Journal of Modern Applied Statistical Methods

Volume 9 | Issue 1 Article 2

5-1-2010

Fisher Was Right

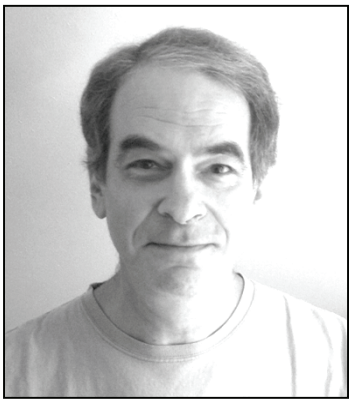
Ronald C. Serlin
University of Wisconsin - Madison, rcserlin@wisc.edu

Journal of Modern Applied Statistical Methods
May 2010, Vol. 9, No. 1, 2-8

1538 – 9472/10/$95.00

INVITED ARTICLES

Fisher Was Right



Ronald C. Serlin
University of Wisconsin-Madison

Invited address presented to the Educational Statistician's Special Interest Group at the annual meeting of the American Educational Research Association, Denver, May 1, 2010.

Key words: Fisher.

Introduction

I would like once again to thank you for awarding me this honor last year. Given the scholars between whom I am sandwiched, the first honoree, Ingram Olkin, and next year's, Joel Levin, I must try very hard to act as though the committee did not make a serious mistake with my nomination. Tonight, I'd like to focus on some of the work of R. A. Fisher, who would have been 120 years old now, to make a couple of points of my own. I hope that some of what I say will give you the same feeling of fun in the discovery of something neat and surprising as I experienced.

Ronald Serlin is Professor Emeritus in the Department of Educational Psychology at the School of Education and is in the Department of Biostatistics and Medical Informatics in the Medical School. Email: rcserlin@wisc.edu.

Early Years: Up to 1922

Fisher held two chairs in genetics, at University College in London and then at Cambridge but, surprisingly, was never a professor of statistics. Regarding Fisher's accomplishments in statistics, Savage (1976) commented that it would be easier to list the few topics in which he was not interested. "In the art of calculating explicit sampling distributions, Fisher led statistics out of its infancy, and he may never have been excelled in this skill" (p. 449).

There is much, of course, about which Fisher was right. Despite his shunning the concept of Type II errors, Fisher (1928) was the first to provide formulas for the noncentral Chi-square, t, and F distributions. (The symbol F was introduced by Snedecor in honor of Fisher, "for which officiousness," according to Savage, "Fisher seems never to have forgiven him" (p. 449)). There once existed a fair amount of disagreement regarding how to count degrees of freedom in a contingency table, with Karl

Pearson (among others) claiming rc − 1 and Fisher (1922) correcting to (r − 1)(c − 1). Fisher, of course, was right here. Fisher was a pioneer in nonparametric statistics, having suggested the use of the sign test in place of the t-test in certain designs, and having introduced what he called exact tests to avoid the assumption of normality in many circumstances.

According to Stigler (2005, p. 33), of Fisher's 97 publications from 1912 to 1920, 91 were in the *Eugenics Review*, two were on genetics related to eugenics, two were papers published in *The Messenger of Mathematics*, and the other two (in 1915 and 1920) were on mathematical statistics. I'll focus briefly on the 1915 and 1920 papers, as described by Stigler (2005, 2006).

Mathematically, the 1915 derivation of the distribution of the sample correlation coefficient was the kind of work to which we all strive. Fisher found the distribution, expressions for moments, transformations (r-to-z) and distributional relationships (including his earlier work on the Student's t-distribution), expressions for the bias of r, and the maximum likelihood estimator of ρ.

Right Nice Stuff

This type of work led Neyman (1951), in his review of Fisher's *Contributions to Mathematical Statistics* (1950), to describe Fisher as "a very able 'manipulative' mathematician" (p. 406). The *Contributions* contain prefatory comments by Fisher on the various papers. For the 1915 paper, Fisher wrote "Here the method of defining a sample by the coordinates of a point in Euclidean hyperspace was introduced..." (p. 87). Unfortunately, according to Neyman (1951), representing the sample by a point in space was used for a similar purpose by Karl Pearson in 1900 and - Neyman suspected - had probably been used even before that; thus, Fisher was wrong in this regard.

