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 analysis 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 χ^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

ANDREWS, D. F. and HERZBERG, A. M. (1985). *Data. A Collection of Problems from Many Fields for the Student and Research Worker.* Springer, New York.

HILL, B. M. (1986). Some subjective Bayesian considerations in the selection of models. *Econometric Rev.* **4** 191–246.

HOLLAND, P. W. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970.

KADANE, J. B. and DICKEY, J. (1980). Bayesian decision theory and the simplification of models. In *Evaluation of Econometric Models* (J. Kmenta and J. B. Ramsay, eds.) 245–268. Academic, New York.

LEAMER, E. E. (1974). False models and post-data model construction. *J. Amer. Statist. Assoc.* **69** 122–131.

SMITH, A. F. M. (1986). Some Bayesian thoughts on modelling and model choice. *Statistician* **35** 97–102.

SMITH, A. F. M. and WEST, M. (1983). Monitoring renal transplants: An application of the Kalman filter. *Biometrics* **39** 867–878.