DEPARTMENT OF STATISTICS

University of Wisconsin-Madison

TECHNICAL REPORT NO. 573

June 1979

SAMPLING AND BAYES' INFERENCE IN THE ADVANCEMENT OF LEARNING

George E. P. Box

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

SAMPLING AND BAYES' INFERENCE IN THE ADVANCEMENT OF LEARNING*

George E. P. Box

June 1979

ABSTRACT

Scientific method is a process of guided learning in which accelerated acquisition of knowledge relevant to some question under investigation is achieved by a hierarchy of iterations in which induction and deduction are used in alternation.

This process employs a developing model (or series of models implicit or explicit) against which data can be viewed. Ideally at any given stage of an investigation, such a model approximates relevant aspects of the studied system and motivates the acquisition of further data as well as its analysis. By the use of a prior distribution it is possible to represent some aspects of such a model as completely known and others as more or less unknown.

Now parsimony requires that, at any given stage, the model is no more complex than is necessary to achieve a desirable degree of approximation and since each investigation is unique we cannot be sure in advance that any model we postulate will meet this goal. Therefore, at the various points in our investigation where data analysis is required, two types of inference are involved: model criticism and parameter estimation. To effect the latter, conditional on the plausibility of the model, and given the data, we can, using Bayes' Theorem, deduce posterior distributions for unknown parameters and so make inferences about them. But, before we can rely on such conditional deduction, we ought logically to check whether the model postulated accords with the data at all and, if not, consider how it should be modified. In practice, this question is usually investigated by inspecting residuals, by other informal techniques, and sometimes by making formal tests of goodness of fit. In any case model criticism, the inferential procedure whereby the need for model modification is induced, is ultimately dependent on sampling theory argument. These principles are formalized by an appropriate analysis of Bayes' formula, and implications for robust estimation are considered.

AMS (MOS) Subject Classifications - 62.02, 62A15, 62F05, 62F10

Key Words - Bayesian inference, Sampling theory inference, Estimation, Models, Predictive distribution

Work Unit Number 4 - Probability, Statistics, and Combinatorics

*A paper read at the International Meeting on Bayesian Statistics, May 28-June 2, 1979 at Valencia, Spain.

Also issued as Technical Report No. 573, by the Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53706

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024. & DAAG29-78-G-0166

SIGNIFICANCE AND EXPLANATION

Sampling theory inference (e.g. inference based on sampling distributions of statistics and in particular on significance tests) and Bayesian inference are usually thought of as rivals and much effort has been spent in propounding their relative merits. In this paper it is argued that both kinds of inference are needed in the scientific iteration whereby knowledge is acquired.

This iteration employs a directed alternation between induction and deduction which uses model criticism on the one hand and parameter estimation on the other. An analysis of Bayes' formula reveals model criticism as a sampling theory concept and parameter estimation as a Bayesian concept. The implications of these ideas for robust estimation are discussed.

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

SAMPLING AND BAYES' INFERENCE
IN THE ADVANCEMENT OF LEARNING

George E. P. Box

Today Statistics appears to be in a somewhat confused state*. The controversy about Bayesian inference and Sampling Theory inference which some believe involves a critical choice is not resolved to most people's satisfaction. Furthermore concepts such as Data Analysis and Robust Estimation are receiving such new emphasis that some advocates of the "new Statistics" are even claiming that all else is useless and old hat.

To some extent the new and admirable emphasis on "looking at the data" is a reaction to previous extremes. On the one hand overemphasis on theory for its own sake (mathematistry) and on the other a knee-jerk approach to statistical analysis (cookbookery)†. Neither of these aberrations was healthy and some adjustment was long overdue. However I think the mistake continues to be made of assuming that different approaches to Statistics are necessarily in an adversary position. I will develop the contrary view and try to show how I believe the pieces fit together.

I start from the idea that Statistics is or should be the art and science of building scientific models which (necessarily) involve probability. Consider then how such stochastic model building should be done.

*What is happening is related to the revolutionary change in computational speed. We need to be deterred less and less by the number of steps required in a calculation although we must correspondingly increase our concern that the human mind is also adequately involved in directing the tactics and strategy of investigation.

†See discussion of "mathematistry" and "cookbookery" in Science and Statistics, (Box 1976).

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024 and DAAG29-78-G-0166.

1. The advancement of learning as an iteration between theory and practice

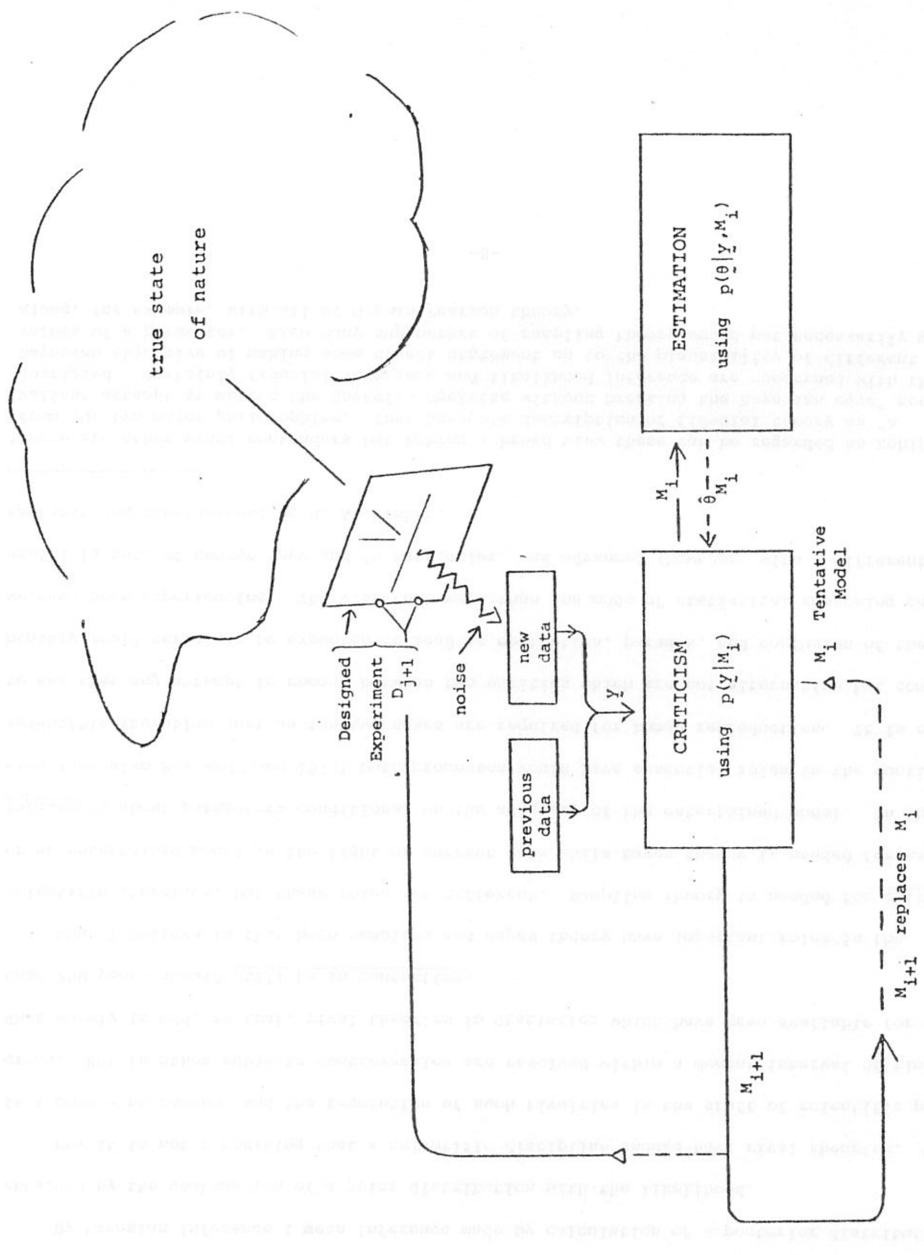
Although the matter was over the centuries debated it seems long ago to have been agreed that scientific knowledge is efficiently advanced, not by mere theoretical speculation on the one hand, nor by the mere accumulation of empirical facts on the other, but by a motivated iteration between these two activities. In this practice-theory interation a tentative theory or model suggests a particular examination and analysis of data already existing or to be acquired*. The results of this examination will then frequently suggest a modified model requiring further practical illumination and so on. The advancement of knowledge thus occurs as the result of an interplay between dual processes of induction and deduction which carry forward an iteration in which the model is not fixed but is continually changing. The statistician's role is to assist this process. In doing so he uses two inferential devices that I will call Criticism† and Estimation. The first can induce model modification, the second leads to estimation of unknown parameters assuming the truth of the model. For illustration, in Figure 1 at some stage of an investigation, model M_i is currently being entertained.

