Science and Statistics

GEORGE E. P. BOX*

Aspects of scientific method are discussed: In particular, its representation as a motivated iteration in which, in succession, practice confronts theory, and theory, practice. Rapid progress requires sufficient flexibility to profit from such confrontations, and the ability to devise parsimonious but effective models, to worry selectively about model inadequacies and to employ mathematics skillfully but appropriately. The development of statistical methods at Rothamsted Experimental Station by Sir Ronald Fisher is used to illustrate these themes.

1. INTRODUCTION

In 1952, when presenting R.A. Fisher for the Honorary degree of Doctor of Science at the University of Chicago, W. Allen Wallis described him in these words.

> He has made contributions to many areas of science; among them are agronomy, anthropology, astronomy, bacteriology, botany, economics, forestry, meteorology, psychology, public health, and—above all—genetics, in which he is recognized as one of the leaders. Out of this varied scientific research and his skill in mathematics, he has evolved systematic principles for the interpretation of empirical data; and he has founded a science of experimental design. On the foundations he has laid down, there has been erected a structure of statistical techniques that are used whenever men attempt to learn about nature from experiment and observation.

Fisher was introduced by the title which he himself would have chosen—not as a statistician but as a scientist, and this was certainly just, since more than half of his published papers were on subjects other than statistics and mathematics. My theme then will be first to show the part that his being a good scientist played in his astonishing ingenuity, originality, inventiveness, and productivity as a statistician, and second to consider what message that has for us now.

2. ASPECTS OF SCIENTIFIC METHOD

A heritage of thought about the process of scientific learning comes to us from such classical writers as Aristotle, Galen, Grossteste, William of Occam, and Bacon who have emphasized aspects of good science and have warned of pitfalls.

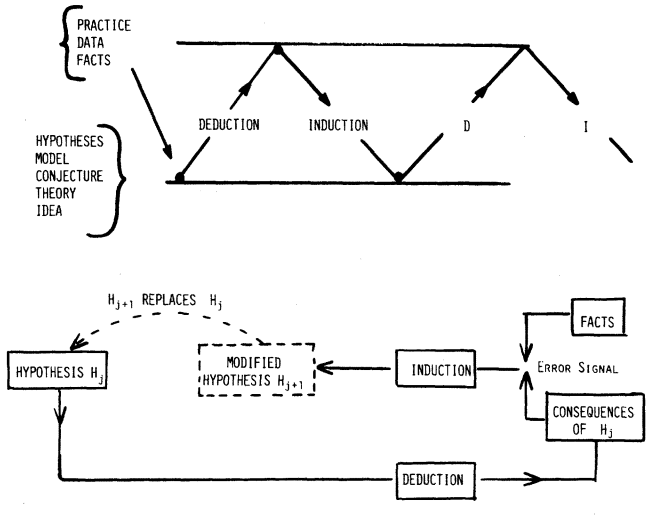
2.1 Iteration Between Theory and Practice

One important idea is that science is a means whereby learning is achieved, not by mere theoretical speculation on the one hand, nor by the undirected accumulation of practical facts on the other, but rather by a motivated *iteration* between theory and practice such as is illustrated in Figure A(1).

A. The Advancement of Learning

A(1) An Iteration Between Theory and Practice

A(2) A Feedback Loop



Matters of fact can lead to a tentative theory. Deductions from this tentative theory may be found to be discrepant with certain known or specially acquired facts. These discrepancies can then induce a modified, or in some cases a different, theory. Deductions made from the modified theory now may or may not be in conflict with fact, and so on. In reality this main iteration is accompanied by many simultaneous subiterations (see, e.g., [1, 2]).

2.2 Flexibility

On this view efficient scientific iteration evidently requires unhampered feedback. The iterative scheme is shown as a *feedback* loop in Figure A(2). In any feedback loop it is, of course, the *error* signal—for example, the discrepancy between what tentative theory suggests *should* be so and what practice says *is* so—that can produce learning. The good scientist must have the flexibility and courage to seek out, recognize, and exploit such errors—especially his own. In particular, using Bacon's

* George E.P. Box is R.A. Fisher Professor of Statistics, University of Wisconsin, Madison, WI 53706. Research was supported by the United States Army under Grant DAHC04-76-G-0010. This is the written version of the R.A. Fisher Memorial Lecture presented at the joint statistical meetings of the American Statistical Association and Biometric Society given at St. Louis in 1974. The author gratefully acknowledges the assistance of his wife Joan who generously shared the results of her research on her father's life and made available the manuscript of her biography of Fisher.

© Journal of the American Statistical Association
December 1976, Volume 71, Number 356
Applications Section

analogy, he must not be like Pygmalion and fall in love with his model.

2.3 Parsimony

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

2.4 Worrying Selectively

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

2.5 Role of Mathematics in Science

Pure mathematics is concerned with propositions like "given that A is true, does B *necessarily* follow?" Since the statement is a conditional one, it has nothing whatsoever to do with the truth of A nor of the consequences B in relation to real life. The pure mathematician, acting in that capacity, need not, and perhaps should not, have any contact with practical matters at all.

