AD-A218 420

(2)

UMENTATION PAGE

Form Approved
OMB No. 0704-0188

ion is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, leting and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this ducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

2. REPORT DATE
October 10, 1989

3. REPORT TYPE AND DATES COVERED
FINAL REPORT, 1 Apr 87 to 31 Mar 89

4. TITLE AND SUBTITLE
A THEORY OF BAYESIAN DATA ANALYSIS

5. FUNDING NUMBERS
AFOSR-87-0192
61102F 2304/A5

6. AUTHOR(S)
Bruce M. Hill

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Department of Statistics
Ann Arbor, MI 48109-1092

8. PERFORMING ORGANIZATION REPORT NUMBER
AFOSR-TR- 90-0213

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
Building 410
Bolling AFB, DC 20332-6448

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFOSR-87-0192

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release;
distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
(see page 1)

Co

14. SUBJECT TERMS

15. NUMBER OF PAGES
23

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

A Theory of Bayesian Data Analysis

Bruce M. Hill*

February 1988
Revised October 10, 1989

1 Introduction

Bayesian data analysis is concerned with the type of data manipulations, transformations, and just plain playing with the data, that any serious scientist engages in during the statistical (or other) analysis of his data. It is largely a post-data procedure, rather than a pre-data procedure, since even when it is desirable to think through such matters quite carefully prior to obtaining the data, in many real world experiments time and other constraints would provide limits on such activities. Compare Hacking ~~(1967)~~, or the discussion in Hodges ~~(1987, p. 291)~~ concerning how much is enough. Bayesian data analysis goes beyond the mere data manipulations, however, and attempts to integrate the theory of subjective probability with such data analysis. In this respect it differs from other data-analytic approaches, which appear, more or less, to abandon probability. In this article ~~I shall~~ attempt further to elucidate the theory of Bayesian data analysis begun in Hill (1985-86, 1987a, 1987b, 1988b). See also Hill (1970a, 1975a) for earlier thoughts on the subject with regard to tests of significance, and Smith (1986). The Bayesian theory of tests of significance that originated with H. Jeffreys (1961), and was developed by Jimmie Savage in Savage (1962), and in the beautiful article "Bayesian statistical inference for psychological research," by Edwards, Lindman and Savage (1963), was the starting point for my own attempts to integrate the Bayesian theory with data analysis. Such an integration could be viewed as a synthesis of the empiricism-pragmatism of John Locke, David Hume, Charles Peirce, and William James, with the rationalistic tradition of Plato, Descartes, Leibniz, Kant, and others.

The purpose of this article is to address some of the basic philosophical and practical issues that arise in attempting to integrate the Bayesian theory with data analysis. Failure to address these issues may have, in the past, led to serious

*This work was supported by the U. S. Air Force under grant AFOSR-87-0192. The US government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

deficiencies in both of these approaches. It will be argued that both conventional data analysis and conventional pre-data Bayesian theory can benefit from one another.

2 Inadequacy of Pre-Data Theories

Perhaps the greatest single source of confusion in modern statistics is due to the failure to distinguish pre-data considerations, such as arise in the design of experiments, from post-data considerations, such as arise in actual decision-making in the light of the data. Sequential analysis provides an excellent example. See Anscombe (1963, p. 381) for a very forceful and convincing analysis of such confusion, and discussion of the waste of time and effort spent on sequential analysis, which he calls a "hoax." [1]

Consider, for example, a sequential stopping rule, N, that depends only upon the observations, as almost all such rules studied do. Suppose that the data of the experiment consists of the fact that one stopped at time N = n, and that the actual observations were $X_1 = x_1, \ldots, X_n = x_n$. If a parametric model is employed, say with parameter θ, then we have

$$\begin{aligned} Pr\{data \mid \theta\} &= Pr\{N = n \mid X_1 = x_1, \ldots, X_n = x_n,\ \theta\} \\ &\quad \times Pr\{X_1 = x_1, \ldots, X_n = x_n \mid \theta\}. \end{aligned}$$

Since the stopping rule depends only upon the observations, and since we did in fact stop at time n, and not before, the first factor on the right-hand side must be identically unity, and thus does not depend upon θ. For example, if the stopping rule were to stop at the first time that the sample mean, $\overline{X}$, exceeded a specified constant, c, then given the actual observations $X_1 = x_1, \ldots, X_n = x_n$, and θ, it would be absolutely certain that we must stop precisely at time n, irrespective of the value of θ. The second factor on the right-hand side is simply the likelihood function for θ, based on a fixed sample size n. This means that on a post-data basis, i. e., given the data, the information obtained from a sequential experiment that actually stopped at time n, is logically equivalent to the information contained in a *fixed* sample size experiment with n observations, together with a logically certain event. Somehow or other, sequential analysts purport to extract information out of this logically certain event, over and above the information contained in the fixed sample size experiment. This appears to have some connection with arguments for perpetual motion, and is perhaps one of the reasons why Anscombe calls the subject a hoax. Savage (1961, 3.23), Savage (1962, p. 18-20), and Edwards, Lindman, and Savage (1963, Section 8) provide further discussion of such matters. Also see Barnard (1947), who apparently first understood the true nature of 'sequential analysis,' and Berger and

[1] It is not the procedure of sequentially observing the data that is being condemned, but rather interpretation of the data according to a body of non-Bayesian statistical technique known as sequential analysis.

Wolpert (1988, 4.2) and Berger (1985, 7.7) for more recent discussions. (To the best of my knowledge, no sequential analyst has yet attempted to answer these 25 year old *logical* objections to their subject. In medical trials the subject has, unfortunately, been used extensively. In practice, often real-world constraints, such as time, patience, and funding, change the true stopping rule, so that it becomes completely unknown. In this case the theory underlying sequential analysis is not even germane to the analysis of the data.)

That the stopping rule is irrelevant for inference and decision-making follows trivially from the likelihood principle, or from the restricted likelihood principle of Hill (1987a, 1988a). However, the analysis that I have given here has some new aspects. It does not depend upon the likelihood principle, but rather upon the willingness of the reader to acknowledge that a logically certain event cannot provide information, *in any meaningful sense*, with respect to empirical questions. See also Hill (1989) for discussion of the concept of an 'analytic' argument in philosophy. All probability based methods of inference, including those violating the restricted likelihood principle, focus attention upon the probability of the data, given the parameter. My argument is that no matter how sequential analysts purport to utilize such distributions, it is necessary for them to pretend to obtain information from a logically certain event, over and above that stemming from the corresponding fixed sample size experiment. Note that on a pre-data basis, the event in question is only *conditionally* logically certain. Indeed, one might not stop at all. However, *given the actual observations,* it follows necessarily that one must stop at precisely time n, so that the event in question is logically certain. This highlights the critical nature of the distinction between pre-data and post-data considerations.

I have used this example to illustrate the difference between a pre-data approach, such as sequential analysis, and a post-data approach, such as the likelihood or Bayesian approach. It shows how a pre-data consideration, such as choice of a stopping rule, can become misleading and/or irrelevant on a post-data basis. Of course, the choice of a stopping rule can be important in the design of a study, and relates to the question 'how much is enough.' It may also be remarked, however, that sequential design, which in principle should have been a serious subject, does not seem to have contributed as much as it might have to real-world problems either, perhaps because of the same underlying confusion. Anscombe (1963, p. 382) says: "In fact, sequential experiments are a most stimulating and provoking topic for the statistical theorist to meditate on. Too little attention of the right sort has been paid to them."

