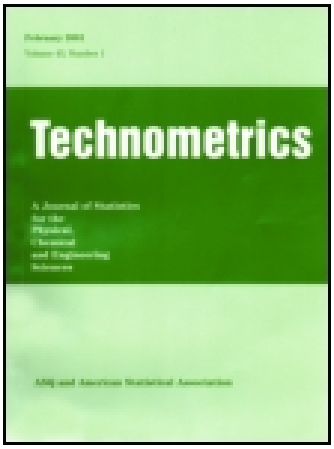

Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

The Importance of Practice in the Development of Statistics

George E.P. Box [a]
[a] Mathematics Research Center University of Wisconsin-Madison , Madison , WI , 53706
Published online: 23 Mar 2012.

Editor's Note: George Box was instrumental in the creation of *Technometrics*. His contributions to the journal and to the practice of statistics in the physical, chemical, and engineering sciences have continued over the 25 years of *Technometrics'* history, so it is a particular pleasure to include an invited paper by him in our 25th anniversary observance. This article is the text of a lecture that was videotaped on 16 November, 1982, for the American Statistical Association Archives under the sponsorship of E. I. DuPont DeNemours and Company in cooperation with the American Statistical Association and its Committee on the Filming of Distinguished Statisticians. The Editors and Management Committee are grateful to these parties for making this article available and possible and to George Box for this and many other contributions to *Technometrics*.

The Importance of Practice in the Development of Statistics

George E. P. Box

Mathematics Research Center
University of Wisconsin-Madison
Madison, WI 53706

The article shows how application and consideration of the scientific context in which statistics is used can initiate important advances such as least squares, ratio estimators, correlation, contingency tables, studentization, experimental design, the analysis of variance, randomization, fractional replication, variance component analysis, bioassay, limits for a ratio, quality control, sampling inspection, nonparametric tests, transformation theory, ARIMA time series models, sequential tests, cumulative sum charts, data analysis plotting techniques, and a resolution of the Bayes-frequentist controversy. It appears that advances of this kind are frequently made because practical context reveals a novel formulation that eliminates an unnecessarily limiting framework.

KEY WORDS: Practice; Theory; Least squares; Ratio estimators; Correlation; Contingency tables; Studentization; Experimental design; Analysis of variance; Randomization; Fractional replication; Variance component analysis; Bioassay; Limits for a ratio; Quality control; Sampling inspection; Nonparametric tests; Transformation theory; ARIMA time series models; Sequential tests; Cumulative sum charts; Data analysis plotting techniques; Bayesian inference; Sampling theory.

1. INTRODUCTION

The importance of practice in guiding the development of statistics hardly needs emphasis, and yet I think it is worth examination. For statistical methods and statistical theory, like so many other things, evolve by a process of natural selection. Least squares, invented at the beginning of the 19th century, is alive and well, but the coefficient of colligation is now seldom used. For development to occur both appropriate tools and motivation are needed. The tools are mathematics, numerical analysis, and computation. An important motivation is the practical need to solve problems. Tools and motivation interact, of course. For example, the existence of fast computers is encouraging the development of new statistical methods that would be impossible without them and that presage further theoretical development. Furthermore, advance must sometimes wait for knowledge of appropriate mathematics. Thus Fisher's ability to solve the distributional problems of correlation and of the linear model rested strongly on his facility with n-dimensional geometry, which his contemporaries lacked.

It would be hard to argue, however, that any one deficiency in the tool-kit is disastrous. Thus least squares—although according to Gauss fully known to him in 1796—could require calculations that were dauntingly burdensome until the onset of modern computers in the 1950's. Galton, Gosset, and Wilcoxon, pioneers respectively in the concepts of correlation, studentization, and nonparameteric tests, did not regard themselves as particularly competent mathematicians. In particular, Gosset's derivation of the sampling distribution of what we now call the t statistic must surely stand as the nadir of rigorous

argument. But he did get the right answer; and he was first.

My theme is to illustrate how practical need often leads to theoretical development. Early examples are the development of the probability calculus, which was closely bound up with the desirability of winning at games of chance, the introduction of least squares by Gauss to reconcile astronomical and survey triangulation measurements, and the invention by Laplace of ratio estimators to determine the population of France (Cochran 1978).

Let us consider some of the children of necessity produced in more modern times.