Criticism involves a confrontation of M_i with available data $\underset{\sim}{y}$ and asks whether M_i is consonant with $\underset{\sim}{y}$ and, if not, how not. It is a process of diagnostic checking. It may be done informally using plotting techniques of various kinds often involving residual quantities and more formally, with tests of goodness of fit. It may suggest that model modification to M_{i+1} is needed. In some instances it will be judged appropriate to now confront M_{i+1} with the same data, in others the nature of the modified model or necessity for independent verification may indicate the need for new data generated by a new design D_{j+1} This will be chosen to explore shadowy regions whose illumination is currently believed to be important to progress.

Estimation. If the process outlined above leads to a verifiable model, that is one which when put to the test appears to provide an adequate approximation to reality, it may logically be used to estimate parameters conditional on its truth. However in practice this

*I shall suppose that data is acquired from a designed experiment but the same argument would apply if data acquisition was from a sample survey or even from a visit to the library.

†The apt naming of model criticism is due to Cuthbert Daniel.



estimation process will be used not only at the termination of the model building sequence but at every stage throughout it. This is because in order to conduct criticism of the model it is often necessary to provisionally estimate parameters at intermediate stages, tentatively entertaining the model as if it were believed true.

I shall argue in this paper that while <u>criticism</u> must ultimately appeal to <u>sampling theory</u> for its justification <u>estimation</u> requires the use of <u>Bayes theorem</u> (or, for the faint-hearted, likelihood). Acceptance of this position provides justification for a specific kind of sampling theory significance tests but none for sampling theory confidence intervals.

2. Rival theories [illegible]nference

The distinction between inferential criticism and parameter estimation has often not been made and proponents both of sampling inference and Bayesian inference* have long sought, mistakenly in my view, for a single comprehensive theory. By sampling theory inference I mean inference made by referring some relevant function of the data to a reference sampling distribution which would be appropriate if some specific hypothetical model M_0 were true.

By Bayesian inference I mean inference made by calculation of a posterior distribution obtained by the combination of a prior distribution with the likelihood.

Now it is not surprising that a scientific discipline should have rival theories. This is a common phenomenon and the resolution of such rivalries is the stuff of scientific progress. But in other subjects controversies are resolved within a decent interval of time. What surely is odd, is that, rival theories in Statistics which have been available for more than 200 years should still be in contention.

What I believe is that both sampling and Bayes theory have important roles in the scientific iteration, but these roles are different. Sampling theory is needed for criticism of an entertained model in the light of current data while Bayes theory is needed for making inferences about parameters conditional on the adequacy of the entertained model. On this view (see also Box and Tiao; 1973) both processes would have essential roles in the continuing scientific iteration just as the two sexes are required for human reproduction. It is easy to see that any attempt to choose between two entities which are not alternative but complementary could certainly be expected to lead to contention, paradox, and confusion of the kind we have been experiencing. The view that more than one mode of statistical reasoning can be useful is not, of course, new and in particular was advanced (however with a different emphasis and conclusions) by R. A. Fisher.

---

*There are other minor contenders but taking a broad view these can be regarded as schisms from the two major philosophies. Thus Savage's description of fiducial theory as "a valiant attempt at making the Bayesian omelette without breaking the Bayesian eggs" seems justified. Certainly fiducial inference and likelihood inference are concerned with the Bayesian objective of making some direct statement as to the plausibility of different values of a parameter. Also many supporters of sampling theory would not necessarily go along, for example, with all of Neyman-Pearson theory.

3. Some remarks on Sampling and Bayes inference

The essence of what I mean by "sampling theory inference" is exemplified by the Shewhart quality control chart. The set of limit lines for the sample mean for example indicates for this function of the data, a reference distribution appropriate for the model M_0 (for the process in control). A single outlying point is surprising because it is associated with unusually low probability density. It thus raises the possibility that M_0 is inappropriate and that an alternative model M_1 might be needed to explain the inadequacy. In the words of Shewhart, the process is out of control in a manner which we may be able to attribute to an assignable cause. A number of different functions of the data may be considered in checking the appropriateness of M_0 and their nature depends on the type of departures from M_0 that are in mind. Thus quality control charts are often kept of both the sample mean and the sample range to indicate departures from M_0 in both level and spread and other functions such as run length of positive deviations might also be considered. Finally patterns which were not foreseen may possibly turn up, invite consideration, and induce possible explanations to be subsequently tested.

Prior probabilities in Bayesian and Sampling inference

In the past the need for prior probabilities has often not been thought of as a necessity for all scientific inference but rather as a feature peculiar to Bayesian inference. Indeed it is often regarded by non-Bayesians as the major point of weakness of Bayes theory and has, therefore, been a focus for attack and sometimes for derision. By contrast a Bayesian proponent might argue (a) that any theory of estimation worthy of the name should make it possible, given a model, to say after data had come to hand what was believed about the values of its parameters and (b) that what was believed after the data was available must surely depend on what was believed before it was available (c) that this would include the possibility of sometimes using non-informative prior distributions either to simulate the actual state of relative ignorance of the investigator or to represent the impact of the data on a hypothetical unbiased observer (or juror). He might argue further that the difficulties and paradoxes that have embarrassed advocates of sampling theory as it has been practiced and their inability to fix up the theory convincingly have come from its past inadequate capability to include prior information.

Sampling theory is of course not free from assumptions of prior knowledge. Instead it is as if only two states of mind have been allowed--complete certainty or complete uncertainty. Whereas in the sampling theory context a parameter had to be treated either as exactly known or as completely unknown, in the Bayesian context a prior could be chosen to approach either of these extremes or any intermediate state.

In this connection it is important to remember that every simple model can be thought of as embedded in a more complex one. For example an outright assumption of normality can be modelled by a suitable parametric family of distributions indexed by a parameter β, which has a sharp prior at the normal value. Independence of errors, so frequently assumed, can similarly be represented by a sharp prior operating on a broader model allowing appropriate dependence. Seen in this way, it appears that, when assumptions of normality and independence are made in sampling theory, it is not that no prior knowledge is used, but rather that implausibly precise prior knowledge is implied.

4. The model is the prior

Such considerations lead me to believe that it is impossible to logically distinguish between the model and the prior distribution. In a real sense the model is the prior. A model is a probability statement of all the assumptions currently to be tentatively entertained a priori. These probability statements can express certainty or various degrees of uncertainty.