Although sequential analysis is a particularly blatant example, there are many other situations where the pre-data considerations are quite different from the post-data considerations. Some examples arising in econometrics are discussed in Hill (1985-86, p. 218). More generally, it can be argued that the conventional classical theory of statistical inference is almost entirely a pre-data theory. For example, the confidence approach of Neyman and others is based upon a pre-data evaluation of the probability of coverage. Does it have any relevance on a post-data basis? If so, how does it allow for situations where, on a post-data basis, it is obvious that no reasonable person could have much confidence in the quoted confidence coefficient? Such a lack of confidence can arise either for logical reasons, such as in the Fieller-Creasy problem, or simply because of common-sense.

The fiducial argument of Sir Ronald Fisher, which appears to have historically preceded the confidence argument, was an important first step towards a genuine post-data approach. Here Fisher formulated the idea that after seeing the data one might wish to retain certain probability statements as still valid. Initially he believed that fiducial probability involved a new concept of probability, but he later acknowledged, in a footnote, Fisher (1959, p.51), that "Probability statements derived by arguments of the fiducial type have often been called statements of 'fiducial probability'. This usage is a convenient one so long as it is recognized that the concept of probability involved is entirely identical with the classical probability of the early writers, such as Bayes. It is only the mode of derivation which was unknown to them." See Zabell (1988) for a fascinating account of the comedy of errors, tragic for statistics in the twentieth century, that took place between the discovery of the innovative fiducial argument by Fisher, and his eventual understanding of its connection with the Bayesian approach, as in the quotation. In Hill (1988b) the role of the fiducial argument in Bayesian nonparametric inference is discussed.

The Bayesian approach provides a framework in which the meaning and validity of an intuitively brilliant argument, such as the fiducial argument, can be critically examined. Some pre-data probability evaluations will remain valid, in the sense that they are also Bayesian post-data evaluations, and some will not. As will be argued in Section 3, there is no objective way to state, on a pre-data basis, which will be retained and which will be dropped. This, in fact, is precisely where data analysis enters the picture.

In the pre-data design situation one may usefully employ a statistical model to get a rough idea of the type of experiment or quantity of data needed to provide a serious answer to a real-world question. In the post-data situation it is necessary to check, in some way or other, the approximate validity of the model, using appropriate diagnostic procedures, if necessary to abandon the original model, and perhaps to replace it with a new one. Such a post-data model might then be used to obtain inferential and decision procedures, given the data. The pre-data considerations, such as initial models and/or prior distributions, may or may not be deemed relevant after exploring the data.

It was argued in Hill (1985-86, Section 2) that conventional pre-data theories of statistical inference, such as the Neyman-Pearson approach, break down completely when considered in the context of real-world data analysis. For in order that confidence coefficients, p-values, etc., have any meaning at all, these would have to be evaluated *conditional* upon all the diagnostics actually used, including their order, and even upon the thoughts that cross one's mind during the analysis of the data. Plainly such conditional probabilities are both unknown and unknowable. Hence even if the conventional theories were not rejected for the many other reasons that Bayesians have put forth, such as incoherency, inadmissibility, the failure to incorporate realistic prior knowledge, etc., they would have to be rejected as being totally inapplicable in the real-world, except in those rare cases where someone is rash enough to give total certainty to the pre-data model that he has selected.

The Bayesian approach also faces some serious challenges in the context of post-data analysis of the data. As argued in Hill (1985-86, p. 223), the saving grace for the Bayesian approach is that the likelihood function, even when it has been formulated through the process of data analysis, remains precisely the same as if it had been specified a priori. The 'prior' distribution is, of course, no longer a prior distribution, since the parameters may not have been even thought of prior to the data. However, in this situation the Bayesian can do a post-data robustness and sensitivity analysis, as in Hill (1980b). See also the related ideas in Berger (1984, 1987). In other words, having perhaps formulated a new model, with new parameters, one can examine the sensitivity of the conclusions to variations in the 'prior' distribution for the parameters of the new model. It may be the case that for decision-making purposes the conclusions are quite clear. If not, it means that reasonable people with different 'priors' would come to quite different conclusions, given the available data, and this is important to know. Compare the discussion in Hill (1985-86, p. 241), and comments by the four discussants.

3 Extreme Data

Suppose that a vector of observation Y, in an n-dimensional Euclidean space, is thought of as being the sum of a parameter vector θ, and a vector of errors ϵ. The use of capital Y indicates here the pre-data status of the observations, i. e., Y has not yet been observed, although it may have already been determined. The data of the experiment is $\{Y = y\}$, so y consists of the observed value of Y. Suppose that a Bayesian regards θ as marginally independent of ϵ. (Here I do not have in mind the conventional assumption of conditional independence, given some other parameters, such as scale parameters, but rather the definition of independence, i. e., the joint distribution factors appropriately.) Let $f(\theta)$ and $g(\epsilon)$ be the marginal prior densities for θ and ϵ, respectively. This means, for example, that if there were an unknown scale parameter, say σ, in connection with the distribution of ϵ, then it would already have been integrated out to obtain $g(\epsilon)$ as the marginal density, as discussed in Hill (1969a).

Clearly the posterior density for θ, given the data $\{Y = y\}$, is

$$f''(\theta) \propto f(\theta) \times g(y - \theta).$$

Similarly, the posterior density for ϵ is

$$g''(\epsilon) \propto g(\epsilon) \times f(y - \epsilon).$$

This model for the realized errors was put forth in Hill (1969a). Hill (1969b) obtained the limiting posterior distributions for extreme data, and was eventually published in Hill (1974a). Note that so far we have only given the ordinary Bayesian posterior distributions in terms of f and g, which are pre-data specifications of the marginal prior distributions for the parameter and error vectors, respectively. The first post-data consideration that arises is that as soon as y is realized (but not necessarily observed), there is complete symmetry with regard to the logical status of θ and ϵ. Thus on a post-data basis, ϵ consists of the ***actual realized*** errors in the experiment. These errors are unknown, but simply form a vector of n numbers, just like θ,, at least in the case where the parameters have a physical meaning.

In conventional non-Bayesian statistical theory it is sometimes argued that after a coin has been flipped, but with the result unknown, that probability is no longer relevant. The result is said to be either heads or tails, and that is supposed to be all there is to say. From a Bayesian point of view, of course, probability remains relevant, since it is used to describe the uncertainty of an individual, and in this example (without further information) the state of uncertainty remains the same. Compare de Finetti (1974, Ch. 2). In the same way, even though the 'errors' have now been realized, for a subjective Bayesian the pre-data density, g, would be just as relevant after y has been realized as before, at least in the absence of further information or thought. Of course such information is often available, in the form of covariates, but we do not consider this case here. It may be noted that there are actually three conceptual stages involved. The first stage is the pre-data stage, before the errors and/or parameters are realized, and Y is determined. The second stage is after both are realized and $Y = y$ is determined, but before y is observed, as in the coin flipping example, after the coin is flipped but before one knows the outcome. The third stage is the post-data stage, after y is observed and the data analysis has taken place.