2. FURTHER EXAMPLES OF THE PRACTICE-THEORY INTERACTION

In the mid-19th century the impact of Darwin's ideas was dramatic. But Darwin, although an intellectual giant, had little mathematical ability. To Francis Galton the challenge was obvious: The rightness and further consequences of Darwin's ideas must be demonstrable using numbers. For example, given that offspring varied about some kind of parental mean, why, with each new branching of a family tree, did variation of species not continually increase? The answer to this practical question lay, he discerned, in the regression toward the mean implied by the bivariate normal surface, which ensures that, on average, sons of six-foot fathers are less than six feet tall (Galton 1886). Again, it was the need he perceived to measure the intensity of the partial similarities between pairs of relatives that led to his introducing the concept of correlation, an idea taken up with great enthusiasm and further developed by Karl Pearson.

Pearson was a man of enormous energy and wide interests, including social reform and the general improvement of the human condition. He was, however, conscious of the fact that, in deciding what kind of reforms ought to be sought, good intentions, although necessary, were not always sufficient. A course of action based on the accepted belief that alcoholism in parents produced mental deficiency in children might be ill advised if, as he demonstrated, that belief was not supported by data (Haldane 1970). Obviously correlation might be useful in such studies, but other measures and tests of association were needed for qualitative variables. Pearson developed such tools, in particular his χ^2 test for contingency tables.

Karl Pearson's methods were developed mainly for large samples and did not meet the practical needs of W. S. Gosset when he came to study statistics for a year at University College, London, in 1906. Gosset had graduated from Oxford with a degree in chemistry and had gone to work for Guinnesses, following the company's policy, begun in 1893, of recruiting scientists as brewers. He soon found himself faced analyzing small sets of observations coming from the laboratory, field trials, and the experimental brewery of which he was placed in charge in 1905 (Gosset 1918).

The general problem Gosset faced was how to deal with unknown nuisance parameters, and specifically the unknown standard deviation in the comparison of means. The method then in use was to substitute some sort of estimate for an unknown nuisance parameter, and then to assume that one could treat the result as if the true value had been substituted. While this might provide an adequate approximation for large samples, it was clearly inadequate when the sample was small. Furthermore, he did not find or expect to find much interest in his problems. He wrote to R. A. Fisher of the t tables, "You are probably the only man who will ever use them" (Box 1978). It must have been clear to him at that time that if anyone was to do anything about small samples it would have to be himself.

Gosset's invention of the t test was a milestone in the development of statistics because it showed how, by studentization, account might be taken of the uncertainty in an estimated nuisance parameter. It thus paved the way for an enormous expansion of the usefulness of statistics, which could not begin to provide answers for agriculture, chemistry, biology, and many other subjects in which small rather than large samples were the rule.

Fisher, as he always acknowledged, owed a great debt to Gosset, both for providing the initial clue as to how the general problem of small samples might be approached, and for mooting the idea of statistically designed experiments.

When Fisher went to Rothamsted in 1919 he was one of several young scientists newly recruited by Russell. He was immediately confronted with a massive set of data on rainfall every day, and harvested yields every year, for 13 Broadbalk plots that had been fertilized in the same pattern for over 60 years. As might be expected his analyses were not routine (Fisher 1921 and 1924); he introduced distributed lag models, orthogonal polynomials, an early form of the analysis of variance, and the distribution of the multiple correlation coefficient. Also, to check the fit of his model he considered the properties of residuals. Furthermore, he devised ingenious methods for lightening the burdensome calculations that had to be made on a desk calculator. But the most important outcome of this "raking over the muck-heap," as he called it, and of analyzing other field experiments that he had had no part in planning, came from the very deficiencies these data presented. The outcome was the invention of experimental design.

How, he was soon led to ask, might experiments be conducted so that they would unequivocally answer the questions posed by the investigator? One can

clearly see the ideas of randomization, replication, orthogonal arrangement, blocking, factorial designs, measurement of interactions, and confounding all developing in response to the practical necessities of field experimentation (Fisher 1926).

Design and analysis came to play complementary roles in Fisher's thinking, so that during the period 1916–1930 we see the analysis of variance first hinted at, and then developed and adapted to accompany the analysis of each new design. In 1923, the analysis of variance first appeared in the tabular form with which we are all familiar (Fisher and Mackenzie 1923). But the object of the investigation was to solve an agricultural problem, and it is typical of Fisher that there is no reference in the title of the article either to the analysis of variance or to the other new statistical ideas it contains. The article, called "The Manurial Response of Different Potato Varieties," introduces us not only to the analysis of variance for a replicated two-way table, but also to its partial justification by randomization theory rather than by normal theory. In addition, it presents methods of analysis (only recently rediscovered; Wold 1966) using models that are nonlinear in the parameters.

