DISCUSSION

W.F. EDDY (*Carnegie-Mellon University*):

I predict that by the end of this century the religious cult of Pure Bayesian Statistics (PBS) will die. There will be no martyrs. Righteousness is not the question; God will not decide in favor of incoherence and destroy Las Fuentes as he destroyed Sodom and Gomorrah. PBS will die the death of the buggy whip, through disuse.

Lest I be misunderstood, by PBS, I mean the belief that finding the distribution of unknown parameters *conditional on the data assuming the truth of the model is the objective of statistics*. The fundamental difficulty with PBS is that all inference is based on the truth of the model. And despite disclaimers I doubt that any practicing statistician believes in the truth of his model.

Professor Box apparently agrees. As I understand his thesis, one first uses sampling theory to find a "true" model and then uses Bayes theory to estimate the parameters in this model. The thrust of his argument is that allowance must be made

for the possibility that the model was not sufficiently broad and thus the prior distribution didn't really account for all uncertainty. On the face of it, this is a valuable thought.

However, Professor Box suggests that one should consequently do diagnostic checking. That is, after finding some unusual aspect of the data one should compute a discrepancy function and compare the observed value with the appropriate reference distribution.

This, I believe, is a mistake. Because the particular discrepancy function was chosen after looking at the data the reference distribution will usually suggest the observed value is unusual; but this is exactly the reason we computed the discrepancy function in the first place. Comparing an observed discrepancy to a reference distribution can only be useful for specific *a priori* departures from the model.

This is not to say that examining residuals and computing discrepancies is worthless. On the contrary, there is no substitute for careful residual analysis. Professor Box and I agree on this point and its implication: Model Building/Data Analysis is subjective. Different people see different things in their data and consequently add different parameters to their models.

I don't believe, however, that Professor Box has solved the fundamental dilemma of statistics: How to generalize from the specific data at hand?

Professor Freeman has presented us with a very practical comparison of several "outlier" linear models. I have been intrigued by the models and their implications but I am puzzled about their Bayesian-ness and thus the quotes around "outlier". In common usage, an outlier is an observation which *appears* to be different than the rest of the data (I emphasize *appears* because it is obviously a subjective matter which aspects of the data one examines). Now the Bayesian is compelled to choose his model(s) before seeing the data and thus, it seems to me, is in a quandry as to how to include the outliers in his model since he doesn't yet know which aspects of the data appear to be different. Since the models here are obviously geared to location and scale shifts (slippage outliers) perhaps the outlier-ness of the models is not to be questioned. The solution to my puzzlement may be that Professor Freeman uses the term "outlier" as shorthand for "what a non-Bayesian would call an 'outlier'." Enough philosophy.

By partitioning the data into outliers and non-outliers he writes the posterior distribution of β as

$$p(\beta \mid y) = \Sigma w_{(r)} p_{(r)}(\beta \mid y)$$

This device has two advantages: first, it allows analysis to proceed conditionally on particular observations being outliers and thus greatly simplifies calculations; second, it allows subsequent inference about which observations are outliers. Professor Freeman considers three specific outlier models: BT, AB, GDF. All three models suppose that the outliers are uniformly distributed over the observations; a more realistic model might distribute them conditionally on X.

The BT model says outliers have the same mean but are scaled by a factor of k. The posterior probabilities ($W_{(r)}$) will be largest when the outliers are observations at one or both extremes. The AB model says all outliers have a different (common) mean.

Consequently, the posterior probabilities will be largest when the outliers are a group of observations at one extreme. The GDF model says outliers each have a different mean. Thus, they are eliminated from the analysis since they contain no information about either β or σ^2; furthermore, the $w_{(r)}$ will be largest when the outliers are two groups, one at each extreme.

