Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange from Nature
Author(s): Karl Pearson, R. A. Fisher, Henry F. Inman
Source: *The American Statistician*, Vol. 48, No. 1 (Feb., 1994), pp. 2-11
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2685077
Accessed: 03/02/2009 13:18

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=astata.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician.*

http://www.jstor.org

Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange From *Nature*

Henry F. INMAN

In 1935, a letter to *Nature* criticizing the logic of statistical tests provoked published responses from Karl Pearson and R. A. Fisher. Their letters illustrate the attitudes of the two men toward the hypothesis-testing problem soon after the Neyman–Pearson formulation and shortly before Karl Pearson's death.

KEY WORDS: Goodness-of-fit test; History of statistics; Philosophy of scientific inference.

1. INTRODUCTION

In 1935 Karl Pearson (1857–1936) and R. A. Fisher (1890–1962) exchanged letters in *Nature* on testing statistical hypotheses. Compared with their other disagreements, public and private, this dispute proved to be relatively mild. These letters demonstrate the differing perspectives these men brought to bear on the problem of testing hypotheses and, more generally, their disparate views on the role of statistical inference in scientific inquiry. Pearson's lengthier letters offer more insight into his position than does Fisher's single letter, but Fisher's letter exhibits several interesting features of his style of argument. This exchange is also significant because it illustrates the conflict between what E. S. Pearson called *Mark I* and *Mark II* statistical inference after the initial development of the Neyman–Pearson approach to tests of statistical hypotheses.

Neither Fisher nor Karl Pearson was a stranger to scientific controversy, and each man had used letters published in *Nature* for public disputation. Pearson had contributed a stream of letters and reviews to *Nature* since 1882. Several of Fisher's disputes were waged in the letters column of *Nature*; his argument with "Student" about the merits of randomized versus systematic agricultural experiments is only one example. Although Pearson and Fisher alluded to some of their previous quarrels in their 1935 letters, the primary issue was the complaint of a working scientist who opined that the statistical tests developed by the two men were largely irrelevant to scientific practice. It is this focus on the confusing relationship between science and statistical inference—confusion frequently shared by statisticians as well as scientific investigators—that calls for our reconsideration of the responses Pearson and Fisher offered to their challenger. Fisher's letter is reproduced in Bennett's compilation of Fisher's papers (Bennett 1973, p. 328–329), and Karl Pearson's letters are cited by E. S. Pearson (1938) and Morant (1939), but here these three letters are for the first time considered together with the letter that initiated the correspondence. In addition to presenting the letters themselves, I provide introductory comments and a concluding discussion of the issues Pearson and Fisher examined and the context they chose for the development of statistical inference.

2. THE CHALLENGE

Pearson's and Fisher's letters were replies to an invitation to "some statistician of international repute" contained in a letter on statistical tests that appeared in *Nature* on August 3, 1935. The letter's author was Hugo John Buchanan-Wollaston (1883–1970), a naturalist on the scientific staff of the fisheries laboratory maintained at Lowestoft, England, by the Ministry of Agriculture, Fisheries and Food. Unable to pursue a university education for financial reasons, Buchanan-Wollaston obtained his scientific training by working as an assistant to marine zoologists at Liverpool and Larne, Northern Ireland. According to Lee (1992), Buchanan-Wollaston joined the Lowestoft laboratory shortly after the Marine Biological Association established it in 1902. By 1909, Buchanan-Wollaston was a member of the laboratory's scientific staff; he was transferred to the (then) Board of Agriculture and Fisheries in 1910, when the Board took control of the Lowestoft fishery laboratory and its investigations (Lee 1992, pp. 67–68, 71; Marine Biological Association 1912, pp. iii, vi). Buchanan-Wollaston soon thereafter moved to the Board's London laboratory. During World War I, Buchanan-Wollaston served as an officer in the Royal Flying Corps. After the war, he rejoined the Ministry of Agriculture and Fisheries; he returned to Lowestoft when the Ministry reestablished its fisheries laboratory there (Lee 1992, p. 114). During the 1920s and 1930s, Buchanan-Wollaston contributed a stream of reports and articles to the Ministry's *Fishery Investigations* and to publications sponsored by the International Council for the Exploration of the Sea. Buchanan-Wollaston remained at Lowestoft until the outbreak of World War II, when the staff of the fisheries laboratory was evacuated. Buchanan-Wollaston was moved to the Freshwater Biological Association's laboratory at Wray Castle, where he quickly became involved in their research program (Freshwater Biological Association 1943, p. 7; 1944, p. 7). Buchanan-Wollaston remained at Wray Castle as statistical advisor for a year after he retired from his position with

Henry F. Inman was Assistant Professor, Department of Mathematical Sciences, Virginia Commonwealth University, Richmond, VA 23284-2014. He is now at 2016 A Park Avenue, Richmond, VA 23220. The letters reproduced here are reprinted by permission from *Nature*, 136, pp. 182–183 (H. J. Buchanan-Wollaston), pp. 296–297 (K. Pearson), p. 474 (R. A. Fisher), and p. 550 (K. Pearson), Copyright 1935 by Macmillan Magazines Ltd. For information concerning H. J. Buchanan-Wollaston, the author thanks his son, Geoffrey Wollaston; Ian Pettman (Head of Library and Information Services), Freshwater Biological Association; and Derek Bate (Librarian) and A. J. Lee (retired Director of Fisheries Research), Ministry of Agriculture, Fisheries and Food, Lowestoft, England. For helpful comments on earlier versions of this article, the author also thanks E. L. Lehmann, an Associate Editor, and two anonymous referees.

 © *1994 American Statistical Association*

the Ministry of Agriculture and Fisheries in 1944 (Freshwater Biological Association 1945, p. 5; 1946, p. 5).

Buchanan-Wollaston's interest in statistical methods appears to have begun with his efforts to determine the distribution of commercial fish stocks in the North Sea on the basis of trawler sampling (Buchanan-Wollaston 1916) and with the use of fish-egg surveys to estimate the extent of fish populations in the North Sea (Buchanan-Wollaston 1911a, 1923, 1926; Lee 1992, p. 108). He was concerned with both the mechanical and statistical problems associated with trawler sampling in the open sea (Buchanan-Wollaston 1911b, 1927, 1929, 1937; Lee 1992, pp. 110–111). In 1923, Buchanan-Wollaston joined D'Arcy Thompson's attack on the use of scale rings to determine the age of herring, but Buchanan-Wollaston's alternative methodology was quickly rejected (Buchanan-Wollaston 1924; Lee 1992, p. 112). In 1924, Buchanan-Wollaston began a study of fish growth rates that led him to propose a graphical method for decomposing mixtures of normal distributions (Buchanan-Wollaston and Hodges 1929) that has occasionally been cited in the statistical literature.