A simple concrete example may be helpful. One walks in a northern city in December, and although it appears quite balmy, say about 60 degrees Fahrenheit, a bank thermometer reads 25 degrees. Let us suppose that it is known from experience that this particular thermometer is usually accurate to within a few degrees. The data is $y = 25$, and we ignore other information, such as the dress of people on the street. Which does one believe? Although the bank thermometer is usually accurate, it may have gone haywire. Also, perhaps one is having a fever or some other form of delusion. The problem is to separate out the component of y due to the true temperature, θ, from the component due to error, ϵ. There is complete symmetry between the status of θ and ϵ, given the data, $\{Y = y\}$, and it is only the character of the distributions determined by f and g that allows one to differentiate the 'errors' from the parameters.

In Hill (1969b, 1974a, Section 4), a basic theorem is proved relating to this situation, in the n-dimensional case. The question that is addressed concerns what happens to the posterior distributions for θ and for ϵ when y is extreme in some sense. My theorem relates to the case in which the Euclidean norm of y is large. The theorem states that only three possible limiting situations can occur, *no matter what the densities f and g may be.* The first possibility is that when $||y||$ goes to ∞, the posterior density of θ - y $= -\epsilon$ converges to the ***prior*** density for $-\epsilon$, i. e., $g(-\epsilon)$. The second possibility is that as $||y||$ goes to ∞ the posterior density of $\epsilon - y$ converges to the ***prior*** density for $-\theta$, i. e., to $f(-\theta)$. By the symmetry between parameters and realized errors alluded to earlier, the second possibility follows automatically from the first. The third possibility is that no limit exists.

A little thought shows that the first case corresponds to both classical non-Bayesian inference, and also to the Bayesian situation when stable or precise estimation in the sense of L. J. Savage occurs, as discussed in Savage (1962, Section 4). In the Bayesian case, the likelihood function for θ is then extremely sharp relative to the prior density for θ, so that one more or less believes the data, and regards y as being approximately equal to θ, with an error whose magnitude is determined by g. This occurs, for example, with normal data and a Cauchy or t prior distribution for θ. A key point to understand, however, is that there is nothing in the logic of the situation that requires this to occur. Indeed, just the opposite can occur, as in the second possibility, in which case one instead regards the ***error***, ϵ, as about equal to y. This second possibility actually corresponds to the intuitive interpretation of ϵ as being an 'outlier.' In the first possibility it is θ - y that has a limiting distribution, while in the second possibility it is ϵ - y that has a limiting distribution. Both have reasonable subjectivistic interpretations.

Remarks 1 through 8 of Hill (1974a, p. 570-573) attempt mathematically and philosophically to characterize these two possibilities in terms of the relative sharpness of f and g. Thus basically what underlies possibility 1 is the view that errors are in some appropriate sense believed to be 'small' relative to the parameters. This may or may not be the case. In the bank thermometer example, I think most of us will ordinarily tend to rely upon our prior judgment. In other examples, where great caution is taken with respect to potential sources of bias, such as in the celebrated Michelson-Morley experiment of physics, we will tend to believe the data. See Hill (1988a) for a discussion of the Michelson-Morley experiment in this context. Of course the Michelson-Morley experiment was exceptional, and in most experiments the possibility of a serious bias must be taken more seriously. Such bias can then be viewed as giving rise to a relatively diffuse distribution for the errors, as perhaps occurs in the bank thermometer example.

The main point that I am making is that it is only careful consideration of one's marginal prior densities, f and g, with respect to their relative sharpness, that can allow one to make a reasoned decision as to which of the two possible limiting distributions one wishes to accept. (It should be emphasized, as in

Remark 5, that a third alternative is that no limiting posterior distribution for θ exists, as occurs when both f and g are of the normal form. I personally regard this case, although possible, and sometimes appropriate, as ordinarily of lesser interest. The assumption that a limiting posterior distribution exists can be viewed as representing a form of stability in one's outlook.) As mentioned earlier classical non-Bayesian inference implicitly concerns only possibility 1, and also some Bayesians have emphasized this case in order to achieve bounded risk.

In the post-data situation, the situation becomes even more complicated. After playing with the data, one may find that either f or g or both have to be regarded as mixtures. For example, it may become apparent that g should include an extremely fat-tailed component, representing an even more diffuse 'outlier' than previously anticipated, even if this was not explicitly thought of prior to obtaining the data. In the bank thermometer example, this component might be used to represent some special form of breakdown in the mechanism. Such a phenomenon will be discussed further in Section 4 to illustrate the process of Bayesian data analysis. Remark 9 of Hill (1974a, p. 572) shows that even when both f and g are represented as mixtures, the basic theorem still holds, but one must compare the fattest-tailed component of f with the fattest-tailed component of g. The one of these that is sharpest wins, in the sense that if the fattest component of g is sharp relative to the fattest component of f, then it is θ - y that has a limiting distribution. Otherwise, it is ϵ - y that has a limiting distribution. Although the 'what if' method, or "device of imaginary results" of Good (1965, p. 19), can be of value, it seems clear that it would not ordinarily be possible, before examining the data, to know which limiting distribution will eventually be thought to be appropriate. This is partly due to the labor involved in trying to assess all the components of f and g a priori, and partly due to the fact that the specific methods of data analysis used may trigger off a chain of thought that leads to some previously unsuspected understanding of the data. Eventually, one must make some decision, and this implicitly involves a post-data assessment of the various components of f and g.

For some related articles, see Dawid (1973), which deals with the limiting posterior expectation when n=1, and Umbach (1978).

4 Methodology for Model Selection

Let us now attempt to characterize Bayesian data analysis as distinct from more conventional forms of data analysis. This will be done for the important problem of model selection. It will be argued that all the ingenious techniques to analyse and display data that have been developed by scientists and others for centuries, automatically become part of Bayesian data analysis. One may perform such searches and displays in any way whatsoever, giving free rein to scientific creativity. The part that is uniquely Bayesian only occurs after, as a result of such techniques, one has reformulated old models, or formulated new models for the data. Such models may have been thought of a priori, or may have arisen partly or entirely through the process of data analysis. Sometimes, in relatively simple examples, such data-instigated models can actually be convincingly obtained through use of Bayes's theorem. However, in realistically complicated examples, the computations would be prohibitive. Furthermore, as I argued earlier, the process of analysing the realized data may trigger off a chain of thought that was not foreseen beforehand. From this point of view the use of data-analytic techniques might be viewed as a computationally efficient way to arrive at new models. Whether or not such models could also have been achieved through use of Bayes's theorem seems to be a moot point.

After the conventional forms of data analysis have been completed, suppose one arrives at one or more models that command some degree of credibility. If one wishes to make a statement with operational content, other than a purely descriptive one about the particular data set that has been analysed, then one must deal in some fashion or other with uncertainty about either models or parameters or both. The work of de Finetti, Savage, and others, makes it abundantly clear that there is no way to do so without bringing in subjective probability explicitly; and of course once this is done, one presumably wishes to avoid so far as possible, Dutch-books and other forms of incoherence and irrationality. There is often no need to make more than a small number of post-data probability assessments, i. e., those that are sufficient for the purpose at hand. This purpose might be to choose amongst a few available decisions, or to specify, approximately, a predictive distribution for future observations, or to make inferences about conventional parameters. De Finetti's fundamental theorem of probability (1974, p. 111) shows that any coherent collection of probability specifications can always be extended coherently, but in practice it is rare for there to be a need to make more than a few such specifications.