In applying mathematics to subjects such as physics or statistics we make tentative assumptions about the real world which we know are false but which we believe may be useful nonetheless. The physicist knows that particles have mass and yet certain results, approximating what really happens, may be derived from the assumption that they do not. Equally, the statistician knows, for example, that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world.

It follows that, although rigorous derivation of logical consequences is of great importance to statistics, such derivations are necessarily encapsulated in the knowledge that premise, and hence consequence, do not describe natural truth. It follows that we cannot know that any statistical technique we develop is useful unless we use it. Major advances in science and in the science of statistics in particular, usually occur, therefore, as the result of the theory-practice iteration.

The researcher hoping to break new ground in the theory of experimental design should involve himself in the design of actual experiments. The investigator who hopes to revolutionize decision theory should observe and take part in the making of important decisions. An appropriately chosen environment can suggest to such an investigator new theories or models worthy to be entertained. Mathematics artfully employed[1] can then enable him to derive the logical consequences of his tentative hypotheses and his strategically selected environment will allow him to compare these consequences with practical reality. In this way he can begin an iteration that can eventually achieve his goal. An alternative is to redefine such words as *experimental design* and *decision* so that mathematical solutions which do not necessarily have any relevance to reality may be declared optimal.

[1] The researcher's purely mathematical ingenuity is likely to be exercised more, not less, by the fact of his dealing with genuine problems.

3. FISHER—A SCIENTIST

With these ideas in mind let us see how Fisher qualifies as a scientist, using for illustration some of the events occurring during his stay at Rothamsted Experimental Station.

3.1 Rothamsted

In 1919, Fisher had rejected the security and prestige of working under Karl Pearson in the most distinguished statistical laboratory in Britain and at that time certainly in the world. Instead, he took up a temporary job as the sole statistician is a small agricultural research station in the country. He was then already 29 years old and he later said that he was aware that he had failed at both the jobs (teacher and actuary) that he had so far attempted.

Sir John Russell, then Director of Rothamsted, later recalled [17, p. 326]

> ... when I first saw him in 1919 he was out of a job. Before deciding anything I wrote to his tutor at Caius college ... about his mathematical ability. The answer was that he could have been a first class mathematician had he "stuck to the ropes" but he would not. That looked like the type of man we wanted.... I had only £200 and suggested he should stay as long as he thought that should suffice.... He reported to me weekly at tea at my house.... It took me a very short time to realize that he was more than a man of great ability, he was in fact a genius.

At the end of a year, Fisher, who had a wife and child, had used up twice the £200, but by that time he had been given a permanent post.

3.2 Weighing the Baby

For the theory-practice iteration to work, the scientist must be, as it were, mentally ambidextrous; fascinated equally on the one hand by possible meanings, theories, and tentative models to be induced from data and the practical reality of the real world, and on the other with the factual implications deducible from tentative theories, models and hypotheses.

Fisher had great interest in practical matters. For example, he begins the real business of his book *Statistical Methods for Research Workers* in Chapter 2, by discussing different ways of plotting data. His first example is introduced as follows [12, p. 25]: "Figure 1 represents the growth of a baby weighed to the nearest ounce at equal intervals from birth." He does not say that this is any particular baby. Recently I was fortunate to see the Fisher family records in which in Fisher's own hand are recorded the weight from birth of every one of his nine

children, weighed by himself, with the results carefully graphed. Comparison shows that the child is his second son, Harry Leonard, who was born in 1923 shortly before the first edition of the book was written. The next leg of the scientific iteration is hinted at as he goes on to discuss how best to plot the data so as to make "a rough examination of the agreement of observation with any (proposed) law of increase."

3.3 Find the Lady

The extraordinary extent to which Fisher's actual every day experience was grist to the mill of his inductive mind is further illustrated in the famous opening lines of Chapter II of Fisher's book *The Design of Experiments* [11, p. 11]: "A lady declares that by tasting a cup of tea ... she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested." Fisher proceeds to use this example to explain and illustrate the basic principles of good statistical design.

There was, of course, a real lady. This incident happened many years before the book was written and just after Fisher came to Rothamsted. The lady was Dr. Muriel Bristol, the algologist, and she had declined the cup of tea that Fisher had offered her because he had added the tea first. Fisher declared it made no difference. To which she replied "Of course it did." Her future husband, William Roach, who was close at hand said "Let's test her," they did, and according to him she made nearly every choice correctly. In this she behaved similarly to the lady in the book who got one wrong.

3.4 From Soil Bacteria to Nonlinear Design

The tea urn was a great catalyst to iteration. There, each afternoon, Fisher conversed with members of the scientific staff and with visitors and became involved in their problems, often with dramatic consequences. One scientist who came to Rothamsted about the same time as Fisher and became his intimate friend was the bacteriologist, Gerard Thornton. It was he who first interested Fisher in improving the time consuming dilution methods for making bacterial counts. This resulted in Fisher's pioneering work on nonlinear design in 1922 mentioned by Cochran [4].