Prior to 1933, Buchanan-Wollaston's statistical approach can be fairly characterized as graphical or mathematical interpolation. Sample estimates generally were taken at face value, although Buchanan-Wollaston at times did concern himself formally with the statistical reliability of his estimates (as in Buchanan-Wollaston 1923, pp. 26–28). Buchanan-Wollaston certainly performed no statistical tests. Like many working scientists since, Buchanan-Wollaston professed a belief that commonly used statistical tests were either obvious or irrelevant to the scientific problem of interest:

> In discussions on statistical tests with various Continental statisticians and users of statistical methods, I have been struck by their universal mistrust of modern statistical tests as developed by Pearson, Fisher and other workers in Great Britain. I have come to the conclusion that the main reason for this attitude is a perfectly sound reason, namely, that a test is used by many workers in Great Britain as a *simultaneous test* of the untruth of one hypothesis and the truth of the reverse hypothesis. There is *in fact* a large region in the distribution of the criterion for which neither a hypothesis nor its reverse can be assumed to be true. One or the other *is* true, of course, but the test cannot help us in coming to a decision on the matter. Judgment must be reserved. For example, we may wish to test whether a given sample differs significantly from a random sample from a normal population. Applying the χ^2 test, after finding the best fitting normal distribution, and using $p = 0.05$, say, as the level of significance, we may find that our sample is just not significantly abnormal.
>
> The χ^2 criterion is perfectly justifiable up to this point. It is quite unjustifiable, however, to assert that the reverse hypothesis is true, namely, that the sample *is likely* to have come from a normal population, unless we have other reasons to believe this, in which case, of course, the χ^2 is not used as a criterion of the truth of the reverse hypothesis. Given an equal possibility of an infinite variety of populations, the most likely group of distributions to have given it contains all those which will give the modal χ^2 value for the appropriate number of degrees of freedom. All these and an infinite number of others may be considered as *likely* to have given the sample, compared to the best-fitting normal distribution, which has indeed comparatively a very small likelihood. This likelihood is sufficient, however, to prevent our assuming abnormal distribution.
>
> It is often of scientific and practical interest to investigate whether Gauss's law or other simple laws of distribution apply to a sample. There is no doubt that the χ^2 test, as usually applied, is quite useless for this purpose, though it may be most useful as a test of *significant heterogeneity*, using a low value of P as a criterion. It seems only reasonable that but a small part of the centre of the χ^2 distribution should be used as a test of *fit*.
>
> I believe the mistrust of British methods on the part of statisticians of other countries to be due partly to their failure to realize that the word 'normal' is usually employed to cover samples which are likely to have arisen from populations the estimates of the mean and other parameters of which have distributions very similar to those of the corresponding normal parameters. The fact that British methods 'work' is due to the prevalence in Nature of distributions *similar* to the Gaussian rather than to any peculiar virtue in the methods themselves. I am writing this in the hope that some statistician of international repute will be tempted to treat the matter fully in some publication such as *Nature* available to statisticians of all countries. (Buchanan-Wollaston 1935a)

Buchanan-Wollaston's criticism of statistical tests is particularly interesting in light of his relationship with Fisher. At some point between 1931 and 1933, Buchanan-Wollaston was permitted by his Ministry to study with Fisher at Rothamsted, England. While visiting Rothamsted, Buchanan-Wollaston carried out an analysis of sample data, "working under the personal supervision of Dr. Fisher," directed toward the possibility of distinguishing "races" of herring on the basis of counts of their vertebrae. Assuming an underlying normal distribution for the observed discrete vertebral counts in each of his samples, Buchanan-Wollaston estimated the parameters of these normal distributions on the basis of grouping by maximum likelihood, determined that differences among these distributions could be explained by differences among the means, and proceeded to an unbalanced analysis of variance through multiple regression with indicator variables. In his written account of this analysis, Buchanan-Wollaston claimed that his main object was "to introduce into fishery research some of the most important methods developed by R. A. Fisher"; Fisher himself contributed a short introductory note (Buchanan-Wollaston 1933). Buchanan-Wollaston also published expositions of statistical analysis that were based heavily on Fisher's (1925a) work (Buchanan-Wollaston 1935b, 1936), and the sophistication of his later statistical work shows Fisher's continuing influence (Buchanan-Wollaston 1935c, 1938, 1945, 1958). According to his son, Buchanan-Wollaston always held Fisher in the highest regard.

There was also a much more indirect link between Buchanan-Wollaston and Karl Pearson. During the Marine Biological Association's operation of the Lowestoft laboratory at the beginning of Buchanan-Wollaston's scientific career, W. F. R. Weldon was one of the governors of the council that supervised the fishery investigations, and Edgar Schuster served as statistical advisor and member of the council (Marine Biological Association 1905, p. iii; 1907, p. iii; 1912, p. iii). Schuster served on the fishery research advisory committee of the Board of Agriculture and Fisheries after the Board assumed responsibility for the Lowestoft investigations (Board of Agriculture and Fisheries 1913, p. 2). Pearson's close personal friend and colleague, Weldon was the initiator of the biometric research program that prompted Pearson's first statistical investigations. Later assistant secretary of the forerunner

of the Medical Research Council of Great Britain, Schuster had been Weldon's student, the first Galton Research Fellow at University College London, and author of two of the first research memoirs issued after Pearson took control of the Galton Laboratory. Schuster had actively defended Pearson during the controversies that developed from Pearson's eugenic investigations.

3. PEARSON'S FIRST RESPONSE

In 1935 Karl Pearson was 78. Two years earlier, he had retired from his professorship at University College London. Despite his objections, Pearson's statistical laboratories were divided into separate departments of statistics and eugenics. His son, E. S. Pearson, became head of the new Department of Statistics. Karl Pearson's successor as Galton Professor of National Eugenies was R. A. Fisher. In office space provided by the Department of Zoology, Pearson continued to edit *Biometrika* until his death in April 1936.

In this letter, published on August 24, 1935, Pearson refers to the χ^2 goodness-of-fit test as the P, χ^2 test; P represents the observed significance level obtained from the χ^2 approximation to the distribution of Pearson's test criterion (K. Pearson 1900; see also K. Pearson 1916, 1922, 1932). The P, λ_n test to which Pearson also referred was Pearson's proposal for a small-sample alternative to the usual χ^2 test that is based on grouping the sample data values. (In this test, λ_n was the product of the individual sample data values transformed by the hypothesized distribution function integral. Pearson noted that the transformed values could be viewed as independent uniform (0, 1) deviates when the null distribution held, and thus $-\log_e(\lambda_n)$ followed a gamma distribution. Pearson argued that large values of this test statistic indicated lack of fit for the hypothesized distribution, and the observed significance level P was calculated accordingly. As in his presentation of the χ^2 test, Pearson chose to ignore the effect of estimating from the sample the parameters necessary for the integral transformation on the subsequent test for uniformity with the transformed data values; see K. Pearson 1933.) Pearson also used *graduation* throughout his discussion to connote a fitted mathematical model for the observed data. In the common case of data supposed to have been obtained from a continuous distribution, the *graduation curve* is the estimated distribution curve or density.

