A RAND NOTE

Uncertainty, Policy Analysis and Statistics

James S. Hodges

August 1988
This research was supported by The RAND Corporation as part of its program of public service.

This Note contains an offprint of RAND research originally published in a journal or book. The text is reproduced here, with permission of the original publisher.

The RAND Publication Series: The Report is the principal publication documenting and transmitting RAND's major research findings and final research results. The RAND Note reports other outputs of sponsored research for general distribution. Publications of The RAND Corporation do not necessarily reflect the opinions or policies of the sponsors of RAND research.
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.
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 Cain, 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.


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 Sudereth, 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 396 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, Rousseauw, 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 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 46), "[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 unsupervisible 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 Rasmussen (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
2.2 Risk

Having tentatively conditioned on some structure for making predictions, which usually includes a class of stochastic models with 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 \( \lambda f_j \), where \( f_j \) is the number of hours F16s flew in month \( j \) and \( \lambda \) is an unknown positive constant, then our data contain information about which values of \( \lambda \) 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 observations \( x_i, j = 1, 2, \ldots, n, n + 1 \), are postulated to be independent normal random variables with mean \( \theta \) and variance 1, where \( \theta \) is an unknown and unobservable parameter. Then the average of the first \( n \) observations, \( \bar{x} \), is a normal random variable with mean \( \theta \) and variance \( 1/n \), so that \( x_{n+1} - \bar{x} \) is normal with mean 0 and variance \( 1 + 1/n \), independently of \( \theta \). 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 \( \theta \) 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 \( \theta \) 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 $f(x | \text{model}, \theta)$ is the probability function or probability density of the observable $x$ given the model and its parameter $\theta$, and $p(\theta | \text{model}, \text{data})$ is the posterior distribution of $\theta$, then the predictive distribution of $x$, conditional on the data, model, and prior distribution, is

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

if $\theta$ has a continuous probability distribution or the obvious analogue if $\theta$ 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 Hoges, 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 non-conjugate 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 Tiabergen (Keynes, 1939, 1940; Tinbergen, 1940) about models of national economies, said that

"Plut 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 countenance 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, Feinstein’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 Rasmussen, 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 pre-positioning 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 $\alpha_3$, the elasticity of the murder rate with respect to the execution rate. (This elasticity is $\frac{\partial Q}{\partial E}(E/Q)$, 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 $\alpha_3$ 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 $\alpha_3$, he then predicted that eight potential murder victims would be spared for each execution. When the endpoints of the 90% approximate confidence interval for $\alpha_3$ were used to calculate this tradeoff, the “expected tradeoffs ... range[d] between limits of 0 and 24” (page 414). This prediction was surmounted 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


Hodges has written a thoughtful and scholarly essay on the strengths and limitations of statistical models in policy analysis. Although there may be some differences in emphasis, his views are quite close to mine. I think the two main points are as follows:

(i) Models are usually chosen on the basis of familiarity and tractability; the degree of correspondence with reality is seldom of primary concern. Hodges' formulation:

"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."

(ii) Policy analysts usually assess only one component of the uncertainty in their results; in effect, they partially away uncertainty about structure. As Hodges dryly says,

"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."

These generalities aside, what really caught my attention in Hodges' paper (as may be only natural) was the sympathetic discussion of my own work. He quotes me—rightly—as saying that in many contexts, ad hoc analyses by experts may be better than modeling (write essays rather than fit models); recognition that some questions cannot be answered at all sensibly by the analysts within their contract performance period may be the wisest course of all.

Hodges then asks, "In what sense are the energy policy models that Freedman attacks not ad hoc expert analyses, albeit elaborate ones?" This is a rhetorical question, but I'll respond to it. Models are often defended—not by Hodges—as being state of the art, objective, scientific exercises, with assumptions made explicit. By contrast, analytical essays informed by data are old fashioned, arbitrary, unscientific, with crucial premises left unstated.

Some may find this defense of models an attractive fantasy, but fantasy it is. My experience includes risk assessment and econometrics. (So I do not comment on the air force logistic models discussed by Hodges.) To make contemporary model-based policy analyses in Washington, the analyst has to introduce dozens if not hundreds of fairly arbitrary assumptions. Rather than being articulated and defended, these are buried in the statistical estimation procedures, or even deeper in the computer code.

Because first versions of models seldom give plausile results, the analyst has to massage inputs, outputs and model innards, until these are more or less in balance. Indeed, one well-known modeling group is famous for the "add factors" that must be applied to regression intercepts in order to get sensible-looking macroeconomic forecasts. Such Rube Goldberg contraptions are models by courtesy only.

Hodges' question implicitly acknowledges the arbitrariness of current policy models and their weakness as formal arguments. He seems to be asking whether I would be more sympathetic to the models if they were relabeled as informal argument. The answer is, a little.

Computer code often functions as a decent veil of technical obscurity covering up some basic silliness. Articulating models in English rather than FORTRAN tends to make their problems—the multiplicity of arbitrary and unreasonable assumptions—more visible. That is why the code was there in the first place.

In summary, essays can be more objective, scientific and explicit than computer models. Consider, for example, the debate on capital punishment. Hodges cites (not approvingly) the model in Ehrlich (1975); for a devastating critique of such models, see Leamer (1983). By contrast, Zeisel (1981) is a fascinating and persuasive essay based on data. For more general discussions of modeling issues see Freedman (1987), Freedman and Zeisel (1987), Kolata (1986) and National Academy of Sciences (1984).

Continuing his review of my position, Hodges also asks, "how should data and data reduction techniques be used to inform all these judgments?" Then his

---

David Freedman is Professor of Statistics, University of California, Berkeley, California 94720.
urkindest remark (which is not so unkind): "It is difficult to see how these questions can even be posed within the frequentist framework." This seems wrong. There is no difficulty in posing the questions, in either the frequentist or Bayesian framework; Hodges just did it. The problem is finding answers.

Now there comes a shade of difference between us. He is a little more optimistic than I am about the potential usefulness of Bayesian techniques for properly integrating judgments about uncertainty. For example, he discusses predictive distributions starting from (i) a prior on models and their parameters and (ii) a likelihood function for the data given the model and parameters.

This is quite sensible, provided there is a sound basis for choosing the prior and the likelihood. Unfortunately, Bayesian policy analysts can be just as sloppily in such matters as us frequentists. For discussion of this issue, see Freedman and Navidi (1986) or Hill (1985).