Of course models are approximations (good ones are artfully chosen approximations which work well in practice). But there is good reason to believe that the "all or none" prior assumptions implied in the traditional use of a sampling theory are frequently too crude even as an approximation. Indeed many of the difficulties of sampling theory which have come to light in recent years may be traced to the primitive means it has available for incorporating prior knowledge and the crippling effect of allowing only probability statement of a certain kind to be included in the model. One illustration of how implied prior knowledge which is implausibly imprecise can lead to trouble in sampling theory in the famous discovery by Stein (1956) of the inadmissibility of normal multivariate mean, and the improved nonlinear shrinkage extimators which give smaller mean square error.

It is however easy to miss the lesson which is to be learned from such examples. To be specific, consider the usual one-way analysis of variance set-up. Here a locally uniform prior distribution for the set of group means $\underset{\sim}{\mu}' = (\mu_1, \mu_2, \ldots, \mu_j, \ldots, \mu_n)$ which would exactly justify the sample averages as estimators makes little sense (see, for example, Box and Tiao (1968), Lindley and Smith (1972)). By contrast the prior essumption which justifies the shrinkage estimator is that the μ_j are random samples from some normal super population having unknown mean and variance. This corresponds to the usual "model II" sampling theory assumption and in appropriate circumstances could be eminently reasonable. It is crucial to notice, however, that there are many circumstances in which this latter assumption would not be sensible either, because, although prior knowledge about $\mu_1, \mu_2, \ldots, \mu_n$ existed, it was of quite a different character. For example, if the μ's were daily batch yields from some production process, it would usually be much more sensible to postulate that the[illegible]llowed some time series model such as a stationary autoregressive process (Tiao and Ali (1971)). The estimators then derived from Bayesian means are not Stein's shrinkage estimators, which would appropriately introduce sample information about σ_μ^2, but alternative estimators allowing incorporation of relevant sample information about the autocorrelation of the batch means.

Some sampling theorists concede that Bayes theorem may be used as a kind of conjuring trick to produce efficient estimators which are then used in a sampling theory context. In this excercise they regard the prior distribution as a convenient prop which is never taken seriously and is quickly discarded. I think the example quoted above is one of many which shows that this idea has no rational status. For it illustrates that there is not one set of "shrinkage estimators", but an infinity of such sets depending (very naturally) on the model (that is the prior) which is appropriate to describe the particular scientific situation under study.

The strength of the explicit statement of prior assumptions is that in the iterative model building process, they make manifest at every stage exactly what assumptions are tentatively entertained and so allow them to be criticized. Some of the nervousness experienced by non-Bayesians confronted with the idea of a prior distribution has perhaps arisen because the iterative nature of scientific process and consequent tentative transitory character of models and all their assumptions, has not been generally understood.

Many of us were taught to think unrealistically in terms of "one shot" procedures. The sequence: frame hypothesis - collect data - test hypothesis/make decision; of course, fails to describe the usual context in which Statistics is applied.

Critics have therefore feared gross mistakes arising from adamantine prior prejudice which ignored "what the data were trying to say." In the iterative context of real scientific enquiry however gross mistakes about the prior or any other aspect of the model will usually be corrected at the criticism phase of the next iteration.

5. Two complementary factors from Bayes formula

If we accept the prior probability distribution of parameters $\underset{\sim}{\theta}$ as an essential part of a model then all aspects of the model, hypothesized at some particular stage of an investigation, are contained in the joint density obtained by combining the likelihood and the prior

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|M) = p(\underset{\sim}{y}|\underset{\sim}{\theta},M) \cdot p(\underset{\sim}{\theta}|M) \qquad (5.1)$$

where $|M$ is understood to indicate conditionality on some aspect of the model and $\underset{\sim}{y}$ is a data vector.

This joint distribution which is a comprehensive statement of the model can also be factored as

$$p(\underset{\sim}{y},\underset{\sim}{\theta}|M) = p(\underset{\sim}{\theta}|\underset{\sim}{y},M)p(\underset{\sim}{y}|M) \qquad (5.2)$$

and can be computed before any data become available. In particular the second factor on the right

$$p(\underset{\sim}{y}|M) = \int p(\underset{\sim}{y}|\underset{\sim}{\theta},M)p(\underset{\sim}{\theta}|M)d\underset{\sim}{\theta}\ , \qquad (5.3)$$

which is the predictive distribution, may be so calculated. It is the distribution of the totality of all possible samples that could occur if the model M were true.

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

$$p(\underset{\sim}{y}_d,\underset{\sim}{\theta}|M) = p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,M)p(\underset{\sim}{y}_d|M). \qquad (5.4)$$

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

$$p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,M) = kp(\underset{\sim}{y}_d|\underset{\sim}{\theta},M)p(\underset{\sim}{\theta}|M) \qquad (5.5)$$

and the second factor

$$p(\underset{\sim}{y}_d|M) = \int p(\underset{\sim}{y}_d|\underset{\sim}{\theta},M)p(\underset{\sim}{\theta}|M)d\underset{\sim}{\theta} = k^{-1}_{\underset{\sim}{y}_d} \qquad (5.6)$$

is the predictive density associated with the data set $\underset{\sim}{y}_d$ actually obtained. Figure 2 illustrates for a single parameter θ and a sample $\underset{\sim}{y}_d$ of $n = 2$ observations.

If the model is to be believed, then the posterior distribution $p(\underset{\sim}{\theta}|\underset{\sim}{y}_d,M)$ allows all relevant estimation inferences to be made about $\underset{\sim}{\theta}$. However even if the model were totally incorrect, this could not be shown by any abnormality in this factor which is conditional on both data and model specification. However plausibility or otherwise of obtaining such a sample if the model were appropriate may be assessed by reference of the density $p(\underset{\sim}{y}_d|M)$ to

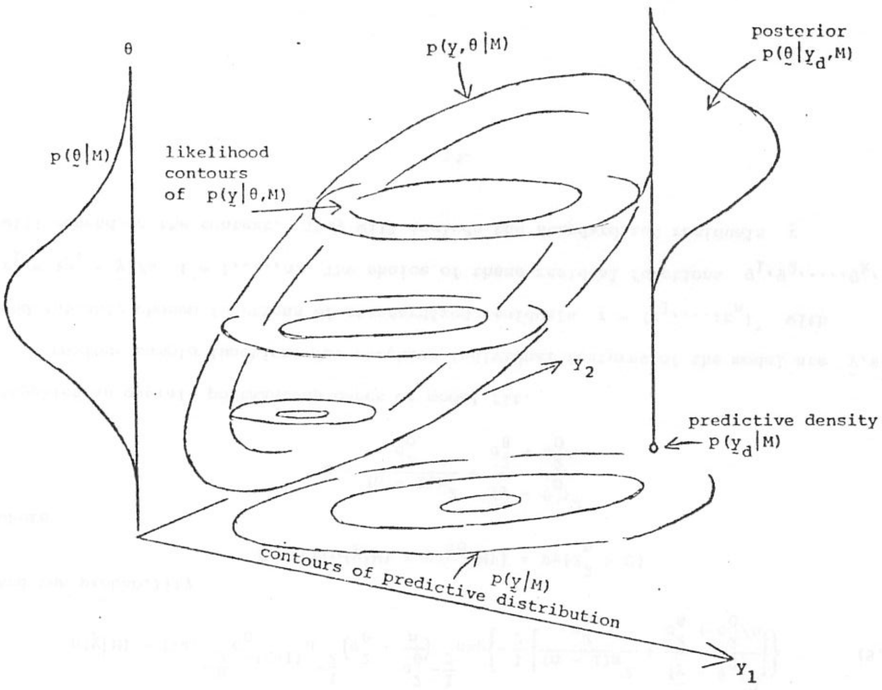


Figure 2. Showing for a single parameter θ and sample $\underset{\sim}{y}_d$ of two observations; the prior distribution, likelihood contours, the posterior distribution and contours of the predictive distribution.