> As the originator of the P, χ^2 test, I should be glad if you can spare me space for some reply to Mr. Buchanan-Wollaston. I should like first to state that I am in no way responsible for all the applications which have recently been made of that test, and do not accept the validity of some of the applications which Prof. R. A. Fisher has made of it in his well-known textbook. I am not concerned with his position and leave him to defend it. My own position is as follows:
>
> (i) I introduced the P, χ^2 test to enable a scientific worker to ascertain whether a curve by which he was graduating observations was a reasonable 'fit'. On this account, and as a measure of success in graduation, I termed it a 'goodness of fit' test. It had no special relation to the normal curve or to any other curve. The scientific worker in the past had chosen any curve he pleased to graduate his observations, but he rarely applied any measure of its aptness, beyond looking at a graph to 'see' whether it was a 'good fit'. The pages of the Royal Society *Transactions* and *Proceedings* are evidence enough of this fact.
>
> (ii) As a measure of 'goodness of fit', the P, χ^2 test does enable one to compare the relative advantages of any two graduation curves. But I personally have never assumed that the better graduation curve was the one from which the material had actually been drawn.
>
> (iii) I have shown both theoretically and experimentally that there is a high correlation between the 'goodness of fit' of a graduating curve to a *sample*, and the 'goodness of fit' of that curve to the parental population from which the sample has been drawn. Accordingly, *if the sample be large*, the graduating curve may be taken as representing reasonably the parent population.
>
> (iv) I have shown that, when dealing with *small* samples, no real distinction can be made between sampling from, say, a normal curve or a rectangle. It requires at least a sample of more than 100 individuals to determine whether it would be best to use a rectangle or some other curve! I have repeatedly insisted that little can be learnt of the superiority of one graduating curve over another, if the sample be not of considerable size, say, well beyond the 100 mark.
>
> All this proves that the P, χ^2 test has no relation to Mr. Buchanan-Wollaston's remark that: "The fact that British methods 'work' is due to the prevalence in Nature of distributions *similar* to the Gaussian [sic] rather than to any peculiar virtue in the methods themselves." It would appear from this remark that my critic and his 'Continental workers' have never gone beyond applying the test to questioning whether the normal curve was a reasonable graduating curve!
>
> (v) The only relation of the P, χ^2 test to the normal curve arises from the use of that curve in the analysis to replace binomials by normal curves. Such replacement is not legitimate *theoretically*, when in the binomial $(p+q)^n$, p is very much larger or very much smaller than q. This has led to the practice of clubbing together small 'tail' groups. But practically there is, *as a rule*, very small difference in the resultant P's, whether we club tail groups and reduce the number of cells, or work P out for the full number after considering outlying individuals which may be anomalous. I do not therefore understand Mr. Buchanan-Wollaston's remark that: "it seems only reasonable that but a small part of the centre of the χ^2 distribution should be used as test of *fit*." In a large percentage of cases to which χ^2 may be applied in biometric and biological investigations, there are no 'tails', that is, no small categories at the terminals. If we wish to avoid the assumption that at such 'tails', where they exist, it is legitimate to replace binomials by normal curves, then the P, λ_n test can be applied.
>
> (vi) From my point of view, the tests are used to ascertain whether a reasonable *graduation* curve has been achieved, not to assert whether one or another hypothesis is true or false. If we narrow ourselves down to asking whether a normal curve will reasonably *graduate* the material and find it does, are we to follow it up by asserting that either the sample or parent-population follows a normal distribution? I should say: Certainly *not*. I have never found a normal curve fit anything if there are enough observations! The astronomical data provided to prove that errors of observation follow normal curves are pitiably scanty, and if proper tests are applied usually show that they do not! The fact is that all these descriptions by mathematical curves in no case represent 'natural laws'. They have nothing in this sense to do with 'hypothesis' or 'reverse of hypothesis'. They are merely *graduation curves*, mathematical constructs to describe more or less accurately what we have observed.
>
> (vii) The reader will ask: "But if they do not represent laws of Nature, what is the value of graduation curves?" He might as well ask what is the value of scientific investigation! A good graduation curve—that is, one with an acceptable probability—is the only form of 'natural law', which the

> scientific worker, be he astronomer, physicist or statistician, can construct. Nothing prevents its being replaced by a better graduation; and ever bettering graduation is the history of science.
>
> What is the use of good graduation curves? Ask the actuary! Such curves enable a mass of details to be summed up with reasonable probability in the knowledge of a few constants, and from those constants we obtain new knowledge of the properties of the mass. Take only the importance of a life table graduated by the Makeham-Gompertz curve and consider, what new knowledge flows from it. But after all, it is only a graduation curve and it is open to anyone to find a better one! If Continental statisticians in the bulk do indeed hold the views of Mr. Buchanan-Wollaston, it can only be that they have not really studied and grasped the fundamental literature of the subject. (Pearson 1935)

4. FISHER RESPONDS

R. A. Fisher was 43 when he became the director of the Galton Laboratory and head of the new Department of Eugenics, fulfilling an ambition he had nurtured since 1919 (see Darwin to Fisher, August 7, 1919, and Fisher to Darwin, February 25, 1929, in Bennett 1983, pp. 70, 96–97). Before formally assuming his new position, Fisher tried to reach an accommodation with the younger Pearson that would have permitted Fisher to lecture on statistical methods, but Fisher reluctantly agreed to E. S. Pearson's proposal that Fisher abstain from teaching statistical theory (Fisher to E. S. Pearson, May 24, 1933, in Box 1978, p. 258). Fisher's relationship with E. S. Pearson quickly deteriorated. Fisher also found himself in an awkward situation regarding the continuing staff of his department, who remained loyal to Karl Pearson (Box 1978, pp. 260–261). The already difficult relations between Karl Pearson and Fisher were aggravated further by the sometimes rocky transition. Shortly before their 1935 *Nature* letters, for example, Fisher's efforts to find space for his blood serological group by removing the material Karl Pearson left behind in his statistical and eugenic museum clearly irritated the elder Pearson (Fisher to K. Pearson, May 7, 1935, and K. Pearson to Fisher, June 13, 1935, both in Box 1978, pp. 346–347; for a description of the museum, see E. S. Pearson 1938, pp. 216–217).

Between 1933 and 1936, Fisher entered into several new public disputes with Neyman, Gosset, and Wishart, among others, about statistical theory and practice; but Fisher did not forget his longstanding quarrels with Karl Pearson. Fisher clearly resented the acclaim and support Pearson had received to maintain his statistical program while, it evidently seemed to Fisher, Pearson frustrated or ignored Fisher's own contributions to the theories of statistics and heredity. Their disagreements had begun during World War I with papers Fisher submitted to Pearson for publication in *Biometrika* concerning the distribution of sample correlation coefficients and the minimization of χ^2 as a criterion for statistical inference (see the exchange of correspondence and commentary in E. S. Pearson's 1968 work). Fisher's challenges to the method of moments and its use with Pearson's system of frequency curves are well known (Fisher 1922a, 1925b), as are Fisher's contributions to the use and interpretation of the χ^2 test (Fisher 1922b, 1923, 1924). Fisher's (1937) infamous article continued his dispute with Pearson after Pearson's death; the fierce attack there was later repeated in Fisher (1956).

Fisher's reference to F. R. Helmert in his letter to *Nature*, which was published on September 21, 1935, illustrates his practice of citing a prior authority to belittle one of Karl Pearson's contributions to mathematical statistics; in the context of the goodness-of-fit test, the reference to Helmert was misleading, but Fisher's tactic was not new. In fact, Helmert derived the χ^2 form of the sampling distribution of the sample variance on the basis of a normal population, and Karl Pearson had publicized Helmert's earlier derivation (Pearson 1931). Of course, Fisher's comments on "errors of the second kind" were veiled references to the Neyman–Pearson formulation of the hypothesis-testing problem.