Good statistical analysis can be done in either the frequentist or the Bayesian framework. However, for either approach to succeed, the analyst has to get the model right, or close enough. That idea may seem ridiculously old fashioned. As policy analysts can be heard to sputter, "Models be right? How can they be right? They're all approximations. Even Newton was wrong. And a mystic besides." Because nothing is perfect, anything goes.

Hodges wants "to bring de Finetti to . . . practitioners." As I understand him, for de Finetti a prior represents a major intellectual commitment to be adopted only after serious investigation of the subject at issue. If policy analysts followed that percept, we would all be better off. The real issues here are of science, not statistical technique.

ADDITIONAL REFERENCES


Comment

Seymour Geisser

Now comes James Hodges to inform us on some of the larger issues of statistics. And what are these issues? They are the ones that statisticians have dealt with—lo these many years—uncertainties from various sources. And there are other issues besides—is it an observational study? a controlled experiment? a retrospective investigation? a haphazard collection of items? Is what is measured or observed actually what one defines as measured? Are there flawed observations? Was the experiment or trial carried out according to the protocols? Is there a temporal imperative with regard to an action or a decision? There is, to say the least, limited interest (other than procedural validation perhaps) in the prediction of events that already have occurred and been observed.

What is the point then? The point is that we have here a lucid and trenchant exposition vividly reminding us of three of the principal sources of uncertainty or variation. What is more novel than most previous explications is that the sources are related to predictivism, which is stressed as the penultimate aim when taking an action is the ultimate goal. Hence, from my point of view, there is really nothing to quarrel with. But it is the job of a discussant if not to be quarrelsome to be at least quibblesome—to coin a neologism.

Hodges intimates that for proper application of statistical methods, the implementation of de Finetti’s approach is required. He also states that the approach “lacks a crucial connection to real problems.” I would like to quibble with both these points. In regard to the latter point, we have only to realize that de Finetti was involved in applications especially in finance,
actuarial mathematics, censuses, football pools, lotteries, etc.—and these problems are just as real in their own realm as spare parts is to the Air Force. In fact, I would surmise that de Finetti’s views were molded to some extent by his actuarial experience.

In regard to the first point there is a fundamental principle in the de Finetti canon that statisticians (even Bayesian predictivists) often ignore. To be fully coherent one’s view about the prior distribution should not depend on the likelihood. This straightjacket is almost always flouted when searching for models in analyzing data. I believe the attitude that many reasonable statisticians take is, if I assume this then I will believe that, and decide, after examining the data in various ways, what it is they eventually believe about values as yet unobserved. Also the de Finetti canon precludes testing one’s predictive methodology against realized further observations or using some other validation technique. But these methods are far too useful in searching for and selecting appropriate models to be summarily dismissed.

Of course the Neyman-Pearson approach is further beset by even greater stringencies. My view of it is to paraphrase a boxing manager’s lament over his fallen fighter “N-P has some good qualities but its bad qualities ain’t so good.”

With regard to the use of predictive distributions for testing and as a diagnostic tool, I believe its most appropriate application is in discordancy testing. In such instances the bulk of the observations are assumed to concur with the postulated model although possibly a few may not and they are to be tested in the absence of any discernible alternative, e.g., Geisser (1980b, 1985, 1987), Johnson and Geisser (1983).

In discussing estimation risk and predictive risk the prediction of future values from a II(θi, 1) distribution is presented in the frequentist framework. Because Xn+1 − X has variance 1 + n−1, Hodges asserts that in terms of variance 1/n is the estimation risk and 1 the prediction risk. This is not so, although it is true that n−1 is the estimation risk, 1 + n−1 is the prediction risk. Estimation risk is inevitably embedded in prediction risk. In fact a deeper interpretation of this example indicates the inappropriateness, even in the classic frequentist setup, in using estimation procedures for predictive purposes, i.e., using N(θ, 1) as the “best” estimate of the distribution function clearly leads to predictive confidence intervals for Xn+1 that are too short for the stated confidence level (Geisser, 1980a). For further discussion of these points see Geisser (1982) and in particular for the relative importance of the two risks see my rejoinder to the discussion by DeGroot (1982) of that paper.

Although it is true that actual applications of predictive distributions to new data sets are relatively rare—much the same could be said of Bayesian applications. And predictive distributions properly exist only in Bayesian contexts. But to state that Duncan and Lambert (1986) is the only application he can find, indicates either myopia or tunnel vision, because there are many others. For example, every Bayesian model selection or classification procedure uses predictive distributions.

In other parts of this otherwise perceptive essay there are some comments alluding to the use of mixing models. It is true that in the context of Bayesian prediction, when the number of entertained models is exhaustive and prior probabilities for them are determined, then the predictive distribution is a mixture, cf., Clayton Geisser, and Jennings (1986) and Geisser and Eddy (1979). Aside from the fact that exhaustiveness is often not the case, it appears that most scientific workers tend to opt for a single preferred model, operating on some principle of parsimony, or some latent loss function that heavily penalizes mixing discrete models. I must admit to exhibiting ambivalence about this issue, or more precisely, I have not really formulated an appropriate loss function that I find satisfactory. There is certainly something elegant about a single reasonable model that appears to be adequate for the data. On the other hand if the data are such that no single model seems to adequately capture all of the features in the data, then I would be more inclined to mixing because it is more likely that my predictions will be better (on some defined measure) than using a single model. A situation not altogether different presents itself in the potential infinite regress on hyperparameters and hyperpriors (Geisser, 1980c, page 466)—when does one stop? True solutions to problems of this sort require a belief in the potential reality of the models and the specified parameters. Serious doubt concerning these and other aspects of the modeling situation should, to a degree, relieve one of the necessity of trying to conjure up distributions for nonexistent entities especially if one is reasonably satisfied with something workable and adequate.

The spread of penetrating predictive applications to social affairs and policy analyses is certainly an important and welcome development and one that requires much thought and considerable effort in dealing with these action-oriented problems. Hodges has made a fine start in delineating some of the principal issues.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation Grant DMS-86-01314 and National Institutes of Health Grant GM-25271.

ADDITIONAL REFERENCES

Comment

Peter J. Huber

This is a very stimulating paper, and the issues raised and discussed in it—how to deal with three distinct types of uncertainty: structural, stochastic and technical—clearly are important not only in applied statistical work, but also beyond.