It is the attempt to integrate the data-analytic search procedures of conventional data analysis with the subjectivistic theory of probability and decision-making that is unique to Bayesian data analysis. Conventional data analysis appears simply to ignore the problem of coherent judgment, and apart from technological advances in the display of the data, returns to the pre-probabilistic frame of mind with regard to the fundamental questions of induction and inference. Such data analysis was, of course, very welcome in the statistical climate of the past quarter century, which had degenerated into a sterile and largely meaningless form of mathematical exercise that was stifling to scientists in many fields. See the article by Salsburg (1985) for an example in the medical area. But it is difficult to see how real progress in the problems of inference and decision-making can be made without some operationally meaningful concept of probability, and at present the subjective Bayesian concept appears to be the only viable one.

How then can the Bayesian approach achieve the integration about which I am speaking? The primary jewel in this crown is the principle of stable estimation of Jimmie Savage, as in Savage (1961, 1962), and its extension to the comparison of hypotheses, as in Edwards, Lindman and Savage (1963). At present these constitute the chief way in which consensus can be achieved other than by fiat. We begin by discussing the role of the likelihood principle in Bayesian data analysis.

Suppose that as a result of data analysis one finds a new model for the data, or modifies a previous one. Suppose this new model is denoted by M, and that it contains a parameter vector θ. It is assumed that, conditional upon M and θ, the probability for the realized data is completely specified. This will indeed be taken as our definition of a model. In other words, in order for M to represent a model in our sense, it must be such that conditional upon M and its parameters, the probability for the data is completely specified. The likelihood function for θ, conditional on the model M, is then $Pr(data \mid M, \theta)$. Note that this likelihood function is precisely the same, whether M has been arrived at through a data-search procedure, or whether M had been specified a priori. This follows immediately from our definition of a model, since given M and θ the probability of the data is uniquely specified by M. The 'data' for this evaluation of the likelihood function can be taken as any subset and/or transformation of the original data, no matter how arrived at.

If only one model appears viable after data analysis, then a question that arises is whether it is worthwhile to use this model for various purposes, such as prediction, inference, or decision-making. This is a fairly subtle question that implicitly involves considerations as to whether the departures from this model are so large as to make its use unwise, even though no alternative model has as yet been formulated. However, one can usefully consider inference and decision-making, conditional upon the truth of M, even when one considers it unlikely that M is true. Similarly, if after data analysis two or more models emerge as being thought worthy of consideration, then the above analysis can be extended to yield posterior odds for each, conditional on the truth of at least one of these models. Suppose, for example, that M_1 and M_2 are two such models, and that the associated parameters are θ_1 and θ_2. These models may be nested or unnested. Let $L_i(\theta_i) = Pr(data \mid M_i, \theta_i)$ be the associated likelihood functions, for i = 1,2. Finally, let $\pi_i(\theta_i) = P(\theta_i \mid M_i)$ be candidate 'prior' densities for θ_i, say relative to Lebesgue measure, for i = 1, 2. As argued above, such likelihood functions do not in any way depend upon the fact that the models may have been developed through data analysis. Note that unlike the case in which there is only a single model under consideration, one must here include all constants of proportionality in the definition of the likelihood function.

The Bayes factor is then $Pr(data \mid M_1)/Pr(data \mid M_2)$, where $Pr(data \mid M_i) = \int L_i(\theta_i) \times \pi_i(\theta_i) d\theta_i$, for i = 1, 2. The posterior odds for M_1 versus M_2 are obtained by multiplying the Bayes factor by the 'prior odds,' i. e., by $Pr(M_1)/Pr(M_2)$. A great deal is known about the behavior of the Bayes factor in cases where the models have been specified in advance. Edwards, Lindman, and Savage (1963) give a number of approximations and bounds for this factor.

Thus once the problem is reduced to this form, we have available a variety of ways to deal with the purely mathematical or computational aspects of the problem. What I would like to discuss here, however, is the question as to the validity of such an analysis, when one or both models have been derived through data analysis. This in fact is the key question that faces one in applying the Bayesian theory in practice.

There are two separate issues. The first concerns the validity of the likelihood principle in this context. In Hill (1987a, 1988a) I have argued that the formulation of the likelihood principle by Birnbaum was incorrect in an essential way, namely in trying to give an abstract objective definition of 'the evidence,' which as the example in these articles shows, cannot be done without greatly delimiting the concept of evidence. Furthermore, even with respect to the limited concept of evidence formulated in my restricted likelihood principle, the evidence is always only relative to a specific model or models. However, the real power of the (reformulated) likelihood principle is best seen in connection with the data-instigated models that I have discussed above. For, as I have argued above, once a model has been formulated, whether pre or post-data, ***the likelihood function for the parameters of that model, conditional upon the truth of that model, does not in any way depend upon the circumstances under which that model was discovered.*** This simple logical fact constitutes, I believe, the only truly 'objective' feature of statistical practice. The second issue concerns how, within the subjectivistic theory, one can live with this fact. Clearly, one must somehow discount some of the adhoc models discovered through data analysis. The subjective Bayesian approach can only do this through the choice of the $Pr(M_i)$, and the $\pi_i(\theta_i)$. It is the logical status, and practical aspects of such evaluations that must now be discussed.

With regard to the $Pr(M_i)$, I believe that when one or both of the models have been formulated through data analysis, then it is ordinarily appropriate to assess or reassess these probabilities ***after*** the process of data analysis that gave rise to them. For example, if only M_1 had been thought of prior to the data analysis, this would suggest that M_2 must have been given negligible a priori probability. However, I think this is largely irrelevant. The results of the data analysis have suggested that one was in error in neglecting M_2, and it would ***now*** be appropriate to give it a non-negligible probability, prior to evaluating the post-data odds. In other words, one should interpret the $Pr(M_i)$ as the probabilities that one would give to the two models, conditional upon the truth of at least one, ***after*** the data analysis that gave rise to M_2, but ***prior*** to the use of the Bayes factor to update the corresponding odds to become the overall post-data posterior odds.

This violates the classical version of Bayes's theorem, but I think it is the sensible way to proceed. What has been suggested may be viewed in the following light. Most of the time in life we do not use Bayes's theorem in updating our opinions. Even if it were thought wise in principle to do so, as for example in the theories of de Finetti and Savage, it would ordinarily be computationally prohibitive. In certain special situations, for example in the narrow but relatively precise problems of science, we attempt so far as possible to follow the Bayesian approach. This is done largely because it is the only approach that even attempts to deal seriously with the problem of rational updating of belief, and rational discourse. But it does not follow that it is sensible *always* to use Bayes's theorem in the updating of opinions. For this reason, I have separated the data-analytic process of discovering or reformulating models, from the formal Bayesian analysis after such models have been formulated. Having done so, I am now in a position to take advantage of the rational and persuasive aspects of the Bayesian approach in updating my odds for the two hypotheses, given the 'data' that is now being explicitly considered. Although the overall data includes experiences with, and results of, the data analysis, it is useful to separate this into two parts, only one of which is dealt with through Bayes's theorem. This may appear to be only an attempt to get the best of both worlds, the empiricist world of the data-analytic school, and the rationalist world of the Bayesian school, but it is difficult to see what alternative there is to such a procedure, other than total subjectivism and adhockery. It should be observed in this context that the Bayesian approach has always had an arbitrary element in it as to the time point at which one proceeds to make a formal Bayesian analysis. I am suggesting that often the appropriate time point is following the process of data analysis.