the predictive reference distribution $p(\underset{\sim}{y}|M)$. An unusually small value of $p(\underset{\sim}{y}_d|M)$ as measured by $\Pr\{p(\underset{\sim}{y}|M) < p(\underset{\sim}{y}_d|M)\}$ casts doubt on the appropriateness of the model M. Now $p(\underset{\sim}{y}|M)$ is an n-dimensional distribution and it will usually be true that if the model is inadequate it is most likely to be deficient in certain directions associated with unusual values of certain specific functions $g_i(\underset{\sim}{y})$ of the data. Examples of such functions are sample averages, variances, moment coefficients, coefficients of serial correlation, and other measures of standardized deviations from a norm. In every case the appropriate reference distribution to which the realized statistic $g_i(\underset{\sim}{y}_d)$ should be referred is the distribution $p\{g_i(\underset{\sim}{y}|M)\}$, when the model M is true, derived by appropriate integration of $p(\underset{\sim}{y}|M)$.

In practice, criticism or diagnostic checking of the model is often conducted by visual inspection of residual displays and other more sophisticated plots. But such a process, although it is informal, still, it seems to me, falls within the logical framework described above. The statistician is looking for "features" in the data which would be surprising or "unusual" if the model M were true. Such a feature can be described by a function $g(\underset{\sim}{y}_d)$ and its unusualness, if formalized, would have to be measured by reference to $p\{g(\underset{\sim}{y})|M\}$.

In addition to possible discrepancies to which we have been alerted by experience, other features may appear pointing to inadequacies of a kind not previously suspected. This possibility has sometimes proved perplexing for statisticians, for while on the one hand the truly unexpected could point the way to precious new knowledge, on the other, associated probabilities will be indeterminate because of the uncountable character of other features that might also have been regarded as surprising. I think the calculation which ignores this difficulty of indeterminate selection should still be made, for while it might lead to the too frequent pursuit of nonexistent assignable causes, the iterative process will quickly terminate this chase and carrying out the exercise will at least eliminate phenomena, which at first sight look surprising but really are not. For example, Feller (1968) shows that for a random group of 30 people, the probability that at least two have coincident birthdays is over 70%, this tells us we need look no further for an explanation when we are surprised to find two such people at a party.

Example: Unknown mean θ, variance σ_0^2 assumed known.

Consider [illegible]ple of n observations drawn randomly from a normal distribution with unknown mean θ and known variance σ_0^2. We express uncertainty about the mean by supposing that a priori θ is distributed normally about θ_0 with variance σ_θ^2

Thus

$$\begin{cases} p(\underset{\sim}{y}|\theta,M) = (2\pi)^{-\frac{n}{2}}\sigma_0^{-n}\exp\left\{-\frac{1}{2}\frac{\Sigma(y_i-\theta)^2}{\sigma_0^2}\right\} & (5.7) \\ p(\theta|M) = (2\pi)^{-\frac{1}{2}}\sigma_\theta^{-1}\exp\left\{-\frac{1}{2}\frac{(\theta-\theta_0)^2}{\sigma_\theta^2}\right\} & (5.8) \end{cases}$$

The posterior distribution from which θ may be estimated conditional on the adequacy of the model is then

$$p(\theta|\underset{\sim}{y},M) = (2\pi)^{-\frac{1}{2}}\left(\frac{1}{\sigma_\theta^2}+\frac{n}{\sigma_0^2}\right)^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_\theta^2}+\frac{n}{\sigma_0^2}\right](\theta-\bar{\theta})^2\right\} \tag{5.9}$$

where $\bar{\theta} = \left(\frac{1}{\sigma_\theta^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{1}{\sigma_\theta^2}\theta_0+\frac{n}{\sigma_0^2}\bar{y}\right)$

The predictive distribution which can act as reference distribution for the observed data vector $\underset{\sim}{y}_d$, thus allowing criticism of the model, is

$$p(\underset{\sim}{y}|M) = (2\pi)^{-\frac{n}{2}}\sigma_0^{-(n-1)}n^{-\frac{1}{2}}\left(\sigma_\theta^2+\frac{\sigma_0^2}{n}\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\frac{(n-1)s^2}{\sigma_0^2}+\frac{(\bar{y}-\theta_0)^2}{\sigma_\theta^2+\sigma_0^2/n}\right]\right\} \tag{5.10}$$

And the probability

$$P = \Pr\{p(\underset{\sim}{y}|M) < p(\underset{\sim}{y}_d|M)\} = \Pr\{\chi_n^2 > C\}$$

where

$$C = \frac{(n-1)s^2}{\sigma_0^2}+\frac{(\bar{y}-\theta_0)^2}{\sigma_\theta^2+\sigma_0^2/n}$$

supplies an overall portmanteau check on model fit.

Obvious sample functions for checking individual features of the model are $\bar{y},s^2$ and suitably chosen functions of standardized residuals $\underset{\sim}{r} = (r_1,\dots,r_n)'$ with $r_i = (y_i-\bar{y})/s$ $i = 1,\dots,n$. The choice of these residual functions $g_1,g_2,\dots,g_k$, will depend on the context. They will include the standardized residuals $\underset{\sim}{r}$

themselves, but might also address the need to apply checks for "bad values", skewness, kurtosis and serial correlation, for example. The standardized residu[illegible]n the form defined above are constrained by the identities $\Sigma r_i = 0$, $\Sigma r_i^2 = n-1$ and can be more conveniently parameterized in terms of $n-2$ independently distributed functions obtained as follows:

Make an orthogonal transformation from $\underset{\sim}{y}$ to $\underset{\sim}{Y} = (Y_1,Y_2,\dots Y_n)'$ with $Y_n = \sqrt{n}\bar{y}$ and then transform to $\bar{y},s^2$ and $\underset{\sim}{u}$ where $\underset{\sim}{u}$ is a vector of $n-2$ residual quantities $\underset{\sim}{u} = (u_1,u_2,\dots,u_{n-2})$ such that

$$u_j = Y_{j+1}\Big/\left\{\sum_{i=1}^{j} Y_i^2/j\right\}^{\frac{1}{2}}.$$

The Jacobian of the transformation from $\underset{\sim}{y}$ to $\bar{y},s^2,\underset{\sim}{u}$ is proportional to $(s^2)^{\frac{n-1}{2}-1}\prod_{j=1}^{n-2}\{1+u_j^2/j\}^{-\frac{1}{2}(j+1)}$. After transformation therefore the predictive distribution contains n elements all of which are distributed independently and becomes

$$p(\bar{y},s^2,\underset{\sim}{u}|M) = p(\bar{y}|M)p(s^2|M)p(\underset{\sim}{u}|M) \tag{5.11}$$

where

$$p(\bar{y}|M) = (2\pi)^{-\frac{1}{2}}(\sigma_\theta^2+\sigma_0^2/n)^{-\frac{1}{2}}\exp(-\tfrac{1}{2}(\bar{y}-\theta_0)^2/(\sigma_\theta^2+\sigma_0^2/n)) \tag{5.12}$$

$$p(s^2|M) = \left\{\frac{1}{2}(n-1)\right\}^{\frac{1}{2}(n-1)}\Gamma^{-1}\left\{\frac{1}{2}(n-1)\right\}(\sigma_0^2)^{-\frac{1}{2}(n-1)}\{s^2\}^{\frac{n-1}{2}-1}\exp\left\{-\frac{1}{2}(n-1)s^2/\sigma_0^2\right\} \tag{5.13}$$

$$p(\underset{\sim}{u}|M) = \Gamma^{-\frac{1}{2}}(n-1)(\pi)^{-\frac{(n-1)}{2}}\Gamma\left\{\frac{1}{2}(n-1)\right\}\prod_{j=1}^{n-2}\left\{1+\frac{u_j^2}{j}\right\}^{-\frac{1}{2}(j+1)} \tag{5.14}$$

