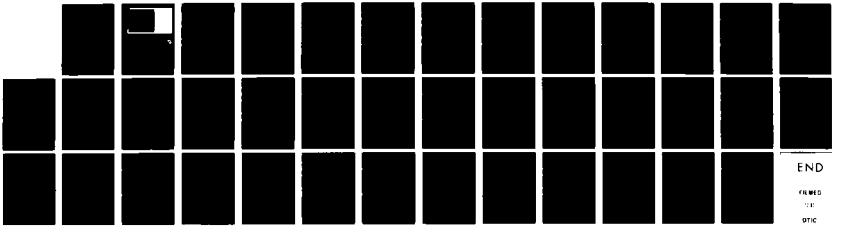
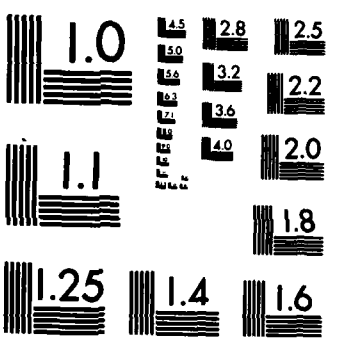
AD-A120 967 AN APOLOGY FOR ECUMENISM IN STATISTICS(U) WISCONSIN 1/1
UNIV-MADISON MATHEMATICS RESEARCH CENTER G E BOX
JUL 82 MRC-TSR-2408 DAAG29-80-C-0041

UNCLASSIFIED F/G 12/1 NL



END

FILMED
DTIC



MRC Technical Summary Report #2408

AN APOLOGY FOR ECUMENISM IN STATISTICS

G. E. P. Box

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

July 1982

(Received May 4, 1982)

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

AN APOLOGY FOR ECUMENISM IN STATISTICS

G. E. P. Box

Technical Summary Report #2408
July 1982

ABSTRACT

Reasons are advanced for the belief that scientific method employs and requires not one, but two kinds of inference - criticism and estimation; once this is understood the statistical advances made in recent years in Bayesian methods, data analysis, robust and shrinkage estimators can be seen as a cohesive whole.

AMS (MOS) Subject Classifications: 62A15, 62A20

Key Words: Bayes inference, Bayes sampling inference, frequentist inference, data analysis, shrinkage estimates, ridge estimates, robust estimates, right and left brain, theory-practice iteration, predictive distribution, binomial model, response surfaces, time series, prior distribution, likelihood principle, significance level, alphabetic optimality, residuals, goodness of fit, statistical criticism, statistical estimation, scientific investigation, iterative investigation, models, normal linear model, subjective probability, contaminated normal model.

Work Unit Number 4 (Statistics and Probability)

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

For many years there has existed a major controversy among statisticians concerning whether Bayesian theory or Sampling (frequentist) theory was appropriate for making statistical inferences. The roles of data analysis and of robust and shrinkage estimators have also been matters of dispute. Building on results from an earlier paper it is here argued that a study of scientific method and of the part played in it by the human brain shows that two different kinds of statistical inference - estimation and criticism are needed from which Bayes and Sampling theory respectively are uniquely appropriate. This point of view also shows how data analysis, robust and shrinkage estimators all have appropriate parts to play in the iterative scheme of scientific enquiry.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

AN APOLOGY FOR ECUMENISM IN STATISTICS

G. E. P. Box

Perhaps I should begin with an apology for my title. These days the statistician is often asked such questions as "Are you a Bayesian?" "Are you a frequentist?" "Are you a data analyst?" "Are you a designer of experiments?" I will argue that the appropriate answer to all these questions can be (and preferably should be) "yes", and that we can see why this is so if we consider the scientific context of what statisticians do.

For many years Statistics has seemed to be in a rather turbulent state and the air has been full of argument and controversy. The relative virtue of alternative methods of inference and, in particular, of Bayes' and Sampling (frequentist) inference has been hotly debated. Recently Data Analysis has rightly received much heavier emphasis, but its more avid proponents have sometimes seemed to suggest that all else is worthless. Furthermore while biased estimators, in particular shrinkage and ridge estimators, which have been advocated to replace the more standard varieties are clearly sensible in appropriate contexts their frequentist justification which ignores context seems unconvincing. Parallel criticism may be made of ad hoc robust procedures the proliferation of which has worried some dissidents who have argued for example that mechanical downweighting of peculiar observations may divert attention from important clues to new discovery.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

Insofar as these debates lead us to progressive change in our ideas they are healthy and productive, but insofar as they encourage polarization they may not be. One remembers with some misgivings Saxe's poem about the six blind men of Hindustan investigating an elephant. It will be recollected that one, feeling only the elephant's trunk, thought it like a snake, another, touching its ear, thought it must be a fan, etc. The poem ends:

And so these men of Hindustan
Disputed loud and long,
Each in his opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong.

Some of the difficulties arise from the need to simplify. But simplification included merely to produce satisfying mathematics or to reduce problems to convenient small sized pieces can produce misleading conclusions. Simplification which retains the essential scientific essence of the problem is most likely to lead to useful answers but this requires understanding of, and interest in, scientific context.

1. SOME QUESTIONABLE SIMPLIFICATIONS.

(a) It has been argued that Bayes' theorem uniquely solves all problems of inference. However only part of the inferential exercises in which the statistical scientist is ordinarily engaged seem to conveniently fit the Bayesian mold. In particular diagnostic checks of goodness of fit involving various analyses of residuals seem to require other justification. In fact I believe (Box [1980]) that the process of scientific investigation involves not one but two kinds of inference: estimation and criticism, used iteratively and in alternation. Bayes completely solves the problem of estimation and can also be helpful at the criticism stage in judging the relative plausibility of two or more models. However because of its necessarily conditional nature, it cannot deal with the most essential part of inferential criticism which requires a sampling (frequentist) justification.

(b) Fisher [1956] believed that the Neyman-Pearson theory for testing statistical hypotheses, while providing a model for industrial quality control and sampling inspection, did not of itself provide an appropriate basis for the conduct of scientific research. This can be regarded as the complement to the objection raised in (a), for statistical quality control and inspection are methods of inferential criticism supplying a continuous check on the adequacy of fit of the model for the properly operating process. I would regard Fisher's comment as meaning that the Neyman-Pearson theory was irrelevant to problems of estimation. Certainly there is evidence in the social sciences that excessive reliance upon this theory alone, encouraged by the mistaken prejudices of referees and editors, has led to harmful distortion of the conduct of scientific investigation in these fields.

(c) In some important contexts the scientific relevance of alphabetic optimality criteria (A,E,D,G etc.) in the choice of experimental designs has been questioned (see discussion of Kiefer [1959], also Box [1982]). Here again there is danger of deleterious feedback since users of statistical design, perhaps dazzled by impressive but poorly comprehended mathematics, may fail to realize the naive framework within which the optimality occurs.

(d) Even Data Analysis, excellent in itself, presents some dangers. It is a major step forward that in these days students of statistics are required more and more to work on real data. Indeed suitable "data sets" have been set aside for their study. But this too can produce misunderstanding. For instance, some examples have become notorious and have been analyzed by a plethora of experts; one finds three outliers, another claims that a transformation is needed and then only one outlier occurs, and so on. Too much exposure to this sort of thing can again lead to the mistaken idea that this represents the real context of scientific investigation. The statistician in his proper role as a member of a scientific team should certainly make such analyses, but realistically he would then discuss them with his scientific colleagues and

present, when appropriate, not one, but alternative plausible possibilities. He need not, and usually should not, choose among them. Rather he should make sure that these possibilities were considered when he and his scientific colleagues planned the next stage of the investigation. Together they would choose the next design so that among other things it could resolve current uncertainties judged to be important. In particular the possible meaning and importance of discrepant values would then be discussed as well as the meaning of analyses which downweighted or excluded them.

The most dangerous and misleading of the unstated assumptions suggested to some extent by all these simplifications concerns the implied static nature of the process of investigation: A Bayesian analysis is made; a hypothesis is tested; one model is considered; a single design is run; a single set of data is examined and reexamined[1].

I believe that the object of statistical theory should be to explain, at least approximately, what good scientists do and to help them do it better. It seems necessary therefore to examine at least briefly the nature of the scientific process itself.

2. SCIENTIFIC METHOD AND THE HUMAN BRAIN.

Scientific method is a formalization of the everyday process of finding things out. For thousands of years, things were found out largely as a result of chance occurrences. For a new "natural law" to be discovered, two

---

[1] While provision is made for adaptive feedback in data analysis, usually the possibility of acquiring further data to illuminate points at issue is not. What we do as statisticians depends heavily on expectations implied by our training. While a previous generation of graduates might have expected to prove theorems, occasionally to test an isolated hypothesis, and perhaps to teach a new generation of students to do likewise, the present generation might be forgiven for believing that their fate is only to explore "data sets" and speculate on what might or might not explain them. We must encourage our students to accept the heritage bestowed by Fisher, who elevated the statistician from an archivist to an active designer of experiments and hence an architect and coequal investigator.

circumstances needed to coincide: (a) a potentially informative experience needed to occur, and (b) the phenomenon needed to be known about by someone of sufficient acuity of mind to formulate, and preferably to test, a possible rule for its future occurrence.

