of θ to be "unbiased," or require an advance specified upper bound to the probability of error of the first kind. That such requirements could lead to absurdities such as randomised "conclusions," assertions with only 90% confidence that a real number lay between $-\infty$ and $+\infty$, etc., was noted by several of the older Berkeley's associates; but the energy, courage, generosity of spirit, brilliant wit, and human warmth of N's character so impressed all those who came into contact with him that the inherent impossibility of the task N had set himself was not stressed, and old Berkeley grew into the over-rigid system which Lindley so mercilessly attacks. Of course, as with the somewhat similar attempts by von Neumann, Birkhoff, and others to "pure-mathematicise" quantum physics, N's programme produced many insights and valuable results in spite of its ultimate failure.

ADDITIONAL REFERENCES

Barnard, G. A. (1982). A new approach to the Behrens-Fisher problem. *Utilitas Math.* **21B** 261–271.

Barnard, G. A. (1989). On alleged gains in power from lower *P* values. *Statistics in Medicine*. To appear.

Chamberlin, S. R. and Sprott, D. A. (1989). The estimation of a location parameter when the scale parameter is confined to a finite range: The notion of a generalised ancillary statistic. *Biometrika* **76** 609–612.

Gardner, M. J. et al. (1989). Cancer near nuclear installations. *J. Roy. Statist. Soc. Ser. A* **152** 305–384.

Sprott, D. A. (1989). Inferential estimation, likelihood, and linear pivotals. *Canad. J. Statist.* To appear.

Comment

James O. Berger

1. INTRODUCTION

There are many reasons to adopt the Bayesian paradigm. Professor Lindley emphasizes the foundational and axiomatic rationales in this paper. Having followed that route to Bayesianism myself, I am particularly appreciative of the job Lindley has done in illuminating the route. I only regret that this paper was not around when I started studying the issues.

I emphasize the foundational nature of Lindley's paper for two reasons. First, it is a common misconception that the arguments for Bayesian statistics are all theoretical, as opposed to practical. To the contrary, an extremely strong case for Bayesian statistics can be made purely on the pragmatic grounds that it is much easier to understand and yields sensible answers with less effort. Lindley has reasonably concentrated on the foundational side, but it is important to note the existence of these very pragmatic rationales. Of course, I completely agree with Lindley that foundational issues can have a profound effect on practice.

The second reason for mentioning the foundational nature of the paper is that, in foundational matters, virtually everyone disagrees in some respect, even (or perhaps especially) Bayesians. Thus the bulk of my discussion focuses on the foundational differences that I have with Lindley, primarily the issue of specification of unique prior probabilities. While this is perhaps a significant issue foundationally, it is much less of an issue in terms of Bayesian statistical practice. Hence, my disagreements with Lindley are actually quite minor from the perspective of statistics in general. Indeed, my motivation for raising the issue (in Section 3) is mainly to argue that uncertainty in probability specifications can be incorporated into the Bayesian paradigm without any major changes being necessary.

2. FREQUENTIST BAYESIANISM

As I read Section 1 of the paper, I agreed with virtually all of the points raised but felt uneasy at the conclusion that coherence is missing from the Waldian paradigm. After all, admissibility is at the heart of the paradigm and, in a sense, admissibility is just a frequentist version of coherence.

Would Wald have disagreed that the correct solution to the mixture problem is to choose a procedure that is Bayesian? Perhaps not. Indeed, there have existed frequentists who consider themselves coherent Bayesians, in the sense that they agree with the use of Bayes' rules, and even utilization of prior information, but still want to base their evaluations of accuracy on frequentist (Bayesian) measures of performance. Such statisticians would presumably disagree with Lindley's statement that "only the Bayesian attitude is coherent ... Consequently the sample space is irrelevant." They would agree with the first part, but disagree with the second because of their insistence that only frequentist measures are meaningful.

James O. Berger is the Richard M. Brumfield Distinguished Professor of Statistics. His address is Department of Statistics, Purdue University, West Lafayette, Indiana 47907.

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

The frequentist Bayesian position cannot be undermined from within. It can only be questioned externally, with notions of conditioning. Thus some version of the conditionality principle (see Birnbaum, 1962, or Berger and Wolpert, 1988) is needed to argue for the posterior Bayesian approach. To illustrate this, consider a modification of one of my favorite examples.

