Inference and Decision Techniques: Essays in Honor of Bruno de Finetti (P. K. Goel and A. Zellner, eds.) 425–439. North-Holland, Amsterdam.

GEISSER, S. (1980b). Discussion of "Sampling and Bayes' inference in scientific modeling and robustness" by G. E. P. Box. *J. Roy. Statist. Soc. Ser. A* **143** 416–417.

GEISSER, S. (1980c). Predictive sample reuse techniques for censored data (with discussion). In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 433–468. University Press, Valencia.

GEISSER, S. (1982). Aspects of the predictive and estimative approaches in the determination of probabilities (with discussion). *Biometrics Suppl.* **38** 75–93.

GEISSER, S. (1985). On the prediction of observables: A selective update. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.) 203–230. North-Holland, Amsterdam.

GEISSER, S. (1987). Influential observations, diagnostics and discordancy tests. *J. Appl. Statist.* **14** 133–142.

GEISSER, S. and EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160. Corrigendum **75** 765 (1980).

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 speaking, what happens is that in large samples the procedure will pick some member of the parametric family close to the true structure (whatever that means) and then try to do the best possible for that member. It depends on the parametric family, on the type of procedure, on the true situation and on an unspecified kind of closeness whether the "best possible" for the model is any good at all for the true situation. Mere intuition can be very misleading.

It would be a delusion to think that a Bayesian approach, that is, the opportunity to choose also a prior, in such situations provides more security than the Neyman-Pearson version—if the family of models chosen by the statistician (i.e., the support of his prior) does not contain the true underlying situation, one has to go outside of the Bayesian framework in order to justify the use of a Bayes procedure.

For me, this infinite regress was a major conceptual difficulty when I first got into statistics in 1961; the stumbling block then turned out to be the stepping stone leading into a theory of robustness. Some personal reminiscences about the struggle preceding my 1964 paper may help to illustrate the point. Somehow, I then wanted to capture situations describable by statements like: "With this kind of data I would expect about 2% grossly wrong observations, but probably not more than 10%; these values could be anywhere." After some stillborn attempts with a nonparametric version of maximum likelihood, I naturally tried Bayesian approaches next, since by then I knew that even large data samples (I had had experiences with nonlinear least squares problems with a few thousand observations) would not allow me to assess distribution tails reliably without using outside information. However, I was unable to invent believable priors. After a while I realized that the problem was not

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 on 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 ψ function) turned out to be asymptotically minimax with respect to ε 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!

ADDITIONAL REFERENCES

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

HUBER, P. J. (1982). Astronomical dating of Babylon I and Ur III. With the collaboration of A. Sachs, M. Stol, R. M. Whiting, E. Leichty, C. B. F. Walker and G. van Driel. *Occasional Papers on the Near East* **1**, (4). Undena, Malibu, Calif.

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 puzzlement 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 brifely 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,

Joseph B. Kadane is the Leonard J. Savage Professor of Statistics and Social Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. These comments were written while the author was on sabbatical leave at the Center for Advanced Study in the Behavioral Sciences, Stanford, California.