Progress was slow because of the rarity of the two necessary individual circumstances and the still greater rarity of their coincidence. Experimental science accelerates this learning process by isolating its essence: potentially informative experiences are deliberately staged and made to occur in the presence of a trained investigator. As science has developed, we have learned how such artificial experiences may be carefully contrived to isolate questions of interest, how conjectures that are put forward may be tested, and how residual differences from what had been expected can be used to modify and improve initial ideas. So the ordinary process of learning has been sharpened and accelerated.

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. To appreciate why this is so, consider how helpless we would be if, each night, all our memories were eliminated, so that we awoke to each new day with no past experiences whatever and hence no models to guide our conduct. In fact, our past experience is conveniently accumulated in models $M_1, M_2, \ldots, M_i, \ldots$. Some of these models are well established, others less so, while still others are in the very early stages of creation. When some new fact or body of facts y_d comes to our attention, the mind tries to associate this 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 in train 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 able

to deduce what facts could be expected as realizations of a particular model and, more difficult, to induce what model(s) are consonant with particular facts.

Thus, it is concerned with two kinds of inference: (A) Contrasting of new facts y_d with a possible model M: an operation I will characterize by subtraction $M - y_d$. This process stimulates induction and will be called criticism. (B) Incorporating new facts y_d into an appropriate model: an operation I will characterize by addition $M + y_d$. This process is entirely deductive and will be called estimation.

I believe then that many of our difficulties arise because, while there is an essential need for two kinds of inference, there seems an inherent propensity among statisticians to seek for only one.

In any case, research which, following the discoveries of Roger Sperry and his associates, has gathered great momentum in the past 25 years shows that the human brain behaves not as a single entity but as two largely separate but cooperating instruments which do two different things (see for example Springer and Deutsch [1981], Blackeslee [1980]).

In most people[2], the left brain is concerned primarily with language and logical deduction, the right brain with images, patterns and inductive processes. The two sides of the brain are joined by millions of connections in the corpus callosum. It is known that the left brain plays a conscious and dominant role while by contrast one may be quite unaware[3] of the working of the right brain.

---

[2] In about one third of left-handed people (about 5% of the population) the roles of the right and left brain are reversed.

[3] For example the apparently instinctive knowledge of what to do and how to do it, enjoyed by the experienced tennis player and by the experienced motorist, comes from the right brain.

The right brain's ability to appreciate[4] patterns in data y_d and to find patterns in discrepancies $M_i - y_d$ between the data and what might be expected if some tentative model were true is of great importance in the search for explanations of data and of discrepant events. This accomplishment of the right brain of pattern recognition is of course of enormous consequence in scientific discovery[5]. However, some check is 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[6]. This is the object of diagnostic checks and tests of fit which, I will argue, require frequentist theory significance tests for their formal justification.

3. THE THEORY - PRACTICE ITERATION.

It has long been recognized that the learning process 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 of an investigation is thus evidenced by a theoretical model, which is not static, but by appropriate exposure to reality continually evolves until

---

[4]Implicit recognition of the need to stimulate the remarkable pattern-seeking ability of the right brain is evidenced by modern emphasis on ingenious plotting devices in the model formulation/modification phases of investigation. In particular Chernoff's representation of multivariate data by faces [1973] and earlier Edgar Anderson's use of glyphs [1960] direct the right brain to the recognition problem at which it excels.

[5]Manifestations of the importance to discovery of unconscious pattern seeking by the right brain have often been noticed. For example, Beveridge [1950] remarks that happenings of the following kind are commonplace: a scientist has mulled over a set of data for many months and then, at a certain point in time, perhaps on a country walk when the problem is not being consciously thought about, he suddenly becomes aware of a solution (model) which explains these data. This point in time is presumably that at which the right brain sees fit to let the left brain know what it has figured out.

[6]See, for example, the King of Heart's rationalization of the poem brought as evidence in the trial of the Knave of Hearts in Lewis Carrol's Alice in Wonderland.

some currently satisfactory level of understanding is reached. At any given stage in a scientific investigation the current model helps us to appreciate not only what we know, but what else it may yet be important to find out and so motivates the collection of new data to illuminate dark but possibly interesting corners of present knowledge. See for example Box and Youle [1955], Box [1976], Box, Hunter and Hunter [1978].

The reader can find illustration of these matters in his 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.

Different levels of adaptation

The adaptive iteration we have described produces change in what we believe about the system being studied, but it can also produce change in how we study it, and sometimes even in the objective[7] of the study. This multiple adaptivity explains the surprising property of convergence of a process of investigation which at first appears hopelessly arbitrary. See for example Box [1957]. To appreciate this arbitrariness, suppose that some scientific problem were being studied by, say, 10 independent sets of investigators, all competent in the field of endeavor. It is certain that they would start from different points, conduct the investigation in different ways, have different initial ideas about which variables were important, on what scales,and in which transformation. Yet it is perfectly possible that they would all eventually reach similar conclusions. It is important to bear this context of multiple iteration in mind

---

[7] If we start out to prospect for silver, we should not ignore an accidental discovery of gold. For example, one experimental attempt to find manufacturing conditions giving greater yield of a particular product failed to find any such, but did find reaction conditions giving the same yield with the reaction time halved. This meant that, by switching to the new manufacturing conditions, throughput could be doubled, and that a costly, previously planned, extension of the plant was unnecessary.

when we consider the scientific process and how it relates to a statistical method.

4. STATISTICAL ESTIMATION AND CRITICISM.

In a recent paper (Box [1980]) a statistical theory was presented which, it was argued, was consonant with the view of scientific investigation outlined above. Suppose at the i^{th} stage of such an investigation a set of assumptions A_i are tentatively entertained which postulate that to an adequate approximation, the density function for potential data $\underset{\sim}{y}$ is $p(\underset{\sim}{y}|\underset{\sim}{\theta},A_i)$ and the prior distribution for $\underset{\sim}{\theta}$ is $p(\underset{\sim}{\theta}|A_i)$. Then it was argued that the model M_i should be defined as the joint distribution of $\underset{\sim}{y}$ and $\underset{\sim}{\theta}$

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|A_i) = p(\underset{\sim}{y}|\underset{\sim}{\theta},A_i)p(\underset{\sim}{\theta}|A_i) \tag{1}$$

since it is a complete statement of prior tentative belief at stage i. In these expressions A_i is understood to indicate all or some of the assumptions in the model specification at stage i. The model of equation (1) means to me that current belief about the outcome of contemplated data acquisition would be calibrated with adequate approximation by a physical simulation involving appropriate random sampling from the distributions $p(\underset{\sim}{y}|\underset{\sim}{\theta},A_i)$ and $p(\underset{\sim}{\theta}|A_i)$.

The model can also be factored as

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|A) = p(\underset{\sim}{\theta}|\underset{\sim}{y},A)p(\underset{\sim}{y}|A) \ . \tag{2}$$

The second factor on the right, which can be computed before any data become available,

$$p(\underset{\sim}{y}|A) = \int p(\underset{\sim}{y}|\underset{\sim}{\theta},A)p(\underset{\sim}{\theta}|A)d\underset{\sim}{\theta} \tag{3}$$

is the predictive distribution of the totality of all possible samples $\underset{\sim}{y}$ that could occur if the assumptions were true.

When an actual data vector $\underset{\sim}{y}_d$ becomes available

$$p(\underset{\sim}{y}_d,\underset{\sim}{\theta}|A) = p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A)p(\underset{\sim}{y}_d|A) \ . \tag{4}$$

The first factor on the right is the Bayes' posterior distribution of $\underset{\sim}{\theta}$ given $\underset{\sim}{y}_d$

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A) \propto p(\underset{\sim}{y}_d|\underset{\sim}{\theta},A)p(\underset{\sim}{\theta}|A) \tag{5}$$

while the second factor

$$p(\underset{\sim}{y}_d|A) = \int p(\underset{\sim}{y}_d|\underset{\sim}{\theta},A)p(\underset{\sim}{\theta}|A)d\underset{\sim}{\theta} , \tag{6}$$

is the predictive density associated with the particular data $\underset{\sim}{y}_d$ actually obtained conditional on the truth of the model and on the data $\underset{\sim}{y}_d$ having occurred.

The posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A)$ allows all relevant estimation inferences to be made about $\underset{\sim}{\theta}$, but this posterior distribution can supply no information about the adequacy of the model. Information on adequacy may be provided, however, by reference of the density $p(\underset{\sim}{y}_d|A)$ to the predictive reference distribution $p(\underset{\sim}{y}|A)$ or of the density $p\{g_i(\underset{\sim}{y}_d)|A\}$ of some relevant checking function $g_i(\underset{\sim}{y}_d)$ to its predictive distribution and in particular by computing the probabilities

$$\Pr\{p(\underset{\sim}{y}|A) < p(\underset{\sim}{y}_d|A)\} \tag{7}$$

and

$$\Pr[p\{g_i(\underset{\sim}{y})|A\} < p\{g_i(\underset{\sim}{y}_d)|A\}] \tag{8}$$

Two illustrative examples follow.