EXAMPLE: Let E and E' be two experiments, each of which consists of observing X and Y, the observations equalling the unknown $\theta \pm 1$ with probability .5 each. In E it is known that X and Y are unequal, while in E' it is known that X and Y are equal. For experiment E, the obvious estimator of θ is $\hat{\theta} = (X + Y)/2$, which is admissible for reasonable losses and can be evaluated by noting that the frequentist probability that $\hat{\theta}$ equals θ is 1. For experiment E', a reasonable estimator of θ is $\hat{\theta} = X + 1$, which is admissible for reasonable losses and can be evaluated by noting that the frequentist probability that $\hat{\theta}$ equals θ is .5.

Consider now the *mixed experiment*, E^*, formed by selecting either E or E' on the basis of a fair coin flip. (This experiment is easily seen to be equivalent to that in which X and Y are independently equal to $\theta \pm 1$ with probability .5.) In E^*, $\hat{\theta}$, as defined above (equal to the average of X and Y if they differ, and equal to their common value plus 1 otherwise), is again admissible and is generalized Bayes. And the frequentist probability that $\hat{\theta}$ is equal to θ is now .75.

The point to be made is that all the above conclusions are compatible with frequentist Bayesian coherency. In each of the three experiments, the estimator used is the Bayes estimator with respect to the same (generalized) prior, so that there is no incoherency there. And the accuracy reports in the three experiments are fully consistent in a frequentist (even a frequentist Bayesian) sense.

What is questionable here is the violation of the conditionality principle, in that the accuracy report for E^* would always be .75, even though the coin flip will lead to actually performing E or E', which have accuracies of 1 and .5, respectively. The intuition of most people is that the reported accuracy should be that from E or E', whichever is actually performed, rather than the average of .75. This example provides especially strong support for this intuition, because it would be rather ludicrous to end up performing E, in which case θ clearly becomes known, and yet report an accuracy of .75. My point here is simply that conditioning, and not just coherency, must be involved to argue against the frequentist position.

There are, of course, many types of conditional coherency that could be employed to argue against the frequentist Bayesian position (such as separate "scoring" of the accuracy reports for each possible observation), but something akin to the conditionality principle can be found in all of them. Also, there have been efforts on the frequentist side to develop coherent frequentist theories that allow some conditioning, so as to escape the silly behavior described in the above example, but few predict success for these theories (and they cannot be fully consistent with the conditionality principle).

3. UNIQUENESS OF PROBABILITY AND MEASUREMENT

Enhancing ability to measure probability is of unquestioned value in Bayesian statistics. In this regard, the proposals of Lindley toward a theory of probability measurement are of great interest. The successes achieved in explaining existing ad hoc guidelines are impressive. The proposals also suggest intriguing new guidelines, such as the suggested use of the predictive form of Bayes' rule in measurement of probability. This is exciting stuff.

The question of the foundational centrality (as opposed to practical usefulness) of such a theory of measurement in the Bayesian paradigm is substantially murkier, however. The foundational role of measurement of probability is strongly related to another issue, that of uniqueness of probability judgement. Lindley's view on this subject is reflected by his comment in Section 3.2:

> "Other axiom systems lead to variants of the probability approach: for example, to upper and lower probabilities . . . These are defective to me because they do not incorporate the notion of a *unique* recommendation."

It is because of his desire to maintain the thesis of unique specification of probabilities, that Lindley introduces the notion that the problem requires only a new theory of probability measurement.

Before addressing this properly as a scientific issue, let me make some sociological observations. First, the axiom systems which lead to the unique-probability Bayesian paradigm all contain the unrealistic axiom that we are capable of arbitrarily fine distinctions in judgment; that, if one thought long enough, one could decide whether one's subjective probability of rain tomorrow was .38792567 or .38792566 (to paraphrase I. J. Good). Lindley's response—that the unique-probability Bayesian paradigm is an ideal that we approach through a theory of probability measurement—may be sensible scientifically, but it surrenders the axiomatic high ground. The opposing position—that the ideal is, in practice, not approachable (i.e., that the measurement problems are insurmountable)—is, logically, a viable escape for non-Bayesians.

