Statistical Science
1987, Vol. 2, No. 3, 259–291

Uncertainty, Policy Analysis and Statistics

James S. Hodges

Abstract. Statistical activity can be divided for descriptive and analytical purposes into (a) discovery/imposition of structure, (b) assessment of variation conditional on structure and (c) execution of techniques. Each of these three areas of activity has an associated type of uncertainty, respectively, structural uncertainty, risk and technical uncertainty. In any statistical analysis, an analyst has limited supplies of time, money, knowhow and computational power and must use these resources to diminish and to characterize better the three main types of uncertainty and the many subtypes that comprise them. No existing school of statistical thinking provides a comprehensive framework for considering the various types of uncertainty and the tradeoffs among them that analysts must make. One result of this is the absence of a system that properly accounts for all of the types of uncertainty. This paper describes the types of uncertainty, catalogues and evaluates current methods as tools for characterizing and diminishing them, considers the types of tradeoffs that analysts must make in applying statistical methods in problems and examines the bias introduced into deliberations by the absence of a proper system of accounting for uncertainty. This paper is an attempt to begin the construction of such a proper system and thus to reduce or eliminate that bias.

Key words and phrases: Bayesian statistics, foundations of statistics, applications of statistics, model, prediction, uncertainty.

1. INTRODUCTION

> "If quantitative precision is demanded, it is gained in the current state of things, only by so reducing the scope of what is analyzed that most of the important problems remain external to the analysis."
>
> John Steinbruner (1974)

For decades, logisticians at the RAND Corporation have helped the Air Force devise methods for predicting the numbers of failures of parts used in its planes. The Air Force uses both short and long term predictions; long term predictions are used in the purchase of spares, and short term predictions are used in algorithms that schedule the repair and subsequent distribution of parts. Statisticians are involved in many phases of the work that culminates in these predictions. This statistical activity—and most if not all other applied statistical work—can be divided for descriptive and analytical purposes into three broad areas of activity. The three areas of activity are (a) discovery/imposition of structure, (b) assessment of variation conditional on structure and (c) execution of the techniques selected.

Structure, at once discovered and imposed, has several elements. Something of value is at stake: in policy problems this is clear, but it is also true in scientific activity, although the payoff is more diffuse. This value is usually captured in a loss or utility function. In the case of predicting parts failures, the payoff function is often the number of planes capable of executing missions on a given day. Several actions are usually available for choice in the pursuit of this payoff. Again, this is clear in policy problems, but even the purest scientific activity requires a choice among possible assertions and possible future observations. The actions can be small or large: predictions can be used to select the next part to fix or to pick a scheme for scheduling the repair of broken parts. Such selections can be made better if they are made using facts and beliefs about the nature of the relevant part of the world. These facts and beliefs often take the form of a model of some process central to the problem—in my example, the process that generates broken parts—and some more or less specific expression of belief about features of that model. The model might simply be a taxonomy of possible states of the world

James S. Hodges is Associate Statistician, Department of Economics and Statistics, The RAND Corporation, Santa Monica, California 90406.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science.*
www.jstor.org

with an assumption that future observables in a cell of the taxonomy are exchangeable with past observables, or it might be a tightly specified stochastic model, perhaps including a subjective probability distribution for the parameters of the model. In Air Force spare parts work, failures are usually treated as arising from a compound Poisson process (see Astrachan and Cahn, 1963; Hillestad, 1982; and references in those papers). Finally, this use of facts and beliefs is usually informed in some way by data. Data themselves have obvious structure—they can be discrete or continuous, for example—but there is more. The process that turns actual events into data can introduce systematic effects that may (and frequently must) themselves be accounted for or modeled. This can be considered a part of the process that is central to the problem, but it is useful and often crucial to consider it separately. Data used for Air Force spare parts predictions come from two main sources, namely systems used for tracking the actions of maintenance personnel and systems used for ordering supplies. These data were once treated as interchangeable and transparent descriptions of the process that generates broken parts, but we now know that the incentives facing the people in these systems shape the data that the systems produce. In some cases, it might not be possible to use data arising from the process of interest; the use of a proxy or analogous data source might be necessary. The error introduced into an analysis by a substitute data source is similar to the effects introduced by data collection systems and can be considered in an analogous fashion.

The second broad type of activity, assessment of variation conditional on structure, is the most familiar—at least, the most publicized—part of statistical work. This assessment can be understood as providing answers to two questions. The first question refers to the past: which of the possible structures (usually values of parameters of models) are more and less plausible? This is often considered under the rubric of estimation. The second question refers to the future: conditional on some structure (usually a model and parameter values), what can be said about future observable features of the modeled process? The link between facts about the past (data) and statements about the future is provided by the structure conditioned upon.

The execution of techniques, the third type of activity, occurs in concert with the other two types of activity, but it is distinct and will be considered separately. First, data must be processed. This includes extraction of items from data bases, conversion of raw numbers or characters into usable quantities, aggregation, counting and like activities, as well as computation of estimates and descriptive or diagnostic quantities. Second, in executing model fitting and prediction techniques, analytical or numerical approximations (or simulations, which are approximations) usually must be considered and often must be employed. In Air Force predictive work with which I am familiar, models and estimators have been chosen to avoid approximations—a choice with implications beyond the computational accuracy it allows.

Each of these areas of activity has an associated type of uncertainty. Along with the discovery and imposition of structure comes *structural uncertainty*: uncertainty about the accuracy of the model as a surrogate for the actual process of interest, about the transparency of the system that turns raw events into data and so on. This type of uncertainty and current approaches and techniques for characterizing and reducing it are discussed in Section 2.1. Assessment of variation conditional on structure is, in an obvious sense, the consideration of uncertainty about the past and the future conditional on a model. This type of uncertainty, *risk*, is considered in Section 2.2. (The terms "structural uncertainty" and "risk" come from Steinbruner, 1974, Chapter 1.) Risk in turn has two aspects corresponding to the forward and backward looking elements of the assessment of variation conditional on structure. Finally, execution of techniques entails uncertainty about inaccuracies introduced by repeated manipulation of raw data items, by numerical instability and by analytical or numerical approximations. This form of uncertainty, *technical uncertainty*, is discussed in Section 2.3. In any statistical analysis, an analyst has limited supplies of time, money, knowhow and computational power, and must use these resources to diminish and to characterize better the three main types of uncertainty and the many subtypes that comprise them. If an analyst is working as part of a team, it might be reasonable for him to concentrate on only one type. But for any sizable analysis to be complete, all three types of uncertainty must be assessed and their effects on the product of the analysis weighed.

No existing school of statistical thinking provides a comprehensive framework for considering the tradeoffs analysts must make in devoting resources to reducing and characterizing the three types of uncertainty, that is, for considering the strategy of statistical analysis. The tradeoffs analysts make, consciously and unconsciously, are considered in Section 3.1. One deficiency created by the lack of a comprehensive framework is the absence of a system that properly accounts for all of the types of uncertainty. I argue in Section 3.2 that, among other things, this creates an inherent tendency for analyses to understate uncertainty about predictions—about what is known—which can lead to invisible biases in policy considerations based on those analyses and can obscure the role of judgment and convention in the conclusions they produce.

Current theoretical approaches provide partial contexts appropriate for subsets of statistical activity. Steinbruner (1974) discussed structural uncertainty and risk without differentiating between the two aspects of risk. DeGroot (1982) and Cyert and DeGroot (1984) differentiated between the two aspects of risk. Fisher (1957) described "a series of meaningful mathematical statements ... each differing from the one before by its greater uncertainty," including statements regarding structural uncertainty and both aspects of risk, although not under those names. Alho and Spencer (1985) decomposed sources of error in population forecasts into specification error, error in parameter estimation and variability conditional on parameter estimates. Vasely and Rasmuson (1984) made a similar decomposition, with their "physical variability" and "parameter uncertainty" in rough correspondence with the forward and backward looking aspects of risk, respectively, and their "modeling uncertainties" and "completeness uncertainties" roughly corresponding to structural uncertainty.

The theory of probability presented in de Finetti (1974, 1975) comes closest to the goal of a complete context for statistical activity, in that de Finetti's approach is intended to be flexible enough to apply to any situation involving uncertainty. But if de Finetti's work is influencing research in statistical methodology and practice, he is not receiving much credit for it in statistical journals. For example, of the 377 papers in the 1985 volumes of *Biometrika*, *Journal of the Royal Statistical Society* (*Series B*), *Journal of the American Statistical Association* and *The Annals of Statistics* (excluding book reviews and corrigenda), the three papers that cited de Finetti (Dawid, 1985; Lane and Sudderth, 1985; Schervish, 1985) were on abstract topics with no obvious implications for statistical practitioners. The purpose of this paper is to bring de Finetti to those practitioners and begin the construction of language and concepts necessary for a more comprehensive framework for the use of data in policy analysis and other applications of statistics. (Leamer (1978) is an insightful pioneering effort in a similar vein, although his approach is quite different from the one taken here.) I take subjective uncertainty as a primitive concept and understand and use probability as a particular mathematical representation of subjective uncertainty, so the language and sensibility in this paper are largely Bayesian. Non-Bayesians need not be deterred, however, for the main thrust of the paper does not depend on Bayesian notions.