4.1. The Binomial Model

As an elementary example, suppose inferences are to be made about the proportion θ of successes in a set of binomial trials.

Suppose n trials are about to be made and assume a beta-distribution prior with mean θ_0. Then

$$p(\theta|A) = [B\{(m\theta_0, m(1-\theta_0)\}]^{-1}\theta^{m\theta_0-1}(1-\theta)^{m(1-\theta_0)-1} \tag{9}$$

$$p(y|\theta,A) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \tag{10}$$

and the predictive distribution is

$$p(y|A)=\binom{n}{y}[B\{m\theta_0, m(1-\theta_0)\}]^{-1}B\{m\theta_0+y, m(1-\theta_0)+n-y\} \tag{11}$$

which may be computed before the data are obtained.

If, now, having performed n trials, there are y_d successes, the likelihood defined up to a multiplicative constant is

$$L(\theta|y_d,A) = \theta^{y_d}(1-\theta)^{n-y_d} \tag{12}$$

the predictive density is

$$p(y_d|A)=\binom{n}{y_d}[B\{m\theta_0,m(1-\theta_0)\}]^{-1}B\{m\theta_0+y_d,m(1-\theta_0)+n-y_d\} \tag{13}$$

and the posterior distribution of θ is

$$p(\theta|y_d,A) = [B\{m\theta_0 + y_d,m(1-\theta_0) + n - y_d\}]^{-1} \times \theta^{y_d+m\theta_0-1}(1-\theta)^{n-y_d+m(1-\theta_0)-1} . \tag{14}$$

In the examples of Figures 1 and 2 full lines are used for items available <u>prior</u> to the availability of data y_d and dotted lines for items available only after the data y_d are in hand. Both Figures 1 and 2 illustrate a situation where the prior distribution $p(\theta|A)$ has mean $\theta_0 = 0.2$ and $m = 20$ and we know that $n = 10$ trials are to be performed. Knowing these facts, we can immediately calculate the predictive distribution $p(y|A)$ which is the probability distribution for all possible outcomes from such a model if we suppose the model is true.

When the experiment is actually performed suppose at first, as in Figure 1, that $y_d = 3$ of the trials are successes. The predictive probability $p(3|A)$ associated with this outcome is not unusually small. In fact $\Pr\{p(y|A) \leq p(3|A)\} = 0.42$ and we have no reason to question the model. Thus for this sample the likelihood $L(\theta|y)$ may reasonably be combined with the prior to produce the posterior distribution shown.

In Figure 2 however it is supposed instead that the outcome is $y_d = 8$ successes so that for this sample $\Pr\{p(y|A) \leq p(8|A)\} = 0.0013$ and the adequacy of the model, and in particular the adequacy of the prior distribution, is now called into question. Inspection of the figure shows how this agrees with common sense; for in the case illustrated the posterior distribution is unlike either the prior distribution or the likelihood which were combined to obtain it.

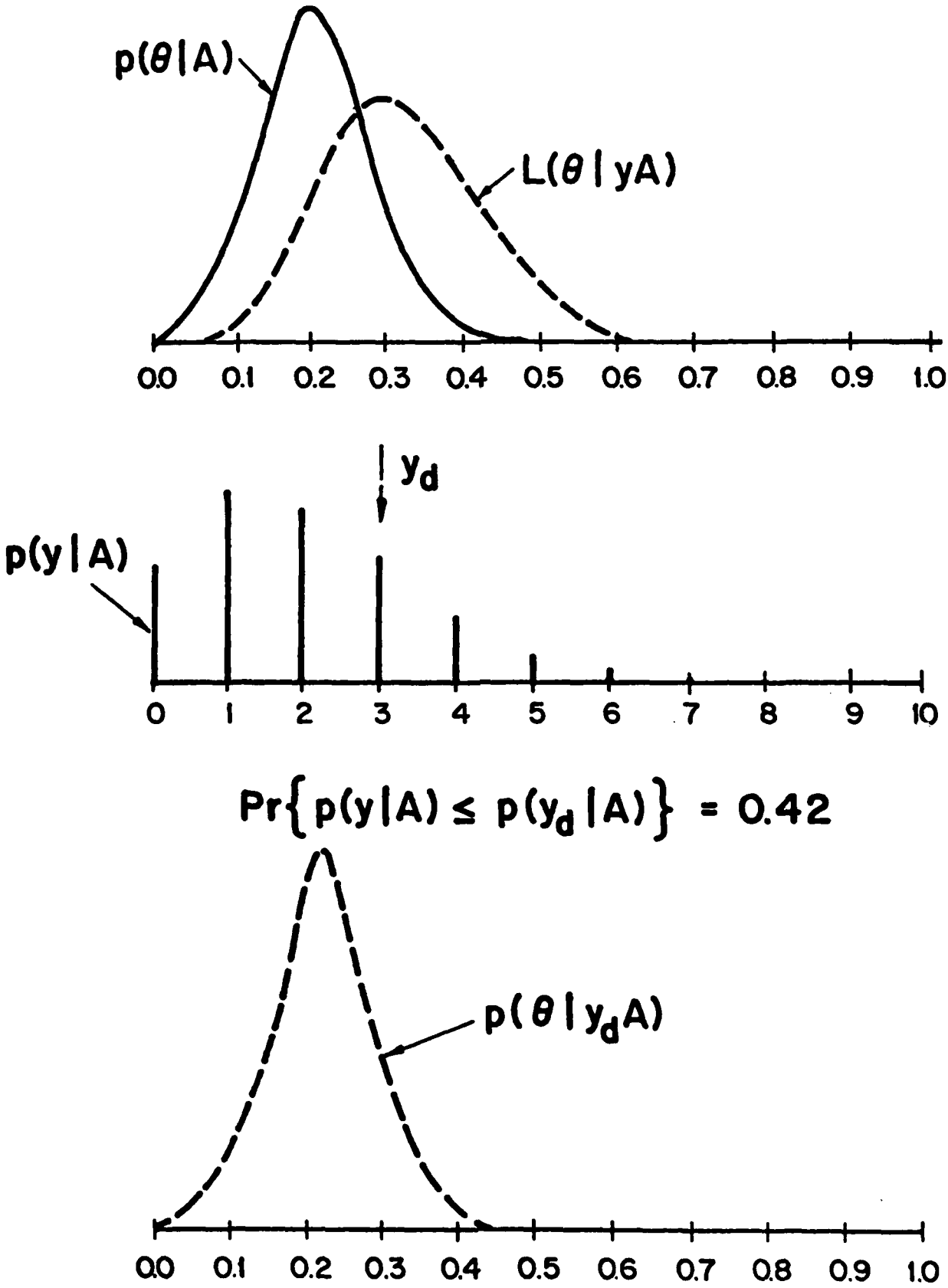
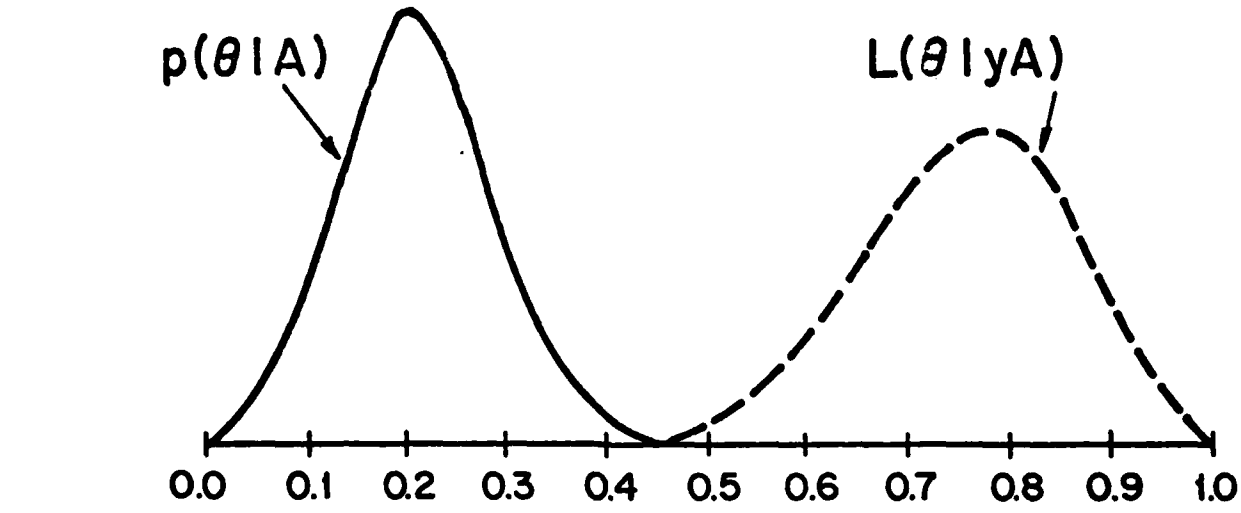
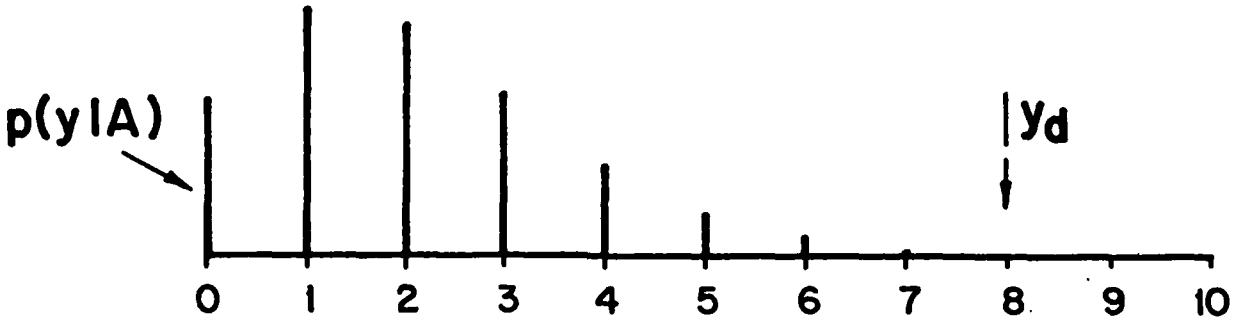