> In a letter to *Nature* of August 23, Prof. Karl Pearson states: "From my point of view, the tests are used to ascertain whether a reasonable *graduation* curve has been achieved, not to assert whether one or another hypothesis is true or false."
>
> This assertion must come as a surprise to many who are familiar with Prof. Pearson's writings. It should not, however, be permitted to divert attention from the points raised in Mr. Buchanan-Wollaston's letter of August 3, for whatever may have been Prof. Pearson's original intention in introducing the term 'goodness-of-fit', and in publishing a table of the distribution of χ^2 (the theoretical form of which had been previously determined by Helmert in 1875), it is certain that the interest of statistical tests for scientific workers depends entirely from their use in rejecting hypotheses which are thereby judged to be incompatible with the observations.
>
> It is certain, too, from many passages which could be cited from Prof. Pearson's own writings, that he has himself used the χ^2 test, not only in connection with the graduation of frequency curves, but also as a means of testing the truth of theories or hypotheses. As one example, I may mention an appendix of five pages entitled "On the Test of Goodness of Fit of Observation to Theory in Mendelian Experiments" (*Biometrika*, 9, pp. 309–314). In this paper he insists very clearly, and quite in accordance with modern usage, taking the extreme case $P = 0$, that either the theory or the observations must be rejected.
>
> Mr. Buchanan-Wollaston's point that the χ^2 test, like the other tests of significance, is cogent for the rejection of hypotheses, but, in the opposite case, by no means cogent for their acceptance, deserves to be widely appreciated. For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning. Yet it does so only too frequently. Indeed, the "error of accepting an hypothesis when it is false" has been specially named by some writers "errors of the second kind". It would, therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as these are contradicted by the data; but that they are never capable of establishing them as certainly true. In fact that "errors of the second kind" are committed only by those who misunderstand the nature and application of tests of significance. (Fisher 1935)

5. PEARSON'S SECOND RESPONSE

It did not take Karl Pearson long to react to Fisher's criticism: Pearson's second letter appeared in *Nature* on October 5, 1935. Here he reemphasized the distinction between judging the value of a proposed model and the acceptance or rejection of hypotheses. Pearson's argument that scientific law was merely a convenient construct that adequately summarized our experience retraced a philosophical position he had propounded half a century before (K. Pearson 1892).

Table 1. Jeans's Star Eccentricity Distribution From Pearson's Second Response (1935b)

Eccentricity	Observed	Theory for 116 stars, $e < 1$	Theory for 83 stars, $e \leq 0.06$
.00– .01	0	4.5	9
.01– .02	11		
.02– .03	9	6.0	12
.03– .04	14	8.0	16
.04– .05	24	10.5	21
.05– .06	25	13.0	25
.06– .07	6	15.0	—
.07– .08	13	17.0	—
.08– .09	7	20.0	—
.09–1.00	7	22.0	—

The example Pearson cites in this letter deals with J. H. Jeans's (1935) use of an investigation of the orbital motions of binary stars to argue for a long (10^{13} years) rather than a short (10^{10} or 10^{11} years) time scale for the age of the universe. Jeans claimed that under conditions consistent with the long time scale (and inconsistent with the short time scale), "the number of orbits of binary stars whose eccentricity is less than $\underline{e}$ will be proportional simply to $\underline{e}^2$" (Jeans 1935). Jeans concluded that the agreement between the observed distribution of orbital eccentricity of 116 binary stars and his derived distribution was "far too good to be accidental," thus providing "strong evidence in favor of the long time scale" (Jeans 1935).

> Prof. Fisher is an apt controversialist, but he knows as well as I do that what I understand by *graduation* is not confined to *curves*; that I should term graduation the fitting of a binomial to a series of observations, or the determining whether a system of correlation coefficients could be reasonably supposed to have arisen from samples of material drawn from a population with a given correlation coefficient. The difference between Prof. Fisher and myself lies in the use (and abuse) of the acceptance and rejection of 'hypotheses'. There is only one case in which an hypothesis can be definitely rejected, namely when its probability is zero. He cites a case which I criticized in the paper he refers to, in which two recessive (say) had produced a dominant, and theory was absolutely contradicted. It did not require an *application* of the (P, χ^2) test to assert that either theory or observations must be rejected! I merely showed that the (P, χ^2) test did not fail in this case. But let us look into what actually happens, and I cannot do better than illustrate it on some statistics provided by Sir James Jeans in *Nature* of September 14, 1935 (p. 432). He is comparing the eccentricities of visual binaries, 116 in number, against a theory of equipartition (not a *curve*, but frequencies are considered). His data expressed by a frequency series run as follows: [Pearson's table is reproduced as Table 1].
>
> If the P, χ^2 test be applied to the total 116 binaries, we have $P < 0.000,0005$. On the other hand, if it be applied to the 83 stars of lowest eccentricity, $P = 0.79$. In neither case can you say the hypothesis is true or false. You reject it in the former case because it is a poor graduation, you say in the latter case that it is a reasonably good graduation because 79 percent of random samples would, were the "hypothesis" true, give a worse result than the observations do. But in accepting it as a working graduation, you do not assert its truth any more than you assert the falsity of the hypothesis applied to the whole 116 stars; you merely say the latter case is a bad graduation, and try for a better. Had Sir James Jeans taken all stars with eccentricity ≤ 0.07 instead of ≤ 0.06, he would have found $P = 0.105$, and if he had proceeded to $e \leq 0.08$, the result would have been $P = 0.00001$, that is, he might have got a worse sample in 100,000 trials. Actually he gives his reasons for cutting off the higher eccentricities. With them I am not concerned, although the exact cutting off at $e = 0.06$ is not discussed; the difficulty of detecting high eccentricity binaries and of then determining their orbits may account for the irregularity of the last four frequency entries, as he holds, or there may be other reasons why the falling off occurs at $e = 0.06$. *Hypotheses non fingo*!
>
> Now Prof. Fisher refers to rejecting hypotheses as a function of the P, χ^2 method, and of accepting them as a logical fallacy. I have in my letter of August 24 stated that the tests are there to ascertain whether a reasonable *graduation* has been reached; not to assert whether one or another hypothesis is true or false. We should accept Sir James Jeans's equipartition as a reasonable graduation for the observed binaries $e \leq 0.06$ ($P = 0.79$) and reject it as a graduation for the observed binaries $e \leq 0.08$ ($P = 0.000,01$). It is not for statisticians to say whether an hypothesis is false except when $P = 0$. All that they can legitimately say is that it gives a poor *graduation*. In particular, it is very unwise in *my* opinion to form tables which provide only the values of $P = 0.01$ and $P = 0.05$, and consider 'hypotheses' which give a value of $P < 0.01$ as 'false', and those with a value between 0.01 and 0.05 as 'doubtful', and for the rest of the scale of P have no descriptive category, for you must not say that such values prove hypotheses to be true. Hence I repeat my assertion, in the face of all the authority of Prof. Fisher and his followers, that all the P, χ^2 test asertains is goodness of graduation, and I hold that 'goodness' of graduation is relative to the nature of the material investigated, our experience of similar material and the purpose to which we intend to put our graduation. The value of P at which we consider goodness or badness of graduation starts cannot be fixed without regard to the special problem under consideration.
>
> There seems somewhere a logical fallacy in the position of both Prof. Fisher and Mr. Buchanan-Wollaston. They both apparently assert that the P, χ^2 test enables one to say an hypothesis is false, yet never to say that an hypothesis is true, but if an hypothesis be *false*, its reverse must be true. If you assert that the hypothesis that a sample is drawn from a normal curve is false, the reverse hypothesis that it is *not* drawn from a normal curve must be true. As a matter of fact, the P, χ^2 has only measured its 'goodness of fit' by a probability coefficient, and it is as idle to say as a result of it, that the hypothesis is 'false', as that the reverse of it is 'true'. The only exception to this rule is when the observations show the existence of individuals in a frequency class which the hypothesis asserts cannot exist.
>
> The 'laws of Nature' are only constructs of our minds; none of them can be asserted to be true or to be false, they are good in so far as they give good fits to our observations of Nature, and are liable at any time to be replaced by a better 'fit', that is, by a construct giving a better graduation.