By the 1930's there existed at Rothamsted a center where careful statistical planning was going into the process of the generation and analysis of data coming from a host of important problems. Fisher left Rothamsted in 1933 and was succeeded by Yates, who had come two years earlier and was not only a mathematician but also had much practical experience in least squares calculations in geodetic survey work. Yates (1970) made many important advances. In particular he further developed factorial designs and confounding, invented many new designs including balanced incomplete block arrangements, and showed how to cope when, as sometimes happened, things went wrong and there were missing data.

These ideas found wide application and inspired much new research. For example, Jack Youden (1937), then working at the Boyce Thomson Institute, was involved in an investigation of the infective power of crystalline preparations of the tobacco-mosaic virus. Some of the difficulties he found were that test plants varied in their tendency to become infected, leaves from the same plant varied depending on their position, and each plant could not be relied upon to provide more than five experimental leaves. In response Youden invented what came to be called the Youden Square, a design that stands in the same relationship to the latin square as the balanced incomplete block does to the randomized block design.

Another important development coming from Rothamsted was fractional replication. Fisher (1935) had pointed out that in suitable circumstances, adequate estimates of error could be obtained in large unreplicated factorials from estimates of high-order interactions that might be assumed to be negligible. Finney (1945), responding to the frequent practical need to maximize the number of factors studied per experimental run, further exploited this possible redundancy by introducing fractional factorial designs. These designs, together with another broad class of orthogonal designs developed independently by Plackett and Burman (1946) in response to war-time problems, have since proved of great value in industrial experimentation. An isolated example of how such a design could be used for screening out a source of trouble in a spinning machine had been described as early as 1934 by L. H. C. Tippett of the British Cotton Industry Research Association (Tippett 1935). The arrangement was a 125th fraction of a 5^5 design!

It seems that whenever a good source of problems existed in the presence of a suitably agile mind new developments were bound to occur. Thus the pressing problem of drug standardization in the hands of Gaddum (1933), Bliss (1935), and (again) Finney (1952) gave rise to modern methods of bioassay using probits, logits, and the like. And in 1940 a study of the standardization of insulin led Edgar Fieller, while working for Boots Pure Drug Company, to a resolution of the problem of finding confidence limits for a ratio and for the solution of an equation whose coefficients were subject to error (Fieller 1940).

Earlier, Henry Daniels, then a statistician at the Wool Industries Research Association, showed how variance component models could be used to expose those parts of a production process responsible for large variations (Daniels 1938). Variance component analysis has since proved of enormous value in the process industries and elsewhere.

Daniels's contribution was one in a series of papers on industrial statistics read in the 1930's before what was then called the Industrial and Agricultural Research Section of the Royal Statistical Society. A leading spirit in getting this section moving was Egon Pearson, whose ideas greatly influenced, and were influenced by, this body. In particular he liked data analysis and graphical illustration and used both effectively to illustrate Daniels's conclusions (Pearson 1938).

An important influence on Pearson was the work of Walter Shewhart on quality control (Shewhart 1931). This work and that on sampling inspection by Harold Dodge heralded more than half a century of statistical innovation coming from the Bell Telephone Laboratories (Dodge 1969, 1970), evidenced most recently by the rekindling of interest in data analysis in a much-needed revolution led by John Tukey (Tukey 1977; Mosteller and Tukey 1977).

Another innovator guided by practical matters was Frank Wilcoxon, a statistician for the Lederle Labs of

the American Cyanamid Company. Just after the Second World War, in the age of desk calculators, he found himself confronted by the need to make thousands of tests on samples from the pharmaceutical research then in progress. He said it was the need for quickness rather than anything else that led to the famous Wilcoxon tests (Wilcoxon 1949), the precursors of much subsequent research on nonparametric methods.

M. S. Bartlett's contributions to statistics are legion. His early contributions to the theory of transformation (Bartlett 1936) had much to do with the fact that, when he was statistician at the Jealotts Hill agricultural research station of Imperial Chemical Industries, he was concerned with the testing of pesticides and so with data that appeared as frequencies or proportions.

Another clear example of the practice-theory interaction is seen in the development of parametric time series models. In 1927 Udny Yule was trying to understand what was wrong with William Beveridge's analysis of wheat price data. The fitting of sine waves of different frequencies by least squares had revealed significant oscillations at strange and inexplicable periods. Yule (1927) suggested that such series ought to be represented, not by a deterministic function subject to error, but by a dynamic system (represented by a linear difference equation) responding to a series of random shocks—this model was likened to a pendulum being periodically hit by peas from a pea shooter. Yule's revolutionary idea, with important further input from Slutsky (1927), Wold (1954), and others, was the origin of autoregressive-moving average models.