All three of the models can be viewed, conditionally on particular observations being outliers, as weighted least squares with the weights depending on the particular outlier model. That is, for all three models

$$\hat{\beta}_{(r)} = (X'V_{(r)}X)^{-1}X'V_{(r)}y$$
$$s^2_{(r)} = (y-X\hat{\beta}_{(r)})'V_{(r)}(y-X\hat{\beta}_{(r)})$$
$$B_{(r)} = (\upsilon_{(r)}/s^2_{(r)})X'V_{(r)}X$$

and

$$w_{(r)} \propto c_{(r)}|X'V_{(r)}X|^{-1/2}s_{(r)}^{-\upsilon_{(r)}}$$

For simplicity suppose the observations are permuted so the r outliers occur first. Then for the BT model

$$V_{(r)} = \begin{bmatrix} k^{-2}I_r & 0 \\ 0 & I_{n-r} \end{bmatrix}$$

$$\upsilon_{(r)} = n-p,$$

$$c_{(r)} = \left[\frac{\alpha}{k(1-\alpha)}\right]^r \text{ and}$$

for the AB model

$$V_{(r)} = \begin{bmatrix} J_r/r & 0 \\ 0 & I_{n-r} \end{bmatrix}$$

where J_r is an $r \times r$ matrix of ones.

$$\upsilon_{(r)} = n-p-1,$$

$$c_{(r)} = (\alpha/(1-\alpha))^r r^{-1/2} \text{ and}$$

for the GDF model

$$V_{(r)} = \begin{bmatrix} 0 & 0 \\ 0 & I_{n-r} \end{bmatrix}$$

$$\upsilon_{(r)} = n\text{-}p\text{-}r$$

$$c_{(r)} = 1$$

The great advantage is that we can now examine the $V_{(r)}$ to see if we really want to use a particular model; we can quickly examine new proposed outlier models.

I personally find the GDF model somewhat disquieting; completely ignoring extreme observations seems dangerous. An alternative I would prefer is a mixed BT-AB model as follows: With probability α_j each observation has mean $X\beta + \delta_j$ and variance $k^2_j\sigma^2$ for $j = 1,2$ and with probability $1\text{-}\alpha_1\text{-}\alpha_2$ each observation has mean $X\beta$. Take $\alpha_1,\alpha_2,k_1,k_2$ known and uniform (improper) priors on β,δ_1,δ_2, and log σ. For r_1 and r_2 outliers, respectively, this yields (in obvious notation)

$$V_{(r_1,r_2)} = \begin{bmatrix} k_1^{-2}(I_{r_1}\text{-}J_{r_1}/r_1) & 0 & 0 \\ 0 & k_2^{-2}(I_{r_2}\text{-}J_{r_2}/r_2) & 0 \\ 0 & 0 & I_{n-r_1-r_2} \end{bmatrix}$$

$$\upsilon_{(r_1r_2)} = n\text{-}p\text{-}2$$

$$c_{(r_1r_2)} = [\alpha_1/k_1(1\text{-}\alpha_1)]^{r_1}\ [\alpha_2/k_2(1\text{-}\alpha_2)]^{r_2}\ \ (r_1r_2)^{-1/2}$$

This model uses either location or scale (or both) information from the outliers; only when the r's are one does it reduce to the GDF expedient of ignoring data.

A. O'HAGAN (*University of Warwick*):

Professor Box argues that sampling theory methods are appropriate in diagnostic checking, and I strongly disagree. But whilst elaborating on this, let me say what a pleasure it is to find that he is actually tackling the right problem in basically the right way. The crucial point is the recognition that every statistical analysis, Bayesian or otherwise, is conditional on the truth of its assumptions. Any analysis which goes no further, which does not challenge these assumptions, is incomplete. So Professor Box is right in pointing to a need for procedures for diagnostic checking. And with the accuracy of an experienced data analyst he chooses the right tool, the predictive density $p(\mathbf{y}/M)$. Then inconceivably he uses the tool in entirely the wrong way. There is a perfectly natural Bayesian approach which uses the predictive density but never lapses into the discredited sampling-theory use of tail area probabilities.