One note of caution: I use the example of Air Force spare parts predictions not because it contains bad examples, but because it is a big, expensive problem with which I am familiar. In fact, RAND's nonstatistical workers in this area are uncommonly sensitive to the many kinds of uncertainty they face that current statistical approaches cannot incorporate, and are leaders in their field in advocating and developing systems that do not depend much on stochastic model assumptions. Contact with their problems provided the initial motivation for the work in this paper.

2. THE THREE TYPES OF UNCERTAINTY

I will continue the example of Air Force spare parts prediction in this section. In doing so I have conceded part of the analyst's problem, as I have specified the context and the things to be predicted. These are not always given, and sometimes are a part of the analysis; for treatments of this by statisticians, see Mallows and Walley (1980) or Freedman (1985).

2.1 Structural Uncertainty

The elements of structure, described in the last section, are the payoff (loss or utility), choices, facts and beliefs organized as a model and the process that turns raw events into data. These correspond to the elements of the classical decision theory problem: loss function, actions, model and prior distribution and data. A large although diminishing fraction of statistical instruction and research treats these elements as if they are known without error. This is seldom, if ever, true.

In the spare parts example, the payoff function is usually the number of planes capable of executing missions on a selected day. In practice, this means the number of planes without "holes" (missing parts) on that day. Leaving aside discounting considerations, it is not obvious how we should regard a plane without holes. Avionics parts are diagnosed on automated test equipment that systematically misses certain types of failures. As a result, malfunctioning avionics parts are regularly treated as serviceable and used, producing planes without apparent holes that cannot execute their missions. It is not hard to imagine a probabilistic scheme for discounting the number of usable planes to account for this, explicitly incorporating the uncertainty about the actual payoff from having a given number of planes available, but no such scheme is used.

As for actions, it is possible to be uncertain about which actions are or will be available. Lateral resupply between airbases—a system of planes that move parts from bases that have them to other bases at which planes are inoperable for lack of those parts—is a way to hedge against inaccurate base-level predictions of part failures, and the United States Air Force has such a system in Europe. Many analyses have among their policy choices different levels of lateral resupply capability, measured by the number of days needed to move a part between bases. But the fleet of resupply planes will be subject to attack during a war; so the actual resupply capability will be subject to uncertainty.

Uncertainty about the accuracy of one's model as a substitute for the process of interest is too familiar to require elaboration. But a model that is satisfactory now might be deficient later, for the period for which predictions are to be made. For example, many spare parts predictions are for wartime; although we are unsure about the accuracy of current model as substitutes for the peacetime part failure process, we are even less sure about how much wartime failure behavior will resemble peacetime behavior. A common approach in Air Force logistical work is to estimate a failure rate for a part from peacetime data and use that rate and projected wartime flying programs to predict mean failures for projected war scenarios. Few workers in the field find this satisfactory, but none of the suggested improvements have garnered wide acceptance. Some of this uncertainty could possibly be captured, for short term predictions, with methods like those in Harrison and Stevens (1971, 1976), which are related to Kalman filters, but these methods do not address the difficulties in making long term predictions for setting stockage policy.

Finally, we are quite unsure of what to make of our Air Force data sources. The incentives facing the people in these data collection systems are coming to be understood, but we do not know how to reconcile the sometimes gross discrepancies between the descriptions of events provided by the two data systems. We must act, and these being the available data, we want to use them somehow, but it is not clear how to do it—in particular, how to allow for biases that the data contain.

The discussion so far has been typical of statistical discussions of models in that it has been oblique about how they are actually built or chosen. In practice, statistical model building is a compromise between plausibility and tractability so thoroughly influenced by personal style that even leaders in research into model building methods have difficulty describing it. For example, in the next to last paragraph of a book full of techniques for checking and elaborating models, Atkinson (1985) admits that "it is hard to see how to answer the question" of how "the overall strategy of model building [is to] be guided," i.e., how to use the techniques in his book.

Certainly the range of possible model selections is strongly conditioned by the set of models the analyst's software can handle and by the analyst's desire or ability to spend time and money developing custom software. Models favored by readily available programs tend to allow only linear causal relationships, and random variables are usually members of exponential families. The dominant position of these models notwithstanding, they are little more than conventions: they have become conventional through constant exposition in service courses and textbooks, through availability in popular software packages and because their mathematical tractability makes them inviting examples for scholars seeking to propagate new theory and methods.

That modeling is an inherently subjective activity has achieved some recognition; that it is constrained by the catalogue of conventional models provided by past researchers is a fact of life. This constraint has implications of particular relevance to policy analysis, which will be discussed in Section 3.2.

The purpose of elaborating on the Air Force example here was to illustrate the pervasiveness of uncertainty about structure; and yet no comprehensive or systematic method exists for characterizing or reducing structural uncertainty. This type of uncertainty *is* perceived as uncertainty, but in statistical research it tends to be handled separately from the other kinds of uncertainty, as if inherently different. (For one prominent but as yet only partially developed exception, see Berger (1984), who includes both likelihoods and priors in his approach to robustness.) Thus, the methods that exist for thinking about structural uncertainty tend to be understood in terms different from those used for the other two kinds of uncertainty. Statistical methods related to structural uncertainty are concerned mostly with models and data structure and those methods fall into two groups: (i) methods for reducing structural uncertainty by discovering structure in data, and (ii) methods for characterizing structural uncertainty to allow it to be propagated through the analysis of risk to the substantive conclusions. After a brief discussion about uncertainty in loss functions, these methods will be discussed in that order.

Uncertainty in the Loss Function. Uncertainty about the loss or utility function has received little attention from statistical researchers. DeGroot (1983) points out that in a Bayesian approach in which one proceeds by maximizing expected utility, the expectation of utility includes expectation with respect to the uncertainty in the utility function itself. If the uncertain aspects of the utility can be given a probabilistic representation, those probability distributions can be integrated out in the computation of the expected utility. DeGroot's emphasis is on sequential experiments for diminishing uncertainty about the utility function as well as about the process of interest, but the general approach is equally applicable to situations in which the actual utility is a stochastic outcome produced by a known mechanism.

Discovering Structure. The gold standard among scientists is carefully controlled experimentation replicated by independent researchers. Tukey and others have emphasized the value of extracting information about structure from data at hand. In many cases, action is needed before this kind of structural information can be superseded by something closer to the gold standard, and in these cases the issues are how best to extract the information and how much to discount it. In any event, methods for describing data and sifting it for structure are indispensable, particularly for understanding the effects of the data collection process. Uncertainty about bias induced by data collection methods can often be reduced or eliminated by consistency checks such as comparison with similar data from other sources (see, for example, Lagakos, Wessen and Zelen, 1986). The rest of this section will concentrate on methods of discovering structure using a given data set. Such methods can be divided for descriptive purposes into two types of approach, namely the data-analytic approach and the diagnostic approach.

The data-analytic approach is often associated with Tukey (Tukey, 1977; Mallows and Tukey, 1982). The object of this approach is to display or describe data and to sift it for patterns in the hope of uncovering relevant, strong, persistent structure or unexpected features that will prompt discovery of structure through means external to the data at hand. This is useful for learning both about the process central to the prediction problem and about the data collection process. In the latter role, these techniques are usually the first applied to a new set of data; they allow the analyst to make judgments about, for example, whether the data have been grouped or rounded, or indeed whether the data are so mangled as to remove any information they might have conveyed. (For an exposition of data descriptive methods emphasizing these judgments, see Chatfield, 1985). When the analyst gets to the stage of learning about the process central to the prediction problem, descriptive techniques can inform judgments about, for example, whether a single simple model will suffice for the data at hand, or whether some particular formulation of the unexplained variability (e.g., a symmetric distribution) is tenable.

Description of data takes the form of words, plots, other graphical summaries and numerical summaries (see Mallows (1983) for a theoretical exposition of data description). The apparently simple task of description is remarkably difficult for higher dimensional data, and it has become the subject of research only recently (see, e.g., Chambers, Cleveland, Kleiner and Tukey, 1982, Chapter 5). The sifting methods include straightforward techniques like smoothing and its generalizations (e.g., Hastie and Tibshirani, 1986), and less immediate techniques like projection pursuit (Huber, 1985) and the ACE algorithm (Breiman and Friedman, 1985). These methods in effect search very large spaces of structural models for one or a small number that capture the strongest relationship in the data between a dependent variable and a preselected collection of explanatory variables.

Clearly, techniques serving these purposes are indispensable, and good software (e.g., the S system, Becker and Chambers, 1984) includes many of them. In recent years, researchers have devised many new descriptive techniques, and in the absence of any widely accepted criteria for judging them, it is difficult to guess how useful these new techniques will be. Mallows (1983) suggests an approach to evaluating data description techniques that avoids probabilistic interpretations. Mallows (1983) and Chatfield (1985) argue that in many cases descriptive techniques are all that is needed or possible for an analysis, with Chatfield extending this argument even to judgment of the statistical significance of an observed effect.