Figure 1. Prior, likelihood, predictive and posterior distributions for n = 10 Bernoulli trials with y_d = 3 successes.





$$\Pr\{p(y|A) \leq p(y_d|A)\} = 0.0013$$

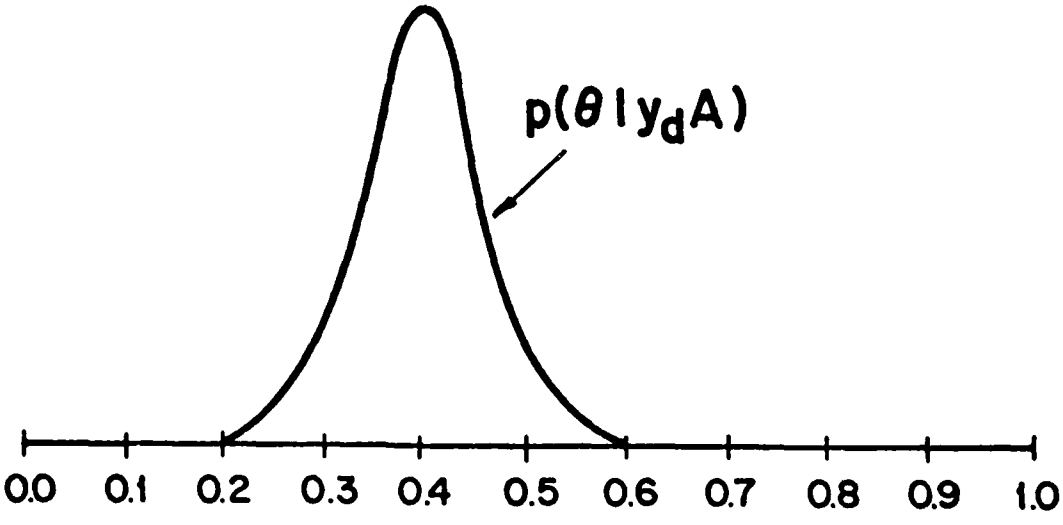


Figure 2. Prior, likelihood, predictive and posterior distributions for n = 10 Bernoulli trials with y_d = 8 successes.

Misgivings about the use of Bayes' theorem which some have expressed in the past are certainly associated with the possibility of distorting the information coming from the data by the use of an inappropriate prior distribution. Without predictive checks, the following objections would carry great weight:

(a) that nothing in the Bayes' calculation of the posterior distribution itself could warn of the incompatibility of the data and the model, and especially the prior; and

(b) that in complicated examples it would not be so obvious when this incompatibility occurred.

A case of particular interest occurs when the prior is sharply centered[8] at its mean value $\theta_0 = 0.2$. This happens in the above binomial setup when m is made very large. Then, <u>if</u> the model is unquestioned, the posterior distribution will be essentially the same as the prior leading to the conclusion that θ is close to θ_0 whatever the data. The predictive distribution in this case is $p(y|\theta_0,A)$, the ordinary binomial sampling distribution, and the predictive check is the standard binomial significance test, which can discredit the model with $\theta = \theta_0 = 0.2$ and hence discredit the application of Bayes' theorem to this case. This, to my mind, produces the most satisfactory justification for the standard significance test.

4.2. <u>The Normal Linear Model and Ridge Estimators</u>

Another example, discussed in Box [1980], concerns the normal linear model. In a familiar notation suppose

$$\underset{\sim}{y} \sim N(\underset{\sim}{1}\mu + \underset{\sim}{X}\underset{\sim}{\theta}, \underset{\sim}{I}_n\sigma^2) \qquad (15)$$

with $\underset{\sim}{1}$ a vector of unities and $\underset{\sim}{X}$ of full rank k and such that $\underset{\sim}{X}'\underset{\sim}{1} = \underset{\sim}{0}$ and suppose that prior densities are locally approximated by

---

[8] Such a model with a prior sharply centered at $\theta_0 = 0.2$ might be appropriate, for instance, if a trial consisted of spinning ten times what seemed to be a properly balanced pentagonal top and counting the number of times the top fell on a particular segment.

$$\mu \sim N(\mu_0, c^{-1}\sigma^2), \underset{\sim}{\theta} \sim N(\underset{\sim}{\theta}_0, \underset{\sim}{\Gamma}^{-1}\sigma^2), \{\sigma^2/\nu_0 s_0^2\} \sim \chi^{-2}(\nu_0) \qquad (16)$$

where $\chi^{-2}(\nu_0)$ refers to the inverted χ^2 distribtuion with ν_0 degrees of freedom and μ and $\underset{\sim}{\theta}$ independent conditional on σ^2.

Given a sample $\underset{\sim}{y}_d$, special interest attaches to $\underset{\sim}{\theta}$ and σ^2 which, given the assumptions, are estimated by $p(\underset{\sim}{\theta}, \sigma^2 | \underset{\sim}{y}_d, A)$ with marginal distributions

$$p(\underset{\sim}{\theta} | \underset{\sim}{y}_d, A) \propto \left\{1 + \frac{(\underset{\sim}{\theta} - \bar{\underset{\sim}{\theta}}_d)'(\underset{\sim}{X}'\underset{\sim}{X} + \underset{\sim}{\Gamma})(\underset{\sim}{\theta} - \bar{\underset{\sim}{\theta}}_d)}{(n + \nu_0)\hat{\sigma}_d^2}\right\}^{-\frac{1}{2}(n+\nu_0+k)} \qquad (17)$$

$$p(\sigma^2 | \underset{\sim}{y}_d, A) \propto \sigma^{-(n+\nu_0+2)} \exp\{-\tfrac{1}{2}(n + \nu_0)\hat{\sigma}_d^2/\sigma^2\} \qquad (18)$$

with

$$\bar{\underset{\sim}{\theta}}_d = (\underset{\sim}{X}'\underset{\sim}{X} + \underset{\sim}{\Gamma})^{-1}(\underset{\sim}{X}'\underset{\sim}{X}\hat{\underset{\sim}{\theta}}_d + \underset{\sim}{\Gamma}\underset{\sim}{\theta}_0) ,$$

$$\hat{\underset{\sim}{\theta}}_d = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y}_d, \quad \nu = n - k - 1 , \qquad (19)$$

$$(n+\nu_0)\hat{\sigma}_d^2 = \nu s_d^2 + \nu_0 s_0^2 + (\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)'\{(\underset{\sim}{X}'\underset{\sim}{X})^{-1} + \underset{\sim}{\Gamma}^{-1}\}^{-1}(\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)$$

$$+ (n^{-1} + c^{-1})^{-1}(\bar{y} - \mu_0)^2 .$$

and

$$s^2 = \{\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\}\underset{\sim}{y}, \quad s_d^2 = \{\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\}\underset{\sim}{y}_d \qquad (20)$$

Now let

$$s_p^2 = (\nu + \nu_0)^{-1}(\nu s^2 + \nu_0 s_0^2) \quad \text{and}$$

$$s_{pd}^2 = (\nu + \nu_0)^{-1}(\nu s_d^2 + \nu_0 s_0^2) . \qquad (21)$$

Then the joint predictive distribution can be factored into independent components for $(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}_0)/s_p$, s^2, and $\nu - 1$ angular elements of the standarized residuals. A predictive check based on the first of these factors

$$\Pr\{p((\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}_0)/s_p | A) < p((\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)/s_{pd} | A)\}$$

$$= \Pr\left\{F_{k,\nu+\nu_0} > \frac{(\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)'\{(\underset{\sim}{X}'\underset{\sim}{X})^{-1} + \underset{\sim}{\Gamma}^{-1}\}^{-1}(\hat{\underset{\sim}{\theta}}_d - \underset{\sim}{\theta}_0)}{k s_{pd}^2}\right\} \qquad (22)$$

is the standard analysis of variance check for compatibility of two estimates $\hat{\underset{\sim}{\theta}}_d$ and $\underset{\sim}{\theta}_0$ and was earlier proposed as a check for compatibility of prior and sample information by Theil [1963].