Consider the basic model M and an alternative M_1. Conditional on M we obtain the basic posterior density $p(\theta/\mathbf{y}_d,M)$. Or conditional on M_1 we could obtain a different posterior density $p(\theta/\mathbf{y}_d,M_1)$. We now widen the analysis by conditioning on the truth of *either M or* M_1. We need extra prior probabilities $P(M/M$ or $M_1)$ and $P(M_1/M$ or $M_1) = 1 - P(M/M$ or $M_1)$, then the posterior analysis is completed by finding the corresponding posterior probabilities $P(M/\mathbf{y}_d, M$ or $M_1)$ and its complement. This can

be done using Bayes' theorem, which gives:

$$\frac{P(M/\mathbf{y}_d, M \text{ or } M_1)}{P(M_1/\mathbf{y}_d, M \text{ or } M_1)} = F. \frac{P(M/M \text{ or } M_1)}{P(M_1/M \text{ or } M_1)}$$

where

$$F = \frac{p(\mathbf{y}_d/M)}{p(\mathbf{y}_d/M_1)}$$

is the so-called Bayes factor, which converts prior odds into posterior odds. This is where the predictive density enters the analysis, but since the approach is Bayesian and obeys the Likelihood Principle, only the predictive density for the observed $\mathbf{y}_d$ is relevant. By looking at tail-area probabilities, involving $p(\mathbf{y}/M)$ for other values of $\mathbf{y}$, Professor Box is making a fundamental departure from the correct Bayesian solution. Why should he do this?

Perhaps the answer is that his approach seems to avoid the need to specify the alternative model M_1. Formally, of course, we cannot discredit M without consideration of alternatives. It is to be discarded if $p(\mathbf{y}_d/M)$is small *not* relative to the value it might have taken had some other sample been observed, but relative to the value it would take under some viable alternative M_1. The word "viable" is to convey the fact that $P(M_1/M \text{ or } M_1)$ should not be extremely small, otherwise a very small value of F need not lead to posterior odds strongly favouring M_1.

In practice we cannot formally consider all the possible alternatives, and if Professor Box has succeeded in avoiding the need for them then this is quite an achievement. He actually refers to the way his procedure might be applied informally, in practice, as follows.

> "In practice... diagnostic checking... is often conducted by visual inspection of residual displays or other more sophisticated plots... The statistician is looking for features in the data which would be surprising or unusual if the model M were true. Such a feature can be described by a function $g(\mathbf{y}_d)$ and its unusualness... measured by reference to $p(g(\mathbf{y})/M)$."

The reason for suddenly introducing $g(\mathbf{y}_d)$ is mentioned in his preceding paragraph, but is much better shown in an example which unfortunately does not appear in the shortened version of the paper. This example was of a sample, acoording to M, from a normal distribution. In diagnostic checking in relation to this example, he clearly has in mind the possibility of outliers as one potentially surprising feature of the data. But the predictive density $p(\mathbf{y}_d/M)$ depends only on the sufficient statistics s^2 and $\bar{y}$. Therefore it registers only weakly the surprise we feel when the data suggest the presence of outliers, for then it is more the pattern of data points than their location or sprend which catches our eye. But clearly Professor Box can choose a $g(\mathbf{y}_d)$ which would register our surprise much more strongly. This is why $g(\mathbf{y})$ is a necessary artefact in his approach, but of course the choice of $g(\mathbf{y})$ is no different from a choice of alternative model.

The correct Bayesian approach makes it clear that surprise is not enough. What a practising statistican does when he looks for surprising and interesting features in his data is more sophisticated than Professor Box supposes. He may have no alternatives in mind explicitly beforehand, and may find it difficult to formulate one afterwards, but *viable* alternatives are implicit in all the ways in which he chooses to look at his data. This is where his skill and experience tell - in what he chooses to look at, in what he registers surprise at. His reaction signifies not only that $p(\mathbf{y}_d/M)$ is small (surprise!) but also that his experience tells him that he will probably be able to find an alternative M_1 such that $p(\mathbf{y}_d/M_1)$ is much larger, i.e. the surprise is removed, and such that $P(M_1/M$ or $M_1)$ is not negligible.

The case of surprising outliers leads neatly to Professor Freeman's paper. He presents three different alternative models, each of which allows a mechanism for the occurrence of outliers. Each would in general greatly reduce the level of surprise we would feel when confronted by data exhibiting outliers, but each mechanism is different. Consider Professor Freeman's analysis of the Darwin data. On the assumption that there are two outliers the Abraham-Box model fails to identify "the most obvious pair (-67, -48)" as the culprits, and he concludes that "The [AB] model is clearly not a good one for identifying outliers". The conclusion is far too strong. The point is that if we believe the AB model to be appropriate then (-67, -48) is *not* a terribly obvious outlier pair, since to accommodate both these as outliers with a single value of the discrepancy parameter δ still necessitates large residuals. The element of surprise is still quite strong. Whereas under the BT model, for example, the Darwin data would be much less surprising. The conclusion is that *if* the BT model were *a priori* viable then the data would favour it through the Bayes factor F, and we would say that the AB model is probably not correct *for these data*.