During the year following the publication of Fisher's article on the correlation coefficient, Kirstine Smith (1916), working at Karl Pearson's laboratory, published an article suggesting that when fitting a frequency curve with grouped data, the constants should be estimated using a minimum Chi-square criterion. She illustrated the use of this criterion through a series of examples. She stated that compared to the use of the minimum Chi-square method of fit, other approaches were arbitrary, including what she termed "the Gaussian 'best' value," (p. 262) the maximum likelihood approach from error theory that Fisher had supported in a paper he wrote as an undergraduate student in 1912. According to Stigler (2005), in response to a letter and manuscript that Fisher submitted to *Biometrika*, Karl Pearson as editor told Fisher that he had to demonstrate the logic of maximum likelihood, to justify it being better than Smith's approach. For a while Fisher could not respond.

The basis for Fisher's reply came, possibly by accident (Stigler, 2005), in the late spring of 1919. Fisher was considering the relative merits of two alternative estimates of the standard deviation of a normal distribution: one was based on the mean absolute deviation, the other the maximum likelihood solution. He had considered combining the two estimates in some way but instead discovered that the whole of the information regarding σ, which a sample provides, is summed up in the value of the maximum likelihood estimator. Not only did it have a smaller standard deviation, it was, in a word, sufficient.

On November 17, 1921, Fisher read a paper to the Royal Society of London entitled *On the Mathematical Foundations of Theoretical Statistics*. The paper opened with a set of definitions that were, in 1921, entirely new to statistical theory, but which are now familiar; they include consistency, efficiency, estimation, likelihood, optimum, and sufficiency. Stigler (2005) pointed out that not in the list is "…another, even more basic statistical concept: It is in this paper of Fisher's that the word 'parameter' is first used in the modern statistical sense" (p. 32). Stigler notes that the word parameter appears 57 times.

According to Fisher, a consistent estimate is called efficient if it is asymptotically normal and if it has the minimum asymptotic variance (Neyman, 1951). In his 1908 paper, however, Edgeworth expressed the idea that maximum likelihood estimates are always efficient and made several attempts to prove his conjecture. The proofs, however, "…of the efficiency of maximum likelihood estimates

offered both by Edgeworth and by Fisher are inaccurate, and the assertion, taken in its full generality, is false" (Neyman, 1951, p. 407). So Fisher was wrong in the assertion, the proof, and in not giving Edgeworth some credit for priority.

Summarizing Fisher's work, Neyman (1951) wrote, "…three major concepts were introduced by Fisher and consistently propagandized by him in a number of publications. These are mathematical likelihood as a measure of the confidence in a hypothesis, sufficient statistics, and fiducial probability," (p. 407) all employed by Fisher in the service of scientific induction.

Inference

Fisher (1947) felt that "the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (p. 16). Regarding the rate of error to assign to an incorrect rejection of the null hypothesis, Fisher wrote (1926) that "it is convenient to draw the line at about the level at which we can say: 'Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials.'" "A scientific fact," he went on, "should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance" (p. 504). Further, Fisher (1973) wrote, "…in the vast majority of cases the work is completed without any statement of mathematical probability being made about the hypothesis or hypotheses under consideration. The simple rejection of a hypothesis, at an assigned level of significance, is of this kind and is often all that is needed, and all that is proper, for the consideration of a hypothesis in relation to the body of experimental data available" (p. 40). This all seems right.

Regarding Type II errors, Fisher (1947) wrote that "the notion of an error of the so-called 'second kind,' due to accepting the null hypothesis 'when it is false' may then be given a meaning in reference to the quantity to be estimated. It has no meaning with respect to simple tests of significance, in which the only available expectations are those which flow from the null hypothesis being true" (p. 17). Thus, Fisher felt that one could not commit a Type II error, because one never drew a conclusion on the basis of a non-rejection of the null hypothesis. As he wrote (Fisher, 1973), "To a practical man, also, who rejects a hypothesis, it is, of course, a matter of indifference with what probability he might be led to accept the hypothesis falsely, for in his case he is not accepting it" (pp. 41-42). Some rightness to this is evident.