Unfortunately, the practical use of these models was for some time hampered by an excessive concern with stationary processes that vary in equilibrium about a fixed mean. The requirement for stationarity is that the characteristic polynomial for the autoregressive part of the model must have all its zeroes outside the unit circle. Many of the series arising in business and economics do not, however, behave like realizations from such a stationary model. Consequently, for lack of anything better, operations research workers led by Holt (1957) and Winters (1960) began in the 1950's to use the exponentially weighted moving average of past data and its extensions for forecasting series of this kind. This weighted average was introduced at first on purely empirical grounds—it seemed sensible to monotonically discount the past and it seemed to work reasonably well. However, in 1960, Muth showed, rather unexpectedly, that this empirically derived statistic was an optimal forecast for a special kind of autoregressive-moving average model (Muth 1960). This model was not stationary. Its autoregressive polynomial had a root on the unit circle. The general class of models with roots on the unit circle, where stationarity would forbid them, later turned out to be extremely valuable for representing many kinds of practically occurring series, including seasonal series.

The Second World War was a stimulus to all kinds of invention. Allen Wallis has described the dramatic consequence of a practical query made by a serving officer about a sampling inspection scheme (Wallis 1980). The question was of the kind "Suppose, from a sample of twenty items, that three is the critical number of duds. If it should happen that the first three components tested are all duds, why do we need to test the remaining seventeen?" Wallis and Milton Friedman were quick to see the apparent implication that "super-powerful" tests were possible! However, their suggestion that Abraham Wald be invited to work on the problem was resisted for some time. It was argued that this would clearly be a waste of Wald's time, because to do better than a most powerful test was impossible. What the objector had failed to see was that the test considered was most powerful only if it was assumed that n was fixed, and what the officer had seen was that n did not need to be fixed. It is well known how this led to the development of sequential tests (Wald 1947). It is heartening that this particular happening even withstood the scientific test of repeatability, for at about the same time and with similar practical inspiration, sequential tests (of a somewhat different kind) were discovered independently in Great Britain by George Barnard (1946).

Nor was this the end of the story. Some years later Ewan Page, then a student of Frank Anscombe, while considering the problem of finding more efficient quality control charts, was led to the idea of plotting the cumulative sum of deviations from the target value (Page 1954).

The concept was further developed by Barnard (1959), who introduced the idea of a V mask to decide when action should be taken. The procedure is identical to a backwards-running two-sided sequential test. Cusum charts have since proved to be of great value in the textile and other industries. In addition, this graphical test has proved its worth in the "post mortem" examination of data where it can point to the dates on which critical events may have occurred. This sometimes leads to discovery of the reason for the events.

A pioneer of graphical techniques of a different kind is Cuthbert Daniel, an industrial consultant who has used his wide experience to make many contributions to statistics. An early user of unreplicated and fractionally replicated designs, he was concerned with the practical difficulty of estimating error. In particular he was quick to realize that higher order interactions sometimes do occur, and when they do it is important to isolate and study them. His introduction of graphi-

cal analysis of factorials by plotting effects and residuals on probability paper (Daniel 1976) has had major consequences. It has encouraged the development of many other graphical aids, and together with the work of John Tukey it has contributed to the growing understanding that at the hypothesis generation or model-modification stage of the cycle of discovery, it is the imagination that needs to be stimulated, and that this can often best be done by graphical methods.

3. SOME INTERIM CONCLUSIONS

Obviously I could go on with other examples, but at this point I should like to draw some interim conclusions.

I think it is possible to see important ingredients leading to statistical advance. They are (a) the presence of an original mind that can perceive and formulate a new problem and move to its solution, and (b) a challenging and active environment for that mind, conducive to discovery.

Gosset at Guinnesses; Fisher, Yates, and Finney at Rothamsted; Tippett at the Cotton Research Institute; Youden at the Boyce Thomson Institute (with which organization Wilcoxon and Bliss were also at one time associated); Daniels and Cox at the Wool Industries Research Association; Shewhart, Dodge, Tukey and Mallows at Bell Labs; Wilcoxon at American Cyanamid; Daniel in his consulting practice: these are all examples of such fortunate conjunctions.

Further recent examples are Don Rubin's work at the Educational Testing Service, Jerry Friedman's computer intensive methods developed at the Stanford linear accelerator, George Tiao's involvement with environmental problems, Brad Efron's interaction with Stanford Medical School, the late Gwilym Jenkins' applications of time series analysis in systems applications, and John Nelder's development of statistical computing at Rothamsted.