Professor Freeman's other examples are similar. What he sees as an outlier may not be the kind of outlier generated typically by one or other of the three models. Performance is inversely related to surprise. The examples are instructive because they tell us something about the different outlier-producing mechanisms of the various models, which in practice will help us to assess prior probabilities.

It is interesting that by focussing his attention on *identifying* outliers Professor Freeman places very different emphasis from Professor Box, who would be more concerned with estimating β. The unstated implication is that all three methods would yield robust inference about β, but this is not true. The AB method simply gives suspected outliers a reduced weight, and if they deviate far enough from the others their influence can be strong. In O'Hagan (1979) I have looked at how robustness can be achieved simply by assuming that the data are sampled from a distribution with a suitably thick tail. Outlier rejection will then take place regardless of our prior distribution for β. It is interesting that, in an earlier paper than their one on outliers, Box and Tiao (1962) examined the Darwin data under thick-tailed alternatives, but that none of their distributions had thick enough tails to guarantee outlier rejection (see O'Hagan (1979)). I hope to publish numerical results soon.

I would like to end by emphasising that I found both papers profoundly stimulating, and that, if I have appeared to be highly critical, this is merely because the

questions they raise are so important and so deep. I would like to congratulate and to thank both authors.

J.M. BERNARDO (*Universidad de Valencia*):

Professor Box's thought provoking paper distinguishes between model criticism and parameter estimation and goes on to advocate a (conditional) Bayesian analysis for the latter but a frequentist-type one for the former. I feel that the division between model and prior is somewhat illusory. What one really needs is the joint distribution $p(x,\theta)$ and it is only tradition which gives $p(x|\theta)$ and $p(\theta)$ a different theoretical status. Indeed when one uses some sort of plot to 'test' empirically $p(x|\theta)$ what one is really 'testing' is rather the predictive $p(x)$. Whether you call $p(x,\theta)$ a 'model' or 'a prior' is unimportant, but it seems to me that empirically testable prediction conditional to $p(x,\theta)$ is often what is precisely needed.

P.J. BROWN, (*Imperial College, London*):

Some of the discussion on outliers so far today does seem a little unreal. In my experience identification of an outlier is just a signal to investigate further. On closer inspection and with more data there may well be good reasons to so regard it. In election night forecasting, for example, 'stringers' waiting at the counting halls are relied upon to telephone in the results as soon as they are declared. It is understandable that a few may take to alcohol to while away the long night. An absurd result, if flagged, will result in further corroboratory telephone calls to the constituency. Thus this outlier problem is sequential.

I would like to see much more precision in the definition of the term 'outlier'. Obviously there are workable definitions outside that of data transmission errors but, without more careful examination of the utility of the concept and its realisation, I think one cannot proceed beyond accepting that there are a number of different possible conclusions, each having some plausibility.

A.P. DAWID (*The City University*):

It is not necessarily true, as Professor Box suggests, that the use of improper priors does not allow model criticism. Suppose our observation is y, with the binomial distribution $\mathbf{B}(n;\theta)$, and we use the improper prior distribution $\beta(0,0)$, viz. $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, considered, say, as a limit of $\beta(\alpha,\alpha)$ with $\alpha \to 0$. The limiting predictive distribution has $P(y = 0) = P(y = n) = \frac{1}{2}$, so that any value $0 < y < n$ discredits this "model-cum-prior". However, if we believe the weak spot to be in the prior specification, rather than the sampling model, we should not be too hasty to discard our assumptions, since our posterior distribution is not likely to be sensitive to the choice of prior. Somewhat paradoxically, it is for the case $y = 0$ (or n), which does *not* discredit the improper prior, that we must be most careful about specifying the "true" prior distribution. This example indicates to me that model-checking using predictive distributions may not always be appropriate.