Fisher always desired to establish a correct theory of statistical inference. According to Kempthorne (1976) "Fisher really did think that one could develop by logical reasoning a probability distribution for one's *knowledge* of a physical constant" (p. 496). Fisher, as Neyman (1951) pointed out, seemed proud to have formulated a measure of rational belief. Thus, Fisher (1973) wrote that the level of significance "in such cases fulfils the conditions of a measure of the rational grounds for the disbelief it engenders" (p. 43). Similarly, Fisher (1925a) had observed that "if the value of P so calculated turned out to be a small quantity such as 0.01, we should conclude with some confidence that the hypothesis was not in fact true of the population actually sampled" (p. 90).

In similar vein, Fisher (1935c) stated "more generally, however, a mathematical quantity of a different kind, which I have termed *mathematical likelihood*, appears to take its place as a measure of rational belief…" (p. 40). In addition, Fisher (1973) commented that "the actual value of P obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis" (p. 80). And finally he also stated (1973) "What has now appeared is that the mathematical concept of probability is, in most cases, inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate…I have used the term 'Likelihood'" (pp. 9-10). There is a whole lot of wrong here, as a measure of rational belief - even if obtainable - provides a theory with no level of epistemological virtue.

Note that even Neyman (1956) was not immune to this inductive probability infection, for he wrote in defense of control of the Type II

error rate "…the numerical values of probabilities of errors of the second kind are most useful for deciding whether or not the failure of a test to reject a given hypothesis could be interpreted as any sort of 'confirmation' of this hypothesis" (p. 290).

Fiducial Probability and Fiducial Intervals

Fisher (1935b) wrote on fiducial probability and fiducial intervals, about which he stated, "This form of argument leads in certain cases to rigorous probability statements about the unknown parameters of the population from which the observational data are a random sample, without the assumption of any knowledge respecting their probability distributions a priori." His argument seems basically the same as that which leads to confidence intervals.

Defining $t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}}$, Fisher noted that the probability statement $P(t > t_\alpha) = \alpha$ can be solved in terms of μ to yield $P(\mu < \bar{x} - t_\alpha s/\sqrt{n}) = \alpha$. Fisher believed that this probability statement holds even after the sample values are substituted. Conversely, Neyman and Pearson contended that at that point, the probability is either zero or one. Neyman (1956) offered a counter-argument in terms of two flips of a fair coin, where the variable Y is the number of heads appearing. So it may be written that P(Y = 1) = 0.5 before the experiment. If Y = 2 is observed, Fisher would say the probability statement holds after substituting, or that P(2 = 1) = 0.5. Fisher appears to be wrong in this case.

To summarize, in Neyman's (1951) words, "Unfortunately, in conceptual mathematical statistics Fisher was much less successful than in manipulatory, and of the three above concepts only one, that of a sufficient statistic, continues to be of substantial interest. The other two proved to be either futile or self-contradictory and have been more or less generally abandoned" (p. 407). As may be observed, it is fiducial probability that Neyman considered self-contradictory, and I agree that a search for a measure of rational belief is futile. Thus, for Fisher, one out of three right will have to do.

Personality

Fisher was not always charming and gracious, and his running battles with Neyman are well known. Regarding Karl Pearson, he wrote, "Pearson's energy was unbounded. In the course of his long life he gained the devoted service of a number of able assistants, some of whom he did not treat particularly well. He was prolific in magnificent, or grandiose, schemes capable of realization perhaps by an army of industrious robots responsive to a magic wand" (1973, p. 2).

In similar vein, in a prefatory note on Fisher's *Contributions to Mathematical Statistics* is a personal attack on Sir Karl: "If peevish intolerance of free opinion in others is a sign of senility, it is one which he had developed at an early age. Unscrupulous manipulation of factual material is also a striking feature of the whole corpus of Pearsonian writings, and in this matter some blame does seem to attach to Pearson's contemporaries for not exposing his arrogant pretensions" (p. 437). On multiple occasions, Fisher (1958) criticized the ability of mathematicians to do science; for example he wrote "…with mathematical symbols, they are of course experts. But it would be a mistake to think that mathematicians as such are particularly good at the inductive logical processes which are needed in improving our knowledge of the natural world, in reasoning from observational facts to the inferences which those facts warrant" (p. 261). Judging by most of those in this audience, I believe that Fisher was wrong in this.