The message seems clear: a statistician who believes himself capable of genuinely original research can find fulfillment in a stimulating investigational environment.

Also, I think it possible to perceive an aspect of the specific nature of the contribution coming from applications—frequently, it is the establishment of a new frame of reference for a problem. This may involve extension, modification, or even abandonment of a previous formulation. It has to be understood that statistical problems are frequently not like, for example, chess problems that may require "white to mate in three moves," given a particular configuration of the pieces. Here a solution based on the pretense that a knight can move like a queen would be unacceptable. Yet the changes in the rules that have sometimes been adopted in reformulation of statistical problems must, at the time of their introduction, have been thought of as little short of cheating. Some examples are:

- Fisher's replacement of the method of moments by maximum likelihood;
- Yates' use of designs in which the number of treatments exceeded the block size;
- Yule's introduction of stochastic difference equations replacing deterministic models;
- Wald's and Barnard's introduction of sequential tests to replace fixed sample tests;
- Page's and Barnard's introduction of quality control charts in which the cumulative sum of the deviations rather than the deviations themselves was plotted;
- Finney's use of fractional, rather than full, factorials;
- Fisher's use of the randomization test to justify normal theory tests as approximations; and
- Daniel's and Tukey's initiation of informal graphical techniques rather than more formal procedures in data analysis.

4. A POSSIBLE RESOLUTION OF THE BAYES CONTROVERSY

One further matter that I think is greatly clarified by the practical context of its application is the problem of statistical inference. Here the consideration of scientific context provides, I believe, a resolution of what is sometimes called the Bayesian controversy. At its most extreme this controversy is a dispute between those who think that all statistical inferences should be made using a Bayesian posterior distribution, and others who believe that sampling theory (that is, frequentist theory) has universal inferential applicability.

I have recently argued (Box 1980, 1983) that the Bayes–Sampling theory controversy arises because of an erroneous tacit assumption that there is only one kind of scientific inference for which there are two candidates, whereas I believe that scientific investigation requires two quite distinct kinds of inference for each of which, one, and not the other, of the Bayes–Sampling candidates is appropriate. One kind of inference that may be called *criticism* involves the *contrasting* of what might be expected if the assumptions A of some tentative model of interest were true with the data $\mathbf{y}_d$ that actually occur. This is conveniently symbolized by subtraction: $y_d - A$. The other kind of inference, which may be called *estimation*, involves the *combination* of observed data $\mathbf{y}_d$ with the assumptions A of some model tentatively assumed to be true. This process is conveniently symbolized by addition: $y_d + A$.

In a statistical context, analysis of residuals, tests of fit, and diagnostic checks, both graphical and numeri-

cal, formal and informal, are all examples of techniques of model criticism intended to stimulate the scientist to model building and model modification, or to the generation of more relevant data should this prove desirable. These techniques must, I believe, ultimately appeal for formal justification to sampling theory.

By contrast, least squares estimation, likelihood estimation, shrinkage estimation, robust estimation, and ridge estimation are all solutions to estimation problems that I think would be better motivated and justified by applying Bayes' theorem with an appropriate model.

There seem to be three distinct considerations supporting this dualistic view of inference: the nature of scientific method, the physiology of the brain, and the mathematics of Bayes' theorem. I consider them in turn.

The Nature of Scientific Method

It has long been recognized that the process of learning is a motivated iteration between theory and practice. By practice I mean reality in the form of data or facts. In this iteration deduction and induction are employed in alternation. Progress is evidenced by a developing model that by appropriate exposure to reality continually evolves until some currently satisfactory level of understanding is reached. At any given stage the current model helps us to appreciate not only what we know, but what else it may be important to find out. It thus motivates the collection of new data appropriate to illuminate dark but possibly interesting corners of present knowledge.

We can find illustration of these matters in everyday experience, or in the evolution of the plot of any good mystery novel, as well as in any reasonably honest account of the events leading to scientific discovery.

Experimental science accelerates the learning process by isolating its essence. Potentially informative experiences are deliberately *staged* and made to occur in the *presence* of a trained investigator.

The instrument of all learning is the brain, an incredibly complex structure, the working of which we have only recently begun to understand. One thing that is clear is the importance to the brain of models where past experience is accumulated. At any given stage of experience some of the models $M_1, M_2, \cdots, M_i, \cdots$, are well established, others less so, while still others are in the early stages of creation. When some new fact or body of facts $\mathbf{y}_d$ comes to our attention, the mind tries to associate the new experience with an established model. When, as is usual, it succeeds in doing so, this new knowledge is incorporated in the appropriate model and can set into motion appropriate action.