I shall confine my comments to two central issues of this paper: the problem of the infinite regress and the question of whether and when to combine different kinds of uncertainty.

The main and obvious difficulty one faces with the structural type of uncertainty is an infinite regress: once one has quantified the structural uncertainty, one also should quantify the uncertainty of this quantification, and so on. The customary (perhaps: the only?) way to cut this regress is to act as if at a certain level there was certainty. Often (although not necessarily), this means that one assumes some parametric family of structural models; if one is a Bayesian, one also posits a fixed prior on the space of parameters. It is somewhat awkward in the case of the Bayesian approach that at this stage of modeling the prior will not reflect a reasonably accurate, objective or subjective probability; it rarely is anything more than a conventional substitute for ignorance (e.g., a flat or a conjugate prior). But what is much worse, and this equally affects all approaches, is that the true structure with practical certainty will lie outside of the parametric family. I am always surprised how glibly a majority of statisticians (especially Bayesians!) are able to talk around these difficulties. Roughly speak-

proaches in the determination of probabilities (with discussion) Biometrics Suppl. 38 75–93.


Peter J. Huber is Professor of Statistics, Harvard University, One Oxford Street, Cambridge, Massachusetts 02138.
merely caused by my lack of imagination or faith, but lay deeper. The set of all probability distributions is very rich, and any genuine prior probability on it lives or a very thin subset (so that entirely reasonable, nearby possibilities are excluded by not belonging to the support of that prior). Once this had become clear, the obvious thing to try was an old recipe of Gauss, namely to take some reasonable-looking procedures and to investigate their properties. Actually, the first candidate I tried in any serious fashion (the $M$ estimate with a truncated-linear $\chi$ function) turned out to be asymptotically minimax with respect to $\epsilon$ contamination. In itself, this does not mean much—optimality theorems are important only because they show that you cannot improve any further by going in the same direction—but it turned out also that the maximum risk was relatively insensitive to misspecification (see Huber, 1964), end of Section 6 and Table I), and that some actually observed distributions were very similar to the least favorable strategies of nature, more so than to the idealized normal model (see, e.g., Huber, 1981, pages 91–94).

The points to be stressed here are: the structural model cutting the infinite regress was not chosen because I believed in it, but because it was relatively safe to act as if it were true and because it was in the ballpark of actually observed error distributions. Hodges’ objection that the minimax procedures would give essentially no protection if the model is far off, is correct in principle, but it nevertheless misses the point: minimaxity is only one of several criteria helping with the selection of a robust procedure and clearly must be supplemented by breakdown and other considerations. Its purpose is to safeguard within a small neighborhood—so small that it is difficult to discriminate in a meaningful fashion between models living inside, but large enough that the differences still matter. This argument is also relevant to Hodges’ criticism of the diagnostic approach: roughly speaking, the proper underlying philosophy is to use diagnostics to catch the larger deviations from the model and robustness to deal with the smaller ones.

The main and undisputed strength of the Bayesian approach is to provide a neat and unified way to combine evidence from different sources. However, it is not always desirable (or even possible) to combine the evidence. Often, it may be more important instead to identify the dominant source of uncertainty, to ascertain to what extent different conclusions may depend on it and what might be done to reduce that uncertainty.

In particular, the combination becomes meaningless and useless, if the lesser sources of uncertainty are comparable in size to the uncertainty in the assessment of the dominant source of uncertainty. A couple of separate and rough back-of-the-envelope calculations then can be more revealing than a black box presentation of the combined effects.

It is particularly embarrassing when structural uncertainty is dominant, but the experts disagree widely about its extent. For example, in a problem of astronomical dating (Huber, 1982), a crucial piece of the evidence was a text with Venus observations. Apart from the fact that the text is replete with gross errors, there is a substantial structural uncertainty: there are doubts about the overall reliability of the text. To mention a specific question: how sure is the attribution of the observations to the reign of king A? One expert might put this probability well above 95%, another below 50% and assertions about the date of king A, based on this text alone, are dominated by this uncertainty.

How should one then combine evidence derived from this text with that derived from others? One possibility is to present a couple of alternative analyses, perhaps using some version of interval arithmetic with lower and upper probabilities. Another is to proceed conditionally, given that the attribution is correct.

Incidentally, by being rather extreme with regard to structural and other uncertainties, this dating problem also happened to put into focus some other delicate aspects of the relative strengths and weaknesses of the different approaches. In a Bayesian framework, it was difficult to go beyond relative likelihoods even with additional, independent data (the main result derivable in this framework was that among four choices suggested by the Venus text, one chronology was favored about 10,000 to 1 over the others by the combined evidence). In the Neyman-Pearson framework, one could use the independent data to test the correctness of the best of the four (a permutation test rejected the hypothesis that all four were wrong on the 1% level).

The third source of uncertainty—technical uncertainty or inadequacy of execution, again raises the question: to combine or not to combine? It is the obligation of a competent professional to keep technical uncertainty small—smaller than the other uncertainties by about an order of magnitude. Of course, he or she will occasionally fail, and in addition there is Murphy’s law. Thus, we can expect a mixture of frequent, but mostly negligible, small errors on one side with rare gross errors on the other side, but we may lack a rational basis for estimating (or even only guessing) the probability of the latter. The problem is certainly worse than in robustness (where it is possible to use a majority of good cases to keep a small minority of bad ones under control, even if one does not know very accurately how often the bad ones occur). The
principal method for checking that kind of failure, at least in the United States, is the malpractice suit, and such suits might even provide some empirical evidence about human frailty of technical professionals, although not necessarily transferable to the cases Hodges has in mind!

Comment
Joseph B. Kadane

I agree with Jim Hodges’ approach to problems of robustness and uncertainty, and congratulate him for his clear exposition of it. I would, however, add a few remarks and references.

Although his paper cites de Finetti as coming “closest to the goal of a complete context for statistical activity,” Hodges does not in this paper bring his analysis very close to de Finetti’s ideas. For example, he does not mention the extreme subjectivity of de Finetti (probabilities represent a person’s opinions; different people may have different opinions). Whose opinions do or should a Rand logistics study represent? Are different experts consulted on different aspects of the problem? If so, by what principles should such opinions be brought together?

A second important aspect of de Finetti’s work is his emphasis on prevision (see Goldstein, 1986). There are important questions about elicitation using de Finetti’s methods when ethical neutrality fails, as it will for most experts most of the time (Kadane and Winkler, 1987a, 1987b).