J.M. DICKEY (*University College Wales, Aberystwyth*):

Professor Freeman has not made the assumption of condition continuity in his paper here on outliers, in the sense that in Section 4 his prior opinion concerning a single outlier is not necessarily the same as if he had been told that of two outliers one had zero disturbance. I am wondering what kind of relationship one would want between these integrable prior distributions conditional on different models. (See my discussion to the paper by Professor A.F.M. Smith in these Proceedings).

I don't like the assumption in Section 1 of a uniform improper prior distribution in as many dimensions as the number of outliers (G.D.F. approach). In principle, the number of dimensions can be as high as the sample size, and constant nonintegrable densities are notoriously troublesome in high dimensions.

I hope Professor Freeman will develop further his interesting integrable-prior methods (Section 4), and report his experience in their use bearing on the important questions of choice of prior distribution.

I.J. GOOD (*Virginia Polytechnic and State University*):

I am pleased to see that so distinguished a statistician as Professor Box has emphasized a Bayes/non-Bayes compromise or synthesis because that has been my philosophical position for a long time, although I regard the Bayesian side of it as more fundamental. One example of such a marriage, especially close to the theme of Professor Box's lecture, is the use of orthodox significance tests for choosing a hyperparameter, and for testing a Bayesian model, for density estimation and bump-hunting. This idea was presented in August 1974 in the invited General Methodology lecture at the annual meetings of the American Statistical Association in St. Louis, Missouri. Practical applications of the method are given in Good and Gaskins, (1980). In the Journal of the American Statistical Association, 75 (1980), 42-73 (with discussion).

By saying that the Bayesian side of the coin is more fundamental I mean that the use of tail-area probabilities can be roughly justified by Bayesian arguments when it can be justified at all. (See my contribution to Professor Barnard's seminar for references).

A.F.M. SMITH (*University of Nottingham*):

Box argues that *criticism* must ultimately appeal to sampling theory for its justification. He may well be correct, but I am not convinced that the development given here succeeds in clearly demarcating an area of critical activity that is out of bounds to a Bayesian. There would seem to be, in broad terms, a one-to-one relationship between any diagnostic checking procedure and an *implicit* family of alternative models. Indeed, Box comes close to conceding the primacy of such implicit alternatives when he turns to "Choosing the diagnostic checks". The ensuing discussion of "Diagnostic checking and Robustification" appears to acknowledge this one-to-one correspondence and thus, surely, to admit that whatever can be probed using a diagnostic check function can also be probed by using Bayes factors against appropriate alternative models. Some of the author's general discussion seems intended

as a defence against this latter accusation, but it has equal force, or rather lack of it, against *both* approaches. Either we attempt no criticism (i.e. *no* diagnostic checks, *no* Bayes factors) or we attempt *some* limited criticism (i.e. apply a *finite* number of diagnostic checks, calculate a corresponding *finite* number of Bayes factors). In neither case can we test against all possible departures (using *all possible* diagnostic checks, or a *totally comprehensive* model).

I am not disposed to think that "it" (the advancement of learning?) can all be done with Bayes, but I do feel that the kinds of *local* model criticism discussed in this paper *can* be carried out within the Bayesian framework and that, at most, we are here discussing rather pragmatic issues and not fundamental questions about inferential paradigms.

REPLY TO THE DISCUSSION

P.R. FREEMAN (*Leicester University*):

Several discussants mentioned the need for a proper definition of an outlier, so that we are all clear what we are talking about. It seems to me impossible to ever get a fully operational definition, although we can all recognise an outlier when we see one, since if we try to model formally all the possible kinds of outlier, we shall end up with something which is far too complex to be of any use. For example, Professor Eddy's suggested model gains in flexibility, certainly, but loses in complexity since we would have to take a double sum over all values of r_1 and r_2, and the combinatorial explosion would defeat us for even very small sample sizes.

I think that outlier identification is important since ideally we want to do the sequential checks just as Dr. Brown describes (and to ensure that the faulty "stringers" are not employed at the next election). There is no real substitute for the hard work of going back over records and finding the exact source of error (or for failing to find any error), and for then re-analysing the data with the suspicious values either corrected, deleted or left unchanged. But in the real world this is just far too much trouble and some robustness of analysis is also desirable so as to save much of this work. It was in this sense that I criticised the AB model. I should have said that it is not flexible enough to detect some kinds of outliers that I think I would like to have detected, namely those occurring at both ends of the data.