The other of the two approaches to discovering structure can be called the diagnostic approach. This approach is sequential: one begins with an off the shelf model (e.g., a linear model with homoscedastic normal errors), then uses the data to test the assumptions of that model and to alter the model and the data until the model's form and assumptions are no longer seriously in conflict with the admitted data evidence. These tests include tests for the appropriate scale of the dependent variable (e.g., logarithmic or square root), and tests of the predictive usefulness of additional explanatory variables (see Weisberg (1985) or Atkinson (1985) for descriptions of these and many other tests). Methods of this type have been most highly developed for linear models with normal errors (e.g., Cook and Weisberg, 1982), although recently some progress has been made in extending them to generalized linear models (Atkinson, 1985, Chapter 11 gives a survey) and to parametric models generally (Cook, 1986).

These methods often take the form of hypothesis tests. In practice, however, they cannot be taken at face value as hypothesis tests with known operating characteristics (as their creators are quick to point out). In any given case they are applied in some unique sequence with other tests and procedures, so their actual frequency properties in that case are unknown. These methods are cast as hypothesis tests for lack of a better way to calibrate them, i.e., to think about what is big and what is small. Thus, although diagnostic methods look more theoretically sound than descriptive techniques, this appearance is not compelling. This is not to say that diagnostic techniques are not useful, only that they, too, lack a theoretical basis that would permit them to be judged. Weisberg (1983) offers the beginning of such a basis, in a collection of principles for the construction of diagnostic methods.

Box (1980), Berger (1984) and others (see Box, 1980, page 386 for citations) suggest using the Bayesian predictive distribution (described in Section 2.2 of this paper) as a general model diagnosis tool: if the observed data fall in the tail of the predictive distribution, then the model and the prior should be reconsidered or replaced. In Box (1980), this suggestion is carried through in one example (pages 386 and 387), in which the prior distribution is a characteristic of a physical process, not a representation of belief. Box appears to suggest its use for priors representing belief as well, and Berger (1984) does so explicitly. For the first of these two kinds of prior distribution, Box's suggested tail area differs little from a P-value in logical content. For priors representing belief, Box's suggestion has very different implications. Taken at face value, it could indicate that a perfectly accurate model should be discarded because the prior beliefs about its parameters happened to be off the mark for the period captured in the data. Implicit in this use of Box's idea, then, is some kind of sensitivity check on the prior. Secondly, if used as a means for evaluating prior distributions, it changes the nature of learning via Bayes' theorem, for it has the user first update his beliefs (prior) by checking the predictive distribution, then update them again using Bayes' theorem. The result is empirical Bayesianism through the backdoor, a convergence that appears not to have been either Box's or Berger's intent.

Diagnostic methods appear in introductory and intermediate statistics courses, and in some statistical packages, and they are used. They address a concern that troubles many people from their first encounter with statistics—what if my assumptions are wrong?—in a straightforward way, although they have not been integrated into either the Bayesian or the frequentist approach to analyzing uncertainty. In addition, Hampel, Ronchetti, Rousseeuw and Stahel (1986), representing the frequentist robustness school, argue that by frequentist standards, the diagnostic approach is inherently flawed because the analyst using it treats the last model he settles on as if it is correct, thus subjecting himself to avoidable risks of losses in accuracy and efficiency that robust methods are intended to minimize.

Characterizing Structural Uncertainty. The second of the two groups of statistical methods includes those used to characterize structural uncertainty to allow it to be propagated through the analysis of risk to the measure of uncertainty attached to the substantive conclusion. To introduce this idea, consider the recent issue of the *New England Journal of Medicine* (October 24, 1985) that contained two apparently sound papers on the effects of the use of postmenopausal hormones, which contradicted each other. In his editorial, Bailar (1985) discussed the two studies and suggested that the contradiction could have arisen because of the differences in the types of women eligible for inclusion in the two studies, or because "... the results of these studies (and by implication the results of countless other observational studies) are subject to a great deal more variability than is captured in the usual kinds of statistical tests and confidence limits" (page 1081). After listing what are, in effect, a number of possible deficiencies of the models underlying the two statistical analyses, Bailar concluded that "[s]uch problems would lead to the improper calculation of error probabilities and confidence limits" (page 1081).

This and Bailar's suggested explanations are examples, described from a frequentist viewpoint, of structural uncertainty that was not propagated through the usual analysis of risk. The results of the analyses of risk (hypothesis tests, confidence intervals) in the two studies were treated as if the models on which they were based were known to be exactly true, with no account taken of the likely deficiencies of those models.

Statisticians have taken several approaches to characterizing and propagating structural uncertainty. Frequentist statisticians have difficulty fitting the idea into their scheme, because their approach is highly dependent on deducing repeated sampling properties from known distributional assumptions. This difficulty is illustrated by the controversy over the appropriate standard error to use for regression coefficients estimated after applying the Box-Cox method for selecting a power transformation (Bickel and Doksum, 1981; Box and Cox, 1982; Hinkley and Runger, 1984). Exponents of this school seem to be more comfortable with sample re-use approaches to characterizing structural uncertainty; for example, see Freedman and Navidi (1986), in which a sample re-use method is used to attack the model that Ericksen and Kadane (1985) propose as a method for adjusting the United States Census, or Efron and Gong (1983), which gives a bootstrap demonstration of the instability of a common variable selection method used in a medical prediction problem. But these methods have only been developed as tools for criticism. As Efron and Gong put it (page 48), "[n]o theory exists for interpreting [this bootstrap demonstration of instability], but the results certainly discourage confidence in the causal nature of the predictors" naively selected by the common method.

In recent years, Freedman has attacked several studies on the grounds that they take insufficient account of structural uncertainty, arguing that those studies are misleading or worse because of that flaw (cf. Freedman, 1981; Freedman, Rothenberg and Sutch, 1983; Freedman and Navidi, 1986). I can hardly agree more with his general position; it is one of the main points of this paper. But Freedman's tactics have provoked substantial resistance, captured in a caricature of his approach to structural uncertainty in Dempster's and Madansky's discussion of Freedman and Navidi (1986). According to this caricature, if a model for the process of interest is acceptable, then (frequentist) statistical methods can be used to make forecasts and to attach a measure of uncertainty to the forecasts. If no model is acceptable (an unspecifiable standard), then statistical methods should not be used to aid thinking about making predictions and choices based on those predictions, and statisticians are obliged to defend their discipline's virtue by denouncing attempts to do so. But Freedman is too subtle for this position. He shows acute awareness of the various types of uncertainty described in this paper—the first two references above are good catalogues of these types of uncertainty in energy policy modeling. And as the conclusion to Freedman (1981) shows, he is well aware of the need to form judgments as a basis for action:

> "When the basic theory is incomplete, or the data sparse, ad hoc analysis by experts may be better than a large scale econometric model. [footnote omitted] In some cases, it may be still better to tell the policymaker that his question is unanswerable. This might prompt a search for policies which do not depend on knowing the unknowable."

But this raises more issues than it resolves. How are the results of these ad hoc expert analyses to be formulated? Surely they must not omit assessments of the uncertainty that the experts attach to their predictions. (In what sense are the energy policy models Freedman attacks *not* ad hoc expert analyses, albeit elaborate ones?) How should the judgments of differing experts be combined? Because no policy can be robust against all possible occurrences, and because robustness costs, how should information or beliefs about the relative plausibility of possible future outcomes be used to evaluate robust policies? Of particular relevance here, how should data and data reduction techniques be used to inform all of these judgments? It is difficult to see how these questions can even be posed within the frequentist framework.

The best known theorists of robust statistical methods have tried to formulate some types of structural uncertainty and incorporate them into an approach to selecting estimation procedures within the frequentist interpretation. Huber's pioneering approach (1964), done within a frequentist decision-theoretic framework, evaluated the properties of decision procedures for all distributions within a neighborhood of a parametric model. Using asymptotic variance as his loss function, Huber took a minimax approach, minimizing the maximum risk over all distributions within the neighborhood, instead of just for the parametric model at the center of the neighborhood. Hampel (Hampel, Ronchetti, Rousseeuw and Stahel, 1986, Chapter 1) began with the idea of neighborhoods around models, but took a different approach, concentrating on first-order characterizations of the effects of model changes on the results produced by decision rules. The key to this characterization is the influence function, from which Hampel and his colleagues derived measures of gross error sensitivity (maximum estimation bias caused by an infinitesimal change in distribution), the asymptotic variance of estimation and similar quantities. In this approach, optimal estimators are found by choosing the measures for which one's estimator must do well, and finding the class of estimators within which any improvement with respect to one measure requires worse performance with respect to another.