A third important aspect of de Finetti’s work is his insistence on finite additivity of probabilities. de Finetti believed that while your probabilities might be countably additive in a given situation, there is no axiom that they must be. Mere finite additivity changes the nature of probability theory, particularly in the failure of conglomerability (Schervish, Seidenfeld and Kadane, 1984). This has a variety of consequences for statistics (Kadane, Schervish and Seidenfeld, 1986; Hill, 1980a). It would be interesting if Hodges would remark on how these aspects of de Finetti’s work may have influenced his work and that of his Rand colleagues, or how they might.

With respect to Bayesian ideas of robustness, there are several important approaches left unmentioned. First, there is the classic paper of Edwards, Lindman and Savage (1963), which introduced the idea of stable estimation. There is a series of papers (Kadane and Chuang, 1978; Chuang, 1984) concerning what happens if the prior, likelihood or utility as assessed is slightly off from “true.” These two papers study conditions under certain topologies in which the achieved expected utility is continuous. There is also important work of Novick and Ramsey (1980) and of Hill (1980b).

Hodges mentions puzzle that so few applications use predictive distributions. In the area of parametric elicitation, these have been used for some time. Predictive distribution in this context have the advantage of being able to present questions to an expert on variables that are familiar, instead of about parameters of an unfamiliar distribution. For papers along these lines, see Kadane, Dickey, Winkler, Smith and Peters (1980), Kadane (1980) and Winkler (1980). The former gives a concrete application in the Appendix. A second use of those programs in a medical context is described briefly in Kadane (1986).

Finally, Hodges might be interested to learn of an explicitly Bayesian effort on the spare parts problems for Naval aircraft almost 20 years ago (Brown and Rogers, 1973). There the problem was that the airplane in question had not yet flown, so priors based on spare part usage of other airplanes were used, together with a judgment about how similar the mechanics (and hence, perhaps, spare parts usage) would be. An additional problem was that spare parts built while the airplanes were being built were much less expensive than spare parts built later, and that spare parts could be partially built, and then completed,

ADDITIONAL REFERENCES


if needed. The uncertainties were so great that the Navy's initial requirement for reliability would have been extremely costly. Hodges' paper is a very welcome addition to the literature.

ADDITIONAL REFERENCES


Comment

Albert Madansky

The best way to referee a mathematics paper is to read only the statement of the theorem and then proceed along the lines of the flow chart given in Figure 1. The process of reading the entire paper through and checking its work in detail is clearly a second-best approach to refereeing. In the same vein, the best way to read a paper whose title is "Uncertainty, Policy Analysis, and Statistics" is to stop at the title and try to construct the list of questions that a paper with such a title should answer. And this I did.

My first question was: "Why is the area of policy analysis different from all other areas in which statistics is applied?"

I next speculated on how the word "statistics" in the title is to be used—to denote "things statisticians know" (i.e., the corpus of knowledge classified by Mathematical Reviews into category 62) or "things statisticians do?" And if the latter, when is it that the statistician crosses the invisible line between "doing statistics" and "doing something else?" Indeed, how is that invisible line defined? Finally, who today is classifiable as a statistician, now that our profession and the computer revolution have jointly made our wares as available as over the counter nonprescription drugs?

As to that word "uncertainty," I mused about whether Hodges is referring to the Knightian use of the term, as contrasted with "risk," to distinguish between subjective and objective probability? Or does he have a different use for that well-worn term?

Consistent with my paradigm for mathematical refereeing, I did not pass the title page until I had constructed answers to these questions, after which I dived into the paper. To my surprise I found none of my questions answered. Instead I found yet another list of the steps in the process that a statistician goes through when dealing with an applied problem, along

Albert Madansky is Professor of Business Administration and Associate Dean for Ph.D. Studies, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, Illinois 60637.
FIG. 1. How to referee a mathematics paper.
with the associated potential sources of error:

<table>
<thead>
<tr>
<th>Step</th>
<th>Source of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery/imposition of structure</td>
<td>Structural uncertainty</td>
</tr>
<tr>
<td>Assessment of variation conditional on</td>
<td>Risk (estimation/prediction)</td>
</tr>
<tr>
<td>structure</td>
<td>Technical uncertainty</td>
</tr>
<tr>
<td>Execution of techniques selected</td>
<td></td>
</tr>
</tbody>
</table>

Then’s mighty big words for the various errors that statisticians can make! I think in more earthy terms. The model created by the statistician (or his client) may be “off-base,” the procedure recommended by the statistician may be “dead wrong” and for a variety of reasons the statistician may “drop the ball” in implementing his recommended procedure even if it (and the model) are quite good. Statisticians have concerned themselves with all these problems, as Hodges has noted. And what his paper does is give us a “scorecard” so that we can tell which statistical players are working on aspects of each of these errors. Others who have put into print their thoughts about aspects of these errors are Kimball (1957), who defined the “error of the third kind” as that of finding the right answer to the wrong problem, and Good (1980), and many other writings both before and since, who tries to construct a philosophy of data analysis. I welcome Hodges’ scorecard, especially as it has led me to a number of interesting papers published in “off the beaten track” places.

But the paper left me with many more questions upon which to muse, some of which are the following:

1. Are these steps the only junctures at which a statistician can make an error? A full taxonomy of where statisticians can err even within the structure discovery step would undoubtedly entail such substeps as “developing a model based on subject-matter theory,” “developing a model based on exploratory data analysis” and “checking a theory-based model against the data.”

2. How much do each of these errors matter? Perhaps a theory of “the professional statistician’s lifetime loss function” is needed. The statistician can no longer get away with such statements as “In my lifetime I will bat .950 with respect to my confidence intervals bracketing the parameters I will estimate and will field .950 with respect to the error of rejecting a null hypothesis when it is true.”

3. How does a statistician remain objective in the face of his everdeeper involvement in exploratory data analysis, pretesting and the setting of “prior” probability distributions on model parameters? And how objective were they in the “good old days”? (“What used to be called prejudice is now called a null hypothesis,” (Edwards, 1971).)

4. How “structural” are the structures found by statisticians? Often the theory with which to model real data is quite well-developed, but our statistical procedures aren’t. For example, multiple linear regression is used to fit linear-in-the parameters approximations to complicated functions dictated by theory only because our statistical technology is not so fully developed that it can work directly with the appropriate functions. The statistician’s error exists because it is inherent in the process the statistician invokes in searching for structure. And the statistician’s structure is at best an approximation to what would be considered by the client as a true structure.