3.5 From Cotton to Extreme Values

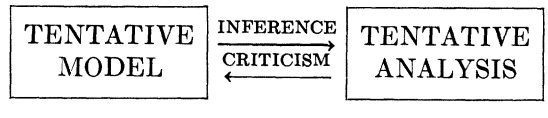
One of the early visitors to Rothamsted was L.H.C. Tippett from the Cotton Research Institute. A matter of great practical concern to him was the strength of cotton yarn. Since the breaking strength of a piece of cotton is the strength of the weakest link, he was faced with what we should now call the extreme value problem. Tippett had first studied with Karl Pearson and had earlier approximated the distribution using the method of moments. In cooperation with Fisher the problem was tackled rather differently. The authors note [14, p. 180] that, "the limiting distribution must be such that the extreme member of a sample of n from such a distribution has itself a similar distribution." This simple but remarkable insight leads to a functional equation which yields as its solution the basic limiting forms. From these forms almost all subsequent work on the subject springs. The theory has applications in such different fields as the design of dams and the reliability of components. Like so many of Fisher's brain children this is now regarded as a distinct field of study.

I will use for further illustration work that Fisher did at Rothamsted between 1919 and 1927 which began with regression analysis and ended with a complete and elegant theory of experimental design which is still the basis for most statistically planned experiments. This work was published in a series of papers having the general title "Studies in Crop Variation" and numbered I, II, III,[2] IV, and VI [7, 13, 8, 5, 6].

3.6 From Dung to Orthogonal Polynomials and Residual Analysis

By 1919 13 plots on Broadbalk wheat fields had received thirteen different manurial treatments uniformly for 67 years. In "Studies in Crop Variation I" [7], Fisher begins by presenting a workmanlike discussion, which lasts for twelve pages, of the responses to the thirteen different manures revealed by his analysis of the Broadbalk data. In particular, he concludes that there is really nothing like plain dung. It gives a high yield with no significant diminution of its effect over the years. He then quite suddenly shifts from manure to mathematics revealing where his analysis has come from. In the next few pages he introduces orthogonal polynomials, presents formulas for their calculation from equispaced data, obtains the distributional properties of the coefficients, and shows how their significance may be judged. Without calling it that he presents the appropriate analysis of variance which he has used in fitting fifth degree polynomials to the annual yields. Most interesting of all, he discusses the properties of the residuals $y - \hat{y}$ from a fitted polynomial of any degree r allowing us to see him in the guise of what some people now call a data analyst.

Data analysis, a subiteration in the process of investigation, is illustrated here.



In the inferential stage, the analyst acts as a sponsor for the model. Conditional on the assumption of its truth he selects the best statistical procedures for analysis of the data. Having completed the analysis, however, he must switch his role from sponsor to critic.[3] Conditional now on the contrary assumption that the model may be

[2] This paper [8] was presented to the Royal Society without the general title but was mysteriously labelled III and had clearly been originally intended for this series.

[3] The apt christening of statistical criticism is due to Cuthbert Daniel.

seriously faulty in one or more suspected or unsuspected ways he applies appropriate diagnostic checks, involving various kinds of residual analysis.

In order to conduct his analysis of the residuals from the fitted polynomials, Fisher obtained

i. the average value of $V(y - \hat{y})$ as $(1 - (r + 1)/n)\sigma^2$,
ii. the individual variances of the residuals $y_j - \hat{y}_j$ for the 67 observations,
iii. the identity

$$\sigma^2 = V(y_j) = V(y_j - \hat{y}_j) + V(\hat{y}_j) \ , \quad j = 1, 2, \ldots, n \ ,$$

iv. an approximate formula for the autocorrelations of residuals from a fitted polynomial of any degree.

The average value of $V(\hat{y})$ from (i) and (iii) is $\sigma^2(r + 1)/n$. Thus, Fisher says if we want to have a small variance for $\hat{y}$ we should keep r small—a demonstration of the value of parsimony, helping to justify his use of polynomials of only fifth degree. Fisher plots the variances $V(y_j - \hat{y}_j)$ for the individual residuals against their time order j. Because of relation (iii) the graph looked at upside down is also a plot of $V(\hat{y}_j)$. Using this he notes the deceptive reduction of $V(y_j - \hat{y}_j)$ at the extremities of the scale and the corresponding increase of $V(\hat{y}_j)$ and says [7, p. 123] "it is a weakness of the polynomial form that the extreme terms should be so much affected."

Finally, mentioning that overfitting and underfitting are both to be avoided he uses the matching of theoretical and empirical *autocorrelations* of residuals to check when a polynomial of sufficiently high degree has been fitted. In particular, he compares theoretical and observed autocorrelations of residuals from polynomials of degree zero and five to show the inadequacy of the former and the satisfactory fit of the latter. This application of serial correlation of residuals to the awkward problem of deciding at what point adequacy of fit has been achieved has great freshness and interest 55 years later.