For this sociological reason, I have always favored axiom systems (such as that of Smith, 1961) that weaken the axiom of complete comparability yet still basically lead to Bayesian analysis. One may have to worry about classes of probabilities (and utilities), but it is still only Bayesian processing that is allowed. (A few other esoteric possibilities might creep in, but nothing in contradiction to Bayesian reasoning can emerge.) One should, in practice, consider classes of probabilities and utilities anyway (through sensitivity studies), so giving up the complete comparability axiom regains the high ground at little cost.

In the scientific domain, much can also be said for the "classes of probabilities" approach. ("Probabilities" is here used to denote both the model and the prior distribution.) The reason is that, because of the difficulty of the probability measurement process, it is often impractical to have a priori a highly accurate assessment of all possibly relevant probabilities. Some probabilities will matter, and some will not, depending on the eventual data and the utility structure. The degree of accuracy necessary in the measurement process will likewise often depend on elements of the problem (such as the data) that are unknown or impossible to assess a priori.

This practical difficulty can be addressed via the "classes of probabilities" approach, in that one can begin with a broad class of prior probabilities (based on a few gross features that can be easily specified) and see if the posterior statistical conclusion is essentially the same for any prior probability in the class. (Traditionally, this is done just by trying disparate members of the class, but there is a large and growing literature concerned with global calculations involving large classes of prior distributions; cf. DeRobertis and Hartigan, 1981, and Berger, 1989.) If it is found that the probabilities in the class give markedly different answers, then further elicitation of probabilities (i.e., refinement of the class) will be needed. Happily, the above process will often indicate where refinements are most needed. In making these refinements, a theory of probability measurement may be very helpful, but consideration of when and where refinement is needed is arguably at least as basic. (A side benefit of this paradigm is that it can accommodate the group consensus problem: the "class of probabilities" could be that arising from a group of different individuals.)

Lindley draws an analogy between the unique-probability Bayesian paradigm and the paradigm for mapping the Earth's surface. The latter became routine when good measurement methods became available, and he hopes that the same will happen for the former. The analogy is good in many respects, but the above comments reflect two possible limitations. First, it can be argued that routine statistical users may never be highly proficient in probability measurement. Even experts may find it difficult to achieve more than, say, first significant digit accuracy. Thus, we may have to accept the frequent presence of a large degree of measurement uncertainty in our Bayesian analyses. The second point is the importance in statistics of interacting with the data in learning where to concentrate measurement efforts. I do not see an analog of this in the mapping problem.

There are, of course, certain concerns with a paradigm that allows refinement of probabilities in light of interaction with data. Everyone does it (in model choice, etc.), but few claim it as a virtue. A Bayesian justification for refinement, and claiming it as a virtue, can be made along the following lines.

Imagine yourself reading an applied Bayesian analysis. The author has considered a variety of models for the data and a variety of prior distributions for the parameters of the models, and perhaps a variety of residuals, Bayesian likelihood and predictive diagnostics, etc., all of which (together with your own knowledge of the subject) convince you that as many reasonable possibilities have been covered as can be expected. Lo and behold all considered models and priors yield essentially the same answer. Would you be happy, even if the development of the models and priors utilized the data?

I would, as long as I felt that the models and priors covered the range of reasonable possibilities. This would be especially so if I knew the researcher was honest, and would not purposefully fail to disclose models or priors that were reasonable and yet supported different conclusions. The point is that I see in front of me the data and the probabilistic descriptions of the situation, and as a Bayesian that is all I feel I need. I care only that all reasonable probabilistic descriptions have been considered. In particular, if the researcher has utilized the data to determine where to focus his elicitation efforts, I would not object.

To be sure, insistence on complete specification of all probabilities (including the model) prior to observation of the data would lessen the risk of purposeful or unintended "cheating" (by which I mean the selection, while looking at the data, of models or priors that are overly special), but this security comes with an impossibly heavy price.