If one adopts this point of view, then some of the more troublesome aspects of the Bayesian approach can be largely bypassed. The Bayesian no longer needs to justify how the pre-data value for $Pr(M_2)$, which must have been negligible, has suddenly become enormously larger. This could happen purely through Bayes's theorem, as discussed in Hill (1970a), but it need not. It could also occur through the informal updating involved in the data analysis. For scientific purposes, as opposed to explicit decision-making purposes, it might now be appropriate to evaluate the 'prior' probability for M_1, given that at least one of the two models or hypotheses is true, as 1/2. It should be understood here that when we speak of a model as being true, ordinarily we mean approximately true. The earth is neither planar nor spherical, but the spherical approximation is generally more useful. In the same way presumably neither of the M_i is literally true, but one or the other may provide a more satisfactory approximation, and it is important to know which.

Similarly, the prior distribution for θ_i, given M_i, can only be truly a priori, if M_i has been specified in advance. However, as I see it this need not necessarily cause any serious problems. For example, as soon as M_i has been stated, it may be abundantly clear that there are things one can say about the value of θ_i based simply on the meaning of M_i and θ_i. For example, it might be clear that there are reasonable grounds for regarding θ_i as diffuse relative to the likelihood function for θ_i, thus setting the stage for stable estimation. Although there are obviously some subtle aspects to such a procedure of speaking about opinions concerning θ_i, given M_i, acting as though one had not already observed and analysed the data, it appears to be the only thing that can usefully be done. Furthermore, the difficulties are primarily psychological, and arise from the fact that having seen the data one must nevertheless attempt to erase it from one's mind. The degree of success that can be achieved will depend upon the circumstances. In the thermometer example, which will be returned to in the next section, I think this can be done quite adequately. It should be noted that the basic difficulty arises not only through data analysis, but to a lesser extent in *any* Bayesian application. The reader of an article making a Bayesian analysis of data will not ordinarily have considered his a priori distributions, although he may have some definite opinions. *Thus in any case the force of a Bayesian analysis of data must depend upon an agreement among scientists that specific prior distributions and likelihood functions are pertinent to the problem, and can be considered on their own merits, even after the data has been observed.* See also Leamer (1978, Ch. 9) and my review of his book in Hill (1980c) for further discussion.

When, for example, it can be agreed that a particular model, M, is relevant for inference about θ, and also that it is reasonable to take the prior distribution for θ, given M, as being diffuse relative to the likelihood function specified by M, then Savage's principle of stable estimation applies, given M, and leads to useful approximations to the post-data distribution for θ. This appears to be the primary method by which consensus can be obtained as to empirical matters, whether for decision-making, inference, or prediction. (It would perhaps be better in this context not to speak of the prior and posterior distributions, but rather of weighting the realized likelihood function by some function $w(\theta)$, as in Barnard, Jenkins and Winston (1962).) From my point of view, the output of a Bayesian analysis of data should include the likelihood function (or, in high dimensions, characteristics thereof), together with a formal Bayesian analysis of the data using one or more prior distributions for the parameters. Through a sensitivity analysis obtained by varying the prior distribution, one can attempt to see what aspects, if any, of the conclusions are robust to the specific form of prior knowledge assumed. See Hill (1980b). The justification would rest in a consensus that on the one hand the Bayesian method for revaluation of probabilities is rational, or at any rate the best we now have, and on the other hand that the specific prior distributions and likelihood functions being employed are plausible and worthy of consideration. Obviously there is no possibility of *proving* that particular distributions are valid for everyone, so the force of the argument must stem from some agreement that the distributions being employed are reasonable for the problem at hand. Conventional classical inference, as interpreted from a Bayesian viewpoint, demands that the prior distribution be taken as diffuse or improper, even when it is ridiculous to do so. This is consensus by fiat, and is a high price to pay for such consensus.

If we agree that none of our models is likely to be true, then the question of real importance is whether, given the data, one thinks the departures from the best model or models one has are so large as to make them not worthy of use. This question cannot be answered without explicit consideration of the purposes for which the model is to be used, and the utility of using the model for such purposes, as opposed to not doing anything. See Dickey and Kadane (1980). One might believe that a model one has is not true, but that it will still get you to the moon. On the other hand, one might regard it as highly improbable that use of the model will be worthwhile. If we don't think that any of the models we now have are adequate to achieve our purposes, then one presumably starts to think hard about formulating new models. It is here, of course, that data analysis and various diagnostic techniques can be of the greatest value. It is my thesis that except in extremely simple situations this process is not usefully viewed as merely an application of Bayes's theorem. But when and if such new models are found, I believe it is entirely appropriate and useful to resume the conventional Bayesian mode of analysis, making use of the knowledge gained through data analysis just as though it were the usual type of background or a priori knowledge. The justification for such a procedure would lie in yet another aspect of the Bayesian paradigm, namely the attempt to maximize ***post-data*** expected utility. In my opinion, this aspect is the more fundamental and overrides even use of Bayes's theorem. Of course, in many situations use of Bayes's theorem follows from such maximization. See Hacking (1967) for a discussion of some related issues.

Some will of course dislike the subjectivity involved in all such considerations. However, I know of no way to avoid it, other than to sweep it under the carpet, SUTC, as Jack Good says. The distinguished philosopher and psychologist, William James (1896, p. 97), puts it quite well:

> Objective evidence and certitude are doubtless very fine ideals to play with, but where on this moonlit and dream-visited planet are they to be found? I am, therefore, myself a complete empiricist so far as my theory of human knowledge goes. I live, to be sure, by the practical faith that we must go on experiencing and thinking over our experience, for only thus can our opinions grow more true; but to hold any one of them–I absolutely do not care which–as if it never could be reinterpretable or corrigible, I believe to be a tremendously mistaken attitude, and I think that the whole history of philosophy will bear me out.

Despite such subjectivity, I believe that Bayesian analyses can have every bit as much impact in obtaining a post-data consensus of opinion as if the model had been specified a priori. Indeed, an M that has been found and confirmed in some sense on the basis of the data, is in many ways a much sounder basis for inference than a speculative a priori M that has not been so founded. When only one such model has been formulated, then all our inferences must be conditional on the truth of that model. One can, of course, add an M_1, etc, if the data support such additions. In this case all inference is conditional upon the truth of one from amongst this finite set of models. One might describe scientific progress as consisting of the refinement of an undifferentiated model, such as the complement of some initial M into specific alternatives such as M_1, M_2, etc. Such alternatives are typically found through the process of creative data analysis and hard thought. The Bayesian approach can then, at any point in time, be used for decisions, inference and predictions, using the models currently taken seriously. In this way the subjective Bayesian approach, as integrated with data analysis, can provide a relatively objective and *reasoned* form of argument with regard to model selection. By contrast, the conventional form of data analysis either does not deal at all with the question of model selection, or else relies upon total subjectivism, since it cannot hope to show that there might be a consensus as to the evaluation of probabilities when it does not even have an operationally meaningful concept of probability to work with.