Obviously, to avoid chaos the brain must be good at allocating data to an appropriate model and at initiating the construction of a new model if this should prove to be necessary. To conduct such business the mind must be concerned with the two kinds of inferences, criticism and estimation, that were mentioned previously.

The Physiology of the Brain

With two kinds of inferences to consider, it seems of great significance that research, which under the leadership of Roger Sperry has gathered great momentum in the past 20 years, shows that the human brain behaves not as a single entity but as two largely separate but cooperating instruments (Blackeslee 1980 and Springer and Peutsch 1981).

In most people the left half of the cerebral cortex is concerned primarily with language and logical deduction, which plays a major role in estimation, while the right half is concerned primarily with images, patterns, and inductive processes, which play a major role in criticism. The two sides of the brain are joined by millions of connections in the corpus callossum, where information exchange takes place. It is hard to escape the conclusion that the iterative inductive-deductive process of discovery is indeed wired into us.

It is well-known that while the left brain plays a conscious and dominant role, one may be quite unaware of the working of the less assertive right brain. For example, the apparently instinctive knowledge of what to do and how to do it enjoyed by an experienced tennis player comes from the right brain. It is significant that this skill may be temporarily lost if we invite the tennis player to explain how he does it, and thus call the left brain into a dominant and interfering mode.

In this context we see the data analyst's insistence on "letting the data speak to us" by plots and displays as an instinctive understanding of the need to encourage and to stimulate the pattern recognition and model generating capability of the right brain. Also, it expresses his concern that we not allow our pushy deductive left brain to take over too quickly and perhaps forcibly produce unwarranted conclusions based on an inadequate model.

While the accomplishment of the right brain in finding patterns in data and residuals is of enormous consequence to scientific discovery, some check is obviously needed on its pattern-seeking ability, for common experience shows that some pattern or other can be seen in almost any set of data or facts. A check that we certainly apply in our everyday life is to consider whether what has occurred is really exceptional in the context of some relevant reference set of circumstances. Similarly, in statistics, diagnostic checks and tests of fit require, at a formal level, frequentist theory significance tests for their justification.

The Mathematics of Bayes' Theorem

It seems reasonable to require that by a statistical model M we mean a complete probability statement of what is currently supposed to be known a priori (that is, tentatively entertained) about the mode of generation of data $\mathbf{y}$ and of the uncertainty about the parameters $\boldsymbol{\theta}$ given the assumptions A of the model. At some stage i of an investigation the current model M_i would therefore be defined as

$$p(\mathbf{y}, \boldsymbol{\theta} \mid A_i) = p(\mathbf{y} \mid \boldsymbol{\theta}, A_i)p(\boldsymbol{\theta} \mid A_i),$$

which can alternatively be factorized as

$$p(\mathbf{y}, \boldsymbol{\theta} \mid A_i) = p(\boldsymbol{\theta} \mid \mathbf{y}, A_i)p(\mathbf{y} \mid A_i).$$

The last factor in the second expression is the predictive distribution. This is the distribution of all possible samples $\mathbf{y}$ that could occur if the model M_i were true.

After the actual data $\mathbf{y}_d$ become available,

$$p(\mathbf{y}_d, \boldsymbol{\theta} \mid A_i) = p(\theta \mid \mathbf{y}_d, A_i)p(\mathbf{y}_d \mid A_i).$$

The first factor on the right is now the posterior distribution of $\boldsymbol{\theta}$, conditional on the proposition that the actually occurring data $\mathbf{y}_d$ are a realization from the predictive distribution that results from the assumptions of the theoretical model M_i. *If we accept this proposition*, all that can be said about $\boldsymbol{\theta}$ must come from this posterior distribution, and the predictive density is without informational content. However, if, as is always in practice the case, the proposition may be seriously wrong, then, correspondingly, residual information may be contained in the predictive density, and this can not only indicate inadequacy but even point to its nature. In particular the relevance of the model may be called into question by an unusually small value of the predictive density for the observed sample $\mathbf{y}_d$ as measured, for example, by

$$\Pr\,[p(y \mid A) < p(y_d \mid A)],$$

or by an unusually small value of the predictive density $p\{g(\mathbf{y}_d) \mid A\}$ of some suitable checking function $g(\mathbf{y}_d)$, as measured by

$$\Pr\,[p\{g(\mathbf{y}) \mid A\} < p\{g(\mathbf{y}_d) \mid A\}].$$

Figure 1 illustrates the idea for a single parameter θ and a single observation y_d. The particular case illustrated is one where, after the data have become available, it seems sensible to investigate the adequacy of the model rather than to proceed with the estimation of θ from its posterior distribution.