Pearson's use of Newton's famous injunction to make (or feign) no hypotheses ("Hypotheses non fingo!") is a characteristic rhetorical flourish. By implication at least, Pearson places his "proper" approach to statistical tests of hypotheses within the inductivist tradition celebrated in their philosophical pronouncements by British scientists since Newton. (For a recent commentary sympathetic to this philosophical position, see Achinstein 1991.)

6. DISCUSSION

How does the χ^2 test measure goodness-of-fit? Here Buchanan-Wollaston makes an obvious point. Assuming that the data do not lead to rejection of the hypothesized model, Buchanan-Wollaston observed that any model in the infinite set of all possible models that yields an acceptable value of the test criterion "fits." Specifically, he suggested that "acceptable" models are those that yield values of the test statistic near the mode of the χ^2 distribution. (If the test is based on ν df, the mode is located

at $\nu - 2$, $\nu > 2$.) But the choice of one of these models as correct is arbitrary, particularly if the sole justification for the choice is the failure to reject it as unacceptable by applying the χ^2 test. In short, Pearson's test is for lack of fit of the proposed model, and the conclusion that a hypothesized model fits well is logically unjustified when the test criterion has not attained statistical significance. Buchanan-Wollaston then extended this criticism to all statistical tests.

To this complaint, Karl Pearson offered small comfort. Pearson observed that his test criterion is a measure of the adequacy of some theoretical mathematical model for the observed data, not a criterion of truth. To drive this point home, Pearson referred to his investigation of what we now call the *power* of his test; his example is the low power of the test to discriminate between a normal and a uniform model when the sample size is small. Pearson's logical argument for goodness of fit is simply that in sufficiently large samples, the power of the χ^2 test can suggest when a hypothesized mathematical form adequately describes the sample; in appropriate circumstances, this model may then be regarded as a reasonable mathematical description of the phenomenon that generated the sample. In Pearson's view, this second inferential step was logical, not statistical. (This is even more evident in Pearson's argument with Fisher over the correct degrees for freedom for the χ^2 test.) Such mathematical descriptions, models, or graduations are the proper objects of scientific investigations, and proper use of the χ^2 test can lead to increasingly better theoretical descriptions of the phenomenon of interest.

Where does the normal distribution fit into all of this? Pearson reacted to Buchanan-Wollaston's claim that the success of British statistical methods could be attributed to the prevalence in nature of distributions similar to the Gaussian rather than any peculiar virtue of the methods themselves by making two observations. First, Pearson reminded his readers that the only link between the χ^2 test and normality was his use of a normal approximation to derive the theoretical distribution of the test criterion; the test itself could be applied to any distribution, continuous or discrete. Second, Pearson noted that his test permitted formal investigation of a hypothesized normal model for the first time, and in fact this model could often be dismissed as inadequate by using the test. On the other hand, Fisher ignored the issue of normality altogether, even though Buchanan-Wollaston's criticism was more legitimately aimed in his direction. Many of Fisher's techniques, like his development of the analysis of variance, tacitly assumed normally distributed observations for the theoretical validity of his exact results, and Fisher was not always careful to distinguish the circumstances when his methods yielded an exact result from the circumstances when his results were valid in some approximate sense. Indeed, E. S. Pearson and W. S. Gosset had in 1929 criticized Fisher for this confusion; see E. S. Pearson (1990, pp. 95–101). Fisher's position could then be restated quite fairly in Buchanan-Wollaston's terms: The success of Fisher's methods often lay in their ready and valid application to data that produced sampling distributions sufficiently close to the exact sampling distribution obtained through the argument of normality (Fisher 1929a).

I now turn to the general logic of statistical tests. Despite their other differences, it is clear that both Pearson and Fisher agreed that failure to reject the null hypothesis does not "prove" its truth. Of course, Fisher said so directly. In his first response, Pearson observed that the model that best fits the sample is not necessarily the best representation of the parent population, and again that with small samples one cannot choose between very different population models by using the χ^2 test. Almost obscured by the polemical hyperbole of his second response is Pearson's assertion that the only hypothesis proved true by a statistical test is the negation of a hypothesis according to which the observed sample outcome has zero probability. Philosophically, Pearson preferred to avoid the question of truth completely, but his position here is compatible with Fisher's statement that "tests of significance, when used accurately, ... are never capable of establishing hypotheses as certainly true" (Fisher 1935, p. 474). The difference between the two statisticians and Buchanan-Wollaston is Buchanan-Wollaston's conclusion that this logic makes statistical tests "quite useless" for the purpose of "scientific and practical interest" (Buchanan-Wollaston 1935, p. 182).

Of particular statistical interest is that both Karl Pearson and Fisher refused to interpret the results of statistical tests within a relative-frequency context for errors associated with a decision that was based on the sample. To a considerable extent, and for obvious reasons, the tradition of the classical test of significance shaped the form and interpretation of statistical tests adopted by the two men. Working in this tradition, Pearson and Fisher stated only one hypothesis—the hypothesis tested, or our null hypothesis—in their statistical tests, because the implicit alternative was simply the negation of this hypothesis. (This point was emphasized in the criticism of Edwards 1992, p. 177; Jeffreys 1961, p. 377.) This null hypothesis, however, is often treated more broadly than a hypothesis limited to the values of some unknown parameters of interest. In such applications, elements of the sampling model are conceived as part of the hypothesis being tested rather than as conditions necessary for the statistical test. The differences between the "Karlovingian" and "Piscatorial" approaches (the terms are George Udny Yule's) and the emerging Neyman–Pearson paradigm can be understood if we look briefly at other comments on statistical tests made by the two statisticians.

Fisher's view of tests of significance begins with his insistence on a nonsampling interpretation for the significance level of the test. Several years before his 1935 letter to *Nature*, Fisher had explained that an investigator's use of the .05 level of significance in statistical tests "does not mean that he allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained" (Fisher 1929b, p. 191). Because Fisher advises us to ignore experimental results when the level of significance has not been attained, it is obvious that one never accepts the null hypothesis. This, then, is the explanation for Fisher's bald assertion in his *Nature* letter that "errors of the second kind" are com-

mitted only by those who misunderstand the nature and application of tests of significance.

Characteristically, however, Fisher's own position in the 1920s and 1930s was not free of ambiguity. For example, immediately after his admonition to ignore the results of nonsignificant experiments (Fisher 1929b) Fisher added that "he [the investigator] should only claim that a phenomenon is experimentally demonstrated when he knows how to design an experiment so that it will rarely fail to give a significant result" (p. 191). From our post–Neyman–Pearson perspective, this injunction can be interpreted as Fisher's recognition that a framework for statistical tests that admits the possibility of Type II errors exists. But Fisher's call for experimental designs with sufficient statistical power does not require such a logical framework, and such an interpretation is certainly consistent with Fisher's more clearly stated argument near the end of his life. Thus Fisher (1973) pronounced "a test of significance contains no criterion for 'accepting' a hypothesis" (p. 45).