5 Examples of Bayesian Data Analysis

Let us return to the bank-thermometer example to illustrate the procedure of Bayesian data analysis. Suppose that five hours later, during which time it appears to have cooled down noticably, one returns and finds the thermometer still reading 25. At this point, whether one had consciously thought about it a priori or not, the thought suggests itself that perhaps the thermometer is simply frozen stuck at 25. Suppose in fact that one had not consciously thought of this hypothesis beforehand. One can nonetheless reason as follows. Let M denote the original model of Section 3, and let M_1 denote the model that states that the thermometer is frozen. Upon reflection, one decides that it would have been reasonable to have attached a non-negligible prior probability, say around .03, to the hypothesis that the bank thermometer would be frozen at some unspecified value. Before seeing the number 25, of course one would not have much information as to what the number would be, but taking a range of say 100 degrees Fahrenheit, one might take the 'a priori' probability of the thermometer being frozen at 25 to have been about .0003. From a post-data point of view, the question of interest is whether or not the thermometer is frozen stuck. If so, then it can only be at 25, and it is not the .0003, but simply the .03 that is the relevant 'prior' probability or weight for the hypothesis under consideration. The datum y = 25 would then be used to revise this 'a priori' probability in accord with Bayes's theorem. This illustrates how careful one must be in dealing with post-data hypotheses or models in the Bayesian framework. Note that in this example, even though the hypothesis M_1 was only thought of after seeing the data, the probability of .03, which would be based on experience and judgment, seems just as compelling as though the evaluation had been made beforehand. It is my opinion that the psychological effect of seeing the data can vary greatly in problems of post-data Bayesian evaluations of probability, and that it will have little effect in problems such as this, where once the model has been formulated, one can easily refer the question to related experience.

Now I think that most people will have a clear post-data preference for M_1 in this case, even without doing the formal Bayesian analysis. Of course, if one returned still again late in the evening when it was even more noticably cooler, and discovered that the thermometer was still at 25, one would then become nearly certain of the truth of M_1. This illustrates the wisdom, sometimes, of the classical statistician's recommendation to take more data. On the other hand, the weakness of this point of view, and the strength of the Bayesian approach, is seen in circumstances where one must act without the possibility of taking more data. The classical statistician who eschews the use of prior knowledge obviously has no basis for even being suspicious of the thermometer, since the value 25 is certainly a possible temperature. Yet even without the confirmatory data of later observations, which would make M_1 almost certain, a Bayesian might see M_1 as highly probable, and be prepared to act accordingly.

The formal Bayesian analysis of the problem sheds some light on how such a conclusion can be arrived at, and how it can be justified to others. Plainly this depends upon the precise specification of f and g in model M, the original model. In accord with my definition of a model, M implies a specific choice of the distribution of errors, represented by g. Because of the symmetry between errors and parameters in the present example, it would be well here to include the specification of f as part of the model as well. Once these have been specified one would simply calculate the posterior odds in favor of M versus M_1 as in the Jeffreys-Savage theory of hypothesis testing. In the case at hand the model M largely discredits itself because of the fact that one's initial opinions about θ and ϵ are such as to make the observed value 25 highly improbable. In other words, for temperatures around 60 or so, which are regarded as highly probable a priori, it would take an improbably low ϵ to yield the result 25. Since M may be interpreted as the hypothesis that the thermometer is functioning normally, in which case the distribution of ϵ is reasonably well known from past experience, it follows that an effective evaluation of the posterior odds can be made. A sensitivity analysis would reveal whether the conclusions are in fact reasonably robust to the precise choice of f and g, and to the probabilities selected for M and M_1.

The model M_1 is a degenerate type of model that explains the data perfectly. The reason that a Bayesian is not necessarily led to adopt such a 'perfect' model is because he may discount such a model due to its low a priori probability. It is one of the great advantages of the Bayesian approach that such discounting can occur, which can prevent one from simply using maximum-likelihood estimates in their most adhoc form. See Hill (1975b) for a striking example. The conventional non-Bayesian theory does not appear to have an adequate way to deal with such things, since it foregoes the use of subjective judgments and a priori probabilities. For example, in the problem of polynomial regression it remains an unsolved problem, within the conventional framework, as to why one does not fit an n th degree polynomial to n−1 data points, thus obtaining a perfect fit. The Bayesian perspective offers a simple answer. Such a fit would require the errors to be identically 0, i.e ϵ to be the 0 vector. This may be viewed as highly improbable a priori. For example, as suggested in Hill (1969a) one might use a spherically symmetrical 'volcanic' prior distribution with density g for ϵ. This is a distribution having a crater centered at the origin, minimum inside the crater at the origin, and with mode along a ridge at some specified positive distance from the origin. For such a prior distribution the most probable value of ϵ is far removed from the origin, and this tends to prevent one from choosing a model for which ϵ is 0. Thus this constitutes another example of how the Bayesian approach can provide a discounting for 'perfect fit' models.

In the thermometer example it might be argued, of course, that the conclusion is self-evident, and the problem hardly worth the effort to make a careful Bayesian post-data analysis. However, with only minor changes, this example would apply equally well to nuclear disasters such as occurred at Three-Mile-Island and Chernobyl, or to the space-shuttle crash. All of these disasters are examples of where there is a conflict between the data and a priori judgements, and where some thought could have averted the disasters. Often engineers cite remarkably low a priori estimates of probabilities for such accidents. These do not appear to be based on experience or sensible forms of data analysis. Confusion as to the meaning of probability versus conditional probability, and predata versus post-data considerations, presumably also plays a role. Although careful data analysis requires substantial expertise, the failure on the part of administrators and others to comprehend even the most basic facts about data analysis and decision-making appears to be partly responsible for many easily preventable foulups.

One of the criticisms that can be made of the conventional Bayesian approach is that it focuses too much attention on the a priori aspects, and not enough on such strategies as 'take more data.' There is a sense in which this is a highly appropriate criticism. Plainly it is foolish to devote an overly large time to the evaluation of a prior distribution. This could never be done perfectly in any case, and there is an important practical question as to when to cease such activity, and simply explore the data, which will often suggest entirely new avenues and hypotheses. Unfortunately there are no hard and fast answers here. In the above examples it is clear that not enough a priori thought had been given so that a quick response could have been made. Decision-makers are sometimes lulled into wishful thinking that certain probabilities are very tiny, when in fact simple Bayesian calculations would reveal otherwise. Similarly, it might pay to consider utilities more carefully than is customary.