3.7 Weeds and the Education Acts

Fisher was perplexed by the shapes of his fitted yield graphs. These showed a pattern of significant slow changes *common* to all the 13 Broadbalk plots. In particular, there was a common tendency for low yields roughly in the period 1870–1880. This common pattern was not due to weather; a similar analysis he conducted for successive yields of experimental wheat at Woburn, wheat averages for the whole of Hertfordshire, and for barley and grass from experimental plots at Rothamsted, failed to show it. He speculates [7, p. 129], "Of all the organic factors which influence the yield of wheat it is probable that weeds alone change sufficiently slowly to explain the changes at Broadbalk."

He goes on to describe, as only a dedicated gardener could, all the various weeds that were found there. He notes that old records show that, in 1853, 211 man-days and 714 boy-days were spent in weeding the field. In particular, the boys probably held in check by hand weeding the slender foxtail grass *Alopecuris agrestis*. But he says [7, p. 131] "it may be remembered that the Education Acts of 1876 and 1880 made attendance at school compulsory." We are left to speculate whether the low wheat yields occurred after that time because the hands of the little boys who pulled the foxtail grass were now covered with ink and not with earth.

3.8 From Rainfall and Wheat Yield to Distributed Lags

In 1924, in the third paper of the series [8], he used the Broadbalk data to demonstrate the influence of rainfall on the wheat yield. At the beginning of the paper he seemed to fear that he might be expected to account for the effects not only of rainfall but also for such other variables as maximum and minimum temperature, dew point, and hours of bright sunshine. But he points out that allowances for the effect of each of these on the final harvested yield would need to be included at least for each month separately. And he says if so many regressors are included a very high proportion of the total variation can seem to be accounted for by chance alone. In case some dissident reader might doubt it, he thereupon outlines the derivation of the distribution of the multiple correlation coefficient in one paragraph flat using n-dimensional geometry and on the next page produces a short table of tail areas for R. He then goes on to discuss the misleading effects of selection in what would now be called step-wise regression.

Fisher's data were as follows:

i. for each of the 13 Broadbalk plots he had the harvested wheat yields for each of 60 years,[4]
ii. for each of these 60 years he had daily rainfall records and for convenience he aggregated these for each year into 61 six-day periods ($6 \times 61 = 366$) beginning immediately after the harvest.

In a remarkable demonstration of parsimonious modeling he first suggests that the yield of wheat in the jth year, w_j say, might be represented by

$$w_j = c + \sum_{t}^{61} a_t r_{jt} \ , \quad j = 1, 2, \ldots, 60 \ . \qquad (3.1)$$

In this model the coefficient a_t provides the average effect on eventual harvested yield of one inch of rain in the tth time period. In modern parlance (3.1) might be called a "transfer function" model expressing the "memory" of the system. Economists later called it a "distributed lag" model but they seem to have been unaware of Fisher's prior work or of his ingenious way of proceeding using orthogonal polynomials.

As it stands (3.1) is highly nonparsimonious. Fisher decided, therefore, to represent the rainfall data r_{jt} by orthogonal polynomials of fifth degree. He now notes that the coefficients a_t should also follow a smooth curve which might be represented in the same way. Thus,

$$a_t = \alpha_0 T_{0t} + \alpha_1 T_{1t} + \ldots + \alpha_5 T_{5t} \ ,$$
$$r_{jt} = \rho_{0j} T_{0t} + \rho_{1j} T_{1t} + \ldots + \rho_{5j} T_{5t} \ .$$

[4] Five years 1890, 1891, 1905, 1906, and 1915 were omitted because the plots in these years had special treatment.

But if the orthogonal functions T_{it} are chosen so that $\sum_t T_{it}^2 = 1$; then, after summing, (3.1) may be written

$$w_j = c + \alpha_0\rho_{0j} + \alpha_1\rho_{1j} + \ldots + \alpha_5\rho_{5j} \ .$$

The α's which determine the lagged weights in the transfer function can thus be obtained by regressing the w_j onto the estimated ρ's.

Having carried through the necessary heavy calculations and graphed his results Fisher conducts a very extensive discussion and comparison of the polynomial distributed lag curves for the differently manured plots from which, in particular, he adduces the predominant effect of rain in reducing soil nitrates. One feels that his love of parsimony was certainly not lessened by the fact that the computations were performed by hand by himself and his assistant. Indeed, much can still be learnt from his discussion about economical processes of calculation and appropriate checks [8, p. 111–3].

3.9 From Fertilizer and Potatoes to the Analysis of Variance

About this time Fisher was getting rather tired of analyzing old records—he later described it as "raking over the muck heap." In "Studies in Crop Variation II," jointly authored with his assistant, Miss W.A. MacKenzie, and subtitled "The Manurial Response of Different Potato Varieties," [13] he tried his hand at analyzing some *experimental* data from Rothamsted. The authors remark that it would be convenient if (contrary to some expert opinion) different varieties of plants did *not* react differently to fertilizers, or as we should say now, if there were no interaction between variety and fertilizer.