Now suppose the $\underset{\sim}{X}$ matrix to be in correlation form and assume $\underset{\sim}{\theta}_0 = \underset{\sim}{0}, \underset{\sim}{\Gamma} = \underset{\sim}{I}_k \gamma_0, \nu_0 \to 0$ so that $s_p^2 \to s^2$. Then the estimates $\bar{\underset{\sim}{\theta}}_d$ are the ridge estimators of Hoerl and Kennard [1970] which, given the assumptions, appropriately combine information from the prior with information from the data. The predictive check (22) now yields

$$\alpha = \Pr\{F_{k,\nu} > \frac{\hat{\underset{\sim}{\theta}}_d' \{(\underset{\sim}{X}'\underset{\sim}{X})^{-1} + \underset{\sim}{I}\gamma_0^{-1}\}^{-1} \hat{\underset{\sim}{\theta}}_d}{k s_d^2}\} \qquad (23)$$

allowing any choice of γ_0 to be criticized.

For example, in their original analysis of the data of Gorman and Toman [1966], Hoerl and Kennard [1970] chose a value $\gamma_0 = 0.25$. But substitution of this value in (23) yields $\alpha = \Pr\{F_{10,25} > 3.59\} < 0.01$ which discredits this choice.

One can see for these examples how the two functions of criticism and estimation are performed by the predictive check on the one hand and the Bayesian posterior distribution on the other.

Thus consider the ridge (Bayes' mean) estimator of the second example. This estimator is a linear combination of the least squares estimate $\hat{\underset{\sim}{\theta}}$ and the prior mean $\underset{\sim}{\theta}_0$, with weights supplied by the appropriate information matrices, and with covariance matrix obtained by inverting the sum of these information matrices. Assuming the data to be a realization of the model, this is the appropriate way of combining the two sources of information.

The predictive check, on the other hand, contrasts the values $\hat{\underset{\sim}{\theta}}$ and $\underset{\sim}{\theta}_0$ with a dispersion matrix obtained by appropriately summing the two dispersion matrices.

The combination of information from the prior and likelihood into the posterior distribution and the contrasting of these two sources of information in the predictive distribution is equally clear in the binomial

example and especially in its appropriate normal approximation.

5. SOME OBJECTIONS CONSIDERED

A recapitulation of the argument and a consideration of some objections is considered in this section.

5.1. Essential elements of the argument

A. Scientific investigation is an iterative process in which the model is not static but is continually evolving. At a given stage the nature of the uncertainties in a model directs the acquisition of further data, whether by choosing the design of an experiment or sample survey, or by motivating a search of a library or data bank. At, say, the i^{th} stage of an investigation all current structural assumptions A_i, including those about the prior, must be thought of, not as being true, but rather as being subjective guesses which at this particular stage of the investigation are worth entertaining. It is consistent with this attitude that when data $\underset{\sim}{y}_d$ become available checks need to be applied to assess consonance with A_i.

B. The statistical model at the i^{th} stage of the investigation should be defined as the joint distribution of $\underset{\sim}{y}$ and $\underset{\sim}{\theta}$ given the assumptions A_i

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|A_i) = p(\underset{\sim}{y}|\underset{\sim}{\theta},A_i)p(\underset{\sim}{\theta}|A_i) \ . \tag{24}$$

C. Not one but two distinct kinds of inference are involved within the iterative process: criticism in which the appropriateness of regarding data $\underset{\sim}{y}_d$ as a realization of a particular model M is questioned; estimation in which the consequences of the assumption that data $\underset{\sim}{y}_d$ are a realization of a model M are made manifest.

This criticism-estimation dichotomy is characterized mathematically by the factorization of the model realization $p(\underset{\sim}{y}_d,\underset{\sim}{\theta}|A_i)$ into the predictive density $p(\underset{\sim}{y}_d|A_i)$ and the posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A_i)$. The predictive distribution $p(\underset{\sim}{y}|A_i)$ provides a reference distribution for $p(\underset{\sim}{y}_d|A_i)$. Similarly the predictive distribution $p\{g(\underset{\sim}{y})|A_i\}$ of any checking function $g(\underset{\sim}{y})$ provides a reference distribution of the corresponding predictive

density $p\{g(\underset{\sim}{y}_d|A_i\}$. Unusually small values of this density suggest that the current model is open to question.

D. If we are satisfied with the adequacy of the assumptions A_i then the posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,A_i)$ allows for complete estimation of $\underset{\sim}{\theta}$ and no other procedures of <u>estimation</u> are relevant. In particular, therefore, insofar as shrinkage, ridge and robust estimators are useful, they ought to be direct consequences of an appropriate model and should not need the invocation of extraneous considerations such as minimization of mean square error.

<u>Objections</u>. Numbered to correspond with the various elements of the argument are responses to some objections that have been, or might be, raised.

A(i) <u>Iterative investigation</u>? Some would protest that their own statistical experience is not with iterative investigation but with a single set of data to be analyzed, or a single design to be laid out and the results elucidated.

Many circumstances where the statistican has been involved in a "one-shot" analysis rather than an iterative partnership, ought not to have happened. Such involvement frequently occurs when the statistician has been drafted as a last resort, all other attempts to make sense of the data having failed. At this point data gathering will usually have been completed and there is no chance of influencing the course of the study. Statisticians whose training has not exposed them to the overriding importance of experimental design are most likely to acquiesce in this situation, or even to think of it as normal, and thus to encourage its continuance.

The statistician who has cooperated in the design of a <u>single</u> experiment which he analyzes is somewhat better off. However one-shot designs are often inappropriate also. Underlying most investigations is a budget, stated or unstated, of time and/or money that can reasonably be expended. Sometimes this latent budget is not adequate to the goal of the investigation, but, for purposes of discussion, let us suppose that it is. Then if a

sequential/iterative approach is possible it would usually be quite inappropriate to plan the whole investigation at the beginning in one large design. This is because the results from a first design will almost invariably supply new and often unexpected information about choice of variables, metrics, transformations, regions of operability, unexpected side-effects, and so forth, which will vitally influence the course of the investigation and the nature of the next experimental arrangement. A rough working rule is that not more than 25% of the time-and-money budget should be spent on the first design. Because large designs can in a limited theoretical sense be more efficient it is a common mistake not to take advantage of the iterative option when it is available. Instances have occurred of experimenters regretting that they were persuaded by an inexperienced statistician to perform a large "all inclusive" design where an adaptive strategy would have been much better. In particular, it is likely that many of the runs from such "all-embracing" designs, will turn out to be noninformative because their structure was decided when least was known about the problem.

Scientific iteration is strikingly exemplified in response surface studies (see, for example, Box and Wilson [1951], Box [1954], Box and Youle [1955]). In particular methods such as steepest ascent and canonical analysis can lead to exploration of new regions of the experimental space, requiring elucidation by new designs which, in turn, can lead to the use of models of higher levels of sophistication. Although in these examples the necessity for such an iterative theory is most obvious, it clearly exists much more generally, for example in investigations employing sequences of orthodox experimental designs and to many applications of regression analysis. It has sometimes been suggested that agricultural field trials are not sequential but of course this is not so; only the time frame is longer. Obviously what is learned from one year's work is used to design the next year's experiments.

However I agree that there are some more convincing exceptions. For example, a definitive trial which is

intended to settle a controversy such as a test of the effectiveness of Laetrile as a cure for cancer. Also the iteration can be very slow. For example, in trials on the weathering of paints, each phase can take from 5-10 years.

A(ii) Subjective probability? The view of the process of scientific investigation as one of model evolution has consequences concerning subjective probabilities. An objection to a subjectivist position is that in presenting the final results of our investigation, we need to convince the outside world that we have really reached the conclusion that we say we have. It is argued that, for this purpose, subjective probabilities are useless. However I believe that the confirmatory stage of an iterative investigation, when it is to be demonstrated that the final destination reached is where it is claimed to be, will typically occupy, perhaps, only the last 5 per cent of the experimental effort. The other 95 per cent - the wandering journey that has finally led to that destination - involves, as I have said, many heroic subjective choices (what variables? what levels? which scales? etc., etc.) at every stage. Since there is no way to avoid these subjective choices which are a major determinant of success why should we fuss over subjective probability?

Of course, the last 5 per cent of the investigation occurs when most of the problems have been cleared up and we know most about the model. It is this rather minor part of the process of investigation that has been emphasized by hypothesis testers and decision theorists. The resultant magnification of the importance of formal hypothesis tests has inadvertently led to underestimation by scientists of the area in which statistical methods can be of value and to a wide misunderstanding of their purpose. This is often evidenced in particular by the attitudes to statistics of editors and referees of journals in the social, medical and biological sciences.