Here are some concrete examples of Bayesian data analysis. Hill (1963) employed the three-parameter log normal distribution to model incubation periods for small pox. Although there was some previous theory suggesting the appropriateness of the log normal model, this obviously could not be taken for granted, and had to be checked for the data at hand. I employed stable estimation, and plotted both the marginal likelihood function for the thresh-hold parameter γ, and what is now called the profile likelihood function. These turned out to be nearly proportional. A table of observed and expected values was also presented, showing some discrepancies, but relatively minor in view of the large sample size (310). The chi-square goodness of fit statistic was obviously going to be misleading because of the large sample size, and was not presented. (There would have been no harm in giving it, other than that it is customarily misinterpreted and used to reject possibly useful models.) I believe that the stable estimation argument is entirely valid in this case. I would describe the thought process as follows. Based upon both other data on incubation periods, and on the observable characteristics of the data set, I made the decision that the log normal model would serve a useful purpose in describing the data set, in making inference about the parameters, and for predictive purposes. This aspect of my inferential procedure was not based on Bayes's theorem per se. Rather it was a result of a fairly complex procedure of data analysis. I did not believe that the log normal model was literally true, but rather was implicitly saying that I regarded the departures as being sufficiently small that this model could be usefully employed. Once having made the decision to employ the model, I believed that the inferential procedure was fairly straight-forward, and used the stable estimation argument. It goes without saying that others would be free to substitute their own 'prior' distribution or weighting function.

Another example of Bayesian data analysis occurs in the variance components problem, as in Hill (1967, 1970a). Here I broadened the conventional one-way model to allow for correlated residuals. This was an attempt to explain the familiar fact that the conventional unbiased estimator for the between variance component is often negative. In this example I evaluated the posterior odds in the Jeffreys-Savage sense, in favor of the original model versus the broadened model. Thus this example is one in which a new model, the one with correlated errors, is formulated on the basis of the data, and is then compared with the original model for the data. This is a fairly subtle comparison. See Chaloner (1987) for a review of some aspects of the components of variance problem.

More complex examples of Bayesian data analysis occur in the Mosteller and Wallace (1964) analysis of the Federalist papers, in the analysis of inference about the tails of distributions in Hill (1975a), in work on Zipf's Law in Hill (1970b, 1974b, 1980a), in Hill and Woodroofe (1975) and in Chen (1980), and in the Bayesian survival analysis of Chen et al (1983). An important example in which the data analysis forms an integral part of the inferential theory occurs with mixtures of distributions, as in Hill (1987b). In all these examples the data are quite complex, and model specification procedures require substantial creative efforts.

There has not been room here to do more than suggest the nature of Bayesian data analysis. However, I believe that a primary stumbling block for both conventional data analysis and for conventional Bayesian statistics in the past has been the failure of each to address the concerns of the other, and to take advantage of the achievements of the other. I see no way to avoid either the data analysis, or the Bayesian approach, if one is to make any form of progress in either the practice or the theory of statistics and decision-making. I have tried to suggest above how each can benefit from the other.

6 Concluding remarks

As the Bayesian approach becomes dominant, which now appears to be only a matter of a short time, it is important that we learn from the mistakes of the past. Bayesian thinking emerged from the writings of some of the greatest thinkers in history, including B. Pascal, D. Bernoulli, T. Bayes, P. Laplace, C. Gauss, H. Poincáre, E. Borel, F. Ramsey, B. de Finetti, H. Jeffreys, and L. J. Savage. In many cases, it was a confused or perhaps even perverted misunderstanding of the writings of these distinguished people, that led to the type of thinking that we Bayesians have been arguing against these many years. At a certain point in history it might have been the state of the Bayesian art to use maximum-likelihood estimates, and in the absence of adequate computational facilities, to derive an asymptotic distribution for the maximum-likelihood estimator. Nowadays we can do much better. In low dimensional cases we simply plot the *realized* likelihood function, weight it, and integrate numerically, if need be. In high dimensional situations we learn concepts and techniques to deal with the display and understanding of the information, such as in Hill (1975a).

Certainly Fisher's fiducial argument, and to some extent even the Neyman-Pearson theory, can be seen as approximations to Bayesian procedures. Compare the discussion in Hill (1988b, Section 4). George Barnard has written often and well on this subject. See Barnard (1985) for references. The present day Bayesian approach, such as has been solidified in some of the textbooks that have been written recently, will perhaps in a few years also be seen as only a crude approximation to more realistic Bayesian procedures. Some of the important problems that we have still to come to grips with include the questions of time coherency, randomization, and other alternatives to the conventional Bayesian approach. See, for example, Diaconis and Zabell (1986), Goldstein (1983), Lane and Sudderth (1985), and Zellner (1988). In my opinion, the real mistake of the past was to take some crude approximation to the Bayesian approach, which may have arisen historically due to a variety of real-world constraints and limitations, to be the final answer. We can do much better.

ACKNOWLEDGMENT

My thanks to Frank Anscombe for a number of helpful comments.

REFERENCES

Anscombe, Frank (1963), "Sequential medical trials," *Journal of the American Statistical Association 58*, 365-383.

Barnard, George (1947), Review of "Sequential Analysis" by A. Wald, *Journal of the American Statistical Association 42,* 658-664.

Barnard, George (1979), "Pivotal inference and the Bayesian controversy," in *Bayesian Statistics,* edited by J. M. Bernardo, M. H. Degroot, D. V. Lindley, and A. F. M. Smith, North-Holland: Valencia University Press. (with discussion).

Barnard, George (1985), Discussion of "In defense of the likelihood principle: axiomatics and coherency," by J. Berger, in *Bayesian Statistics 2,* edited by J. M. Bernardo, M. H. Degroot, D. V. Lindley, and A. F. M. Smith, North-Holland: Valencia University Press.

Barnard, G., Jenkins, G., and Winsten, C., (1962), "Likelihood inference and time series," *Journal of the Royal Statistical Society A,* 125, 321-372 (with discussion).

Berger, James (1984), "The robust Bayesian viewpoint," in *Robustness of Bayesian Analysis,* J. Kadane, ed., North-Holland: Amsterdam, 63-124 (with discussion).

Berger, James, and Wolpert, Robert (1988a), *The Likelihood Principle, Second Edition,* Institute of Mathematical Statistics, Lecture Notes-Monograph Series, Volume 6.

Berger, James (1985), *Statistical Decision Theory and Bayesian Analysis, Second Edition,* Springer-Verlag.

Berger, James (1987), "Robust Bayesian analysis: sensitivity to the prior," Purdue University Technical Report No. 87-10.

Berkson, J. (1938), "Some difficulties of interpretation encountered in the application of the chi-square test," *Journal of the American Statistical Association,* 33, 526-542.

Chaloner, K. (1987), "A Bayesian approach to the estimation of variance components for the unbalanced one-way random model," *Technometrics,* 29, 323-337.

Chen, Wen-Chen (1980), "On the weak form of Zipf's Law," *Journal of Applied Probability,* 17, 611-622.

Chen, Wen-Chen, Hill, Bruce M., Greenhouse, J., and Fayos, J. (1985), "Bayesian analysis of survival curves for cancer patients following treatment," *Bayesian Statistics 2,* edited by J. M. Bernardo, M. H. Degroot, D. V. Lindley, and A. F. M. Smith, North-Holland: Valencia University Press. (with discussion).

Dawid, A. P. (1973), "Posterior expectations for large observations," *Biometrika,* 664-666.

De Finetti, Bruno (1974), *Theory of Probability, Volume 1,* John Wiley & Sons.