In later statements of his position, Fisher stressed the context of scientific research for statistical tests; in this context, Fisher linked the significance level of a statistical test to the null hypothesis tested rather than to a decision framework grounded in repeated sampling from some specified population. "A typical test of significance is based on a probability statement derived from the hypothesis to be tested [our null hypothesis], and therefore leads to no probability statement in the real world, but to a change in the investigator's attitude toward the hypothesis under consideration" (Fisher to N. Keyfitz, November 21, 1955, in Bennett 1990, p. 186). Fisher (1973) emphasized this point. In contrast to industrial settings, in which an interpretation of the significance level in terms of repeated sampling may be sensible, Fisher insists that in scientific applications, "the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination through the hypotheses he has decided to test" (Fisher 1956, p. 77; 1973, p. 81). To Fisher, the significance level of a statistical test served as a "well-defined measure of reluctance to the acceptance [!]" of the null hypothesis tested (Fisher 1956, p. 44), a "measure of the rational grounds for the disbelief" in the null hypothesis that the statistical tests "engenders" in rational minds (Fisher 1973, p. 46). As a measure of reluctance applied to the null hypothesis rather than a probability that is based on any meaningful model of repeated sampling, the significance level connotes no probability of erroneous decisions due to rejecting or accepting the null hypothesis. Moreover, Fisher observed that the scientific context of a statistical test often leads the investigator to test a null hypothesis he or she believes is false. In such circumstances, the significance level fails to correspond to the probability of an erroneous decision to reject the null hypothesis, "supposing such a phrase to have any meaning" (Fisher 1956, p. 42; 1973, p. 45; see also the useful discussion of Fisherian significance tests by Seidenfeld 1979, pp. 70–102).

In Fisher's later criticism of the Neyman–Pearson theory of testing hypotheses, he would claim that "in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses" to attack the relative frequency interpretation of the significance level of a statistical test (Fisher 1956, p. 42; 1973, p. 45). Yet as R. L. Plackett and G. A. Barnard noted in their commentary in the work of E. S. Pearson (1990, p. 116), Fisher (1925a) had encouraged the practice of fixed-level tests of significance by Fisher's form of the chi-squared and t tables, in which critical values of these distributions are given for fixed tail probabilities. Thus Karl Pearson (1935b) could criticize Fisher's presentation of these tabled quantiles as a de facto (and inappropriate) decision rule for statistical tests in language that anticipated Fisher's verbal assaults on Neyman–Pearson theory.

In his letters, Karl Pearson advanced a view of statistical tests grounded in the philosophical framework he erected for scientific investigations before he began his statistical career and an approach to statistical inference that Pearson often explicitly or implicitly expressed in terms of "inverse probabilities." According to Pearson, scientists are not in the business of searching for truth; rather, they seek to construct verbal or mathematical summaries of relevant perceptual data. These mental constructs and concepts are the material of science, whereas the scientific method consists of "the careful and often laborious classification" of sense impressions, the comparison "of their relationships and sequences," and at last "the discovery by aid of the disciplined imagination of a brief statement or *formula* which in a few words resumes the whole range of facts" (Pearson 1892, pp. 92–93). These scientific formulas "*describe*, they never *explain* the routine of our perceptions, the sense-impressions we project into an 'outside world'" (Pearson 1892, p. 119). Pearson translated this philosophical predisposition into a representational structure in his statistical investigations—the goal was an adequate distributional description of the observed data. Once this descriptive model was achieved, the mathematical implications of the model could be used to draw structural or relational inferences about the phenomenon represented by the model. For example, the correlation coefficient determined from a bivariate distribution fitted to the heights of fathers and sons could be used to describe the process of heredity (and make predictions and draw conclusions about it) without any specification of a biological mechanism. This approach defined the biometric investigations undertaken by W. F. R. Weldon and Karl Pearson (Norton 1975). As an intellectual construct, the descriptive model had no claim to truth; as Pearson concluded his second letter, such models "are good in so far as they give good fits to our observations of Nature" (Pearson 1935b). (Although Pearson rejected explanation as a goal of statistical models, Fisher was uncharacteristically quiet on the matter of models; see Lehmann 1990. The implications of his own scientific work suggest that Fisher did not share Pearson's extreme position.)

Compared with the elaborate mathematical structures he introduced to represent data, Karl Pearson's probabilistic framework for statistical inference seems underdeveloped. Jeffreys (1961) observed that although Pearson "always maintained the principle of inverse probability, ... he seldom used it in actual applications, and usually presented his results in a form that appears to identify a probability with a frequency" (p. 385). Indeed, Jeffreys claimed that

Pearson is unique among advocates for Bayesian probability calculations because Pearson insisted that prior probability distributions have a frequency interpretation that is based on previous experience (Jeffreys 1961, p. 404). Pearson's attachment to inverse probability antedates his statistical investigations, as the discussion of scientific induction and the Bayes–Laplace "rule of succession" in Pearson's (1892) work demonstrates. There Pearson cited Edgeworth (1884) to provide an experiential justification for uniform prior distributions, whereas Pearson himself argued that an observed data sequence can be augmented by other data sequences from analogous phenomena in scientific applications of the rule (Pearson 1892, pp. 177–178.) As Pearson began his statistical work, he extended his interpretation of probability as a degree of belief, measured "*in a rough and approximate way* by the statistics of past occurrence" (Pearson 1941, p. 93). On the other hand, the level of inference in Pearson's early statistical work is restricted to comparing computed values of estimated parameters with their probable errors. But Pearson's sense of the theory of probable errors had a Bayesian spin, and Pearson's later formulation of inverse probability can be understood as his attempt to counter the criticism of inverse probability calculations advanced by Boole and Venn, influenced in part by Edgeworth's counterargument for prior distributions on the basis of experience. Pearson seems to have followed Edgeworth too in developing his practical approach to tests of significance (Pearson 1941; Stigler 1986, pp. 327–329, noted Edgeworth's importance in the development of Pearson's early statistical attitudes).

Although Jeffreys is correct that Pearson often offered what seems to be a frequentist interpretation of an observed significance level, as when Pearson (1935b) explained that "$P = 0.79$" in his second letter with "79 percent of random samples would, were the 'hypothesis' true, give a worse result than the observations do" (p. 550). Pearson's interpretation was based on random sampling at least as hypothetical as Fisher's. Pearson appeared to adopt random sampling as a convenient framework for examining the observed data rather than the actual mechanism by which the data were obtained. Random sampling provided Pearson the chance framework for computing an "objective" probability, with a corresponding frequency interpretation. This probability, in turn, approximated the degree of belief appropriate for outcomes of future observations like those in the sample. In effect, the observed sample data were the sequence of past sense impressions on which predictions of the future were to be based, and the observed objective probability was taken "as the basis of belief as to the future" (Pearson 1941, p. 96). In terms of a test of significance, this distinction appears to make the assumption of random sampling part of the hypothesis under test instead of a necessary condition for the statistical test. Pearson could and did apply statistical tests to data that were not the product of random sampling or where no sensible sampling context existed. Thus Pearson used random sampling as the logical basis for computing a probability P in his test for goodness of fit to judge the probability of the observed data, thereby obtaining a measure of evidential support for the hypothesized model without an inverse probability calculation. Given his philosophy of science, it is not surprising that Pearson also emphasized that differences in the goodness-of-fit test criterion measured the relative descriptive adequacy of competing scientific models for the observed data. As he argued (Pearson 1936), "the P, χ^2 criterion gives if not an absolutely accurate, still for practical purposes an excellent measure of what most statisticians need to know, namely the relative superiority of one graduating curve over another" (p. 49). Of course, differences in values of χ^2 computed from the same sample data for different models correspond asymptotically to differences of the multimonial log-likelihood, and thus such comparisons produce measures of relative support for the various models considered in the sense of Hacking (1965) and Edwards (1992). (For examples of this use of χ^2 by Karl Pearson and Fisher, see Fisher 1925a, pp. 81–82; Pearson 1936, p. 59.)