Analysis of Variance

It is not clear why Neyman did not include analysis of variance among Fisher's major accomplishments. Perhaps, as seems possible, it was due to personal enmity. Fisher's first paper on this subject, with W. A. Mackenzie, was published in 1923. According to Cochran (1980), "two aspects of this paper are of historical interest. At that time, Fisher did not fully understand the rules of analysis of variance—his analysis is wrong—nor the role of randomization" (p. 17), but by the time *Statistical Methods for Research Workers* came out in 1925, he was back on top of his game.

Fisher was the first to discuss Neyman's 1935 paper regarding analysis of variance in randomized blocks and Latin Square designs, *Statistical Problems in Agricultural Experimentation*, presented to the Royal Statistical society. In this paper, Neyman formulated a model that allowed each treatment to respond differently in each plot, making no assumption that treatment effects were fixed and additive in the plots. As noted by Holschuh (1980), "the null hypothesis he [Neyman] considered was that the average treatment response over the entire experimental area was the same for all treatments. Under this null hypothesis, he found that the z-test for the randomized block design was unbiased" (p. 43) but that the test for the Latin square design was, in general, not unbiased (z is one-half the natural log of the F-statistic). If it is assumed that the correlation of plot errors is unity, the z-test is unbiased.

Fisher (1935) began his comments by writing, "…he [Fisher] had hoped that Dr. Neyman's paper would be on a subject with which the author was fully acquainted, and on which he could speak with authority…Since seeing the paper, he had come to the conclusion that Dr. Neyman had been somewhat unwise in his choice of topics" (p. 154). Fisher focused primarily on Neyman's analysis of the z-test for treatment effects. Fisher scolded Neyman for obtaining the wrong result for the Latin square design and said that he may have been "misled by his excessive use of symbolism" (Holdschuh, 1980, p.43).

Fisher, however, had ignored Neyman's null hypothesis. The null hypothesis Fisher entertained was that in any plot the treatments have the same effect. In that case the correlation of plot errors is unity and Neyman's conclusion is correct: the z-test is unbiased. In the course of the discussion, Neyman (1935) exposed Fisher's error, but Fisher then claimed that the z-test was only intended to test the null hypothesis of identical treatment effects. Neyman replied that he was "considering problems which are important from the point of view of agriculture" (p. 173).

Neyman (1935) began his written response sarcastically, writing:

> I am grateful to Professor Fisher for a sentence in the third part of his contribution…: 'I suggest that before criticizing previous work it is always wise to give enough study to the subject to understand its purpose…' The sentence I have quoted applies to its author, Professor Fisher, himself, who not only criticized my paper, but blamed me for a variety of sins of which I am not guilty—all this before apparently taking the trouble to discover what my paper is about and what are the results. According to him: I was unwise in the choice of my topics, I have been speaking of things with which I am not fully acquainted, I deceived myself on so simple a question, I forgot the meaning of the facts, I confuse the questions of estimation and the tests of significance and I am apparently not able to grasp the very simple argument!" (p. 174)

Here, again, Fisher seems to have been wrong.

It is in his book *Design of Experiments* (1935a) that Fisher described a method that all have come to know to be defective, except in special cases, that being Fisher's Least Significant Difference (LSD) procedure. Fisher wrote (1935a) that if the F test is not significant in comparing yields of different varieties, "…they will not often need to be considered further," whereas if the test was significant, he continued,

> …the null hypothesis has been falsified, and may therefore be set aside. We shall thereafter proceed to interpret the differences between the varietal yields as due at least in part to the inherent qualities of the varieties, as manifested on the conditions of the test, and shall be concerned to know with what precision these different yields have been evaluated. …In either case the square root of the variance gives the standard deviation, and provides therefore a means of judging which of the differences among our varietal yield values are sufficiently great to be

> regarded as well established, and which are to be regarded as probably fortuitous. If the experiment leaves any grounds for practical doubt, values may be compared by the t test… (pp. 64-65)

He implied that these t tests would each be conducted with a Type I error rate of five percent.