In practice, this Bayesian paradigm involving nonunique probabilities and refinement would not necessarily operate through formal consideration of classes of probabilities. In particular, simplifications of the probability structure will often be made, for pragmatic reasons. A simplification should only be made, however, if it seems likely to yield the same answer that all reasonable Bayesian analyses would have yielded (i.e., if the answer is robust over the class of full probabilities consistent with the simplification).

As an example, if a plot of real-valued data closely follows a normal histogram with no outliers, one will usually feel confident in assuming normality, confident in the sense that alternative more general analyses (compatible with prior beliefs about the smoothness of the situation) will likely yield very similar conclusions. Or suppose there are clear "outliers." Depending on the situation, one might try outlier contamination models or densities with fatter tails, but in any case one is now probably willing to completely forget about the normal model. (The predictive likelihood of the data under a normal model will be so small that a full blown Bayesian analysis, incorporating prior probabilities of a range of models, would give essentially no posterior weight to the normal component.)

Of course, if the data does not look normal enough that one feels confident that the assumption of normality would be innocuous, but normality is not completely ruled out by a diagnostic such as predictive likelihood, then one has to carry along both normal and alternative models. Often I might carry along separate models until the end, hoping that the answer will turn out to be insensitive to the various models. If the model does turn out to matter, however, I will retroactively assign prior probabilities to the models or, more likely, try to embed them in a larger class of models (by introducing more parameters) and place a prior distribution on the class (or on the defining parameters). Of course, I've seen the data and this may contaminate my thinking about the prior, but again all that matters is whether or not the final probability structure is judged reasonable. (In this sense, of course, it is only the Bayesian who can even attempt to proceed. The classical statistician has no internally justifiable mechanism to go back and retroactively incorporate alternative models in a unified analysis. There is nothing systematically wrong with the Bayesian doing this; one can question the probabilities assigned to the models, but one can always question these.)

Note that the flexible Bayesian paradigm I am advocating does not license non-Bayesian methods, even in the simplification phases. For instance, I would argue strongly against choosing between two models based on a chi-squared significance test, since this has almost no relationship to whether a Bayesian analysis would allow simplification to one of the models or require consideration of both. Of course, approximate Bayesian methods may be useful. For instance, consideration of likelihoods alone (as opposed to posteriors) may often suffice to carry out a variety of judgements concerning simplification; the evidence from the likelihood can be strong enough that one feels confident that the posterior would reflect the same thing. Asymptotics is another potential tool in this regard. But all such tools have to justify themselves as being reliable in attaining the Bayesian goal. As an example, I feel that significance tests are so unreliable in reflecting Bayesian judgements, that I would argue against their being part of the simplification toolkit.

4. MINOR COMMENTS

(i) Some would argue with the implication at the beginning of Section 5.1, that the modern Bayesian paradigm has only developed in the last 50 years. Much of Laplace's work would be hard to differentiate from modern Bayesian work. And many statistical advances throughout the nineteenth century were initially made using "*inverse probability*."

(ii) I would quibble with the implication in Section 5.1 that it is "ridiculous" to think about a prior for a Greek letter θ, divorced from the reality it represents. At one level I certainly approve of the sentiment, but I also happen to feel that there is a very valuable role for noninformative priors in Bayesian statistics. And because noninformative priors are meant to be automatic priors that can be used for their associated statistical models, they will, almost by definition, be priors for Greek letters rather than for real quantities. Of course, not all Bayesians are enamored with noninformative priors. (My advocacy of noninformative priors may seem somewhat paradoxical, in that I argued in Section 3 against basing the foundation of statistics on using a single prior. How can I now say that it is okay to use a single automatic prior, one that requires no thought? Well, Bayesian analysis with a noninformative prior is very good; Bayesian analysis with a subjective single prior is usually even better; and doing it with a class of priors is best.)

(iii) Multiple comparisons and related ranking problems (cf. Berger and Deely, 1988) are indeed wonderful domains for showing the intuitive sensibility of the Bayesian approach. As another illustration of this, suppose, in an unbalanced model, that the largest observed treatment effect has associated with it a much larger variance than the other treatment effects. Intuition would argue against choosing this as the best treatment, but it is hard to do anything else within the classical paradigm. On the other hand, common hierarchical Bayesian analyses naturally lead to a substantial downweighting of the largest observed treatment effect in such a situation.