The novelty in these two approaches was the introduction into the estimation problem of some uncertainty about the accuracy of the model, particularly about gross errors in recording or processing observations but also about other features of distributional shape. Both approaches require judgment in the selection of the parametric model whose neighborhood is to be considered. Huber's approach requires another judgment of the appropriate size of the neighborhood. Hampel's approach requires two other judgments, namely which measures are to be used in the optimization, and which member of the resulting optimal class is to be used. As with any frequentist procedure, there is no way within these approaches to account for the uncertainty one attaches to these judgments—and after all, the neighborhoods permitted by these methods are not capacious in general, so that the model could be so far off that the minimax procedure would give essentially no protection. Also, these methods do not obviate the need for diagnostic model checking, so the theoretical problems mentioned for diagnostic methods are present here as well. An equally important difficulty is that these methods are thoroughly oriented toward parameter estimation. I know of nothing in this literature about applying these methods to predictions, in particular, about introducing the effects of model uncertainty into the measure of uncertainty attached to a prediction. Given the awkwardness with which frequentist theory accommodates predictions (see Section 2.2), extending this approach to predictions will be difficult if it is possible.

Berger (1984) arrives at a conclusion similar to Huber's, beginning from a Bayesian decision-theoretic viewpoint. His main concern is that prior information cannot be specified with perfect precision, and that the preferred decision rule can depend on features of the prior distribution chosen by convention or for convenience. (Berger mentions that the choice of a model is an expression of prior belief, so that uncertainty about models would be included in his approach, but he does not develop this and concentrates on the more traditional kind of prior distribution.) Like Huber, Berger proposes defining a neighborhood of distributions around a prior that captures as much information as the user can specify, and finding the minimax Bayes rule within that class of priors. This approach is similar to Huber's, differing in the emphasis on prior distributions (as opposed to likelihoods) and on introducing substantive information into the selection of the neighborhood around the nominal prior. In contrast to the frequentist robustness approach, however, this Bayesian approach can be extended to predictions trivially, in theory, via the Bayesian predictive distribution (see Section 2.2). Instead of concentrating on parameter estimation, the method would concentrate on the predictive decision problem, but otherwise the approach would be identical.

Both of these minimax approaches are special cases of a much broader technique, namely sensitivity analysis. Of all the techniques related to model uncertainty, this is the only one applicable to elements of structure other than the model and data structure. In a sensitivity analysis, the analyst varies the details of his specification of structure to see if the conclusions depend on those details. Many statisticians prefer to cast their treatment of structural uncertainty explicitly as sensitivity analysis. For example, in Cook's (1986) scheme, the effects of a broad range of model expansions (including differential case weighting and case deletion) can be examined by considering how they change the maximum likelihood estimate of the parameter of interest. Changes in the maximum likelihood estimate are calibrated by inserting the estimate under the expanded model into the log likelihood under the original model and subtracting the new log likelihood from the original maximum. Johnson and Geisser (1982, 1983) and McCulloch (1985) use Bayesian predictive distributions for a similar calibration. These researchers concentrate on abstract, nonsubstantive measures of change, like curvature and change in log likelihood (Cook) or Kullback-Liebler divergence (Johnson-Geisser and McCulloch). Dempster (1975) emphasizes the substantive context of the sensitivity analysis: "judgments [of sensitivity] can only be made relative to assessments of ranges of failures in the accuracy of the data and in the validity of the models (i.e., failures deemed plausible enough to rate concern) and relative to assessments of the effect of such failures on the import of the analysis" (page 1).

Dempster's approach to sensitivity analysis differs from the other approaches in emphasizing the substantive judgment. But like the others, he offers little advice about what to do if the analysis is sensitive to changes in the specification. Berger (1984) says that if the outcome of your analysis is sensitive to which of the priors in your neighborhood you use, you should seriously consider not drawing any conclusion. Vasely and Rasmuson (1984) take this position for other uncertain features of the specification as well, arguing that the sensitivity analysis is the appropriate form for the product of the quantitative analysis. This brings us back to the question raised in the discussion of Freedman's approach: how is this information to be used, that is, how is this manifest uncertainty to be used in the eventual decision?

One straightforward approach motivated by Bayesian thinking is to place a probability distribution on the area of remaining uncertainty (e.g., over Huber's neighborhood of distributions) and integrate it out. In Box and Tiao (1962), this is accomplished by expanding the model (in their example, by adding a kurtosis parameter in a two-sample location parameter) to capture the range of plausible model uncertainty. Regression diagnostic techniques often produce a range of distinct models having different scales and collections of regressors; Box and Tiao's model expansion approach can accommodate such a range of models, but does so awkwardly. An alternative suggested by Leamer (1978), Zellner (1984) and others is to assess prior probabilities for the distinct models, update those probabilities via Bayes' theorem if appropriate and use these probabilities to mix the predictive distributions from the distinct models. Special cases of this approach are suggested by Box (1980), Harrison and Stevens (1976) and Smith (1983). Berger (1984) mentions it, but dismisses it as generally intractable.

This solution repels many people because, in the words of an anonymous referee, the apparent result is that we will "build layer on layer of ever more remote concepts of uncertainty." (Shafer (1986) discusses this as a manifestation of "the conditional probability fallacy," which he considers a telling flaw in the Bayesian scheme.) It is an unavoidable feature of the Bayesian method that at some point the user must express a *judgment* as a probability distribution without further qualifications about its uncertainty. The issue is then the level at which to do so. This decision cannot be made without considering the context of the particular substantive problem, specifically the importance of the remaining structural uncertainty relative to the other types of uncertainty, conditioned by the constraints of budget and time that are present in any analytic exercise (but especially policy analytic exercises). It is clear that failing to propagate structural uncertainty biases a policy choice in favor of policies that rely on more certain information. At this point in the development of statistical method, the practical issue is whether this bias ever matters, and if so, when.

This is an empirical question, and it can be approached empirically. This empirical approach, reminiscent of Tukey (1962), can be begun with tools that are available now, by re-doing analyses, allowing a range of models and mixing the corresponding range of predictive distributions according to judgments (updated by the data) about the likely accuracy of the various models. Researchers in several fields—in Air Force and Army logistics at RAND, in nuclear missile accuracy (Bennett, 1980), in broader matters related to nuclear war (Blackett, 1962), in quantitative risk assessment (Hattis and Kennedy, 1986; Hattis and Smith, 1985; Vasely and Rasmuson, 1984)—express the belief that structural uncertainty matters. These beliefs have been reached without tools that permit the effect of ignoring structural uncertainty to be expressed explicitly. Draper and Hodges (1987) pursue the approach of mixing predictive distributions for oil price predictions.

2.2 Risk

Having tentatively conditioned on some structure for making predictions, which usually includes a class of stochastic models with some unspecified parameters, the analyst can then use his data to differentiate among members of the class of models as more and less plausible by estimating the parameters or computing confidence regions or posterior distributions for them. Armed with this information about the plausibility of different parameter values, he must then account for the future stochastic behavior of the model conditioned upon.

These are the two elements of risk, the second broad type of uncertainty. (This distinction was made, using different terms, in DeGroot (1982) and Cyert and DeGroot (1984).) One element refers to the past: if we assume, in my running example of predicting airplane parts failures, that the monthly counts of F16 radar failures are realizations of independent Poisson random variables with mean λf_j, where f_j is the number of hours F16s flew in month j and λ is an unknown positive constant, then our data contain information about which values of λ are more and less plausible. This element of risk will be called "estimation risk." The second element of risk refers to the future: even if next month's count of radar failures is a Poisson random variable with a known mean, the number of failures next month is still uncertain because it is a random variable. This element of risk will be called "prediction risk."

Estimation risk is the most familiar of the kinds of uncertainty discussed so far. It is the central concern of most statistics and econometrics instruction and research. Huge bodies of theory and method have been developed for using data to make statements about parameters. Most of what analysts report—parameter estimates, hypothesis tests, confidence regions—addresses this element of risk.

But very little effort is devoted to prediction risk. Many students see it in an elementary class, in an example like the following. Suppose that the observables x_j, $j = 1, 2, \ldots, n, n + 1$, are postulated to be independent normal random variables with mean θ and variance 1, where θ is an unknown and unobservable parameter. Then the average of the first n observations, $\bar{x}$, is a normal random variable with mean θ and variance $1/n$, so that $x_{n+1} - \bar{x}$ is normal with mean 0 and variance $1 + 1/n$, independently of θ. This variance reflects the two elements of risk: $1/n$ and 1 allow for estimating and predicting, respectively. From this modest sleight of hand, it follows that the absolute value of $x_{n+1} - \bar{x}$ will be less than 1.96 $(1 + 1/n)$ with probability 0.95, using the frequentist interpretation of probability.

This example is useful because it conveys the difference between the measures of uncertainty to be attached to an estimate of θ and to a prediction of a new value of y, the latter being larger. In spite of the importance of this distinction, even some of the best statistics departments do not teach it. For example, a colleague of mine got his Ph.D. from the Berkeley statistics department—the flagship of the so-called American school of statistics—in 1981 without ever hearing of it. This is a reflection of the discipline's orientation toward inference about parameters, a topic to be discussed further in Section 3.2.

This orientation notwithstanding, policy analysts (in particular, but many other users of statistics as well) don't need to know if θ is greater than 3, they need some information about the likely value of some future observable. Statistical researchers have provided few practical tools for assembling this information—for combining information about estimation and prediction risk into a single statement of information about the future, given the past—so users must improvise. In the software used at RAND to study the Air Force spare parts system, estimates of parameters of assumed failure models are used as if known without error, thus ignoring the substantial estimation uncertainty and systematically overstating the certainty of predictions of failures. This introduces a systematic bias into the policy deliberations, as I will argue further in Section 3.2.