B(i) The Statistical Model? The statistical model has sometimes been thought of as the density function $p(\underset{\sim}{y}|\underset{\sim}{\theta},A)$ rather than the joint density $p(\underset{\sim}{y},\underset{\sim}{\theta}|A)$ which reflects the

influence of the prior. However only the latter form contains all currently entertained beliefs about $\underset{\sim}{y}$ and $\underset{\sim}{\theta}$. It seems quite impossible to separate prior belief from assumptions about model structure. This is evidenced by the fact that assumptions are frequently interchangeable between the density $p(\underset{\sim}{y}|\underset{\sim}{\theta})$ and the prior $p(\underset{\sim}{\theta})$. As an elementary example, suppose that among the parameters $\underset{\sim}{\theta} = (\underset{\sim}{\phi},\beta)$ of a class of distributions β is a shape parameter such that $p(\underset{\sim}{y}|\underset{\sim}{\phi},\beta_0)$ is the normal density. Then it may be convenient, for example in studies of robustness, to define a normal distribution by writing the more general density $p(\underset{\sim}{y}|\underset{\sim}{\phi},\beta)$ with an associated prior for β which can be concentrated at $\beta = \beta_0$. The element specifying normality which in the usual formulation is contained in the density $p(\underset{\sim}{y},\underset{\sim}{\theta})$ is thus transferred to the prior $p(\underset{\sim}{\theta})$.

B(ii) Do we need a prior? Another objection to the proposed formulation of the model is the standard protest of non-Bayesians concerning the introduction of any prior distribution as an unnecessary and arbitrary element. However, recent history has shown that it is the omission in sampling theory, rather than the inclusion in Bayesian analysis, of an appropriate prior distribution, that leads to trouble.

For instance Stein's result [1955] concerning the inadmissibility of the vector of sample averages as an estimate of the mean of a multivariate normal distribution is well known. But consider its practical implication for, say, an experiment resulting in a one-way analysis of variance. Such an experiment could make sense when it is conducted to compare, for example, the levels of infestation of k different varieties of wheat, or the numbers of eggs laid by k different breeds of chickens or the yields of k successive batches of chemical; in general, that is, when a priori we expect similarities of one kind or another between the entities compared. But clearly, if similarities are in mind, they ought not to be denied by the form of the model. They are so denied by the improper prior which produces as Bayesian means the sample averages, which are in turn the orthodox estimates from sampling theory.

Now the reason that k wheat varieties, k chicken breeds or k batch yields are being jointly considered is because they are, in one sense or another, comparable. The presence of a specific form of prior distribution allows the investigator to incorporate in the model precisely the kind of similarities he wishes to entertain. Thus in the comparison of varieties of wheat or of breeds of chicken it might well be appropriate to consider the variety means as randomly sampled from some prior super-population and, as is well known, this can produce the standard shrinkage estimators as Bayesian means (Lindley [1965], Box and Tiao [1968], Lindley and Smith [1972]). But notice that such a model is likely to be quite inappropriate for the yields of k successive batches of chemical. These mean yields might much more reasonably be regarded as a sequence from some autocorrelated time series. A prior which reflected this concept led Tiao and Ali [1971] to functions for the Bayesian means which are quite different from the orthodox shrinkage estimators.

In summary, then, both sampling theory and Bayes theory can rationalize the use of shrinkage estimators, and the fact that the former does so merely on the basis of reduction of mean square error with no overt use of a prior distribution, at first seems an advantage. However, only the explicit inclusion of a prior distribution, which sensibly describes the situation we wish to entertain, can tell us what is the appropriate function to consider, and avoid the manifest absurdities which seem inherent in the sampling theory approach which implies, for example, that we can improve estimates by considering as one group varieties of wheat, breeds of chicken, and batches of chemical.

C(i) Is there an iterative interplay between criticism and estimation? A good example of the iterative interplay between criticism and estimation is seen in parametric time series model building as described for example by Box and Jenkins [1970]. Critical inspection of the plotted time series and of the corresponding plotted autocorrelation function, and other functions derivable from it, together with their rough limits of error, can suggest a

model specification and in particular a parametric model. Temporarily behaving as if we believed this specification, we may now estimate the parameters of the time series model by their Bayesian posterior distribution (which, for samples of the size usually employed, is sufficiently well indicated by the likelihood). The residuals from the fitted model are now similarly critically examined, which can lead to respecification of the model, and so on. Systematic liquidation of serial dependence brought about by such an iteration can eventually produce a parametric time series model; that is a linear filter which approximately transforms the time series to a white noise series. Anyone who carries through this process must be aware of the very different nature of the two inferential processes of criticism and estimation which are used in alternation in each iterative cycle.

C(ii) Why can't all criticism be done using Bayes posterior analysis?

It is sometimes argued that model checking can always be performed as follows: let $A_1, A_2, \ldots, A_k$ be alternative assumptions; then the computation of

$$p(A_i|\underline{y}) = \frac{p(\underline{y}|A_i)p(A_i)}{\sum_{j=1}^{k} p(\underline{y}|A_j)p(A_j)} \qquad (i = 1,2,\ldots,k) \qquad (25)$$

yields the probabilities for the various sets of assumptions.

The difficulty with this approach is that by supposing all possible sets of assumptions known a priori it discredits the possibility of new discovery. But new discovery is, after all, the most important object of the scientific process.

At first, it might be thought that the use of (25) is not misleading, since it correctly assesses the relative plausibility of the models considered. But in practice this would seem of little comfort. For example suppose that only $k = 3$ models are currently regarded as possible, and that having collected some data the posterior probabilities $p(A_i|\underline{y})$ are 0.001, 0.001, 0.998 $(i = 1,2,3)$. Although

in relation to these particular alternatives $p(A_3|\underline{y})$ is overwhelmingly large this does not necessarily imply that in the real world assumptions A_3 could be safely adopted. For, suppose unknown to the investigator, a fourth possibility A_4 exists which given the data is a thousand times more probable than the group of assumptions previously considered. Then, if that model had been included, the probabilities would be 0.000,001, 0.000,001, 0.000,998, and 0.999,000.

Furthermore, in ignorance of A_4 it is highly likely that a study of the components of the predictive distribution $p(\underline{y}|A_3)$ and in particular of the residuals, could (a) have shown that A_3 was not acceptable and (b) have provided clues as to the identity of A_4. The objective of good science must be to conjure into existence what has not been contemplated previously. A Bayesian theory which excludes this possibility subverts the principle aim of scientific investigation.

More generally, the possibility that there are more than one set of assumptions that may be considered, merely extends the definition of the model to

$$p(\underline{y},\underline{\theta},A_j) = p(\underline{y}|\underline{\theta}A_j)p(\underline{\theta}|A_j)p(A_j) \qquad (j = 1,2,\ldots,k)$$

which in turn will yield a predictive distribution. In a situation when this more general model is inadequate a mechanical use of Bayes theorem could produce a misleading analysis, while suitable inspection of predictive checks could have demonstrated, on a sampling theory argument, that the global model was almost certainly wrong and could have indicated possible remedies.[9]

C(iii) An abrogation of the likelihood principle? The likelihood principle holds, of course, for the estimation aspect of inference in which the model is temporarily assumed true. However it is inapplicable to the criticism process in which the model is regarded as in doubt.

---

[9] I am grateful to Dr. Michael Titterington for pointing out that in discriminant analysis the atypicality indices of Aitchison and Aitken [1976] use similar ideas.

If the assumptions A are supposed true, the likelihood function contains all the information about $\underset{\sim}{\theta}$ coming from the particular observed data vector $\underset{\sim}{y}_d$. When combined with the prior distribution for $\underset{\sim}{\theta}$ it therefore tells all we can know about $\underset{\sim}{\theta}$ given $\underset{\sim}{y}_d$ and A. In such a case the predictive density $p(\underset{\sim}{y}_d|A)$ can tell us nothing we have not already assumed to be true, and will fall within a given interval with precisely the frequency forecast by the predictive distribution. When the assumptions are regarded as possibly false, however, this will no longer be true and information about model inadequacy can be supplied by considering the density $p(\underset{\sim}{y}_d|A)$ in relation to $p(\underset{\sim}{y}|A)$. Thus for the Normal linear model, the distribution of residuals contains no information if the model is true, but provides the reference against which standard residual checks, graphical and otherwise, are made on the supposition that it may be importantly false.

In the criticism phase we are considering whether, given A, the sample $\underset{\sim}{y}_d$ is likely to have occurred at all. To do this we must consider it in relation to the other samples that could have occurred but did not.

For instance in the Bernoulli trial example, had we sampled until we had r successes rather than until we had n trials, then the likelihood, and, for a fixed prior, the posterior distribution, would have been unaffected, but the predictive check would (appropriately) have been somewhat different because the appropriate reference set supplied by $p(\underset{\sim}{y}|A)$ would be different.

C(iv) How do you choose the significance level?