The standardized residual quantities of interest $g_1,g_2,\dots,g_k$ can then be expressed equally as functions $f_1(\underset{\sim}{u}),f_2(\underset{\sim}{u}),\dots,f_k(\underset{\sim}{u})$ of the u's. So that, in particular, unusual features of $\bar{y},s^2$, and $g_1,\dots,g_k$ given the model could be assessed by computing

(i) $\Pr\{p(\bar{y}|M) < p(\bar{y}_d|M)\}$

(ii) $\Pr\{p(s^2|M) < p(s_d^2|M)\}$

(iii) $\Pr\{p(g_j|M) < p(g_{jd}|M)\}$ $j = 1,2,\dots,k$.

These are the (two tail area) probabilities associated with reference of

(i) $(\bar{y}_d - \theta_0)/(\sigma_\theta^2 + \sigma_0^2/n)^{\frac{1}{2}}$ to the Normal table

(ii) $(n - 1)s_d^2/\sigma_0^2$ to the χ^2 table

(iii) g_{jd} to the reference distribution obtained by appropriate integration of the distribution $p(\underline{u}|M)$.

They yield checks on the adequacy of the model which we denote by $c(\bar{y}), c(s^2), c(g_j)$.

For example suppose the yield of a batch process was under study and that a sample $\underline{y}$ was available of n observations all from a single batch having unknown mean θ. Suppose at this stage of the investigation that the tentative model assumed that, because of process variation, batch means varied Normally and independently about some value θ_0 with variance σ_θ^2 and, because of testing variation, the ith observation y_i varied about θ normally and independently with variance σ_0^2. Then the model would be that discussed above and, if this model could be believed, the batch mean θ would be estimated by the posterior distribution $N\{\bar{\theta}, (I_\theta + I_{\bar{y}})^{-1}\}$ where $I_{\bar{y}} = n\sigma_0^{-2}$, $I_\theta = \sigma_\theta^{-2}$. And, if we write $w = I_{\bar{y}}/(I_\theta + I_{\bar{y}})$ for the proportion of the information coming from the sample, then $\bar{\theta} = w\bar{y} + (1 - w)\theta_0$.

Before drawing such a conclusion however a prudent statistician would question the model. In particular applying the checks $c(\bar{y}), c(s^2), c(g_j)$,

(i) an unusually small value of $p(\bar{y}|M)$ could call into question the choice of some or all of $\theta_0, \sigma_\theta^2$ and σ_0^2.

(ii) an unusually small value of $p(s^2|M)$ could call into question the choice of σ_0^2.

(iii) an unusually small value of $p(g_j|M)$ could suggest departures from the assumed distributional form $p(y|\theta,M)$ produced by serial correlation, bad data values, non-normality, etc.

Only after the investigator had found that the evidence offered by the data did not invalidate the model should he proceed to make the conditional deductive inference supplied by Bayes theorem.

6. Some Implications

Consider the problem of making inferences about θ in the previous example. If we assume the model true then we can estimate θ from a normal posterior distribution with mean $\bar{\theta} = w\bar{y} + (1 - w)\theta_0$ and variance $(I_\theta + I_{\bar{y}})^{-1}$ where $w = I_{\bar{y}}/(I_\theta - I_{\bar{y}})$ is the fraction of the information coming from the sample. First however we require to check the model using the predictive distribution. In particular the check $c(\bar{y})$ requires a reference of $(\bar{y}_d - \theta_0)/(\sigma_\theta^2 + \sigma_0^2/n)^{\frac{1}{2}}$ to the normal table.

Significance test. Suppose σ_θ^2 is assumed small compared with σ_0^2/n, then w, the relative amount of information, supplied by the data is small and $1 - w$ is close to unity. Then, if this model can be relied upon, the posterior distribution is essentially the same as the prior and is sharply centered at θ_0. (A practical context is one where the statistician is told that process variation is negligible compared with testing variation and the process mean is known to be θ_0.) If this model is assumed, then information from available data $\underline{y}$ can add very little to what is known already. However, it can deny the relevance of this model. In particular $c(\bar{y})$ involves the reference of $(\bar{y} - \theta_0)/(\sigma_0/\sqrt{n})$ to normal tables; the failure of this check means that the model is discredited and therefore the operation that leads to a sharp posterior distribution centered at θ_0 may not logically be undertaken.

The above most satisfactorily explains to me the rationale of a significance test.

(i) The tentative model (null hypothesis) implies that $\theta = \theta_0$.

(ii) A check on this aspect of the model is provided by reference of $(\bar{y} - \theta_0)/(\sigma_0/\sqrt{n})$ to the Normal Table.

(iii) If the tail area probability is not small we do not question the model. The application of Bayes theorem then produces a posterior distribution which is a delta function at θ_0. We have "no reason to question the null hypothesis".

(iv) If the tail area probability is small we conclude that the model which postulates that $\theta = \theta_0$ is discredited by the data and that some other model is appropriate. The "null hypothesis is rejected."

(v) Notice too that although the failure of this check would most immediately proscribe the use of Bayes theorem, the failure of other checks (and of $c(s^2)$ in particular) would also indicate the necessity of model modification before proceeding further.

A difficulty that this removes for me is that, as usually formulated, significance tests seem to provide no basis for belief. On the above argument, if we accept the model, we believe a priori that θ is close to θ_0. We must therefore believe that $\theta = \theta_0$ very nearly a posteriori. The availability of data provides however an opportunity to assess the concordance of data and model.

The significance test itself provides a means only of discrediting the model. Our belief in the proposition $\theta = \theta_0$ comes from an application of Bayes theorem for a model which there is no reason to question (as a reasonable approximation to truth).

In particular this underscores the illogicality of testing a null hypothesis which is not credible to begin with. Thus the Durbin-Watson test for serial correlation, for which the null hypothesis is that errors are distributed independently, is frequently misapplied to test serial data which a priori can be expected to be autocorrelated.

Precise measurement and improper priors

Suppose now that σ_θ^2 was very large compared with σ_0^2/n. The predictive check $c(\bar{y})$ now approaches $(\bar{y}_d - \theta_0)/\sigma_\theta$ implying that for sets of data having widely different sample averages the model would not be called into question. The situation where such a non-informative prior distribution was relevant was referred to by L. J. Savage as that where the theory of precise measurement applied. The invocation of this principle might, at first, seem a license to use Bayes theorem without any restraining checks of the model. But this idea makes no sense either from an applied or a theoretical point of view.

The practical situation is that the sample information coming from y must be evaluated in a context where there is relatively very little prior information about the value of θ.