Although they are rarely implemented in software, many techniques have been devised for producing statements that combine information about the two kinds of risk. Analytical frequentist tools, like the example above, have been developed for situations where the same trick or an analogue to it can be used (Geisser, 1980a). This is not a large collection of situations. Sample re-use methods, such as cross-validation (described in Stone (1974) and Geisser (1980a), Section IV, and implemented in the Classification and Regression Tree [CART] program of Breiman, Friedman, Olshen and Stone, 1984) and the bootstrap (see Efron and Gong, 1983, or Efron and Tibshirani, 1986) are much more widely applicable, although I do not know of many predictive applications of them.

The Bayesian approach, in theory, permits predictive statements to be made regardless of the model. The vehicle is the predictive probability distribution (Geisser, 1971, 1980a) for the future observation. If $f(x \mid \text{model}, \theta)$ is the probability function or probability density of the observable x given the model and its parameter θ, and $p(\theta \mid \text{model, data})$ is the posterior distribution of θ, then the predictive distribution of x, conditional on the data, model, and prior distribution, is

$$f(x \mid \text{data, model, prior}) = \int f(x \mid \text{model}, \theta) p(\theta \mid \text{model, data})\, d\theta$$

if θ has a continuous probability distribution or the obvious analogue if θ has a discrete probability distribution.

Predictive distributions are not novel, but they are rarely used in applications; the paper by Duncan and Lambert (1986) is the only example I can find. This has happened at least partly because, although in theory they can be computed for any model, in practice the necessary integral or sum is usually intractable. A first order approximation for mixed or unmixed predictive distributions, akin to the usual large sample likelihood approximation and as easily computed, can be derived without difficulty (Draper and Hodges, 1987). Some better but more elaborate approximations (e.g., Lee and Geisser, 1972; Tierney and Kadane, 1986) and numerical methods (Smith, Skene, Shaw, Naylor and Dransfield, 1985) have been developed, but they are not readily available yet and their performance in practice has not been widely tested.

2.3 Technical Uncertainty

The third of the broad areas of statistical activity is the execution of technique. As noted in the introduction, execution has two subsidiary areas of activity, namely the processing of data and the application of approximations.

Data processing can, in turn, be broken into two activities: manipulating data for input to substantive algorithms, and application of those algorithms. Careful file manipulation requires familiarity with the data collection systems, to understand the nature and meaning of the data items, and sound practices of data handling. The object of expending resources on these practices is to gain some confidence that the product of the processing is not garbage—a real concern when working with massive data files collected for an administrative purpose. The selection of substantive algorithms is important both for the cost of using them (including the time and effort they require of the user) and for their numerical stability when applied to data (i.e., the certainty one has that they produce meaningful results). I consider the latter under the application of approximations.

Each of these aspects of data processing leads to a substantial field of inquiry, but for the purposes of this paper a few things are immediate. First, if data aren't processed properly, subsequent work with the processed data is pointless or excessively difficult. Second, people who work with real data on real problems devote tremendous resources to turning raw data into usable files. Within RAND, statisticians who specialize in this area are in constant demand. Finally, although statisticians and subject matter researchers must and do make tradeoffs between devoting resources to careful data handling and devoting resources to the other, better publicized kinds of uncertainty, we have little theory to guide these tradeoffs. Relles (1986) covers some of these issues; Spencer (1985), although not directly related, addresses problems similar to those discussed here.

The second subsidiary activity within execution is the application of approximations. Approximations and numerical methods are unavoidable. But these technical aids introduce their own uncertainty. Analytical approximations are inexact, usually by an unknown amount; many numerical optimization routines find local optima and may not find global optima; optimization routines can, particularly for higher dimensions, "get lost" in subspaces or in flat spots of the function being optimized. From the analyst's point of view, uncertainty about models or about the values of the parameters of those models, and uncertainty about the accuracy of an approximation produce the same effect: they diminish the confidence with which he can predict.

But although it is standard practice to impress on statistical consumers the measure of uncertainty attached to a parameter estimate, it is not standard to do the same for the uncertainty attached to an approximation. Consider the popular software packages. Except for the normal linear model, maximum likelihood estimates are actually local maxima of the likelihood, and the "t statistics" for those estimates are computed from the usual large sample normal approximation. It is not common practice for documentation to inform users of either the local optimization or the normal approximation. But statistical consumers do consider technical uncertainty when selecting techniques. A few years ago, RAND's computation center changed its billing algorithm to make off-hours computing free, and users responded with much greater willingness to replace analytical approximations with Monte Carlo approximations, for which operational assessment of accuracy is usually simple. Given that these users could have used Monte Carlo approximations when night computing was not free, this change of behavior does reveal a belief that approximation error is a lesser concern than other things on which a computing budget can be spent. Now that a nominal charge has been introduced at RAND for night computing, however, the persistent interest in Monte Carlo approximations suggests that knowledge about approximation error is worth something.

The technical barriers that force the use of approximations have been besieged by a vigorous and fruitful research effort in recent years. Saddlepoint approximations and higher-order Edgeworth expansions have been used to improve on workhorses like the usual large-sample normal approximation (e.g., Durbin, 1980). Differential geometry is being used for many purposes related to curved exponential families in general and nonlinear regression in particular (e.g., Bates and Watts, 1980; Cook and Goldberg, 1986). New numerical integration techniques should make possible the routine use of Bayesian methods for nonconjugate priors (Smith, Skene, Shaw, Naylor and Dransfield, 1985, Section 3; Tierney and Kadane, 1986), and others are emerging (e.g., Tanner and Wong, 1987). But except for the Bates-Watts and Cook-Goldberg work on nonlinear regression models, papers by Minkin (1983), Jennings (1986) and Hodges (1987) on normal approximations, and the dissertation by Jones (1986) on computation errors, this work has produced better and more costly approximations, not better ways to think about the tradeoffs involved in the use of approximations.

Bates and Watts (1980) give two measures for assessing the accuracy of inferences made using the usual normal approximation to the distribution of the maximum likelihood estimate in nonlinear regression. Minkin's methods give bounds on the nominal confidence coefficient of elliptical approximations to likelihood regions, and Hodges' approach gives two measures of the accuracy of a wider range of normal approximations (one of these is a refinement of Jenning's method). Although the methods of Hodges and Minkin use nominal confidence coefficients and probabilities, invoking probabilistic intuitions about uncertainty, neither they nor the Bates-Watts methods provide an explicit way to evaluate the importance of technical uncertainty, either relative to the other kinds of uncertainty or in the ultimate policy choice. Jones' method (1986) would allow this incorporation, but as yet it has only been applied to the computation of sums.

I frequently hear the opinion that these kinds of inaccuracy are not important, that they are an order of magnitude smaller than the other kinds of uncertainty. This is an empirical proposition—one without the support we would demand for a scientific assertion. This proposition could be assessed by an exercise similar to the one proposed in Section 2.1, by gathering a collection of real problems, on which standard models and common approximations were used, and evaluating how far off the approximations really were. With such a collection of problems, we could begin to classify situations (a situation being a combination of a model, sufficient statistics and an approximation) in some manner useful for routine operational assessment of the accuracy of approximations. This would be nothing but a more formal version of what practicing statisticians now do informally, but it would be available to all and it would put a more dependable basis under rules of thumb.

2.4 Persistence of Structure and the Value of Effort

So far, my discussion has been typical in that it relies on the idea of an unchanging true mechanism out there in the world, that generated the data and will generate future observables and whose nature is at least partially discernable with available data. On this foundation, we can build assessments of the uncertainty of predictions, assembling uncertainty about models (within a class capturing important doubt about the true mechanism), about parameter values given a model, about future observables given a model and parameter values, about inaccuracies of approximation and about noise introduced by data processing. But as a framework for assessment of prediction uncertainty, this construction is incomplete and potentially misleading. The true mechanism can change or the mechanism can remain unchanged, but some condition that was fixed when the data were generated could change before or during the period being predicted. In our Air Force work, we have data from peacetime, and must make predictions for wartime.

We know that the failure processes can differ substantially under these two conditions. Data collected during intensive exercises are sometimes useful, but they do not solve our problem: the circumstances of an exercise, e.g., whether the accuracy of dropped bombs is measured, have a strong effect on whether part failures are reported during the exercise, and thus on the data produced by the exercise.

Implicit in the above framework is an assumption of *persistence of structure*, i.e., that the elements of structure, particularly the element captured in the model, persist through time. To illustrate the idea further, if a breakthrough is made in the understanding of a chemical process, that breakthrough will decrease uncertainty in the prediction of outcomes involving that process wherever and whenever the appropriate conditions can be created. The structure captured by this new understanding persists. But a breakthrough in the understanding of the U.S. economy in 1825 does not necessarily reduce the uncertainty of any predictions anywhere; in particular it does not necessarily improve our ability to predict how any part of the U.S. economy will perform in 1990, say. The output from a given set of inputs to the U.S. economy would be different in 1990 than in 1825 because the relevant structure does not persist, it changes. This is more than a matter of changing parameter values; the structure itself changes.