An experiment had recently been run by Thomas Eden, a crop ecologist at Rothamsted, in which each of twelve varieties of potatoes were tested with six different combinations of manure. This experiment was analyzed as if it were a thrice replicated and randomized 12×6 factorial. (It wasn't, but we return to that later.)

From the analysis of variance which is presented, the answer to the question, "Is there significant interaction between varieties and manures?" appears to be No!

There are some remarkable things about this paper, however:

i. The analysis of variance, hinted at earlier, appears here for the first time in its completeness. It arrives quite suddenly and unannounced in the middle of the paper after the discussion of agricultural questions. It is, of course, not even mentioned in the title.

ii. After the algebraic identity between the total sum of squares and the within and between treatments sum of squares has been written down, the statement is made [13, p. 315] "If all the plots were undifferentiated, as if the numbers had been mixed up and written down in random order, the average value of each of the two parts is proportional to the number of degrees of freedom in the variation of which it is compared." Thus, at the very beginning, randomization, an important flag under which Fisher will sail, is firmly nailed to the mast.

iii. The analysis is wrong, because in fact the trial was actually run as what is now called a split plot design. Feedback in the form of the appropriate correction came quickly in the first edition of *Statistical Methods* in 1925 (see [12, p. 238]). Using part of the same data, Fisher there gives the correct analysis and points out that it is essential to use separate error variance estimates (for between and within plot comparisons) and shows that one is indeed significantly larger than the other.

iv. In this very first paper on the analysis of variance, Fisher demonstrated the flexibility of his thought by questioning the linear model (which almost everybody else has ever since accepted as representing received truth). The authors say [13, p. 316], "the above test is only given as an illustration of the method; the summation formula for combining the effects of variety and manurial treatment is evidently quite unsuitable for the purpose. No one would expect to obtain from a low yielding variety the same actual increase in yield which a high yielding variety would give ... a far more natural assumption is that the yield should be the product of two factors one depending on the variety and one on the manure." With the possibility of transformation so much a part of Fisher's everyday thought, we might expect him now to proceed along that route but in fact he derives the appropriate nonlinear analysis, devising methods which have only recently been rediscovered [18].

3.10 Mice, Tigers, and Randomization

A man in daily muddy contact with field experiments could not be expected to have much faith in any direct assumption of *independently* distributed normal errors. While the supposition of marginal normality for the errors might be regarded as innocuous, the idea that errors from adjacent plots of land could be treated as independent would be obviously absurd and dangerous. This was one important reason for Fisher's insistence (i) on the physical act of randomization as a necessary condition for the validity of any experiment and (ii) that given that randomization had been carried out inferences should be made from the appropriate randomization distribution; to which, however, standard normal theory often provided an adequate approximation.

To guarantee the exact validity of the usual null tests made with the standard linear model it is not, of course, necessary that the density function of the error vector $\mathbf{e}$ be spherically normal, it is necessary only that it be spherically symmetric,[5] i.e., the density function be of the form $f(\mathbf{e}'\mathbf{e})$. The fact that standard normal theory often provides an adequate approximation to that given by randomization theory is not because the density for randomized errors is necessarily approximated by that of independent normal deviates. It is rather because, in the appropriate vector space, the symmetry induced by randomization is approximated by spherical symmetry.

Fisher showed some irritation with later workers who saw only a rich source of purely mathematical development in his work. In particular, workers on what has come to be called "distribution-free" tests have often failed to emphasize and sometimes perhaps even to realize the limitations imposed by the necessary assumption of symmetry of the joint error distribution. The

[5] Obviously, this must be true for any criterion which is a homogeneous function of the data of degree zero.

validity of this assumption could, of course, only be guaranteed by randomization. Otherwise, the derived procedures, far from being distribution free, would be almost as restrictive as those derived on the assumption of normal independent errors. It is true that long usage has seemed to sanctify the proposition that density functions are of the form $p(\mathbf{y}) = \prod_i f(y_i)$ or at least that $p(\mathbf{y}) = S(\mathbf{y})$, where S is some symmetric function of the elements of $\mathbf{y}$. These propositions have come to be treated almost as natural laws or at least as rules of the game that no sportsman would question.[6] In fact, of course, experiments where errors cannot be expected to be independent are very common.

These points are not new but if we are to appreciate Fisher's point of view they need to be brought together and illustrated together. For this latter purpose the results of a simple sampling experiment are shown in the table. Two samples of 10 observations from identical populations of the forms indicated were taken and subjected to a t-test (t) and a Mann-Whitney test (MW). The sampling was repeated 1,000 times and the number of results significant at the 5 percent point was recorded. Ideally, this number should be 50 (that is, 5 percent of the total) but it has a standard deviation of about 7 because of sampling errors. More accurate results may be obtained by taking larger samples or by analytical procedures, however, since there is no practical difference between a significance level of say 4 percent and 6 percent, the present investigation suffices for illustration. Autocorrelation between adjacent values was introduced by generating observations from a moving average model of the form $y_t = u_t - \theta u_{t-1}$. In this model the u_t were independently and identically distributed about zero in the forms indicated in the table. Values of θ were chosen so that ρ_1, the first serial correlation, had values of -0.4 and $+0.4$.