5. Is the statistician’s search for “structure” a search for “reality substitutes” or a search for “perspectives” (cf. Strauch (1983) for a discussion of this issue, especially in light of policy analysis)? Should statistical practice differ in these two contexts? And if so, how?

6. Where is mention made of the time-honored (but of anonymous authorship, hence not bibliographically approachable) approach to the search for structure involving use of a “hold-out sample”? Has this become a casualty of the bootstrap?

Hodges devotes a great deal of attention to “prediction risk.” This is quite reasonable, as the statistician (and also the soothsayer) are called upon to answer the question “Based on the past, what can I say about the future?” But look carefully at the “prediction” example given by Hodges. Why should one use \( \hat{x} \) as the basis for predicting the next observation?

Let us step back from the fact that \( \mu \) is unknown and ask what the prediction would be if \( \mu \) were in fact known. Moreover, let us ask the question more generally, for the case where \( x \) is drawn from an arbitrary distribution with known mean \( \mu \) and unit standard deviation. What is lacking in the problem formulation is a loss function. If the loss function is \( E(x - p)^2 \), where \( p \) is the prediction, then, because this loss function is the moment of inertia of \( f(x) \) about \( p \), it is minimized by taking the population mean as the prediction. But suppose the loss function were 0 if the prediction is “close” and \( c > 0 \) if it is not “close” to the new observation. Then, the population mode would be the best prediction, based not merely on the mathematics that leads to this conclusion but on the more intuitive dictum of Damon Runyon, “The race is not always to the swift, but that's the way to bet.” Worse yet, for certain loss functions a “bold play” prediction, i.e., predicting a low probability observation (a “long shot”) is the optimal prediction (cf. Dubins and Savage, 1965).

Now for the normal distribution the population mode is equal to \( \mu \), so when \( \mu \) is unknown an estimate of the population mean (aka mode) is needed. Hence,
\( \bar{x} \) is the prediction in Hodges' example. But in general, when the population mean is not equal to the population mode, it's moot whether one should use \( \bar{x} \) or a sample-based estimate of the population mode as the prediction. Perhaps this "prediction" is not taught at Berkeley because it might gull statisticians into the bad habit of thinking that they should always predict the "average."

Enough of these questions. Since I so sorely missed answers to my original set of questions prompted by the title of Hodges' paper, I will use my remaining space to provide my answers to these questions.

**ON POLICY ANALYSIS**

There are a number of ways in which policy analysis is unique in its use of statistics. First of all, as Hodges say, in policy analysis "(w)e must act." Contrast this with other areas in which statistics is useful: as an extreme instance, in the discovery of theoretical constructs. (See the exchange between Freedman and Fienberg in Mason and Fienberg (1985) for some caricatures of the role of statistics in that context. I believe these are caricatures in the sense in which Hodges uses the term.) Secondly, because so much of policy analysis rests on cost/benefit analysis, there is a good deal of explicit concern given to loss functions here. Indeed, it is here more than in any other area of statistical application that the full panoply of concepts originating in economics and incorporated into statistics (e.g., expected utility maximization) come into explicit use. Unfortunately, the utility function is somewhat murky, ill-defined, not easily measurable and often measured either by surrogate measures or in indirectly observable quantities such as "opportunity costs."

**ON STATISTICS**

If the statistician were to circumscribe his domain merely to making statements such as "if the data \( x_1, \ldots, x_n \) are independently drawn from a normal distribution with unknown mean \( \mu \) and unit variance, then in repeated samples from this distribution the statistics \( \bar{x} \pm 1.96/\sqrt{n} \) will bracket \( \mu \) 95% of the time," there would be scant need for the kind of introspection given in Hodges' paper (and others, many of them cited by Hodges). It is precisely because the statistician's wares are used beyond the confines of the limitations expressed in such statements as quoted above, and by people we have trained in a limited fashion, that we as a profession must begin to "break mental set" in the way in which we teach applied statistics. It can no longer be taught as a "watered-down" parallel to a first course in mathematical statistics (as exemplified by both classical elementary textbooks as Dixon and Massey (1957) and modern elementary textbooks as Freedman, Pisani and Purves (1978)). Rather, it should take the student along Hodges' steps, searching for structure (perhaps including exercises in adducing utility functions and/or subjective probability distributions in addition to teaching exploratory data analysis), followed by training in assessment of variation, selection of techniques and their execution.

As to the related issue of the statistician's hubris to tread beyond the confines of the data, contrast the following quote from Chernoff and Moses (1959), "Years ago a statistician might have claimed that statistics deals with the processing of data . . . Today's statistician will be more likely to say that statistics is concerned with decision making in the face of uncertainty" with that of Kerridge (1968), "It is the statistician's job to inform, not to decide." To the extent that the statistician crosses the line and is more than a reporter and interpreter of data is it relevant to consider the issue about the role of statistics raised by Freedman in the papers cited in Hodges' references.

**ON UNCERTAINTY**

I find myself wishing that Tukey, with his penchant for inventing new words, had edited this paper. Given that Hodges felt the need to "give names to all the animals," the least he could have done is invented totally new names, not used old names, such as "risk" and "uncertainty," with pre-existent meanings, to mean new things.

**ON THE PAPER ITSELF**

I've spent about half my allotted space discoursing on the title of Hodges' paper, but what of the paper itself? Had I not been convinced from prior experience that indeed policy analysis is different from all other areas in which statistics is applied, I would not have found Hodges' arguments persuasive. Indeed, I would welcome a convincing essay confirming my priors about statistics in policy analysis. Had I not seen other lists of the steps in the applied statistician's process and the associated errors, I would have welcomed that of Hodges; given that I have seen such lists, I can find nothing in Hodges' paper that commends it above the others.

Finally, Hodges leaves me with a great deal of "uncertainty" as to where he comes out on the issue raised by the Freedman papers. Sure, no one (not even Dempster or I) would disagree with the thesis that
Comment

Adrian F. M. Smith

On the one hand, this paper claims that, both in theory and in practice, statisticians currently fail to acknowledge and incorporate important aspects of uncertainty in their modeling and analysis methodology, thus potentially distorting the inference and decision making processes in many areas of application. On the other hand, it claims that the subjectivist approach of de Finetti provides the most promising general framework for developing a language and methodology that might overcome the defects of current approaches. I am entirely in agreement with these views and therefore naturally welcome Hodges' paper, both in its own right and as a focus for a general discussion of the issues raised.