This is not a new idea. Keynes, in his exchange with Tinbergen (Keynes, 1939, 1940; Tinbergen, 1940) about models of national economies, said that

> "[p]ut broadly, the most important condition is that the environment in all relevant respects, other than fluctuations in those factors of which we take particular account [by including them in the model], should be uniform and homogeneous over a period of time. We cannot be sure that such conditions will persist in the future, even if we find them in the past. But if we find them in the past, we have at any rate some basis for an inductive argument."

In policy analysis—where we cannot wait a hundred years for the right theory—and with the easy availability of "curve fitting" methods, which some users misinterpret as absolving them of the obligation to understand what they are modeling (see Hattis and Smith, 1985), persistence is more deserving of attention. Without an assumption of persistence of structure, many popular and useful techniques—including Box-Jenkins modeling and smoothers, for example—are not very interesting. But further, the degree of persistence one is willing to assume places a limit on one's ability to reduce structural uncertainty and risk in predicting—one of Keynes' points in his critique of Tinbergen. If the relation of future structure to present and past structure is highly uncertain, perfect knowledge of present and past structure will not give much certainty to predictions. Even if data on the present and the past were perfectly collected and unambiguously interpretable, if structure does not persist little could be gained by applying elaborate, costly techniques to those data.

Degrees of persistence of structure can be represented in the Bayesian approach. As suggested in Section 2.1, it is possible to mix predictive distributions from different structures that capture the range of plausible possibilities for the future period of interest. A version of this idea is presented in Harrison and Stevens (1971, 1976), and West (1986), and the idea could be partly captured within an expansive model like the vector autoregressive models of Litterman (1986).

3. STATISTICS AND POLICY ANALYSIS

When I describe the scheme in Section 2 to researchers at RAND, it is well received because it corresponds to their experience: the available data are usually seriously deficient and distressingly scarce, models in the literature are not particularly plausible but are imbued with respectability by customary usage, and approximations are ubiquitous. But time is short, and clients are often analytically unsophisticated and not particularly interested in the elaborate brand of equivocation in which statisticians specialize. A theoretical scheme to incorporate the three kinds of uncertainty is interesting, but how can a practicing analyst use it?

The scheme presented in Section 2 is valuable for at least two reasons. First, it provides an explicit framework for considering the strategy of an analysis, for weighing the tradeoffs made in devoting resources to diminishing or characterizing the different types of uncertainty. This is examined in Section 3.1. Second, it provides the basis for a proper system of accounting for uncertainty, and extends statistical language to include many problems that currently must be handled outside it. This is examined in Section 3.2.

3.1 Strategy of Analysis: Making Tradeoffs

People who use statistical methods to extract information from data constantly make tradeoffs among the three kinds of uncertainty. I do not suggest that anybody try to construct a formal scheme for capturing these tradeoffs (although such a scheme might have a place in the economics of information), mainly because it immediately creates a problem of infinite regress. Rather, the point is that it is beneficial to be aware of tradeoffs that must be made, and preferable to be explicit about them.

For example, until recently the models of the part failure and repair processes used at RAND were chosen to permit analytical calculations. This was an explicit acceptance of diminished realism of the models (more structural uncertainty) in return for increased economy and precision in calculation (less technical uncertainty). If you prefer, a larger risk of a prediction error induced by the representation of structure was accepted in return for a smaller risk of error induced by the use of approximations. Software under development reflects a reversal of this choice: the structures needed for analytical calculations are too restrictive to counterbalance the diminishing computing cost needed to get precise answers from Monte Carlo simulations.

But either of the choices in the last paragraph can be justified, depending on the circumstances. Software plays a crucial role here. The GLIM system (Baker and Nelder, 1978), for example, permits the following kind of choice: by restricting himself to generalized linear models, an analyst gains the ability to consider a large range of models easily and cheaply. That analyst might accept an increment of structural uncertainty by restricting himself to the models handled by GLIM, but the ease and thrift of GLIM allow him to consider a greater range of models, thus reducing the increment.

Many times at RAND we have data sets that are so enormous (e.g., the entire Medicare case file for several years) that expensive logistical problems can be avoided with little loss by using samples of the data set. This is obvious for exploratory work—who wants a scatter plot of ten million points?—but it also holds for the products of the analysis, say, for confidence intervals. A data set can be so large that a 20% sample of it will still give confidence regions small enough that the increase in size over full sample regions cannot be detected among the other uncertainties in the problem (i.e., the region is still too small to be believed). In such a case, the substantially reduced cost and risk of processing errors can justify using the sample instead of the whole data set.

Similarly, an analyst might choose to avoid iterative methods. This might mean accepting estimation methods with larger standard errors or approximation error in return for fewer problems with iterative procedures (e.g., convergence) and lower cost. This tradeoff is implicit in the methods of West, Harrison and Migon (1985) and West (1986). This choice becomes particularly important when the data set is large and the iterative procedure requires one pass through the data for each iteration.

Anyone who applies statistics to real problems can add examples to this list—compromises like these are part of the statistical common sense that practitioners develop. But in the absence of a common measure for all the types of uncertainty, it is difficult to apply to these tradeoffs what we know about allocating resources among competing demands. The scheme in Section 2 offers a possibility, with the common measure being predictive uncertainty expressed as probability.

3.2 Accounting for Uncertainty

The absence of a system of accounting for uncertainty and for analytical tradeoffs creates several problems that do not fit into current statistical theories. With the scheme presented in Section 2, some of these problems can be treated as statistical problems, as they should be.

To begin with something familiar, as part of an analysis a statistician might be inclined to expand his model by adding parameters. If he uses the usual statistical framework and tools, he converts previously uncounted structural uncertainty into counted estimation risk, reducing his unexplained residuals in an exercise restrained only by consideration of the vague evil of "overfitting." I have never seen a serious attempt to define overfitting, but the notion is operationalized in some smoothing techniques through penalized likelihoods, in which a penalty for the roughness of the smoothed fit is added to the log likelihood (as entry points to this large literature, see Leonard, 1978; O'Sullivan, Yandell and Raynor, 1986). The notion of overfitting that is almost explicit in Leonard's paper is that past a certain point, choosing a fitted model (in his case, a density estimate) closer to the observed data means choosing a model that is less probable according to the prior distribution of densities in the space of continuous functions. The extension of this idea is natural within the scheme presented in Section 2, and hardly needs to be reworded at all. Overfitting occurs when movement from a less to a more elaborate model means moving to a model with lower posterior probability (in spite of smaller residuals, for example). This fits naturally into the mixing approach suggested in Section 2.1, in which the predictive distributions corresponding to the various models are mixed according to the probability assigned to the models. A model believed to be overfitted would be assigned a relatively smaller probability for that reason, and accordingly it would contribute less to the mixture. This is analogous to the effect produced by shrinkage estimators, some of which have explicit Bayesian interpretations.

Another problem permitted by the absence of an accounting system is model stereotyping: some fields develop stereotypical models or modeling approaches. For example, Builder (1986) contrasts the modeling styles of the United States armed forces: Army modelers prefer highly detailed models, whereas Air Force modelers let the level of detail depend on the problem. Economists, for another example, work hard to express ideas in operational forms, to divine behavioral relationships and to gather data, and all too often dump this work into a linear or log-linear regression with scarcely a second thought. As an illustration, Feldstein's (1975) policy for eliminating wealth effects in local education spending depends entirely on his unexamined assumption of a log-linear regression. If the log-linear form is incorrect, his recommendation is as likely to produce perverse wealth effects on local education spending as the competing policies he criticizes. In a similar fashion, military logistics modeling has been stalled for decades, unable to move beyond compound Poisson models. Clearly, structural uncertainty deserves more attention, and within the scheme presented in Section 2, it gets it.

There is another more subtle effect like this. A model can develop momentum: people working on a problem become accustomed to it, larger models are built on it, computer programs are written for it and the language of workers in the field can even come to be defined in its terms, which may not be meaningful if the model is badly inaccurate. All of these things have happened in the forecasting of Air Force spare parts requirements. A similar effect has occurred with models of conventional forces in the United States Army, which are used to make decisions about purchasing, organization, doctrine and training (Stockfisch, 1975). Again, the inappropriateness of this tendency is clear in a scheme that treats structural uncertainty like it treats risk.

Finally, the absence of a proper system of accounting for uncertainty makes it difficult or impossible to attach a believable measure of uncertainty to a prediction. In the Air Force work I have described in this paper, structural uncertainty, estimation risk and technical uncertainty are ignored in numerical calculations. Hattis and Smith (1985, Section 3.1.3) describe a similar practice in quantitative risk assessment. In probabilistic risk assessment for nuclear power plants, uncertainty about the consequences of an accident is not propagated—is treated as if it doesn't exist—precisely because it is so great (Vasely and Rasmuson, 1984). Ignoring these kinds of uncertainty amounts to acting as if more is known than actually is. This introduces a consistent bias into policy considerations based on these calculations, because the efficacy of some policy options—like prepositioning of spare parts and repair facilities—depends on knowing where and when part failures are going to occur, although others—such as making a heavy investment in lateral resupply capability—do not.