I take Dr. O'Hagan's point that we need some automatic protection against very extreme observations. The GDF model does this by ignoring them completely, but I agree that models with thick tails should be used in many situations where we dangerously use normal tails at present. Dr. Eddy finds this aspect of GDF unattractive, but I would justify it by saying that the overall effect is somewhat comparable to that of the jack knife with the more sensible refinement of taking a weighted average of the results obtained by dropping one or more observations at a time. The extremely deviant values only get ignored completely when they are so far out that one subset attracts all the posterior weight to itself.

I thank Dr. Eddy for unifying the notation of the 3 models. I only wish I had thought of doing so when I wrote the paper.

Professor Dickey comes close to the heart of the problematic area of my paper-the choice of priors. I, too, am perturbed by the improper priors in the GDF model, though they do in practice give beautifully robust results for parameter estimation. I am not too worried by the lack of condition continuity in my priors as I can see no intuitively compelling reason to obey that condition and it is not, as far as I can tell, an essential requirement for coherence. The dependence on the exact form of the conditioning again makes me sceptical of its usefulness.

The proper-priors section of my paper still seems to me to contravene what was enunciated verbally at the conference as Lindley's principle - that if you take a problem, treat it coherently and use sensible priors you will always get a sensible answer. It is not clear to me what part of the conditions I am violating, thought the answers I get are disappointingly misleading. Perhaps the attemps to discriminate among members of a nested family of hypotheses is doomed to failure due to lack of enough data, whatever the priors. Only further work and deeper consideration will tell.

G.E.P. BOX (*University of Wisconsin*):

It is perhaps hardly surprising that I have not been totally successful in convincing a conference of Bayesians of the auxilliary need for Sampling Theory and I have sympathy with some of my critics.

In response to Professors Smith, O'Hagan and Eddy, my main point is that since Bayes is conditional, if it is to be used exclusively in the pursuit of an adequate model, we inevitably find ourselves engaged in a game of "Yes but". It is rather as if, when I was preparing for my early morning dash to the airport on leaving Los Fuentes, my conversation with the hotel manager had gone as follows:

> Do you think I can catch my plane?
> Yes, if the taxi is on time.
> Do you think the taxi will be on time?
> Yes if the taximan gets up early enough.
> Do you think he will get up early enough?
> Yes if his wife remembers to wake him.
> etc., etc.

More specifically, *however* far the model building process had been carried by Bayesian methods the final model would still be

$$p(\mathbf{y},\theta \mid M_k) = p(\theta \mid \mathbf{y},M_k)p(\mathbf{y} \mid M_k)$$

and there remains the n-dimensional space of the marginal predictive distribution $p(\mathbf{y} \mid M_k)$ which has not yet been explored and which can, on a sampling theory argument, discredit the relevance of the assumptions on which the Bayesian analysis is conditional.

I grant that, as soon as we start to consider specific alternative models, then

Bayesian versions of diagnostic checks are available. In particular for the case of a discrepancy parameter β taking the value $\beta = \beta_0$ for an ideal model M, one way in which this duality may be formalized is as follows. A natural function of the data to consider for making diagnostic checks is

$$g_\beta(\mathbf{y}) = \left. \frac{\partial \log p(\mathbf{y}\,|\,\beta)}{\partial \beta} \right|_{\beta = \beta_0}$$

But since $p_u(\beta\,|\,\mathbf{y}) = p(\beta\,|\,\mathbf{y})/p(\beta)$ we see that $p_u(\beta\,|\,\mathbf{y}) = p(\mathbf{y}\,|\,\beta)$ so that $g_\beta(\mathbf{y})$ is Fisher's score function for the parameter β. So it may be argued why not just look at the distribution $p_u(\beta\,|\,\mathbf{y})$?

The amount of effort that can be expended on any particular analysis is finite and we may not want to expend a full Bayesian analysis on every discrepancy that occurs to us. In many cases the model builder would be satisfied with graphical checks. Even so such checks need not be entirely ad hoc and indeed it is possible to show that $g_\beta(\mathbf{y})$ defined above is often valuable in showing the form that graphical checks should take.