However, the structuring of the paper left me a little unclear as to what particular emphasis was intended in various of its sections. Sometimes, the emphasis seemed to be on drawing a pragmatic boundary between those problems and activities that can and cannot be approached by using some kind of more or less formal statistical modeling and analysis. At other times, the emphasis seemed to be on drawing attention to the unique merits of the Bayesian approach in providing a natural and unified framework for the development of precisely those tools that Hodges seems to consider so desirable, including predictive forms of uncertainty statements and between-as well as within-model uncertainty evaluations, both as outputs in themselves and as the basis for sensitivity analysis. Policy analysis applications seemed to fall somewhat between these two tools. Were we supposed to see policy analysis as an archetypal area where the boundary problem is particularly acute? Or as an archetypal area where Bayesian methods particularly come into their own? I fully realize that Hodges is attempting a grand overview of a large number of conceptual and practical problems that are all too rarely discussed together, but I would welcome some clarification from him of the main messages he was hoping we would extract from all this.

What I certainly do recognize from Hodges' running example and his general discussion is the total inadequacy of any view of modeling and analysis that does not appreciate the sociologic and institutional dimensions of dealing with large, messy systems in large, messy organizations. In an unpublished joint study undertaken for a major government agency in the United Kingdom, Dr. Ray Paul, of the London School of Economics, and I considered similar broad issues of model building and validation in representing and summarizing uncertainties in the context of very large scale problems. I shall briefly describe some of our general perceptions and conclusions and would very much welcome Hodges' views as to whether and to what extent we are thinking along the same lines.

Arian F. M. Smith is Professor of Mathematical Statistics, Department of Mathematics, University Park, Nottingham, NG7 2RD, England.
closely related view is set out in Landry, Malouin and Oral (1983). Model development and use typically involves a progression through five broad stages: perceived problem: situation, conceptual model, formal model, technical solution and summary output of some kind. In this process, the problem situation is defined and elaborated by the perceptions and behavior of a number of participants, who may include subject matter experts and their administrative or managerial overseers, as well as statisticians and other quantitative analysts. The conceptual model that emerges is a broad-brush mental image of the problem situation, corresponding to the participants’ identification of the features to be included, their interrelationships, the level of detail of required description and the form of inference or decision required. The formal model then consists of the translation of this conceptual model into a formal language, typically mathematics, usually converted into some form of computer code. The technical solution adopted will depend on the model and the form of output required, but will typically be a combination of algebraic manipulations, analytic approximations and numerical algorithms. The output, which may include both numerical and graphic summaries, might be directed toward an immediate decision problem or might provide a variety of conditional uncertainty statements directed more toward diagnostics and sensitivity analysis and a further iteration of the modeling and analysis process. In addition, it should be recognized that this entire process is predicated on three types of data base, mental, written and numerical, and that in major studies, which span long time periods, there might well be changes in the participants, their perceptions and the relevance of some or all of the data bases.

Taking the above as a schematic for the various processes involved in dealing with large, messy problems, there are a number of points at which steps could and should be taken to increase participants’ confidence in the modeling, analysis and reporting by systematically heightening awareness about the uncertainties implicit in these processes. In particular, such awareness is needed in respect of the uncertainties involved in arriving at a conceptual model of the problem, the potential distortions involved in passing from the conceptual to a formal model, the static and dynamic relevance, integrity and coherence of the databases, the appropriateness and adequacy of the technical machinery applied in order to obtain answers and, finally, the selection of a form of output, which is both practically useful and intellectually honest in its incorporation of the necessary conditional caveats and sensitivity studies.

It seems to me that this kind of analysis corresponds closely to Hodges’ plea for a less narrowly based approach to uncertainty on the part of statisticians. However, Paul and I drew one strong conclusion from all this, which Hodges does not seem to emphasize to the same extent, namely, that the quantitative analyst cannot and should not be acting in splendid technical isolation in such studies, but instead should form part of a team, whose members represent a range of substantive and statistical expertise and liaise closely and regularly on the basis of informed mutual respect. It is here that sociologic, institutional and, of course, budgetary, constraints inevitably enter the picture.

So far as Hodges’ advocacy of the multimodel Bayesian approach is concerned, this seems to me an excellent way of dealing with the intra- and inter-personal hesitations and disagreements, which will doubtless then result in a team context from the systematic heightening of awareness discussed above. In addition to the references cited, he will find further support for this general approach in Dickey (1973) and Smith (1978, 1984, 1986).

ADDITIONAL REFERENCES


Rejoinder

James S. Hodges

Several themes emerged in the comments; this rejoinder is organized as a discussion of those themes. In the sequel, section numbers refer to the paper under discussion.

1. REPRESENTATION

The main idea that I gleaned from de Finetti (1974, 1975)—and the idea that was to be brought to practitioners—was the idea that all uncertainty can and should be represented as probability. This idea has practical and normative implications. On the practical side, it and the taxonomy of uncertainty in Section 2 suggest a strategy of allocating resources in analysis (to which Huber alluded, and which was discussed in Section 3.1). In this regard, what matters is not how large the various kinds of error are over one’s lifetime (Madansky), but how large they are in the problem at hand—if you take care of the latter, the former will take care of itself. De Finetti’s idea and Section 2 also provide a framework for communication among members of a team (Smith, 1986, and his comments above).

On the normative side, this central idea requires practitioners to acknowledge all of the sources of uncertainty in an analysis and to incorporate them explicitly in choices made in the course of the analytical work and in the products that arise from it. In the Air Force example (to use the expression of a RAND colleague, Jim Quinlivan), the noise is the signal, and it must be reported and used in decisions. This imperative does not imply a “black-box presentation” (Huber), or that one cannot form an attachment to some particularly elegant model (Geisser); nor is it clear that an exhaustive list of models is necessary (Geisser) for an adequate representation of predictive uncertainty. What is clear is that when the time comes for betting on what the future holds, one’s uncertainty about that future should be fully represented, and model mixing is the only tool around.

In this sense, I am “more optimistic” about the Bayesian framework than Freedman: in the Bayesian approach all of the types of uncertainty can be represented and discussed in the same language and thus acquire the same importance. In the frequentist framework, this is not the case. But with this Bayesian advantage comes a disadvantage. Taken at face value, the approach generates an infinite regress (Huber, Geisser, Section 2.1) in that expressions of uncertainty are themselves often somewhat indefinite; at some level a Bayesian must make an assertion without further qualification (Huber, Section 2.1).