Frequency in 1,000 Trials of Significance at the 5 Percent Level Using the t-Test (t) and the Mann-Whitney Test (MW) with No Randomization (NR) and Randomization (R)

ρ_1	Test	Parent distribution: Rectangular NR	Rectangular R	Normal NR	Normal R	Chi-square[a] NR	Chi-square[a] R
		Independent observations					
0.0	t	56	60	54	43	47	59
	MW	43	58	45	41	43	44
		Autocorrelated successive observations					
−0.4	t	5	48	3	55	1	63
	MW	5	43	1	49	2	56
+0.4	t	125	59	105	58	114	54
	MW	110	46	96	53	101	43

[a] The parent chi-square distribution has four degrees of freedom and is thus highly skewed.

[6] Except in the study of time series.

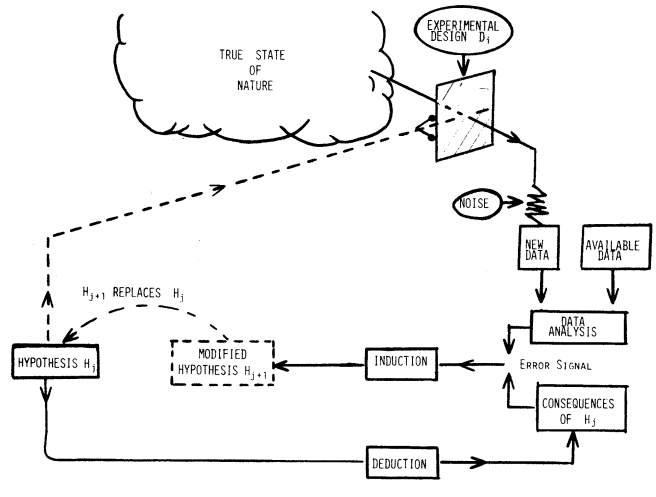
The frequencies shown under NR are those obtained for a nonrandomized test. The frequencies under R are those obtained when the observations were randomly allocated to the two groups.

As is to be expected the significance level of the t-test is affected remarkably little by the drastic changes made in the marginal parent distribution—changes for which the distribution-free test provides insurance. Unfortunately, of course, both tests are equally impaired by error dependence unless randomization is introduced when they do about equally well. The point is, of course, that it is the act of randomization that is of major importance here not the introduction of the distribution-free test function.

3.11 From Muck Raking to Group Theory

Eden's potato data served to illustrate the method of analysis of variance but Fisher appears to have had no hand in planning that experiment. The design is not randomized nor blocked and its very deficiencies call for appropriate remedies. When Fisher's friend Gosset saw the paper, he wrote to Fisher [15, Letter No. 29], "The experiment seems to me to be quite badly planned, you should give them a hand in that ...." Fisher later notes Gosset's "suggesting that I should start designing experiments" [15, summary of Letter No. 29]. This he proceeded to do. The iterative process including the design aspect is sketched in Figure B.

B. Data Analysis and Data Getting in the Process of Scientific Investigation[a]



[a] The experimental design is here shown as a movable window looking onto the true state of nature. Its positioning at each stage is motivated by current beliefs, hopes, and fears.

Between 1919 and 1928 an iterative sequence occurred that went through three main stages, each leading logically to the next via interaction of theory and practice. The analysis of existing records led to the analysis of *experimental trials* which then led to the *design* of experimental trials.

There were different but interactive aspects to this development. We can see (i) sequential evolution of the

new methods in response to unfolding realizations of need, (ii) the persuading of practitioners to try the new techniques, and (iii) the changing role of the statistician implied by the development.

3.12 Evolution of the New Methods

Fisher's attempts to analyze experimental data quickly led him to the essential principles of experimental design. The need for randomization to achieve validity; for replication to provide a valid estimate of error; for blocking extraneous sources of disturbance to achieve accuracy. Blocking in two directions simultaneously (by randomized Latin squares) was particularly appealing. Fisher would have been brought to see the enormous advantages of the unorthodox factorial arrangements as an economical way to assess the effects of variables in combination by, for example, his early attempts to impart meaning to the differences associated with the 13 differently manured Broadbalk plots to which fertilizers had been applied in a highly nonbalanced manner. However, while the efficiency of factorial designs could be increased by packing in more factors, larger factorial designs required bigger blocks and hence produced greater inhomogeneity in the experimental material, giving larger experimental errors. The answer which quickly followed was confounding.