This problem also manifests itself in the pervasive orientation toward parameter estimation mentioned in Section 2.2, and the resulting widespread use of statements about parameters for predictive purposes. For example, Ehrlich (1975) postulated utility maximizing behavior by murderers to motivate a Cobb-Douglas "production function" for the rate of murders in the United States as a function of the rate of executions, among other things. (For a description of Cobb-Douglas production functions, see Mansfield, 1979, page 150.) In this specification, the deterrent effect of capital punishment is captured in α_3, the elasticity of the murder rate with respect to the execution rate. (This elasticity is $(\partial Q/\partial E)(E/Q)$, where Q is the murder rate and E the execution rate. See Mansfield, 1979, page 24.) Using aggregate data for the United States for the years 1933–1969, and some minor variations on his specification, Ehrlich got estimates for α_3 ranging between -0.039 and -0.074, with ratios of estimates to approximate standard errors ranging between -1.59 and -3.82.

Ehrlich interpreted this as evidence that capital punishment has a deterrent effect. Using one of his estimates for α_3, he then predicted that eight potential murder victims would be spared for each execution. When the endpoints of the 90% approximate confidence interval for α_3 were used to calculate this tradeoff, the "expected tradeoffs ... range[d] between limits of 0 and 24" (page 414). This prediction was surrounded by verbal qualifications. For example, Ehrlich acknowledged that it was inherently weak because it "may be subject to relatively large prediction errors" (page 414)—even though his article contains no evaluation of the predictive power of his result, not even the R^2 values for his regressions—and that the "validity" of his estimated tradeoff "is conditional upon that of the entire set of assumptions underlying the econometric investigation" (page 414).

This paper spawned a substantial scholarly literature (at least 168 citations in the Social Science Citation Index up to August 1986) and was a central part of the Solicitor General's argument to the Supreme Court in *Gregg* v. *Georgia* (see Justice Marshall's dissent, 1977, page 909), which reaffirmed the constitutionality of capital punishment (Glenn, 1978). Taking into account Ehrlich's paper and the subsequent criticisms of it, the majority opinion of Justices Stewart, Powell and Stevens (1977) in *Gregg* v. *Georgia* found the evidence of a deterrent effect to be "inconclusive" (page 881) and Justice Marshall declared Ehrlich's study "of little, if any, assistance in assessing the deterrent impact of the death penalty" (page 909). Both the court majority (page 881) and Marshall (page 909) cited papers criticizing Ehrlich's use of data aggregated across the entire United States to draw inferences about the effect of state laws, the sensitivity of Ehrlich's finding to the choice of time period used in his regressions, the quality of his data, the choice of explanatory variables, the absence of consideration of the collinearity of the explanatory variables and the choice of a functional form for the regression. In effect, these critics—who for the most part used Ehrlich's own data—performed the assessment of structural uncertainty that Ehrlich omitted. It is reasonable to wonder how much less influential his paper would have been had it included an accounting of predictive risk and a proper consideration of structural uncertainty.

This inability to capture structural uncertainty, particularly uncertainty about the model, has a deeper effect. Consider the statistical techniques used in setting radiation exposure standards. The effects of long term exposure to low levels of radiation are beyond the reach of laboratory methods. Radiation standards are set by subjecting animals to large doses of radiation, observing cancer rates, estimating a model for the relationship between dose and cancer rate, and extrapolating back to the low doses (Kalbfleisch and Prentice, 1980, page 69; Hattis and Kennedy, 1986; Hattis and Smith, 1985). Leaving aside the issue of extrapolating from animals to humans, the low dose extrapolation is purely conventional. Without observations on animals subjected to the lower doses, there are no data against which to check the adequacy of the model. Kalbfleisch and Prentice give two models that have plausible substantive rationalizations, but note that "... [d]ifferences in tail shape for these models generally lead to completely different low dose risk estimates" (page 69). Thus, radiation standards are largely determined by the choice of a model, i.e., by the selection of a convention.

The selection of this convention—that determines standards that can affect large numbers of lives—is made perhaps by a few experts, or perhaps by a few people who know how to run logistic regression programs and little about the subject area (Hattis and Smith, 1985). This selection can have a huge effect on the eventual policy choice, yet there might be no evidence in the record that any selection has occurred. An analogy can be drawn to the larger scale debate over the terms and ethical presuppositions that will be used to construe issues. This latter goes on constantly in attempts by contending parties to construe, say, abortion as a legitimate choice a woman must have or as a heinous crime. The choice of the terms in which issues will be construed is at the core of democratic politics. But most models, in terms of which technical issues are construed, are chosen quietly by experts and accepted with little public contention: models for nuclear reactor safety, for safe radiation levels, for food inspection schemes, for pension projections, for the accuracy of nuclear missiles and so on. As the low-dose extrapolation example illustrates, our tools can induce substantive outcomes by their inability to propagate model uncertainty, that is, by the information inserted into deliberations by their use (Section 3.1.2 of Hattis and Smith (1985) is of particular interest in this regard).

4. CONCLUSION

The goal of this paper was to make a beginning at devising language and ideas that would allow a more complete context for quantitative empirical analysis, particularly policy analysis. The main thrust of the attempt was to distinguish among types of uncertainty that an analyst faces, to describe their natures and catalog statistical methods used to analyze them and to apply this construction to the strategy of analysis and to the problem of properly accounting for uncertainty. Clearly, much work remains to be done in developing the context described here. In particular, I know of no explicit system embodying the scheme of three uncertainties. The theory of de Finetti (1974, 1975) comes closest, but his theory lacks a crucial connection to real problems, and this paper is an attempt to provide the connection. I have suggested, in an echo of Tukey (1962), an empirical approach to the matter of whether the ordinarily omitted types of uncertainty are important, and to constructing an operational understanding of the accuracy of approximations. These two tasks can be undertaken without new theory, and work is in progress.

ACKNOWLEDGMENTS

This work was supported in part by RAND Corporation funds. I had the benefit of comments and suggestions from Morris DeGroot, Maureen Lahiff, Robert Levine, John Rolph and anonymous referees. I owe particular thanks to my RAND colleague David Draper for his careful scrutiny and discussion. Any remaining errors of omission, commission or interpretation are mine alone.

REFERENCES

ALHO, J. M. and SPENCER, B. D. (1985). Uncertain population forecasting. *J. Amer. Statist. Assoc.* **80** 306–314.

ASTRACHAN, M. and CAHN, A. S. (1963). *Proc. of RAND's Demand Prediction Conference, January 25–26, 1962.* RAND Corp. RM-3358-PR.

ATKINSON, A. C. (1985). *Plots, Transformations, and Regression.* Clarendon Press, Oxford.

BAILAR, J. C. III. (1985). When research results are in conflict. *New England J. Med.* **313** 1080–1081.

BAKER, R. J. and NELDER, J. A. (1978). *The GLIM System, Release 3.* Numerical Algorithms Group, Oxford.

BATES, D. M. and WATTS, D. G. (1980). Relative curvature measures of nonlinearity (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 1–25.

BECKER, R. A. and CHAMBERS, J. M. (1984). *S—An Interactive Environment for Data Analysis and Graphics.* Wadsworth, Belmont, Calif.

BENNETT, B. (1980). How to assess the survivability of U. S. ICBM's. RAND Corp. R-2577-FF.

BERGER, J. O. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analysis* (J. B. Kadane, ed.) 63–144. North-Holland, New York.

BICKEL, P. and DOKSUM, K. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76** 296–311.

BLACKETT, P. M. S. (1962). Critique of some contemporary defence thinking. In *Studies of War* (P. M. S. Blackett, ed.) 128–146. Hill and Wang, New York.

BOX, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430.

BOX, G. E. P. and COX, D. R. (1982). An analysis of transformations revisited, rebutted. *J. Amer. Statist. Assoc.* **77** 209–210.

BOX, G. E. P. and TIAO, G. C. (1962). A further look at robustness via Bayes' theorem. *Biometrika* **49** 419–432.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, Calif.

BUILDER, C. H. (1986). On the Army style in analysis. RAND Corp. P-7267.

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B. and TUKEY, P. A. (1982). *Graphical Methods for Data Analysis*. Wadsworth, Belmont, Calif.

CHATFIELD, C. (1985). The initial examination of data (with discussion). *J. Roy. Statist. Soc. Ser. A* **148** 214–253.

COOK, R. D. (1986). Assessment of local influence (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 133–169.

COOK, R. D. and GOLDBERG, M. L. (1986). Curvatures for parameter subsets in nonlinear regression. *Ann. Statist.* **14** 1399–1418.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

CYERT, R. M. and DEGROOT, M. H. (1984). The maximization process under uncertainty. *Adv. Inform. Proc. Org.* **1** 47–61.

DAWID, A. P. (1985). Calibration-based empirical probability (with discussion). *Ann. Statist.* **13** 1251–1285.

DE FINETTI, B. (1974, 1975). *Theory of Probability* **1, 2**. Wiley, New York.

DEGROOT, M. H. (1982). Comments on the role of parameters in the predictive approach to statistics. *Biometrics* **38** 86–91.