2. ADDING INFORMATION TO DATA

At this point one can no longer avoid a question that has been glossed over so far: what is being represented? Information—but plainly not just the information in the data (whatever that might mean). A data set, by itself, refers only to itself; in uttering a predictive or inferential statement, we necessarily add assertions to the data set. For one, we assert the relationship of the seen to the unseen, e.g., the relationship of the observations on experimental units to the properties of some unseen mechanism that produces the effect of a treatment. This assertion is usually slipped in implicitly, and it is justified (when it can be) by the design of the experiment, by knowledge of the experimental apparatus and protocols, and so on. But without the addition to the data of this assertion or something like it, any computations done using those data produce only descriptions of the data, not inferences or predictions about anything distinct from the data. (Holland (1986) gives an excellent discussion of different types of such assertions.)

For another example, we represent the relationship of the unseen future to the recorded past, usually with an explicit model. The data themselves do not and cannot support an assertion that future events will arise from the same mechanism as past observations or be otherwise comparable. This assertion must be added to the data; it is a judgment, perhaps difficult to criticize, but a judgment nonetheless. (Holland (1986) addresses this as well.) I think this explains de Finetti’s argument (alluded to by Geisser) that it is unfair to criticize someone’s predictions after the predicted events have passed; you can test predictions, but the legitimacy of the test as a gauge of future predictive power depends on an unverifiable assertion that the past—as represented by the collection of earlier predictions and the standard against which they are evaluated—is relevant to the future, for which a prediction is to be made. Even predictive validation is necessarily subjective.

Thus, I do not suggest dumping cross-validation (Geisser, Madansky), but I do suggest that cross-validators stop kidding themselves about getting something for nothing and figure out what information a cross-validation adds to the data on which it is performed. I am not sure what this information is, but it must involve exchangeability of future and past observables conditional on explanatory variables, for stratified cross-validations, and unconditional exchangeability, for unstratified cross-validations.
If models are understood as information added to data, several implications follow. For example, it doesn't matter how a given model was obtained or whether some substantive justification exists for it (Madansky). The model is a particular assertion of the relation of future observations to past observations, regardless of how it was obtained. This idea can probably be developed formally as an extension of Bayesian work on the irrelevance of stopping rules. (For a brief concurring discussion, see Hill, 1986; for a more fully developed differing view, see Leamer, 1974, 1978.)

No one will dispute that a prior distribution is an addition of information to data. But the nature of the added information is not so obvious. I agree with Geisser that we lack a satisfactory explanation of what prior distributions are supposed to mean in a data analytic context, that is, when the analyst specifies his prior after having poked through the data extensively. Leamer (1974, 1978) develops the idea that the sequence of models examined in a data analytic reveals prior beliefs about the parameters of those models (in much the same way that economic choices are supposed to reveal preferences) and thus introduces constraints on prior distributions for their parameters. This might be an accurate description of the behavior of econometricians, but it is not a convincing portrayal of, for example, the users of statistics turned out by the service courses at the University of Minnesota (my alma mater). People trained in these courses have very small data analytic and modeling repertoires, so limited that their behavior reveals more about training than about prior beliefs. The diagnostic approach (Section 2.1) includes a preference for simplicity; but again, simplicity is determined as much by the state of mathematical and computational art as by substantive considerations. Thus, in ingenious as it is, Leamer's approach is not satisfying.

In Huber's intriguing reminiscences about his path to the robustness theory, he said that he hoped to arrive at methods that would allow him to assert only vague information and still get a procedure that "worked well." But as the foregoing suggests, Huber solved this problem by a very important narrowing of scope. If the problem is extended from making inferences about parameters to making predictions—even predictions subsequent to the inferences that Huber's theory treats—the original difficulty looms even larger. In assessing the uncertainty to be attached to a prediction, how much of the variability associated with outliers should be counted? Certainly recording errors should not be counted; the object is to predict the actual values that will occur. But some theorists of robustness (e.g., Hampel, Ronchetti, Rousseeuw and Stahel, 1986) would have us lump together such recording errors with, for example, large residuals in econometric modeling, which have resulted from real effects, and thus must be counted in attaching a "give or take" to a prediction. If the goal is to predict and to assess the uncertainty of the prediction, Huber's path will not take us there.

This notion of inference and prediction as the addition of information to data is obviously too broad and deep to be treated here, but a few more general points are readily available. Using a specific mathematical assertion of information is like buying a dog; you won't really know whether you've got a good watchdog until you actually have a prowler—a predictive test—and if your dog turns out to have been a bad choice, it is too late to choose another. Moreover, your dog brings surprises with it; it might carry ticks or chew the furniture or the neighbors' children. So while it is bracing to see Kadane and others probe for ticks in the fur of the finitely additive dog, for the present I prefer to stick with more familiar hounds whose shortcomings are better understood.

In response to Madansky, though (and echoing Smith, 1986), I can say this: objectivity is a hoax. The best a statistician can do is to know what information is added to the data by an analysis, to strive to understand the nature of the added information, and to be explicit about it.

3. THE STATISTICIAN'S ROLE

What then is the statistician's role? In short to suggest legitimate ways and forms in which information can be added to data and to assist in their use, to identify the information added by procedures, to identify the information that must be added to get from a given set of data to a desired form of inferential or predictive statement and to advocate candor and rigor in the evaluation, selection and reporting of the information to be added.

I agree completely with Smith that quantitative analysts should operate as part of a team in all of his five phases of the analytical enterprise—that is the way we generally do business at RAND. The role I have described is, in large part, one of elicitation of information from collaborators, and it is difficult to fill that role without participating fully enough to ask the right questions. Smith's point is particularly important in view of the plethora of otherwise useful and well-meant publications like Andrews and Herzberg (1985), Atkinson (1985) or Hastie and Tibshirani (1986), which foster the mistaken notion that a pint of technique added to a quart of numbers yields a data analysis (as Brillinger pointed out in his discussion of Hastie and Tibshirani). I would add to Smith's five phases the notion that in longer term work, research teams often iterate through his scheme. Current Air Force work on spare parts supply is a result of problems perceived in the conceptual and formal models adopted in the late 1940s and early 1950s, which were
real achievements in their day. The importance of this iteration is emphasized by stories of ossification like those given in Section 3.2.