3.13 Persuading Practitioners

The blessings of feedback were only available if scientists would try out his designs but, not surprisingly, Fisher at first did not have an easy job selling his revolutionary ideas at Rothamsted. Indeed, the first design run to his specification (in 1924) was not done at Rothamsted at all. It was a randomized Latin Square design run at Bagshot for the Forestry Commission who had asked for and acted on his advice. But between 1924 and 1929, as described in "Studies in Crop Variation IV and VI" [5, 6], there is a rapid development of ideas which were quickly put into practice. It is clear that Eden had become a convinced disciple during this period and it is refreshing, but alas unfamiliar, to see publication of new designs simultaneous with data obtained from their successful use. By the end of this period data were being collected from designs of great accuracy and beauty which included all of Fisher's ideas.

In spite of all this in 1926 the Director of Rothamsted, Sir John Russell, wrote a paper [16] in the *Journal of the Ministry of Agriculture* about agricultural experimentation which almost totally ignored the ideas of his protegé. However, in the next issue [9] in a paper notable for its brevity and clarity, Fisher outlined *his* philosophy on the subject, setting his boss to rights and anyone else who would listen.

3.14 A New Heritage for Statisticians

The original concept that the research station needed a statistician was revolutionary, but certainly the role initially envisaged in 1919 for the statistician was a passive and possibly even a temporary one. Russell wondered if anything more could be extracted from the existing records.

Fisher's work gradually made clear that the statistician's job did not begin when all the work was over—it began long before it was started. The statistician was not a curator of dusty relics. His responsibility to the scientific team was that of the architect with the crucial job of ensuring that the investigational structure of a brand new experiment was sound and economical. The latter role is much more fun than the former. He himself relished it and we should thank him for bequeathing it to us. It calls for abilities of a high order. It requires among other things the wit to comprehend complicated scientific problems, the patience to listen, the penetration to ask the right questions, and the wisdom to see what is, and what is not, important. Finally, it requires from the statistitian the courage to wager his reputation each time an experiment is run. For the time must come when all the data are in and conclusions must be drawn; at this stage oversights in the design, if they exist, will become embarrassingly evident.

4. PERILS OF THE OPEN LOOP

We have seen some examples of the extraordinary progress made in our science over a brief ten-year period as a result of feedback between theory and practice. Feedback requires a closed loop. By contrast, when for any reason the loop is open, progress stops. Such stagnation can occur with the (normally iterative) cycle stuck either in the practice mode or in the theory mode.

4.1 Cookbookery and Mathematistry

The maladies which result may be called *cookbookery* and *mathematistry*. The symptoms of the former are a tendency to force all problems into the molds of one or two routine techniques, insufficient thought being given to the real objectives of the investigation or to the relevance of the assumptions implied by the imposed methods. Concerning the latter, Fisher's apparently bivalent attitude towards mathematicians has often been remarked and has been the cause of perplexity and annoyance. He himself was an artist in the use of mathematics and emphasized the importance of mathematical training for statisticians—the more mathematics known the greater the potential to be a good statistician. Why then did he sometimes seem to refer so slightingly to mathematicians? The answer I think is that his real target was "mathematistry." It is to make the distinction that the word is introduced here.

Mathematistry is characterized by development of theory for theory's sake, which since it seldom touches down with practice, has a tendency to redefine the problem rather than solve it. Typically, there has once been a statistical problem with scientific relevance but this has long since been lost sight of. Fisher felt strongly about

this last point, particularly when he himself had produced the originally useful idea. I have cited already the development of distribution-free tests which, he felt, misused ideas initiated in Chapter III of his book *Design of Experiments* [11, p. 48]. Another annoyance was the generalization to what he felt was absurdity of his applications of group theory and combinatorial mathematics to experimental design.

The penalty for scientific irrelevance is, of course, that the statistician's work is ignored by the scientific community. But this does not come to the notice of a statistician who has no contact with that community. It is sometimes alleged that there is no actual harm in mathematistry. A group of people can be kept quite happy, playing with a problem that may once have had relevance and proposing solutions never to be exposed to the dangerous test of usefulness. They enjoy reading papers to each other at meetings and they are usually quite inoffensive. But we must surely regret that valuable talents are wasted at a period in history when they could be put to good use.

Furthermore, there is unhappy evidence that mathematistry is not harmless. In such areas as sociology, psychology, education, and even, I sadly say, engineering, investigators who are not themselves statisticians sometimes take mathematistry seriously. Overawed by what they do not understand, they mistakenly distrust their own common sense and adopt inappropriate procedures devised by mathematicians with no scientific experience.

An even more serious consequence of mathematistry concerns the training of statisticians. We have recently been passing through a period where nothing very much was expected of the statistician. A great deal of research money was available and one had the curious situation where the highest objective of the teacher of statistics was to produce a student who would be another teacher of statistics. It was thus possible for successive generations of teachers to be produced with no practical knowledge of the subject whatever. Although statistics departments in universities are now commonplace there continues to be a severe shortage of statisticians competent to deal with real problems. But such are needed.