DEGROOT, M. H. (1983). Decision making with an uncertain utility function. In *Foundations of Utility and Risk Theory with Applications* (B. P. Stigum and F. Wenstop, eds.) 371–384. Reidel, Dordrecht.

DEMPSTER, A. P. (1975). A subjectivist look at robustness. Research Report S-33, Dept. Statistics, Harvard Univ.

DRAPER, D. and HODGES, J. S. (1987). Mixing predictive distributions: the example of oil price predictions. Unpublished manuscript.

DUNCAN, G. T. and LAMBERT, D. (1986). Disclosure-limited data dissemination. *J. Amer. Statist. Assoc.* **81** 10–28.

DURBIN, J. (1980). Approximations for densities of sufficient estimators. *Biometrika* **67** 311–333.

EFRON, B. and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Statist.* **37** 36–48.

EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). *Statist. Sci.* **1** 54–77.

EHRLICH, I. (1975). The deterrent effect of capital punishment: A question of life and death. *Amer. Econ. Rev.* **65** 397–417.

ERICKSEN, E. P. and KADANE, J. B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *J. Amer. Statist. Assoc.* **80** 98–131.

FELDSTEIN, M. S. (1975). Wealth neutrality and local choice in public education. *Amer. Econ. Rev.* **65** 75–89.

FISHER, R. A. (1957). The underworld of probability. *Sankhyā* **18** 201–210.

FREEDMAN, D. A. (1981). Some pitfalls in large econometric models: A case study. *J. Business* **54** 477–500.

FREEDMAN, D. A. (1985). Statistics and the scientific method (with discussion). In *Cohort Analysis in Social Research* (W. M. Mason and S. E. Fienberg, eds.) 343–390. Springer, New York.

FREEDMAN, D. A. and NAVIDI, W. C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statist. Sci.* **1** 3–39.

FREEDMAN, D. A., ROTHENBERG, T. and SUTCH, R. (1983). On energy policy models (with discussion). *J. Bus. Econ. Statist.* **1** 24–36.

GEISSER, S. (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 456–469. Holt, Rinehart, and Winston, Toronto.

GEISSER, S. (1980a). A predictivistic primer. In *Bayesian Analysis in Econometrics and Statistics* (A. Zellner, ed.). North-Holland, New York.

GLENN, J. A. (1978). Annotation: Supreme Court's views on constitutionality of death penalty and procedures under which it is imposed. *U.S. Supreme Court Reports*. Lawyers Edition Second Series **51** 886–909 (51 LEd2d 886).

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

HARRISON, P. J. and STEVENS, C. F. (1971). A Bayesian approach to short-term forecasting. *Oper. Res. Q.* **22** 341–362.

HARRISON, P. J. and STEVENS, C. F. (1976). Bayesian forecasting (with discussion). *J. Roy. Statist. Soc. Ser. B* **38** 205–247.

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models (with discussion). *Statist. Sci.* **1** 297–318.

HATTIS, D. and KENNEDY, D. (1986). Assessing risks from health hazards: An imperfect science. *Technol. Rev.* May/June 60–71.

HATTIS, D. and SMITH, J. A. (1985). What's wrong with quantitative risk assessment? Presented at the Conference on Moral Issues and Public Policy Issues in the Use of the Method of Quantitative Risk Assessment, Georgia State Univ., September 26–27, 1985. To appear in *Biomed. Ethics Rev.*

HILLESTAD, R. J. (1982). DYNA-metric: Dynamic multi-echelon technique for recoverable item control. RAND Corp. R-2785-AF.

HINKLEY, D. V. and RUNGER, G. (1984). The analysis of transformed data (with discussion). *J. Amer. Statist. Assoc.* **79** 302–320.

HODGES, J. S. (1987). Assessing the accuracy of normal approximations. *J. Amer. Statist. Assoc.* **82** 149–154.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525.

JENNINGS, D. E. (1986). Judging inference accuracy in logistic regression. *J. Amer. Statist. Assoc.* **81** 471–476.

JOHNSON, W. and GEISSER, S. (1982). Assessing the predictive influence of observations. In *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh, eds.) 343–358. North-Holland, New York.

JOHNSON, W. and GEISSER, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *J. Amer. Statist. Assoc.* **78** 137–144.

JONES, A. C. (1986). A stochastic analysis of the propagation of rounding error in floating point computations. Unpublished Ph.D. thesis, Yale Univ.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

KEYNES, J. M. (1939). Professor Tinbergen's method. *Econom. J.* **49** 558–568.

KEYNES, J. M. (1940). Comment [on Tinbergen's rejoinder]. *Econom. J.* **50** 154–156.

LAGAKOS, S. W., WESSEN, B. J. and ZELEN, M. (1986). An analysis of contaminated well water and health effects in Woburn, Massachusetts (with discussion). *J. Amer. Statist. Assoc.* **81** 583–614.

LANE, D. and SUDDERTH, W. (1985). Coherent predictions are strategic. *Ann. Statist.* **13** 1244–1248.

LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley, New York.

LEE, J. C. and GEISSER, S. (1972). Growth curve prediction. *Sankhyā Ser. A* **34** 393–412.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

LITTERMAN, R. B. (1986). A statistical approach to economic forecasting. *J. Bus. Econ. Statist.* **4** 1–4.

MALLOWS, C. L. (1983). Data description. In *Scientific Inference, Data Analysis, and Robustness* (G. E. P. Box, T. Leonard and C.-F. Wu, eds.) 135–152. Academic, New York.

MALLOWS, C. L. and TUKEY, J. W. (1982). An overview of data analysis, emphasizing its exploratory aspects. In *Some Recent Advances in Statistics* (J. Tiago de Oliveira and B. Epstein, eds.) Chap. 7. Academic, New York.

MALLOWS, C. L. and WALLEY, P. (1980). A theory of data analysis? *Proc. Amer. Statist. Assoc. Bus. Econ. Statist. Sec.* 8–14.

MANSFIELD, E. (1979). *Microeconomics: Theory and Applications,* shorter 3rd ed. Norton, New York.

MARSHALL, T. (1977). Dissenting opinion in *Gregg v. Georgia. U.S. Supreme Court Reports.* Lawyers Edition Second Series **49** 907–912 (49 LEd2d 907).

MCCULLOCH, R. E. (1985). Model influence in Bayesian statistics. Unpublished Ph.D. thesis, Univ. Minnesota.

MINKIN, S. (1983). Assessing the quadratic approximation to the log likelihood function in non-normal linear models. *Biometrika* **70** 367–372.

O'SULLIVAN, F., YANDELL, B. S. and RAYNOR, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

RELLES, D. A. (1986). Allocating research resources: The role of a data management core unit. RAND Corp. N-2383-NICHD.

SCHERVISH, M. J. (1985). Comment on "Self-calibrating priors do not exist." *J. Amer. Statist. Assoc.* **80** 341–342.

SHAFER, G. (1986). Comment on "Combining probability distributions: A critique and an annotated bibliography." *Statist. Sci.* **1** 135–137.

SMITH, A. F. M. (1983). Bayesian approaches to outliers and robustness. In *Specifying Statistical Models: From Parametric to Non-Parametric, Using Bayesian or Non-Bayesian Approaches.* (J. P. Florens, M. Mouchart, J. P. Raoult, L. Simar, and A. F. M. Smith, eds.). Springer, New York.

SMITH, A. F. M., SKENE, A. M., SHAW, J. E. H., NAYLOR, J. C. and DRANSFIELD, M. (1985). The implementation of the Bayesian paradigm. *Commun. Statist. A—Theory Methods* **14** 1079–1102.

SPENCER, B. D. (1985). Optimal data quality. *J. Amer. Statist. Assoc.* **80** 564–573.

STEINBRUNER, J. D. (1974). *The Cybernetic Theory of Decision.* Princeton Univ. Press, Princeton, N. J.

STEWART, P., POWELL, L. and STEVENS, J. P. (1977). Majority opinion in *Gregg v. Georgia. U.S. Supreme Court Reports.* Lawyers Edition Second Series **49** 866–893 (49 LEd2d 866).

STOCKFISCH, J. A. (1975). Models, data, and war: A critique of the study of conventional forces. RAND Corp. R-1526-PR.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 111–47.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86.

TINBERGEN, J. (1940). On a method of statistical business cycle research. A reply. *Econom. J.* **50** 141–154.

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67.

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

VASELY, W. E. and RASMUSON, D. M. (1984). Uncertainties in nuclear probabilistic risk analysis. *Risk Anal.* **4** 313–322.

WEISBERG, S. (1983). Some principles for regression diagnostics and influence analysis. *Technometrics* **25** 240–244.

WEISBERG, S. (1985). *Applied Linear Regression,* 2nd ed. Wiley, New York.

WEST, M. (1986). Non-normal multi-process models. Research Report No. 81, Dept. Statist., Univ. Warwick.

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Assoc.* **80** 73–97.

ZELLNER, A. (1984). Posterior odds ratios for regression hypotheses: General considerations and some specific results. In *Basic Issues in Econometrics* (A. Zellner, ed.) 275–305. Univ. Chicago Press, Chicago.