There are many conclusions that flow from this approach, which are discussed and illustrated elsewhere. The most important in the present context is that the investigational background against which statistics is applied seems to require that when Bayes' procedure is employed, the proposition on which it is conditioned ought to be considered in the light of the data. This can be done by appropriate consideration of the predictive density associated with the data $\mathbf{y}_d$. Such an approach can, for example, justify and suggest appropriate analyses of residuals, and at a more formal level produce sampling theory significance tests.

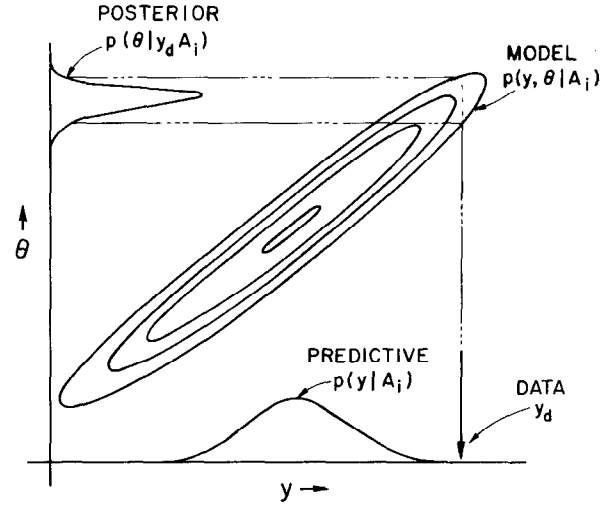

Figure 1.

5. CONCLUSION

In summary, then, I have tried to show how application and consideration of the scientific context in which statistics is used can initiate important advances such as least squares, ratio estimators, correlation, contingency tables, studentization, experimental design, the analysis of variance, randomization, fractional replication, variance component analysis, bioassay, limits for a ratio, quality control, sampling inspection, nonparametric tests, transformation theory, ARIMA time series models, sequential tests, cumulative sum charts, data analysis plotting techniques, and a resolution of the Bayes–frequentist controversy.

It appears that advances of this kind are frequently made because practical context reveals a novel formulation that eliminates an unnecessarily limiting framework.

ACKNOWLEDGMENT

The research for this article was sponsored by The U.S. Army under Contract DAA C29-80-C-0041.

[*Received March 1983.*]

REFERENCES

BARNARD, G. A. (1946), "Sequential Tests in Industrial Statistics," *Journal of the Royal Statistical Society*, Ser. B, 8, 1–21.

——— (1959), "Control Charts and Stochastic Processes," *Journal*

of the Royal Statistical Society, Ser. B, 21, 239- 271.

BARTLETT, M. S. (1936), "The Square Root Transformation in Analysis of Variance," *Supplement to the Journal of the Royal Statistical Society*, 3, 68–78.

BLACKESLEE, T. R. (1980), *The Right Brain*, Garden City, New York: Anchor Press/Doubleday.

BLISS, C. I. (1935), "The Calculation of the Dosage-Mortality Curve," *Annals of Applied Biology*, 22, 134–137.

BOX, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society*, Ser. A, 143, 383–430.

——— (1983), "An Apology for Ecumenism in Statistics," in *Scientific Inference, Data Analysis and Robustness*, eds. G. E. P. Box, T. Leonard, and C. F. Wu, New York: Academic Press.

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

COCHRAN, W. G. (1978), "Laplace's Ratio Estimator," in *Contributions to Survey Sampling and Applied Statistics, Papers in Honor of H. O. Hartley*, ed. H. A. David, New York: Academic Press.

DANIEL, C. (1976), *Applications of Statistics to Industrial Experimentation*, New York: John Wiley.

DANIELS, H. E. (1938), "Some Problems of Statistical Interest in Wool Research," *Journal of the Royal Statistical Society*, Ser. B, 5, 89–112.

DODGE, H. F. (1969 and 1970), "Notes on the Evolution of Acceptance Sampling Plans," *Journal of Quality Technology*, 1, 77–88; 2, 155–162; 3, 225–232; 4, 1–8.

FIELLER, E. C. (1940), "The Biological Standardisation of Insulin," *Journal of the Royal Statistical Society, Supplement*, 7, 1–64.