Diaconis, P., and Zabell, S. (1986), "Some alternatives to Bayes's rule," Proceedings of the Second University of California, Irvine, Conference on Political Economy (B. Grofman and G. Owen, eds.), Greenwich, Conn.: JAI press, 25-38.

Dickey, J., and Kadane, J. (1980), "Bayesian decision theory and the simplification of models," in *Evaluation of Econometric Models,* J. Kmenta and J.

Ramsey, eds., Academic Press, 245-268.

Edwards, W. Lindman, H., and Savage, L. J. (1963), "Bayesian statistical inference for psychological research," *Psychological Review 70*, 193-242. Reprinted in *Robustness of Bayesian Analysis*, J. Kadane, ed., North-Holland: Amsterdam, 1-62, (with discussion).

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference*, Second Edition, New York: Hafner Publishing Co.

Freedman, David (1986), "A case study in nonresponse: plaintiff vs California state board of equalization," *Journal of Business and Economic Statistics*, 4, 125-126.

Goldstein, Michael (1983), "The prevision of a prevision," *Journal of the American Statistical Association, 78*, 817-819.

Good, I. J. (1965), *The Estimation of Probabilities*, MIT Research Monograph No. 30.

Hacking, I., (1967), "Slightly more realistic personal probability," *Philosophy of Science, 34*, 311-325.

Hill, Bruce M., (1963), "The three-parameter log normal distribution and Bayesian analysis of a point-source epidemic," *Journal of the American Statistical Association*, 58, 72-84.

Hill, Bruce M. (1967), "Correlated errors in the random model," *Journal of the American Statistical Association*, 62, 1387-1400.

Hill, Bruce M., (1969a), "Foundations for the theory of least squares," *Journal of the Royal Statistical Society, Series B*, 31, 89-97.

Hill, Bruce M., (1969b), "Inference about many parameters in the theory of least squares," unpublished manuscript, written at Lancaster University, UK.

Hill, Bruce M., (1970a), "Contrasts between Bayesian and classical inference in the analysis of variance and testing of models," *Bayesian Statistics*, ed. by R. Meyer and R. Collier, F.E. Peacock, Itasca, Illinois, 1970.

Hill, Bruce M. (1970b), "Zipf's law and prior distributions for the composition of a population," *Journal of the American Statistical Association, 65*, 1220-1232.

Hill, Bruce M. (1974a), "On coherence, inadmissibility, and inference about many parameters in the theory of least squares," in *Studies in Bayesian Econometrics and Statistics in Honor of L. J. Savage*, ed. by S. Fienberg and A. Zellner, North-Holland, 555-584.

Hill, Bruce M. (1974b), "The rank frequency form of Zipf's law," *Journal of the American Statistical Association, 69*, 1017-1026.

Hill, Bruce M. (1975a), "A simple general approach to inference about the tail of a distribution," *Annals of Statistics 3*, 1163-1174.

Hill, Bruce M. (1975b), "Aberrant behavior of the likelihood function in discrete cases," *Journal of the American Statistical Association, 70*, 717-719.

Hill, Bruce M. (1979), "Posterior moments of the number of species in a finite population, and the posterior probability of finding a new species," *Journal of the American Statistical Association, 74*, 668-673.

Hill, Bruce M. (1980a), "Invariance and robustness of the posterior distribution of characteristics of a finite population, with reference to contingency tables and the sampling of species." In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys,* ed A. Zellner, North-Holland, 383-395.

Hill, Bruce M., (1980b), "Robust analysis of the random model and weighted least squares regression," in *Evaluation of Econometric Models*, ed. by J. Kmenta and J. Ramsey, Academic Press, 197-217.

Hill, Bruce M., (1980c), Review of *Specification Searches,* by E. Leamer, *Journal of the American Statistical Association,* 75, 252-253.

Hill, Bruce M. (1984), Discussion of "The robust Bayesian viewpoint," by J. Berger. In *Robustness of Bayesian Analysis,* J. Kadane, ed., North-Holland: Amsterdam, 63-124 (with discussion).

Hill, Bruce M., (1985-86), "Some subjective Bayesian considerations in the selection of models," (with discussion), *Econometric Reviews*, 4, Number 2, 191-288.

Hill, Bruce M., (1986), "Comment: A Bayesian approach to the nonresponse problem, using covariates à la Freedman," *Journal of Business and Economic Statistics*, 4, 125-126.

Hill, Bruce M., (1987a), "The validity of the likelihood principle," *The American Statistician 41*, 95-100.

Hill, Bruce M. (1987b), "Parametric models for A_n: splitting processes and mixtures," Unpublished technical report, The University of Michigan.

Hill, Bruce M. (1988a), "On the validity of the likelihood principle," in *Statistical Decision Theory and Related Topics* IV, Volume 1, S. S. Gupta and J. O. Berger, eds., Springer-Verlag, 119-132.

Hill, Bruce M. (1988b), "De Finetti's theorem, induction, and A(n), or Bayesian nonparametric predictive inference," in *Bayesian Statistics 3,* J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), Oxford University Press, 211-241 (with discussion).

Hill, Bruce M. (1989), Comment on "An alternative interpretation of Freedman's nonresponse case study," by R. K. Steinhorst and C. Randall Byers, *Journal of Business and Economic Statistics, 7,* 27-33 (with discussion).

Hill, Bruce M. and Woodroofe, M. (1975), "Stronger forms of Zipf's law," *Journal of the American Statistical Association, 70,* 212-219.

Hodges, J. (1987), "Uncertainty, policy analysis and statistics," *Statistical Science,* 2, (with discussion), 259-291.

Hodges, J. (1988), "Can Bayesians do pure tests of significance," Unpublished manuscript, The Rand Corporation.

James, William (1896), "The will to believe," in *Essays in Pragmatism,* The Hafner Library of Classics, Number 7, ed. by Alburey Castell, 1966.

Jeffreys, H. (1961), *Theory of Probability,* Third Edition, Oxford University Press.

Lane, D., and Sudderth, W., (1985), "Coherent predictions are strategic," *Annals of Statistics, 13,*, 1244-1248.

Leamer, E., (1978), *Specification Searches*, New York: John Wiley.

Mosteller, F., and Wallace, D. L., (1964), *Inference and Disputed Authorship: The Federalist*, Addison-Wesley Publishing Co. Inc., Reading, Mass.

Salsburg, D. S. (1985), "The religion of statistics as practiced in medical journals," *The American Statistician*, 39, 220-223.

Savage, L. J. (1961), *The Subjective Basis of Statistical Practice*, Unpublished manuscript, Department of Mathematics, The University of Michigan.

Savage, L. J. (1962), *The Foundations of Statistical Inference, A Discussion*, London: Methuen & Co. Ltd.

Smith, A. F. M. (1986), "Some Bayesian thoughts on modelling and model choice," *The Statistician, 35*, 97-102.

Umbach, D. (1978), "On the approximate behavior of the posterior distribution for an extreme multivariate observation," *Journal of Multivariate Analysis, 8*, 518-531.

Zabell, S. (1988), "R. A. Fisher on the history of inverse probability," Technical Report, Departments of Mathematics and Statistics, Northwestern University. To appear in *Statistical Science*, (1989).

Zellner, Arnold, (1987), "Optimal information-processing and Bayes' theorem," *The American Statistician*, (with discussion), 1988.