Although Fisher and Karl Pearson regarded the significance level attained by a statistical test to be hypothetical probability, Karl Pearson and the other two *Nature* correspondents disagreed completely on which outcomes of the test supported the hypothesized model in the test for goodness of fit. Buchanan-Wollaston looked to the center of the sampling distribution of the χ^2 statistic for values that supported the hypothesized model; his argument that values of the test statistic near the center of the χ^2 distribution offer positive evidence for the hypothesized model anticipated the position advanced by Berkson (1942). As is clear, Fisher argued that tests of significance produce no support for the null hypothesis under any circumstance. Regarding the test for goodness of fit, Fisher would state the following in every edition of Fisher (1925a): "If P is between .1 and .9, there is no reason to suspect the hypothesis being tested" (p. 71). Pearson's sampling framework was the multivariate normal approximation to the multinomial cell frequencies, in which the more probable sample outcomes were those nearer the cell expectations. In Pearson's exposition of the χ^2 test, the single most probable sample outcome was one in which the cell frequencies equaled their expectations exactly, because this sample exhibited the highest multivariate probability density possible. This most probable sample offered the maximum support for the hypothetical model, with $\chi^2 = 0$ and $P = 1$. Samples that produced greater discrepancies between observed and expected cell frequencies were less probable according to the null model and thus provided less evidential support for it; such samples yielded increasingly large values of the test statistic and smaller values of Pearson's probability coefficient, P. For Pearson, values of his criterion closest to zero, with corresponding values of P nearest to unity, indicated the highest degree of support for the hypothesized model, and Pearson frequently attached the value of P to the hypothesis tested, which is consistent with this sense of support. Fisher (1925a, p. 80) argued that this belief was fallacious, because if the hypothesized model were correct a value of $P = .999$ was just as improbable as $P = .001$ and just as surely would lead to rejection of the model. Pearson, however, would never have rejected a hypothetical model because the data were too good to be true, and he clearly prized values of P near unity in applications of his test. (See Pearson 1932, pp. 358–359, for examples; Plackett 1983 briefly discussed Pearson's

interpretation of P; Seidenfeld 1979, pp. 84–86, argued for an interpretation consistent with Pearson's position.)

In their *Nature* letters, neither Pearson nor Fisher addressed Buchanan-Wollaston's complaint that statistical tests were scientifically irrelevant because by construction they did not produce "appropriate evidence for affirmative conclusions" (Berkson 1942, p. 326). Now Buchanan-Wollaston certainly did not reject the use of statistical tests, as his own later work demonstrates. Shortly before his challenge to Pearson and Fisher in *Nature*, Buchanan-Wollaston presented his own approach to statistical tests (Buchanan-Wollaston 1935b): He proposed that the null hypothesis in statistical tests "should be such that it is acceptable on *apriori* grounds if the data do not show it unlikely to be true" (p. 254). From the multitude of possible null hypotheses that might have accounted for the data, the "chosen hypothesis is merely that which is considered by the scientist to be the simplest and most acceptable" (Buchanan-Wollaston 1935b, p. 254). By asserting a nonstatistical justification of the null hypothesis in terms of its scientific utility, Buchanan-Wollaston argued that the result of some statistical test still served a limited scientific purpose when it proved to be nonsignificant in the statistical sense. Unlike Fisher, Buchanan-Wollaston did not want to ignore sample outcomes that did not attain statistical significance. Unlike Karl Pearson, Buchanan-Wollaston was reluctant to banish the search for truth from the scientific enterprise. The apparent conflict between the objectives of statistical tests and scientific inference prompted Buchanan-Wollaston's appeal to Pearson and Fisher. Despite occasional assertions that statistical inference and scientific inference are identical, this conundrum still continues to shape the dialogue between statisticians and scientists.

[*Received September 1992. Revised May 1993.*]

REFERENCES

Achinstein, p. (1991), *Particles and Waves: Historical Essays in the Philosophy of Science*, New York: Oxford University Press.

Bennett, J. H. (ed.). (1973), *Collected Papers of R. A. Fisher*, 3, 1932–36, Adelaide, Australia: University of Adelaide.

——— (ed.). (1983), *Natural Selection, Heredity, and Eugenics, Including Selected Correspondence of R. A. Fisher With Leonard Darwin and Others*, Oxford, England: Oxford University Press.

——— (ed.). (1990), *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*, Oxford, England: Oxford University Press.

Berkson, J. (1942), "Tests of Significance Considered as Evidence," *Journal of the American Statistical Association*, 37, 325–335.

Board of Agriculture and Fisheries. (1913), *First Report of the Committee Appointed by the President of the Board of Agriculture and Fisheries to Advise the Board on Questions Relating to the Elucidation Through Scientific Research of Problems Affecting Fisheries*, London: Darling & Son.

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

Buchanan-Wollaston, H. J. (1911a), "Report on the Fish-Egg Cruise of the 'Huxley' in June 1909," in *International Fishery Investigations: Third Report (Southern Area) on Fishery and Hydrographical Investigations in the North Sea and Adjacent Waters, Conducted for His Majesty's Government by the Marine Biological Association of the United Kingdom 1906–1908*, London: Darling & Son, pp. 207–234b.

——— (1911b), *On the Calculation of the "Filtration Coefficient" of a Vertically Descending Net, and on the Allowance to Be Made for Clogging: On a New Form of Plankton-Net, Designed to Make Truly Vertical Hauls in Any Weather* (International Council for the Exploration of the Sea, Occasional Publications Nos. 58 and 59), Copenhagen: Høst.

——— (1916), "Report on the Spawning-Grounds of the Plaice in the North Sea," *Fishery Investigations* (Board of Agriculture and Fisheries, Ser. 2, Vol. 2, No. 4), London: H. M. Stationery Office.

——— (1923), "The Spawning of Plaice in the Southern Part of the North Sea in 1913–14," *Fishery Investigations* (Ministry of Agriculture and Fisheries, Ser. 2, Vol. 5, No. 2), London: H. M. Stationery Office.

——— (1924), "Growth-Rings of Herring Scales," *Nature*, 114, 348–349.

——— (1926), "Plaice-Egg Production in 1920–21, Treated as a Statistical Problem, With Comparison Between the Data From 1911, 1914, and 1921," *Fishery Investigations* (Ministry of Agriculture and Fisheries, Ser. 2, Vol. 9, No. 2), London: H. M. Stationery Office.

——— (1927), "On the Selective Action of a Trawl Net, With Some Remarks of Selective Action of Drift Nets," *Journal of the International Council for the Exploration of the Sea*, 2, 343–355.