Here computational convenience and logic must of course be carefully distinguished. Replacing "relatively very little" by zero can be justified computationally in those circumstances where to do so provides a good numerical approximation but not otherwise. However in either case zero remains infinitely smaller than any small quantity. In this example, substitution of an improper uniform prior will produce a normal posterior distribution having mean $\bar{y}$ and variance σ_0^2/n, also obtained as the limit when, in our model, the fraction of information w supplied by the data tends to unity. But not only that, the specification of the prior for θ as $N(\theta_0, \sigma_\theta^2)$ is obviously overly specific, and the improper prior could provide an appropriate limit for disperse priors which were widely different in structure and/or much less specific.

All statistical results, in so far as they relate to reality, are approximations. Those obtained from improper priors do in many important examples provide excellent approximations. I hasten to add of course that limiting processes can be tricky and theoretical statisticians are right to worry about them.

Notice however that the situation is different for the predictive check. To say that w is close to unity is only to say that σ_θ^2 will dominate the denominator in $(\bar{y} - \theta_0)/\left(\sigma_\theta^2 + \frac{\sigma_0^2}{n}\right)^{\frac{1}{2}}$. But to say that it is equal to unity implies that σ_θ^2 is infinite and the check cannot be made, which implies that there are absolutely no values of $\bar{y}$ which could discredit the model - a situation which I cannot imagine as practically possible.

Consider for example, a physical chemist who runs experiments to determine the activation energy θ for a particular chemical reaction about which little is known. It would usually be true that his initial uncertainty about θ would be large compared with the anticipated standard deviation $\sigma/\sqrt{n}$ of the experimental procedure, the theory of precise measurement would apply therefore and the limiting result obtained from the usual improper prior would supply a good approximation. Nevertheless the chemist may know that activation energies for compounds of the kind being tested are usually measured in tens of kilo calories per gram mole. If the statistician, who has perhaps misplaced a decimal point, presents him with an estimate of say 0.1 kilo calories per gram mole he will rightly reject it. In doing so he will be informally conducting a check formalized by $c(\bar{y})$. In practice then, checks such as $c(\bar{y})$ can never really be dispensed with. The non-informative prior used in practice must to make practical sense always be proper, but nevertheless the appropriate posterior distribution can, in suitable circumstances, be numerically approximated by the device of substituting an improper prior. I labour this point because although it has been made earlier (see for example Box and Tiao 1973, p. 28) critics seem to have misunderstood earlier discussions. Explicit consideration of predictive checks makes the situation even clearer.

Choosing the diagnostic checks

Frequently the checking functions $g(\underline{y})$ which are to be used formally or informally for checking various features of a model M are chosen on an ad hoc basis.

One formal basis for selection of such functions follows essentially the route explored by Neyman and Pearson. Suppose a basic model M_0 is given and an alternative model M_1 represents some discrepancy from M_0 which is of interest. Then a function of the data suitable for detecting such discrepancies may be obtained from the ratio[†]

$$p(\underline{y}_d|M_0)/p(\underline{y}_d|M_1)$$

Parsimony: Diagnostic checks versus Robustification

A question which confronts[*] the statistician at every stage of an investigation is "How complex a model should I use?" The possibilities for model elaboration are of course limitless. For instance a commonly used model assumes errors to be Independently, Identically and Normally distributed (IIN). It is easy to imagine a sequence of fall-back models which might begin like this

$$\begin{array}{ccccccccc} M_0 & \rightarrow & M_1 & \rightarrow & M_2 & \rightarrow & M_3 & \rightarrow & \ldots. \\ IIN & & II\cancel{N} & & \cancel{I}I\cancel{N} & & \cancel{I}\cancel{I}\cancel{N} & & \end{array}$$

†Model criticism cannot logically be conducted by the study of the magnitude of such ratios however, for even if this ratio were very high the predictive check could still show the favored model to be highly implausible.

*An apparently different question is "Should I use a robust procedure?", but I will argue that this is subsumed by the broader question.

At each stage of elaboration there are many forms the modified model could take and most require additional parameter values either to be given from prior knowledge or to be estimated from the data. Obviously compromise is necessary; for, on the one hand, simpler models can allow better scientific understanding and better estimation, while, on the other hand, more complex ones can, but need not, be closer to the truth. A consolation is that, realistically, model building is iterative, so that mistakes can be rectified.

This fact of necessary compromise raises the dilemma of where should the compromise be made, that is to say, of what should be left out and what be included. In particular suppose some deviation from an "ideal" model M_0 can be parametrized by a discrepancy parameter β or a vector of such parameters.

For illustration M_0 might be the usual normal model and β could measure

(i) possible serial correlation of errors $(e_{t-1}, e_t, e_{t+1}, \ldots)$; for instance, the serial correlation might be generated by a first order autoregressive process $e_t = \beta e_{t-1} + a_t$ where a_t was a source of discrete white noise.

(ii) possible deviation from error normality; for example[*] according to $p(e|\sigma,\beta) = \text{const}\ \sigma^{-1}\exp[-\frac{1}{2}\{e^2/\sigma^2\}^{1/(1+\beta)}]$

(iii) need for parametric transformation; for example the normal linear model would be valid not for y but for y^β.

(iv) need to allow for bad values; for example with probability β (close to unity) the error variance was σ^2, with probability $1-\beta$ it was $k^2\sigma^2$.

In each case there are two ways to handle the possible model discrepancy, depending on whether the parameter β is omitted from, or included in, the model. We call these diagnostic checking and robustification.

Diagnostic checking. If the discrepancy parameter is omitted from the model then an appropriate diagnostic check can be made. Formally this would be done by referring

*Here and elsewhere other functional forms might be found more appropriate. These examples are intended only to illustrate essential principles; not, of course, to be comprehensive.

some suitable function $g(\underset{\sim}{y})$ of the data to a reference distribution derived for the predictive distribution $p(y|M_0)$.

Robustification. If the discrepancy parameter is included then robust estimation* of $\underset{\sim}{\theta}$ is provided by the posterior distribution

$$p(\theta|\underset{\sim}{y}) = \int p(\underset{\sim}{\theta}|\beta,\underset{\sim}{y})p(\beta|\underset{\sim}{y})d\beta \qquad (6.1)$$

If we write

$$p_u(\beta|\underset{\sim}{y}) = p(\beta|\underset{\sim}{y})/p(\beta) \qquad (6.2)$$

$$p(\theta|\underset{\sim}{y}) = \int p(\theta|\beta,\underset{\sim}{y})p_u(\beta|\underset{\sim}{y})p(\beta)d\beta \qquad (6.3)$$

In this last expression

(i) $p(\beta)$ can be chosen to represent approximately the probability of occurrence of different values of β in the real world

(ii) the function $p_u(\beta|\underset{\sim}{y})$ is a pseudo-likelihood which reflects information about β supplied by the data

(iii) considered as a function of β, $p(\theta|\beta,\underset{\sim}{y})$ reflects the sensitivity of estimation to the choice of the discrepancy parameter.

The omission of the parameter β is equivalent to setting it equal to the value β_0 which it takes in the ideal model M_0. Table 1 shows some examples of diagnostic checks and corresponding robust estimation methods. A fuller discussion is given elsewhere (Box 1979).

Discussion. There may be Bayesians who would deny the need for diagnostic checks based on sampling theory. They may feel that "they can do it all with Bayes". I do not believe this position can be sustained because it implies either

(i) that they know what the model is in advance or

(ii) that they are prepared to make the model so comprehensive that nothing could possibly be overlooked.

*Numerous authors (Huber, Tukey, Andrews, Hampel, etc.) have proposed ad hoc methods of robust estimation relying on the empirical modification of classical estimation procedures. It seems more logical to me to modify the model which is presumably at fault rather than the method of estimation which is not. Furthermore this has the advantage of clearly revealing the assumptions which are being made.