It has been argued that if significance tests are to be employed to check the model, then it is necessary to state in advance the level of significance α which is to be used and that no rational basis exists for making such a choice.

While I believe the ultimate justification of model checking is the reference of the checking function to its appropriate predictive distribution, the examples I have given to illustrate the predictive check may have given a misleading idea of the formality with which this should be done. In practice the predictive check is not intended as a

formal test in the Neyman-Pearson sense but rather as a rough assessment of signal to noise ratio. It is needed to see which indications might be worth pursuing. In practice model checks are frequently graphical, appealing as they should to the pattern recognition capability of the right brain. Examples are to be found in the Normal probability plots for factorial effects and residuals advocated by Daniel [1959], Atkinson [1973] and Cook [1977]. Because spurious patterns may often be seen in noisy data some rough reference of the pattern to its noise level is needed.

D. As might be expected the mistaken search for a single principle of inference has resulted in two kinds of incongruity:

attempts to base estimation on sampling theory, using point estimates and confidence intervals; and

attempts to base criticism and hypothesis testing entirely on Bayesian theory.

The present proposals exclude both these possibilities .

Concerning estimation, we will not here recapitulate the usual objections to confidence intervals and point estimates but will consider the latter in relation to shrinkage estimators, ridge estimators, and robust estimators. From the traditional sampling theory point of view these estimators have been justified on the ground that they have smaller mean square error then traditional estimators. But from a Bayesian viewpoint, they come about as a direct result of employing a credible rather than an incredible model. The Bayes' approach provides some assurance against incredibility since it requires that all assumptions of the model be clearly visible and available for criticism.

For illustration, emphasized below by underlining, are the assumptions that would be needed for a Bayesian justification of standard linear least squares. We must postulate not only the model

$$y_u = \underset{\sim}{\theta}' \underset{\sim}{x}_u + e_u \qquad u = 1,2,\ldots,n \tag{26}$$

with the e_u's <u>independently and normally</u> distributed with

constant variance σ^2, but also postulate an improper prior for $\underset{\sim}{\theta}$ and σ^2.

(a) Consider first the choice of prior. As was pointed out by Anscombe [1963], if we use a measure such as $\underset{\sim}{\theta}'\underset{\sim}{\theta}$ to gauge the size of the parameters, a locally flat prior for $\underset{\sim}{\theta}$ implies that the larger is the size measure $\underset{\sim}{\theta}'\underset{\sim}{\theta}$ the more probable it becomes. The model is thus incredible. From a Bayesian viewpoint shrinkage and ridge estimators imply more credible choices of the model, which, even though approximate are not incredible.

(b) For data collected serially (in particular, for much economic data) the assumption of error independence in equation (26) is equally incredible and again its violation can lead to erroneous conclusions. See for example Coen, Gomme and Kendall [1969] and Box and Newbold [1971].

(c) The assumption that the specification in (26) is necessarily appropriate for every subscript $u = 1,2,\ldots,n$ is surely incredible. For it implies that the experimenter's answer to the question "Could there be a small probability (such as 0.001) that any one of the experimental runs was unwittingly misconducted?" is "No; that probability is exactly zero."

So far as the last assumption is concerned a more credible model considered by Jeffreys [1932], Dixon [1953], Tukey [1960] and Box and Tiao [1968] supposes that the error e is distributed as a mixture of Normal distributions

$$p(e|\underset{\sim}{\theta},\sigma) = (1 - \alpha)f(e|\underset{\sim}{\theta},\sigma) + \alpha f(e|\underset{\sim}{\theta},k\sigma) \ . \qquad (27)$$

This model was used by Bailey and Box [1980] to estimate the 15 coefficients in the fitted model

$$y = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{4} \sum_{j>i}^{4} \beta_{ij} x_i x_j + \sum_{i=1}^{4} \beta_{ii} x_i^2 + e \qquad (28)$$

using data from a balanced incomplete 3^4 factorial design. Table 1 shows some of their Bayes' estimates (marginal means and standard deviations of the posterior distribution). For simplicity, only a few of the coefficients are shown; the behaviour of the others is similar. Table 1a uses data from

(a) BOX BEHNKEN DATA
ONE OR TWO SUSPECT VALUES

ϵ	Zero	.001	.005	.010	.015	.020
α	Zero	.005	.024	.048	.070	.091
Estimates	Least Squares	Robust				
β_4	-3.7	-3.2	-3.1	-3.1	-3.1	-3.1
	(.5)	(.3)	(.2)	(.2)	(.2)	(.2)
β_{44}	-2.6	-3.0	-3.1	-3.1	-3.1	-3.1
	(.7)	(.4)	(.4)	(.4)	(.4)	(.4)
β_{13}	-3.8	-3.8	-3.8	-3.8	-3.8	-3.8
	(.8)	(.4)	(.4)	(.4)	(.3)	(.3)
β_{14}	1.0	-.5	-.5	-.5	-.5	-.5
	(.8)	(.9)	(.9)	(.9)	(.9)	(.9)

(b) BACON DATA
NO SUSPECT VALUES

ϵ	Zero	.001	.005	.010	.015	.020
α	Zero	.005	.024	.048	.070	.091
Estimates	Least Squares	Robust				
β_4	4.7	4.7	4.7	4.7	4.7	4.7
	(.3)	(.3)	(.3)	(.3)	(.3)	(.3)
β_{44}	.9	.9	.9	.9	.9	.9
	(.5)	(.5)	(.5)	(.5)	(.5)	(.5)
β_{13}	.8	.8	.8	.8	.8	.8
	(.6)	(.6)	(.5)	(.5)	(.5)	(.5)
β_{14}	-.4	-.4	-.4	-.4	-.4	-.4
	(.6)	(.6)	(.6)	(.6)	(.7)	(.7)

Table 1. Bayesian means with standard deviations (in parentheses) for selected coefficients using various values of (ϵ,α) in the contaminated model (with $k = 5$).

a paper by Box and Behnken [1960]. These data (see Figure 3) apparently contain a single bad value (y_{10}), with a small possibility of a second bad value (y_{13}). Table 1b shows the same analysis for a second set of data arising from the same design and published by Bacon [1970], which (see Figure 4) appears to contain no bad values. It was shown by Chen and Box [1979] that for $k > 5$ the posterior distribution of $\underline{\beta}$ is mainly a function of the single parameter $\varepsilon = \alpha/(1-\alpha)k$ and the results obtained for $k = 5$ are labelled in terms of ε as well as α. The analysis is based on locally noninformative priors on $\underline{\beta}$ and on $\log \sigma$ so that the estimates in the first columns of the tables ($\varepsilon = \alpha = 0$) are ordinary least squares estimates. The important point to notice is that for the first set of data which appears to contain one or two bad values, a major change away from the least squares estimates can occur as soon as there is even a slight hint ($\varepsilon = 0.001$, $\alpha = 0.005$) of the possibility of contamination. The estimates then remain remarkably stable for widely different values of ε over a plausible range.[10] But for the second set (Bacon's data), which appears to contain no bad values, scarcely any change occurs at all as ε is changed.

It has been objected that while the Normal model is inadequate, the contaminated model (27) may be equally so, and that "therefore" we are better off using ad hoc robust procedures such as have been recommended by Tukey and others and justified on the basis of their sampling properties. This argument loses force, however, since it can be shown by elementary examples (Chen and Box [1979], Box [1980]) that the effect of the Bayes' analysis is also to produce downweighting of the observations with downweighting functions very similar to those proposed by the empiricists. However, the Bayes' analysis has the advantage of being based on a visible model which is itself open to criticism and has greater adaptivity, doing nothing to

---

[10] They are however (see reply to the discussion of Box [1980]) considerably different from estimates obtained by omitting the suspect observation and using ordinary least squares.

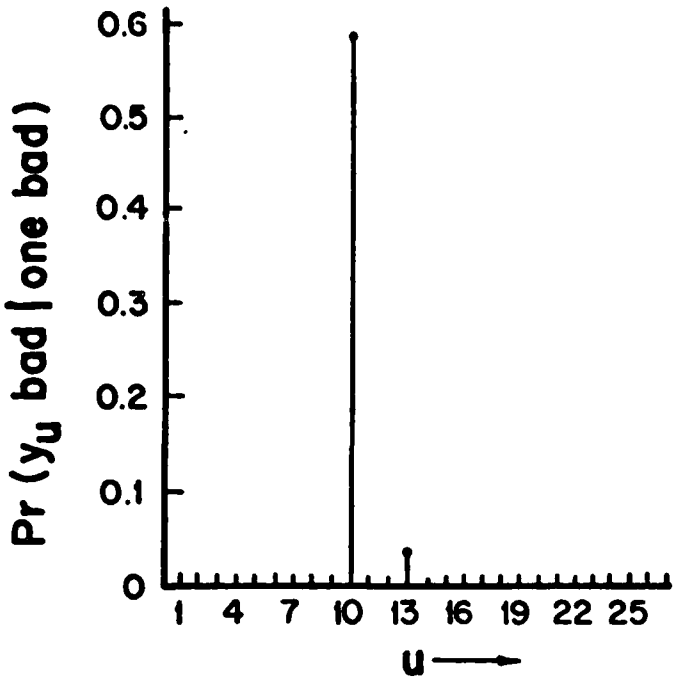


Figure 3. Posterior probability that y_u is bad given that one observation is bad (Box-Behnken data).