——— (1929), "The Selective Action of the Gelder Cod-End and that of Other Cod-Ends Compared," *Journal of the International Council for the Exploration of the Sea*, 4, 300–308.

——— (1933), "Some Modern Statistical Methods: Their Application to the Solution of Herring Race Problems," *Journal of the International Council for the Exploration of the Sea*, 8, 7–47.

——— (1935a), "Statistical Tests," *Nature*, 136, 182–183.

——— (1935b), "The Philosophic Basis of Statistical Analysis," *Journal of the International Council for the Exploration of the Sea*, 10, 249–263.

——— (1935c), "On the Component of a Frequency Distribution Ascribable to Regression," *Journal of the International Council for the Exploration of the Sea*, 10, 81–98.

——— (1936), "The Philosophic Basis of Statistical Analysis," *Journal of the International Council for the Exploration of the Sea*, 11, 7–26.

——— (1937), "Two Technical Notes," *Journal of the International Council for the Exploration of the Sea*, 12, 333–334.

——— (1938), "On the Application of the Statistical Theory of Space Distributions to Hydrographic and Fishery Problems," *Journal of the International Council for the Exploration of the Sea*, 13, 173–186.

——— (1945), *On the Statistical Treatment of the Results of Parallel Trials With Special Reference to Fishery Research* (Scientific Publication No. 10), Ambleside, England: Freshwater Biological Association.

——— (1958), "Statistical Tests for Significance Applicable to Distributions in Space," *Journal of the International Council for the Exploration of the Sea*, 33, 161–172.

Buchanan-Wollaston, H. J., and Hodgson, W. C. (1929), "A New Method of Treating Frequency Curves in Fishery Statistics, With Some Results," *Journal of the International Council for the Exploration of the Sea*, 4, 207–225.

Edgeworth, F. Y. (1884), "The Philosophy of Chance," *Mind*, 9, 223–235.

Edwards, A. W. F. (1992), *Likelihood*, Baltimore: Johns Hopkins University Press.

Fisher, R. A. (1922a), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society*, Ser. A, 222, 309–368.

——— (1922b), "On the Interpretation of χ^2 From Contingency Tables, and the Calculation of P," *Journal of the Royal Statistical Society*, 85, 87–94.

——— (1923), "Statistical Tests of Agreement Between Observation and Hypothesis," *Economica*, 3, 139–147.

——— (1924), "The Conditions Under Which χ^2 Measures the Discrepancy Between Observation and Hypothesis," *Journal of the Royal Statistical Society*, 87, 442–450.

——— (1925a), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

——— (1925b), "Theory of Statistical Estimation," *Proceeding of the Cambridge Philosophical Society*, 22, 700–725.

——— (1929a), "Statistics and Biological Research," *Nature*, 124, 266–267.

——— (1929b), "The Statistical Method in Psychical Research," *Proceedings of the Society for Psychical Research*, 39, 185–189.
——— (1935), "Statistical Tests," *Nature*, 136, 474.
——— (1937), "Professor Karl Pearson and the Method of Moments," *Annals of Eugenics*, 7, 308–318.
——— (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.
——— (1973), *Statistical Methods and Scientific Inference* (3rd ed.), New York: Hafner.
Freshwater Biological Association. (1943), *Eleventh Annual Report*, Ambleside, England: Freshwater Biological Association.
——— (1944), *Twelfth Annual Report*, Ambleside, England: Freshwater Biological Association.
——— (1945), *Thirteenth Annual Report*, Ambleside, England: Freshwater Biological Association.
——— (1946), *Fourteenth Annual Report*, Ambleside, England: Freshwater Biological Association.
Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge, England: Cambridge University Press.
Jeans, J. H. (1935), "Age of the Universe," *Nature*, 136, 432.
Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, England: Oxford University Press.
Lee, A. J. (1992), *The Directorate of Fisheries Research: Its Origins and Development*, Lowestoft, England: Ministry of Agriculture, Fisheries and Food, Directorate of Fisheries Research for England and Wales.
Lehmann, E. L. (1990), "Model Specification: The Views of Fisher and Neyman, and Later Developments," *Statistical Science*, 5, 160–168.
Marine Biological Association. (1905), *International Fishery Investigations: First Report on Fishery and Hydrographical Investigations in the North Sea and Adjacent Waters (Southern Area), Conducted for His Majesty's Government by the Marine Biological Association of the United Kingdom, 1902–1903*, London: Darling & Son.
——— (1907), *International Fishery Investigations: Second Report (Southern Area) on Fishery and Hydrographical Investigations in the North Sea and Adjacent Waters, Conducted for His Majesty's Government by the Marine Biological Association of the United Kingdom, 1904–1905, Part I*, London: Darling & Son.
——— (1912), *International Fishery Investigations: Fourth Report (Southern Area) on Fishery and Hydrographical Investigations in the North Sea and Adjacent Waters, Conducted for His Majesty's Government by the Marine Biological Association of the United Kingdom, 1909*, London: Darling & Son.
Morant, G. M. (1939), *A Bibliography of the Statistical and Other Writings of Karl Pearson*, London: *Biometrika* Office.
Norton, B. J. (1975), "Biology and Philosophy: The Methodological Foundations of Biometry," *Journal of the History of Biology*, 8, 85–93.
Pearson, E. S. (1938), "Karl Pearson, An Appreciation of Some Aspects of His Life and Work, Part II: 1906–1936," *Biometrika*, 29, 161–248.
——— (1968), "Some Early Correspondence Between W. S. Gosset, R. A. Fisher and Karl Pearson, With Notes and Comments," *Biometrika*, 55, 445–457.
——— (1990), *"Student": A Statistical Biography of William Sealy Gosset* (eds. R. L. Placket and G. A. Barnard), Oxford, England: Oxford University Press.
Pearson, K. (1892), *The Grammar of Science*, London: Scott.
——— (1900), "On the Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen From Random Sampling," *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science* 50, Ser. 5, 157–175.
——— (1916), "On a Brief Proof of the Fundamental Formula for Testing the Goodness of Fit of Frequency Distributions and on the Probable Error of P," *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science* 31, Ser. 6, 369–378.
——— (1922), "On the χ^2 Test of Goodness of Fit," Biometrika, 14, 186–191.
——— (1931), "Historical Note on the Distribution of the Standard Deviations of Samples of Any Size Drawn From an Indefinitely Large Normal Parent Population," *Biometrika*, 23, 416–418.
——— (1932), "Experimental Discussion of the (χ^2, *P*) Test for Goodness of Fit," *Biometrika*, 24, 351–381.
——— (1933), "On a Method of Determining Whether a Sample of Size *n* Supposed to Have Been Drawn From a Parent Population Having a Known Probability Integral Has Probably Been Drawn at Random," *Biometrika*, 25, 379–410.
——— (1935a), "Statistical Tests," *Nature*, 136, pp. 296–297.
——— (1935b), "Statistical Tests," *Nature*, 136, pp. 550.
——— (1936), "Method of Moments and Method of Maximum Likelihood," *Biometrika*, 28, 34–59.
——— (1941), "The Laws of Chance, in Relation to Thought and Conduct," *Biometrika*, 32, 89–100.
Plackett, R. L. (1983), "Karl Pearson and the Chi-Squared Test," *International Statistical Review*, 51, 59–72.
Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference: Learning From R. A. Fisher*, Dordrecht, The Netherlands: D. Reidel.
Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge, MA: Harvard University Press.