	M_0		M_1
Example	Make Inferences Using $p(\theta|y_d,\beta=\beta_0)$	Check Using $g(y_d)$	Make Robust Inference† Using $p(\theta|y_d)=\int p(\theta|\beta,y_d)p(\beta|y_d)d\beta$
Serial Correlation	Normal Linear Model	e.g. Durbin-Watson (1950, 1951) check	e.g. Zellner and Tiao (1964)
Kurtic Error Distribution	Normal Linear Model	e.g. Anscombe and Tukey (1963) checks for kurtosis	e.g. Box and Tiao (1962)
Transformation	Normal Linear Model	e.g. Tukey's (1949) one degree of freedom for transformation	e.g. Box and Cox (1964)
Bad Observations	Normal Linear Model	Tests for outliers e.g. Grubbs (1950), Dixon (1950), Ferguson (1961), David, Hartley and Pearson (1954)	e.g. Box and Tiao (1968)*

TABLE 1 SOME EXAMPLES OF DIAGNOSTIC CHECKS AND ROBUSTIFICATION

*It is of course only in relation to this one problem that robust estimation is usually considered and that usually from the empirical non-Bayesian approach of Huber, Tukey, Andrews, etc.

†The robust model would, of course, also be subjected to checks.

Both positions are grandiose and unrealistic and the second if attempted could lead to unnecessarily complicated models which would impede scientific progress.

In this connection it must be realized that looking at residuals is essentially a sampling theory procedure and is an acknowledgement of the often happy fact that an experiment might reveal more than was bargained for. To put it another way, every Bayesian statement is conditional and somewhere there has to be an anchor.

An acceptance of my theme implies of course that what is tentatively included in a model is a matter of judgement.* However we can still look for guidelines for model building on what to tentatively include (robustify for) and what to tentatively omit (and later check for).

Obviously the need for special features in the model depends on the context, e.g.: (a) serial data (in particular most economic and business data) cannot reasonably be expected to be represented by a model with uncorrelated errors, autocorrelation is virtually certain (temporary changes in mean and variance are also very likely in serial data), (b) data for which y_{max}/y_{min} is large is likely to need transformation before any simple model could apply, (c) most experimental data are liable to occasional bad values. Elaborations which are primary candidates for robustification (inclusion in the model) <u>reflect features which might easily elude diagnostic checks and could then invalidate subsequent analysis</u>.

Although the ad hoc robustifiers seem to have given all their attention to possible non-normality of (assumed independent) observations, an even greater source of serious trouble is autocorrelation in serial data. See for example Coen, Gomme and Kendall (1968), Box and Newbold (1971), Pallesen (1977), Box and Jenkins (1970).

---

*This idea that a statistician has to use scientific judgement is not a universally popular one. The objectivity of statistics like that of science does not of course mean that all statisticians (or scientists) even though capable of using the same set of tools will do equally well when using them. Just as there are good lawyers and bad lawyers, there are good statisticians and poor ones.

APPENDIX

<u>Another Example: θ known, σ^2 unknown</u>

Suppose now we have a known mean θ but unknown variance σ^2. Also suppose we express uncertainty about the variance by assuming a priori that σ^2 is distributed about s_0^2 in a scaled χ^{-2} distribution having ν_0 degrees of freedom. This is equivalent to assuming that a supposedly relevant estimate s_0^2 of σ^2 having ν_0 degrees of freedom is available from past data and has been assessed against a non-informative reference prior (i.e. prior to the first sample the distribution of log σ was flat in the neighborhood of the likelihood). Then for a prospective sample of $n = \nu + 1$ observations

$$\left\{\begin{array}{ll} p(\underline{y}|\sigma^2,M) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left[\dfrac{-\frac{1}{2}\nu s^2 + n(\bar{y}-\theta)^2}{\sigma^2}\right] & \text{(A.1)} \\[2ex] p(\sigma^2|M) \propto (\sigma^2)^{-\left[\frac{\nu_0}{2}+1\right]} (s_0^2)^{\frac{\nu_0}{2}} \exp\left[\dfrac{-\frac{1}{2}\nu_0 s_0^2}{\sigma^2}\right] & \text{(A.2)} \end{array}\right.$$

The complete prospective statement about the model is thus

$$p(\underline{y},\sigma^2|M) \propto (s_0^2)^{\frac{\nu_0}{2}} (\sigma^2)^{-\left[\frac{n+\nu_0}{2}+1\right]} \exp\left[\frac{-\frac{1}{2}(\nu_0+n)\hat{\sigma}^2}{\sigma^2}\right] \quad \text{(A.3)}$$

where $\hat{\sigma}^2 = (n(\bar{y}-\theta)^2 + \nu s^2 + \nu_0 s_0^2)/(n+\nu_0)$.

When actual data $\underline{y}_d$ becomes available then <u>conditional on the acceptance of this model</u> inferences about σ^2 must be made from the posterior distribution

$$p(\sigma^2|\underline{y}_d,M) \propto (\sigma^2)^{-\left[\frac{n+\nu_0}{2}+1\right]} (\hat{\sigma}_d^2)^{\left[\frac{n+\nu_0}{2}\right]} \exp\left[\frac{-\frac{1}{2}(\nu_0+n)\hat{\sigma}_d^2}{\sigma^2}\right] \quad \text{(A.4)}$$

However rational acceptance of the relevance of this model for the situation in which $\underline{y}_d$ is generated requires that relevant aspects of $\underline{y}_d$ are not surprising when assessed against a reference distribution derived from the predictive distribution.

$$p(\underline{y}|M) \propto \frac{(s_0^2)^{\frac{\nu_0}{2}}}{(\hat{\sigma}^2)^{\frac{n+\nu_0}{2}}} \quad \text{(A.5)}$$

Pertinent features of the sample are $\bar{y}_d = \Sigma y/n$, $s_d^2 = \Sigma(y - \bar{y})^2/\nu$ and functions of $(n - 2)$ residual quantities $u_1, u_2, \ldots, u_{n-2}$ defined as before. These must be considered against their relevant reference distributions derived from $p(y|M)$. The Jacobian of the transformation from y to $\bar{y}, s^2, \underline{u}$ is proportional to

$$(s^2)^{\frac{\nu}{2}-1} \prod_{j=1}^{n-2} \left(1 + \frac{u_j^2}{j}\right)^{-\frac{1}{2}(j+1)}$$

Thus

$$p(\bar{y}, s^2, \underline{u}|M) = p(\bar{y}|s^2, M)\, p(s^2|M)\, p(\underline{u}|M) \tag{A.6}$$

and

$$p(\bar{y}|s^2, M) \propto \frac{1}{s_p}\left\{1 + \frac{n(\bar{y} - \theta)}{\nu_p s_p^2}\right\}^{-\left(\frac{\nu_p+1}{2}\right)} \quad \text{where } s_p^2 = (\nu s^2 + \nu_0 s_0^2) \,|\, (\nu + \nu_0) \text{ and } \nu_p = (\nu + \nu_0) \tag{A.7}$$

$$p(s^2|M) \propto \frac{1}{s_0^2} \frac{F^{\frac{\nu}{2}-1}}{\left(1 + \frac{\nu}{\nu_0} F\right)^{\frac{\nu_p}{2}}} \quad \text{where } F = \frac{s^2}{s_0^2}$$

$$p(\underline{u}|M) \propto \prod_{j=1}^{n-2} \left(1 + \frac{u_j^2}{j}\right)^{-\frac{1}{2}(j+1)}$$

Unusual features of $s^2, \bar{y}$ and $u_1, \ldots, u_{n-2}$ would thus be assessed by computing

(i) $\Pr\{p(s^2|M) < p(s_d^2|M)\}$

(ii) $\Pr\{p(\bar{y}|s^2, M) < p(\bar{y}_d|s_d^2, M)\}$

(iii) $\Pr\{p(g_j|M) < p(g_{jd}|M)\}$

These are two tailed probabilities associated with reference of

(i) s^2/s_0^2 to an F distribution with ν and ν_0 degrees of freedom

(ii) $\sqrt{n}(\bar{y} - \theta)/s_p$ to a t table.