I, of course, agree with Dr. O'Hagan that the predictive ratio $p(\mathbf{y}\,|\,M_1)/p(\mathbf{y}\,|\,M_0)$ can be used not only to indicate the appropriate form for diagnostic checking functions, but also in the direct Bayesian assessment of the relative evidence of any one model versus another. Notice, however, that the inherent Bayesian limitation of conditionality ensures that, however large this ratio may be, the preferred model M_1 can still be manifestly implausible because $Pr\{p(\mathbf{y}\,|\,M_1) < p(\mathbf{y}_d\,|\,M_1)\}$ is small.

I am grateful to Professor Good for his encouraging comments and references.

Consider Professor Dawid's example when the limit $\alpha = 0$ is *not* approached, remembering to make due allowance for the fact that while θ is continuous y is discrete. The choice of prior $\beta(\alpha,\alpha)$ is equivalent to supposing a uniform prior in $\phi = \int^\theta t(1\text{-}t)^{\alpha-1}dt$. If we take $\alpha = 1$ the predictive distribution $p(y\,|\,M)$ is such that $p(y/N\,|\,M) = (N+1)^{-1}$, $(y = 0,1,2,\ldots N)$ and the predictive cumulative distribution plots as a linear "staircase function" against y/N. Thus supposed indifference about θ itself results in no predictive critical ability for y/N. But suppose following Jeffreys we set $\alpha = \frac{1}{2}$, then $\phi = \sin^{-1}\sqrt{\theta}$, $0 \le \phi \le \pi/2$. The corresponding predictive distribution for $\sin^{-1}\sqrt{(y/N)}$ is, of course, unequally spaced but again the cumulative distribution even for small samples approximates a straight line and supposed indifference about $\sin^{-1}\sqrt{\theta}$ results in no predictive critical ability for $\sin^{-1}\sqrt{(y/N)}$. The approximation holds for other non-zero values of α, however, as we go to the limit $\alpha = 0$ the range for ϕ goes from $-\infty$ to $+\infty$ and consequently the discrete predictive distribution is dominated by values corresponding to $y = 0$ and $y = N$ which are infinitely removed from other realizations. I would argue, therefore, that this example reconfirms the unsuitable nature of this particular prior, the unsuitability of which as Professor Dawid says is not clear from consideration of the posterior distribution which over the range considered is sensitive to the changes discussed. In choosing prior distributions we must clearly consider their predictive consequences.

Although I much enjoyed this Bayesian Conference, there was for me an eerie feeling that something important was missing. Bayesian inference is an instrument for

use in scientific enquiry. But except for a couple of rather distant echos we seemed to have talked for a week securely insulated from the world of real investigation. It has been said that

> "Theory and Practice are like man and wife in a happy marriage; each complements and inspires the other and without interaction between them there can be no new life".

Certainly the work of such practicioners as Gauss, Laplace, Daniel Bernoulli, Fisher and Jeffreys provides no reason to doubt this aphorism.

I believe it is agreed that scientific iteration employs in alternation the dual processes of model criticism on the one hand and exploitation of the tested model on the other. Suppose we accepted, as I suggested in my paper, that two different kinds of inference are needed to conduct these two different activities conveniently. Suppose it was agreed that the first activity (which subsumes model specification/identification and tests of fit) although often conducted informally under the name of Exploratory Data Analysis ultimately requires Sampling Theory for its justification, while the second requires Bayesian Theory. Then it would be understandable why a purely Bayesian conference would have little to say about any real scientific investigation (and perhaps a conference entirely devoted to "Exploratory Data Analysis" might be equally disappointing).

It is rather as if we called a conference of airplane pilots* who knew everything about landing a plane but nothing about how to take off (or vice versa). At such a conference there should be little surprise if in a welter of papers viewing from every angle the finer theoretical points of landing an airplane the discussion seldom turned on going anywhere or on interesting voyages experienced.

* They might more properly be called "landers" rather than pilots, just as some of us are called Bayesians rather than Statisticians.

REFERENCES IN THE DISCUSSION

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

GOOD, I.J. and GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scatering and meteorite data. *J. Amer. Statist. Assoc.* **75**, 42-73 (with discussion).

O'HAGAN, A. (1979). On outlier rejection phenomena in Bayes inference. *J. Roy. Statist. Soc. B.* **41**, 358-367.