 1907–1908; delivered February 22, 1908; Ministry of Finance, Survey Department, Egypt). National Printing Department, Cairo.

Bennett, T. L. (1920). The theory of measurement of changes in cost of living. *J. Roy. Statist. Soc.* **83** 455–462.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *Ann. Statist.* **8** 457–487.

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

Chauvenet, W. (1891). *A Manual of Spherical and Practical Astronomy*, 5th ed. (two volumes). Lippincott, Philadelphia. The appendix on least squares was also published separately by the same publisher in 1884 with the title *A Treatise on the Method of Least Squares, or the Theory of Probabilities in the Combination of Observations*.

Crathorne, A. R. (1926). The course in statistics in the mathematics department. *Amer. Math. Monthly* **33** 185–194.

Eddington, A. S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, London.

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

Edwards, A. W. F. (1994). R. A. Fisher on Karl Pearson. *Notes and Records Roy. Soc. London* **48** 97–106.

Edwards, A. W. F. (1997a). Three early papers on efficient parametric estimation. *Statist. Sci.* **12** 35–47.

Edwards, A. W. F. (1997b). What did Fisher mean by “inverse probability” in 1912–1922? *Statist. Sci.* **12** 177–184.

Fienberg, S. E. and Hinkley, D. V., eds. (1980). *R. A. Fisher*: *An Appreciation*. *Lecture Notes in Statist.* **1**. Springer, New York.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41** 155–160. Reprinted as paper 1 in R. A. Fisher (1974). Also reprinted in A. W. F. Edwards (1997a). *Statist. Sci.* **12** 35–47.

Fisher, R. A. (1913). Applications of vector analysis to geometry. *Messenger of Mathematics* **42** 161–178. Reprinted as paper 2 in R. A. Fisher (1974).

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521. Reprinted as paper 4 in R. A. Fisher (1974).

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinburgh* **52** 399–433. Reprinted as paper 9 in R. A. Fisher (1974). Also reprinted with commentary in P. A. P. Moran and C. A. B. Smith (1966). *Eugenics Laboratory Memoirs* **XLI**.

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 Roy. Astronomical Soc.* **80** 758–770. Reprinted as paper 12 in R. A. Fisher (1974).

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368. Reprinted as paper 18 in R. A. Fisher (1974).

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725. Reprinted as paper 42 in R. A. Fisher (1974).

Fisher, R. A. (1935). Review of *Biomathematics*, by W. M. Feldman. *Annals of Eugenics* **6** 252.

Fisher, R. A. (1937). Professor Karl Pearson and the method of moments. *Annals of Eugenics* **7** 303–318. Reprinted as paper 149 in R. A. Fisher (1974).

Fisher, R. A. (1974). *Collected Papers of R. A. Fisher* (5 volumes.) (J. H. Bennett, ed.). Univ. Adelaide.

Geisser, S. (1992). Introduction to “On the mathematical foundations of theoretical statistics,” by R. A. Fisher. In *Breakthroughs in Statistics* **1** (S. Kotz and N. L. Johnson, eds.) 1–10. Springer, New York.

Hald, A. (1998). *A History of Mathematical Statistics From 1750 to 1930*. Wiley, New York.

Hinkley, D. V. (1980). Theory of statistical estimation: The 1925 paper. *R. A. Fisher*: *An Appreciation. Lecture Notes in Statist.* **1** 85–94. Springer, New York.

Kiefer, J. C. (1959). Optimum experimental designs. *J. Roy. Statist. Soc. Ser. B* **21** 272–319.

Kruskal, W. (1980). The significance of Fisher: A review of *R. A. Fisher*: *The Life of a Scientist* (by Joan Fisher Box). *J. Amer. Statist. Assoc.* **75** 1019–1030.

Merton, R. K. (1967). *On Theoretical Sociology*. Free Press, New York.

Moran, P. A. P. and Smith, C. A. B. (1966). Commentary on R. A. Fisher’s paper on “The correlation between relatives on the supposition of Mendelian inheritance.” *Eugenics Laboratory Memoirs* **XLI**. Published for the Galton Laboratory, University College London, by Cambridge Univ. Press.

Neyman, J. (1951). Review of R. A. Fisher’s *Contributions to Mathematical Statistics*. *The Scientific Monthly* **72** 406–408.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* **20A** 175–240.

Norton, B. and Pearson, E. S. (1976). A note on the background to, and refereeing of, R. A. Fisher’s 1918 paper “On the correlation between relatives on the supposition of Mendelian inheritance.” *Notes and Records Roy. Soc. London* **31** 151–162.

PEARSON, E. S. (1938). *Karl Pearson*: *An Appreciation of Some Aspects of His Life and Work*. Cambridge Univ. Press, Cambridge. Originally published in 1936, 1938 in parts: *Biometrika* **28** 193–257, **29** 161–248.

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, 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. London Ser. A* **191** 229–311.

PORTER, T. M. (2004). *Karl Pearson*: *The Scientific Life in a Statistical Age*. Princeton Univ. Press.

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

RAO, C. R. (1961). Asymptotic efficiency and limiting information. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 531–546. Univ. California Press, Berkeley.

RAO, C. R. (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **24** 46–72.

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

SILLS, D. L. and MERTON, R. K. (1991). *Social Science Quotations*. Macmillan, New York. First appeared in 1991 as Vol. 19 of the *International Encyclopedia of the Social Sciences*.

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

SMITH, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* **12** 1–85.

SMITH, K. (1922). The standard deviations of fraternal and parental correlation coefficients. *Biometrika* **14** 1–22.

STIGLER, S. M. (1973). Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika* **60** 439–445. Reprinted in M. G. Kendall and R. L. Plackett, eds. (1977). *Studies in the History of Statistics and Probability* **2** 271–277. Griffin, London.

STIGLER, S. M. (1976). Contribution to discussion of "On rereading R. A. Fisher," by L. J. Savage. *Ann. Statist.* **4** 498–500.

STIGLER, S. M. (1978). Mathematical statistics in the early states. *Ann. Statist.* **6** 239–265.

STIGLER, S. M. (1986). *The History of Statistics*: *The Measurement of Uncertainty Before 1900*. Harvard Univ. Press.

STIGLER, S. M. (1999a). *Statistics on the Table*: *The History of Statistical Concepts and Methods*. Harvard Univ. Press.

STIGLER, S. M. (1999b). The foundations of statistics at Stanford. *Amer. Statist.* **53** 263–266.

STIGLER, S. M. (2001). Ancillary history. In *State of the Art in Probability and Statistics. Festschrift for Willem R. van Zwet* (M. de Gunst, C. Klaassen and A. van der Vaart, eds.) 555–567. IMS, Beachwood, OH.

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

ZABELL, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.* **7** 369–387.