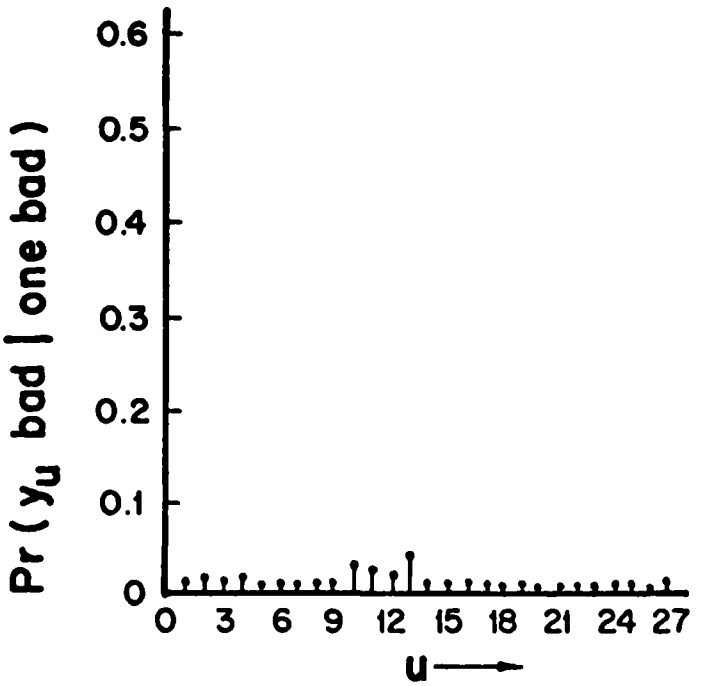


Figure 4. Posterior probability that y_u is bad given that one observation is bad (Bacon data).

samples that look normal, and reserving robustification for samples that do not. A further advantage of the present point of view is that when an outlier occurs, while the posterior distribution will discount it, the predictive distribution will emphasize it, so that the fact that a discrepancy has occurred is not lost sight of.

CONCLUSION.

In summary I believe that scientific method employs and requires not one, but two kinds of inference - criticism and estimation; once this is understood the statistical advances made in recent years in Bayesian methods, data analysis, robust and shrinkage estimators can be seen as a cohesive whole.

REFERENCES

1. Aitchison, J. and C. G. G. Aitken (1976), "Multivariate Binary Discrimination by the Kernel Method", Biometrika, 63, 413-420.
2. Anderson, E. (1960), "A Semigraphical Method for the Analysis of Complex Problems", Technometrics, 2, 387-391.
3. Anscombe, F. J. (1963), "Bayesian Inference Concerning Many Parameters with Reference to Supersaturated Designs", Bull. Int. Stat. Inst. 40, 721-733.
4. Atkinson, A. (1973), "Testing Transformations to Normality", J. Royal Statis. Soc. B, 35, 473-479.
5. Bacon, D. W. (1970), "Making the Most of a One-shot Experiment", Industrial and Engineering Chem., 62 (7), 27-34.
6. Beveridge (1950), The Art of Scientific Investigation, New York: Vintage Books.
7. Blackeslee, T. R. (1980), The Right Brain, Garden City, New York: Anchor Press/Doubleday.
8. Bailey, S. P. and G. E. P. Box (1980), "The Duality of Diagnostic Checking and Robustification in Model Building: Some Considerations and Examples", Technical Summary Report #2086, Mathematics Research Center, University of Wisconsin-Madison.

9. Box, G. E. P. (1954), "The Exploration and Exploitation of Response Surfaces: Some General Considerations and Examples", Biometrics, 10, 16-60.
10. Box, G. E. P. (1957), "Integration of Techniques in Process Development", Transactions of the 11th Annual Convention of the American Society for Quality Control.
11. Box, G. E. P. (1976), "Science and Statistics", J. Amer. Statis. Assoc., 71, 791-799.
12. Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness", J. Royal Statis. Soc. A, 143, 383-430 (with discussion).
13. Box, G. E. P. (1982), "Choice of Response Surface Design and Alphabetic Optimality", Yates Volume.
14. Box, G. E. P. and D. W. Behnken (1960), "Some New Three-level Designs for the Study of Quantitative Variables", Technometrics, 2, 455-475.
15. Box, G. E. P., W. G. Hunter and J. S. Hunter (1978), Statistics For Experimenters, New York: John Wiley and Sons.
16. Box, G. E. P. and G. M. Jenkins (1970), Time Series Analysis: Forecasting and Control, San Francisco: Holden-Day.
17. Box, G. E. P. and P. Newbold (1971), "Some Comments on a Paper of Coen, Gomme and Kendall", J. Royal Statis. Soc. A, 134, 229-240.
18. Box, G. E. P. and G. C. Tiao (1968), "A Bayesian Approach to Some Outlier Problems", Biometrika, 55, 119-129.
19. Box, G. E. P. and G. C. Tiao (1968), "Bayesian Estimation of Means for the Random Effect Model", J. Amer. Statis. Assoc., 63, 174-181.
20. Box, G. E. P. and K. B. Wilson (1951), "On the Experimental Attainment of Optimal Conditions", J. Royal Statis. Soc. B, 13, 1-45 (with discussion).
21. Box, G. E. P. and P. V. Youle (1955), "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, 287-323.

22. Chen, G. G. and G. E. P. Box (1979), "Further Study of Robustification via a Bayesian Approach", Technical Summary Report No. 1998, Mathematics Research Center, University of Wisconsin-Madison.
23. Chernoff, H. (1973), "The Use of Faces to Represent Points in k-Dimensional Space Graphically", J. Amer. Statis. Assoc., 68, 361-368.
24. Coen, P. G., E. D. Gomme and M. G. Kendall (1969), "Lagged Relationships in Economic Forecasting", J. Royal Statis. Soc. A, 132, 133-152.
25. Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression", Technometrics, 19, 15-18.
26. Daniel, C. (1959), "Use of Half-normal Plots in Interpreting Factorial Two-level Experiments", Technometrics, 1, 311-341.
27. Dixon, W. J. (1953), "Processing Data for Outliers", Biometrics, 9, 74-89.
28. Fisher, R. A. (1956), Statistical Methods and Scientific Inference, Edinburgh: Oliver and Boyd.
29. Gorman, J. W. and R. J. Toman (1966), "Selection of Variables for Fitting Equations to Data", Technometrics, 8, 27-51.
30. Hoerl, A. E. and R. W. Kennard (1970), "Ridge Regression: Applications to Non-orthogonal Problems", Technometrics 12, 69-82.
31. Jeffreys, H. (1932), "An Alternative to the Rejection of Observations", Proc. Royal Society, A, CXXXVII, 78-87.
32. Kiefer, J. (1959), Discussion in "Optimum Experimental Designs", J. Royal Statis. Soc. B, 21, 272-319.
33. Lindley, D. V. (1965), Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2, Inference, Cambridge University Press.
34. Lindley, D. V. and A. F. M. Smith (1972), "Bayes Estimates for the Linear Model", J. Royal Statis. Soc. B, 34, 1-41 (with discussion).
35. Springer, S. P. and G. Deutsch, Left Brain, Right Brain, San Francisco: W. H. Freeman.

36. Stein, C. (1955), "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution", Proceedings of the Third Berkeley Symposium, 197-206.

37. Theil, H. (1963), "On the Use of Incomplete Prior Information in Regression Analysis", J. Amer. Statis. Assoc., 58, 401-414.

38. Tiao, G. C. and M. M. Ali (1971), "Analysis of Correlated Random Effects Linear Model with Two Random Components", Biometrika, 58, 37-52.

39. Tukey, J. W. (1960), "A Survey of Sampling from Contaminated Distributions", in Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, 448-485, Stanford: Stanford University Press.

GEPB/ed

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2408	2. GOVT ACCESSION NO. AD-A120 967	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* AN APOLOGY FOR ECUMENISM IN STATISTICS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) G. E. P. Box		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE July 1982 13. NUMBER OF PAGES 34
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*)		15. SECURITY CLASS. (*of this report*) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. Key Words: Bayes inference, Bayes sampling inference, frequentist inference, data analysis, shrinkage estimates, ridge estimates, robust estimates, right and left brain, theory-practice iteration, predictive distribution, binomial model, response surfaces, time series, prior distribution, likelihood principle, significance level, alphabetic optimality, residuals, goodness of fit, statistical criticism, statistical estimation, scientific investigation, iterative investigation, models, normal linear model, subjective probability, contaminated normal model.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Reasons are advanced for the belief that scientific method employs and requires not one, but two kinds of inference - criticism and estimation; once this is understood the statistical advances made in recent years in Bayesian methods, data analysis, robust and shrinkage estimators can be seen as a cohesive whole.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*