4.2 Meeting the Challenge

As long ago as 1950, Fisher, delivering the Eddington Memorial Lecture at Cambridge, said [10, p. 22]

> For the future, so far as we can see it, it appears to be unquestionable that the activity of the human race will provide the major factor in the environment of almost every evolving organism. Whether they act consciously or unconsciously human initiative and human choice have become the major channels of creative activity on this planet. Inadequately prepared we unquestionably are for the new responsibilities, which with the rapid extension of human control over the productive resources of the world have been, as it were, suddenly thrust upon us.

One by one, the various crises which the world faces become more obvious and the need for hard facts on which to take sensible action becomes inescapable. The demand for competent statisticians who can tease out the facts by analyzing data, planning investigations, and developing the necessary new theory and techniques will, therefore, continue to increase.

4.3 Training of Statisticians

Competent statisticians will be front line troops in our war for survival—but how do we get them? I think there is now a wide readiness to agree that what we want are neither mere theorem provers nor mere users of a cookbook. A proper balance of theory and practice is needed and, most important, statisticians must learn how to be good scientists; a talent which has to be acquired by experience and example. To quote Fisher once more, in 1952, in a letter concerning a proposed Statistics Center to be set up in Scotland he said: "I have no hesitation in advising that such a centre as you have under discussion should plan to integrate teaching closely with project work in which practical experience can be gained by those who are capable of learning from it; in contradistinction to the ruinous process of segregating the keener minds into a completely sterile atmosphere" [3]. It is encouraging that at more and more statistical centers such advice is now being taken seriously.

5. CONCLUSION

We may ask of Fisher

Was he an applied statistician?
Was he a mathematical statistician?
Was he a data analyst?
Was he a designer of investigations?

It is surely because he was all of these that he was much more than the sum of the parts. He provides an example we can seek to follow.

[*Received May 1976.*]

REFERENCES

[1] Box, G.E.P. and Tiao, G.C., *Bayesian Inference in Statistical Analysis*, Reading, Mass.: Addison-Wesley Publishing Co., 1973.

[2] ——— and Youle, P.V., "The Exploration and Exploitation of Response Surfaces: An Example of the Link Between the Fitted Surface and the Basic Mechanism of the System," *Biometrics*, 11, No. 3 (1955), 287–323.

[3] Box, Joan Fischer, *Fisher, The Life of a Scientist*, New York: John Wiley & Sons, Inc. In press.

[4] Cochran, W.G., "Experiments for Nonlinear Functions," *Journal of the American Statistical Association*, 68, No. 344 (1973), 771–81.

[5] Eden, T. and Fisher, R.A., "Studies in Crop Variation IV. The Experimental Determination of the Value of Top Dressings with Cereals," *Journal of Agricultural Science*, 17 (1927), 548–62.

[6] ——— and Fisher, R.A., "Studies in Crop Variation VI. Experiments on the Response of the Potato to Potash and Nitrogen," *Journal of Agricultural Science*, 19 (1929), 201–13.

[7] Fisher, R.A., "Studies in Crop Variation I. An Examination of the Yield of Dressed Grain from Broadbalk," *Journal of Agricultural Science*, 11 (1921), 107–35.

[8] ———, "Studies in Crop Variation III. The Influence of Rainfall on the Yield of Wheat at Rothamsted," *Philosophical Transactions of the Royal Society of London*, B, No. 213 (1924), 89–142.

[9] ———, "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture*, 33 (1926), 503–13.

[10] ———, "Creative Aspects of Natural Law," The Eddington Memorial Lecture, Cambridge, Eng.: Cambridge University Press, 1950.

[11] ———, *The Design of Experiments*, (8th Ed.), Edinburgh: Oliver & Boyd, Ltd., 1966.

[12] ———, *Statistical Methods for Research Workers*, (14th Ed.), Edinburgh: Oliver & Boyd, Ltd., 1970.

[13] ——— and MacKenzie, W.A., "Studies in Crop Variation II. The Manurial Response of Different Potato Varieties," *Journal of Agricultural Science*, 13 (1923), 311–320.

[14] ——— and Tippett, L.H.C., "Limiting Forms of the Frequency Distribution of the Largest and Smallest Member of a Sample," *Proceedings of the Cambridge Philosophical Society*, 24 (1928), 180–90.

[15] Gosset, W.S., *Letters from W.S. Gosset to R.A. Fisher, 1915–1936*, with summaries by R.A. Fisher and a foreword by L. McMullen, (2nd Ed.), Privately circulated, 1970.

[16] Russell, E. John, "Field Experiments: How They Are Made and What They Are," *Journal of the Ministry of Agriculture*, 32 (1926), 989–1001.

[17] ———, *A History of Agricultural Research in Great Britain*, London: Allyn and Unwin, Ltd., 1966.

[18] Wold, H., "Nonlinear Estimation by Iterative Least Squares Procedures," in F.N. David, ed., *Research Papers in Statistics, Festschrift for J. Neyman*, New York: John Wiley & Sons, Inc., 1966, 411–44.