Fisher went on in the next paragraph to describe a method introduced to the literature 26 years later by Dunn. He explained that when the test is not significant, and yet the researcher goes on to examine comparisons suggested by the data, much caution should be used. He wrote (1935a),

> …for if the variants are numerous, a comparison of the highest with the lowest observed value, picked out from the results, will often appear to be significant, even from undifferentiated material. Properly, such unforeseen effects should be regarded only as suggestions for future experimentation, in which they can be deliberately tested…Thus, in comparing the best with the worst of ten tested varieties, we have chosen the pair with the largest apparent difference out of 45 pairs, which might equally have been chosen. We might, therefore, require the probability of the observed difference to be as small as 1 in 900, instead of 1 in 20, before attaching statistical significance to the contrast." (p. 66)

Although testing contrasts, even with a Dunn-Bonferroni adjustment, after a non-significant F test inflates the Type I error rate, it is of interest to discuss the LSD procedure. Fisher maintained that significance tests reveal facts. Sometimes these facts are used to falsify hypotheses, and at other times, many such revealed facts can serve as the genesis of a conjecture intended to explain them. Multiple comparison procedures attempt to control family-wise Type I error rates across a number of comparisons, which comprise a family of comparisons in the sense that a false rejection of any one of them would lead a researcher to claim as false a statement at a higher conceptual level, and which one would not like to do in error at a rate higher than the adopted alpha. But the LSD method does just that. If the F test is not significant, the experiment is stopped. If it is significant in error, it holds the error rate at the appropriate level in falsifying the higher-level proposition, and any contrasts examined afterward and found significant erroneously do not contribute to the overall error rate, because it is already wrong at an acceptable rate. If the F is correctly significant, one cannot make an error in declaring the higher-level statement false, and one is thus in fact-generating mode for the next attempt at an improved explanation. So Fisher was right after all.

References

Cochran, W. G. (1980). Fisher and the analysis of variance. In *R. A. Fisher: An appreciation*, S. E. Fienberg & D. V. Hinkley (Eds.), pp. 17-34.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52-64.

Edgeworth, F. Y. (1908). On the probable errors of frequency-constants (cont'd). *Journal of the Royal Statistical Society*, *71*, 499-512.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly notices of the Royal Astronomical Society, 80,* 758-770.

Fisher, R. A. (1925a). Applications of Student's distribution. *Metron*, *5*, 90-104.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507-521.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables and the calculation of *p*. *Journal of the Royal Statistical Society*, *85*, 87-94.

Fisher, R. A. (1950). *Contributions to mathematical statistics*. New York: Wiley.

Fisher, R. A. (1935a). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1947). *The design of experiments (4th Edition)*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935b). The fiducial argument in statistical inference. *Annals of Eugenics*, *6*, 391-398.

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society*, *Series A*, *121*, 654-673.

Fisher, R. A. (1935c). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*, 39-82.

Fisher, R. A. (1958). The nature of probability. *Centennial Review*, *2*, 261-274.

Fisher, R. A. (1973). *Statistical methods and scientific inference*. New York: Hafner Press.

Fisher, R. A. (1925b). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Fisher, R. A. & MacKenzie (1923). Studies in crop variation (ii)—The manurial response of different potato varieties. *Journal of Agricultural Science*, *13*, 311-320.

Holschuh, N. (1980). Randomization and design I. In *R. A. Fisher: An appreciation*, S. E. Fienberg & D. V. Hinkley (Eds.), 35-45.

Kempthorne, O. (1976). Discussion of On rereading R. A. Fisher. *The Annals of Statistics*, *4*, 495-497.

Neyman, J. (1951). Fisher's Collected Papers: Contributions to Mathematical Statistics. *Scientific Monthly*, *72*, No. 6, 406-408.

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *Journal of the Royal Statistical Society (B)*, *18*, 288-294.

Neyman, J., Iwaszkiewicz, K., & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, *2*, 107-180.

Savage, L. J. (1976). On rereading R. A. Fisher. *The Annals of Statistics*, *4*, 441-500.

Smith K. (1916). On the 'best' values of the constants in frequency distributions. *Biometrika*, *11*, 262-276.

Stigler, S. M. (2005). Fisher in 1921. *Statistical Science*, *20*, 32–49.

Stigler, S. M. (2006). How Ronald Fisher became a mathematical statistician. *Mathematics and Social Sciences*, *44*, 23-30.