ADDITIONAL REFERENCES

BERGER, J. (1989). Robust Bayesian analysis: Sensitivity to the prior. In *The Proceedings of the Conference in Honor of I. J. Good* (K. Hinkelman, ed.). To appear.

BERGER, J. and DEELY, J. (1988). A Bayesian approach to ranking and selection of related means with alternatives to AOV methodology. *J. Amer. Statist. Assoc.* **83** 364–373.

BERGER, J. and WOLPERT, R. (1988). *The Likelihood Principle: A Review and Generalizations*, 2nd ed. IMS, Hayward, Calif.

BIRNBAUM, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* **57** 269–326.

DEROBERTIS, L. and HARTIGAN, J. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **9** 235–244.

Comment

José M. Bernardo

I was delighted to be asked to contribute to the discussion of this article by the man whom I have always proudly considered my *maestro*. I will limit my comments to a couple of issues.

1. Professor Lindley has long been arguing for the indirect assessment of probabilities, suggesting that one should always try to "extend the conversation" to include other related events, and then combine the results by simple use of probability theory. It is hard to overestimate the importance of such advice, and the work he reports on conditions under which improvement is guaranteed is especially welcome.

I would like to illustrate this procedure with a suggestive example drawn from my recent work in election forecasting. Trying to predict the outcome in Valencia of the recent European Parliamentary elections, I designed a survey where 1000 people over 18 randomly chosen from the census were asked to state not only the party they intended to vote for, but *also* the party they voted for in the last election. By only using the numbers $\{n_i,\ i = 1, \cdots, 6\}$, of the people expressing their intention to vote for party i, I got the estimates of the percentages of the vote to be obtained by each party which are reproduced in the first row of Table 1.

Alternatively, using the numbers $\{n_{ij},\ i = 1, \cdots, 6,\ j = 1, \cdots, 6\}$, of the people expressing their intention to vote for party i given that they voted j last time, and then using the probability equation

$$p(i \mid \text{data}) = \sum_{j=1}^{6} p(i \mid j, \text{data}) p(j),$$

I obtained the estimates reproduced in the second row. Note that the $p(j)$'s, the proportion of people who voted for party j last time, *are known*, for those are the results from the past elections.

In both cases I used a hierarchical Multinomial-Dirichlet model, with a reference prior for the Dirichlet (hyper)parameters, and allocated nonresponse by means of a probabilistic classification procedure (Bernardo, 1988) based on the social profiles (age, sex, level of education) of the nonrespondents, which are known from the census.

Comparison of these estimates with the final results, reproduced in the third row of Table 1, is striking. The direct estimates are rather poor, probably due to the bias induced by people's propensity to relatively overstate their alignment with the party in power (the socialists in Spain). The indirect estimates, however, are surprisingly good, with an average absolute error of about 0.4%, to be compared with the standard deviations of about 1.5% which would correspond to the naïve analysis of the sample of size 1000. It is important to note that I had no need to invent some form of "bias correction"; probability theory did it all "automatically."

2. Any review is invariably biased by his author's preferences, and Lindley's account is no exception. I would like to draw attention to one of my own biases, the role and use of reference "noninformative" priors, which he has chosen not to mention.

In Section 5.1, Lindley recognizes the need for robust procedures with respect to the choice of the prior $\pi(\theta)$, to the point of considering this necessary for the change of paradigm to take place; surprisingly however, he blames Berkeley for not taking on the job. But, if Berkeley has not, Bayes has made some progress. Indeed, reference priors (Bernardo, 1979; Berger and Bernardo, 1989) are best seen as robust

TABLE 1
European parliamentary elections. Percentage of valid votes in the province of Valencia

	Socialist	Conservative	Nationalist	Communist	Liberal	Other
Direct	53.9	15.7	7.2	8.0	5.5	9.8
Indirect	41.1	20.0	10.4	7.3	6.4	14.8
Final	41.0	20.7	11.0	6.5	6.3	14.5

José M. Bernardo is Personal Advisor to the President of the State of Valencia and is on leave of absence from his position as Professor at the University of Valencia. His mailing address is Departamento Estadística, Presidencia de la Generalitat, Caballeros 2, E-46001 Valencia, Spain.