FINNEY, D. J. (1945), "Fractional Replication of Factorial Arrangements," *Annals of Eugenics*, 12, 291–301.

——— (1952), *Statistical Method in Biological Assay*, London: Charles Griffin & Co.

FISHER, R. A. (1921), "Studies in Crop Variation I. An Examination of the Yield of Dressed Grain from Broadbalk," *Journal of Agricultural Sciences*, 11, 109–135.

——— (1924), "Studies in Crop Variation III. The Influence of Rainfall on the Yield of Wheat at Rothamstead," *Philosophical Transactions of the Royal Society of London*, Ser. B, 213, 89–142.

——— (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture*, 33, 503–513.

——— (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

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

GADDUM, J. H. (1933), *Reports on Biological Standards III. Methods of Biological Assay Depending on a Quantal Response*, Medical Research Council, Special Report Series, No. 183.

GALTON, F. (1886), "Family Likeness in Stature," *Proceedings of the Royal Society of London*, 40, 42–63.

GOSSET, W. S. (1970), *Letters from W. S. Gosset to R. A. Fisher, 1915–1936* (2nd ed.), privately circulated.

HALDANE, J. B. S. (1970), "Karl Pearson, 1857–1957" in *Studies in the History of Statistics and Probability*, eds. E. S. Pearson and M. G. Kendall, Darien, Conn.: Hafner.

HOLT, C. C. (1957), "Forecasting Trends and Seasonals by Exponentially Weighted Moving Averages," *O. N. R. Memorandum, No. 52*, Carnegie Institute of Technology.

HUNTER, J. STUART (1983), "The Birth of a Journal," *Technometrics*, 25, 3–7.

MOSTELLER, F., and TUKEY, J. W. (1977), *Data Analysis and Regression*, Reading, Mass.: Addison-Wesley.

MUTH, J. F. (1960), "Optimal Properties of Exponentially Weighted Forecasts of Time Series with Permanent and Transitory Components," *Journal of the American Statistical Association*, 55, 299–306.

PAGE, E. S. (1954), "Continuous Inspection Schemes," *Biometrika*, 41, 100–115.

PEARSON, E. S. (1938) "Discussion of H. E. Daniels' article: Some Problems of Statistical Interest in Wool Research," *Journal of the Royal Statistical Society*, Ser. B, 5, 89–112.

PLACKETT, R. L., and BURMAN, J. P. (1946), "The Design of Optimum Multifactorial Experiments," *Biometrika*, 33, 4, 305–325.

SHEWHART, W. A. (1931), *Economic Control of Quality of Manufactured Product*, New York: D. Van Nostrand.

SLUTSKY, E. (1927), "The Summation of Random Causes as the Source of Cyclic Processes" (Russian), Problems of Economic Conditions, *3*; English translation in *Econometrica*, (1937) 5, 105.

SPRINGER, S. P., and DEUTSCH, G. (1981), *Left Brain, Right Brain*, San Francisco: W. H. Freeman.

TIPPETT, L. H. C. (1935), "Some Applications of Statistical Methods to the Study of the Variation of Quality of Cotton Yarn," *Journal of the Royal Statistical Society*, Ser. B, 1, 27–55.

TUKEY, J. W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley.

WALD, A. (1947), *Sequential Analysis*, New York: John Wiley.

WALLIS, W. A. (1980), "The Statistical Research Group 1942–45," *Journal of the American Statistical Association*, 75, 320–335.

WILCOXON, F. (1949), *Some Rapid Approximate Statistical Procedures*, Stanford, Conn.: American Cyanamid Company.

WINTERS, P. R. (1960), "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Science*, 6, 324–342.

WOLD, H. (1954), *A Study in the Analysis of Stationary Time Series*, Uppsala, Sweden: Almquist Wiksell Book Co.

——— (1966), "Nonlinear Estimation by Iterative Least Squares Procedures" in *Research Papers in Statistics, Festschrift for J. Neyman*, ed. F. N. David, New York: John Wiley.

YATES, F. (1970), *Experimental Design: Selected Papers of Frank Yates*, Darien, Conn.: Hafner.

YOUDEN, W. J. (1937), "Use of Incomplete Block Replications in Estimating Tobacco-Mosaic Virus," *Contributions from Boyce Thomson Institute IX*, 91–98.

YULE, G. U. (1927), "On a Method of Investigating Periodicities in Disturbed Series," *Philosophical Transactions of the Royal Society of London, A226*, 207.