In this connection, I wonder whether Huber had real problems in mind when he found himself unable to specify "believable priors." It is difficult to imagine how one could assess the believability of any assertion of information, including a prior, outside of the context of a particular real problem. (Huber generalizes too quickly about where practicing Bayesians get their priors; the literature contains many examples of informative priors, e.g. Litterman, 1986, or Smith and West 1983.)

Huber also opined that it is the statistician's job to keep technical uncertainty "smaller than the other uncertainties by about an order of magnitude." Not necessarily; if technical uncertainty is bigger by an order of magnitude than the other types of uncertainty, but in aggregate the uncertainty is small enough to allow an unambiguous decision for the problem at hand, reducing the technical uncertainty is a waste of effort.

In any case, before we can make operationally useful assessments of the relative importance of technical uncertainty in specific problems, we must understand the extent of technical uncertainty for important classes of procedures. Currently, we do not. For example, the GLIM computer package relies heavily on the large sample normal approximation to the distribution of the maximum likelihood estimate and on related approximations. Its manual (Baker and Nelder, 1978) says of t tests made using the approximate standard errors: "No general results are known about the adequacy of this approximation for [non-normal] models covered by GLIM, so the standard errors provided must be regarded as only a general guide to the accuracy of the estimates ..." (Part I, Section 6, page 1); and of the $\chi^2$ approximation to the scaled deviance: "rather little is known about how good the asymptotic approximation is for small sets of data" (Part I, Section 6, page 2). Let the user beware!

A final note on the statistician's role. At the end of his comment, Freedman says "The real issues here are of science, not statistical technique." I do not agree entirely. Technique constrains; the less unnecessarily constraining, the better. Many things are natural within a Bayesian framework and difficult or awkward in a frequentist setting. For example, partly as a result of the paper under discussion, one of RAND's spare parts researchers is now willing to postulate a few possible functional forms relating the flying program to the expected number of failures of each of a collection of parts. One of these functional forms will hold (approximately), but which one is not known ahead of time. It is difficult to conceive of the actual functional form as the outcome of some stochastic mechanism, but perfectly natural to treat it as simply uncertain, in the Bayesian fashion. By considering the frequency with which each of the possible functional forms obtained in the past for this and comparable parts, this researcher is willing to postulate a few possible prior distributions across the functional forms, that is, to mix models. Bayesian technique has assisted his formulation of the problem. Moreover, it will assist him in evaluating the possible resupply options he considers, for it gives him a structure for simulations that incorporates a source of uncertainty that he previously omitted.

4. HOW POLICY ANALYSIS IS DIFFERENT

The purpose of this paper was not to argue that policy analysis is a unique environment for applying statistics (Madansky); it happens to be the area in which I do most of my work. I face similar problems in my ornithological work.

But policy analytic applications of statistics do raise considerations that scientists can often neglect (Huber, Geisser):

(i) Usually some decisions must be made.

(ii) The analysis is usually done under time and budget pressure. In the Army work in which I collaborate, we are under pressure, not to produce any particular result, but to produce some result, and analytical niceties are expected to yield. This makes it difficult to get funding for model criticism and improvement, and fosters an atmosphere in which scrupulous model criticism is viewed as vaguely traitorous.

(iii) The problems are often of mind-boggling scale and complexity (e.g., those addressed by macroeconomic modeling) and thus ill understood and not susceptible to the minute dissection and isolation of causal factors possible in the physical sciences.

(iv) Strategies for hedging against substantial uncertainty are often available, but

(v) Application of techniques motivated by laboratory sciences—especially the practice of picking a model and making statements conditional on it—presents a real danger in that it prejudices the process against the hedging strategies.

The latter three points in particular (to respond to Smith) make the Bayesian approach uniquely suited to understanding and characterizing where the action is in policy analytic situations. These problems have indicated to me the nature of the boundary between where formal and informal methods can be applied, although I find that boundary to be less well-defined than does Smith (1986).
5. FREEDMAN VS. MADANSKY

Madansky would like me to choose sides in this tiff. I decline, because I am ambivalent.

A substantial portion of the Army’s analytical effort goes into setting up models (usually the elaborate computer variety) and running dozens of “what if’s” through them. This can be a useful activity if it produces a better heuristic understanding of the modeled situation; as such, it provides an exercise in combining assumptions to see what they imply. But these models are regularly used as if their runs provided experimental replications of the modeled situations, which can lead to serious mistakes. This truth receives a lot of lip service within the Army but little serious regard. Admittedly, it is extremely difficult to apply empirical methods to the criticism and improvement of these models, because the data are scarce and problematic at best. But in the case of one close combat model, the institutional parents of the model do not want its users to understand the details of its algorithms, and actively discourage them from doing so. This is perverted in a very fundamental sense.

In a world where this happens, I can only applaud self-appointed and intentionally truculent critics like Freedman, even though I strongly disagree with him about substance and tactics in particular instances. But Freedman’s crusade has an irritating disingenuousness about it. He insists that models must be “right,” but never tells us what that is supposed to mean. Is Newtonian mechanics “right?” Of course not, but the bridges near Berkeley were built by engineers who acted as if it were, and I’ll wager that Freedman’s trips across them are not troubled by his knowledge of this fact. We know that in his own applied work Freedman must make similar judgments of when a model is good enough; so why doesn’t he tell us how he does it, instead of just kicking other people when they’ve done it particularly badly? (For discussions of “good enough,” see Hill (1986) and Leamer’s discussion of that paper, and for a simple formal approach to “good enough,” see Kadane and Dickey, 1980.)

The operative question cannot be whether the model is “right,” but whether the model’s users, and the consumers of their analyses, understand the assertions of information of which the model consists, and the nature of the justification (or lack thereof) for those assertions. The consumer of the analysis must ultimately judge whether the model is “good enough” for his purposes, for in any circumstance, regardless of the technique applied, validation is founded on subjective judgments and thus is necessarily subjective. You can fault analysts for being negligent or less than candid, but you can’t fault them for doing the best they can with what’s available. Beyond that, Freedman’s preference for qualitative over quantitative ways of doing this limited “best” can perhaps be justified on the grounds that it focuses the consumer’s attention more appropriately—many people are stupefied by computer models—but otherwise the preference seems stylistic.

ACKNOWLEDGMENT

In writing this rejoinder, I had the benefit of comments and suggestions from David Draper.

ADDITIONAL REFERENCES