(iii) g_j to the reference distribution obtained by appropriate integration of $p(u)$.

<u>Inferences about the variance</u>

(a) <u>Suppose $\nu_0 \to 0$.</u>

This limit corresponds to usual noninformative Jeffereys' prior. Again the values of ν_0 that could represent real situations could approach zero but not reach it. Since in practice there could always be values of s^2 which would be surprising even though $p(\sigma^2|M)$ was disperse, this would correspond to the situation where a very small value of $p(F_{\nu\nu_0})$ was found even though ν_0 was very small.

(b) <u>Suppose ν_0 is very large</u>

Then s_0^2 and $s_p^2 = (\nu s^2 + \nu_0 s_0^2)/(\nu + \nu_0)$ are very precisely known and if <u>we believe the model</u> the posterior distribution $p(\sigma^2|y, M)$, is sharply concentrated about s_0^2 and our belief a posteriori is the same as that a priori. However for $p(\underline{y}|M)$ we obtain

$$p\left(\frac{\bar{y} - \theta}{s_p/\sqrt{n}} \,\Big|\, M\right) = p\left(z = \frac{\bar{y} - \theta}{s_p/\sqrt{n}}\right) \tag{A.10}$$

where z is a unit normal deviate and

$$p\left(\frac{s^2}{s_0^2} \,\Big|\, M\right) = p\left\{\frac{\chi_\nu^2}{\nu} = \frac{s^2}{s_0^2}\right\} \tag{A.11}$$

So that it is only after applying the checks $c(\bar{y})$ and $c(s^2)$ as well as $c_j(\underline{u})$ that we could logically use Bayes theorem.

REFERENCES

[1] Anscombe, F. J. and Tukey, J. W. (1963), The examination and analysis of residuals. Technometrics, 5, p.141.

[2] Box, G.E.P. (1976), Science and Statistics, JASA, 71, p.791-799.

[3] Box, G.E.P. and Cox, D. R. (1964), An analysis of transformations. JRSS, Series B, 26, p.211.

[4] Box, G.E.P. and Jenkins (1976), Time Series Analysis: Forecast and Control. Holden-Day.

[5] Box, G.E.P. and Tiao, G. C. (1962), A further look at robustness via Bayes' theorem. Biometrika, 49, p.419.

[6] Box, G.E.P. and Tiao, G. C. (1968), A Bayesian approach to some outlier problems, Biometrika, 55, p.119.

[7] Box, G.E.P. and Tiao, G. C. (1973), Bayesian Inference in Statistical Analysis. Addison-Wesley.

[8] Coen, P. J., Gomme, E. D. and Kendall, M. G. (1969), Lagged Relationships in Economic Forecasting. JRSS, Series A, 132, p.133.

[9] David, H. A., Hartley, H. O. and Pearson, E. S. (1954), The distribution of the ratio, in a single normal sample, of range to standard deviation. Biometrika, 41, p.482.

[10] Dixon, W. J. (1950), Analysis of extreme values. Ann. Math. Statist., 21, p.27.

[11] Durbin, J. and Watson, G. S. (1950), Testing for serial correlation in least square regression I. Biometrika, 37, p.409.

[12] Durbin, J. and Watson, G. S. (1951), Testing for serial correlation in least square regression II. Biometrika, 38, p.159.

[13] Ferguson, T. S. (1961), On the rejection of outliers. Proceedings of the Fourth Berkeley Symposium, I, p.253.

[14] Feller, W. (1968), An Introduction to Probability Theory and its Applications. Vol. 1, Wiley.

[15] Grubbs, F. E. (1950), Sample criteria for testing outlying observations. Ann. Math. Statist., 21, p.27.

[16] Lindley, D. V. and Smith, A.F.N. (1972), Bayes' Estimates for the Linear Model, (w/ discussion). JRSS, B, 34, p.1-41.

[17] Pallesen, L. C. (1977), Studies in the analysis of serially dependent data. Ph.D. thesis, University of Wisconsin.

[18] Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium, 1, p.197.

[19] Tiao, G. C. and Ali, M. M. (1971), Analysis of correlated random effects: linear model with two random components, Biometrika, 58, p.37.

[20] Tukey, J. W. (1949), One degree of freedom for non-additivity, Biometrics, 5, p.232.

[21] Zellner, A. and Tiao, G. C. (1964), Bayesian analysis of the regression model with autocorrelated errors. JASA, 59, p.763.

[22] Box, G.E.P. and Tiao, G. C. (1968), Bayesian analysis of means for the random effect model, J. Amer. Statist. Assoc., 63, p.179.

[23] Box, G.E.P. and Newbold, Paul (1971), Some comments on a paper of Coen, Gomme, and Kendall, JRSS, A, 134, p.229.

[24] Box, G.E.P. (1979), Robustness in the strategy of scientific model building, Proceedings of A.R.O. workshop at Durham, N.C., April 1978, on Robustness in Statistics, Academic Press (to appear).

GEPB/ck/ed

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 1969
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): SAMPLING AND BAYES' INFERENCE IN THE ADVANCEMENT OF LEARNING
5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): George E. P. Box
8. CONTRACT OR GRANT NUMBER(s): DAAG29-75-C-0024 & DAAG29-78-G-0166
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 - Probability, Statistics, and Combinatorics
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709
12. REPORT DATE: June 1979
13. NUMBER OF PAGES: 28
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 (Continue on reverse side if necessary and identify by block number): Bayesian inference, Sampling theory inference, Estimation, Models, Predictive distribution
20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

Scientific method is a process of guided learning in which accelerated acquisition of knowledge relevant to some question under investigation is achieved by a hierarchy of iterations in which induction and deduction are used in alternation.

This process employs a developing model (or series of models implicit or explicit) against which data can be viewed. Ideally at any given stage

(continued)

of an investigation, such a model approximates relevant aspects of the studied system and motivates the acquisition of further data as well as its analysis. By the use of a prior distribution it is possible to represent some aspects of such a model as completely known and others as more or less unknown.

Now parsimony requires that, at any given stage, the model is no more complex than is necessary to achieve a desirable degree of approximation and since each investigation is unique we cannot be sure in advance that any model we postulate will meet this goal. Therefore, at the various points in our investigation where data analysis is required, two types of inference are involved: model criticism and parameter estimation. To effect the latter, conditional on the plausibility of the model, and given the data, we can, using Bayes' Theorem, deduce posterior distributions for unknown parameters and so make inferences about them. But, before we can rely on such conditional deduction, we ought logically to check whether the model postulated accords with the data at all and, if not, consider how it should be modified. In practice, this question is usually investigated by inspecting residuals, by other informal techniques, and sometimes by making formal tests of goodness of fit. In any case model criticism, the inferential procedure whereby the need for model modification is induced, is ultimately dependent on sampling theory argument. These principles are formalized by an appropriate analysis of Bayes' formula, and implications for robust estimation are considered.