Discussion by Professor Bruce M. Hill

(University of Michigan)

I should like to congratulate Berger and Wolpert on their lucid and informative presentation of the history and substance of the likelihood principle, and their extension of the likelihood principle. Although I found their extension interesting, and hope that it may resolve some doubts concerning the status of the likelihood principle in the infinite case, my own view is that the likelihood principle really stands or falls in the finite case. The part of their article that I would like to discuss is that concerning the various examples that have been presented against the likelihood principle, where my views are perhaps different from those of Berger and Wolpert (BW), and in the course of the discussion my approach to the infinite case should become clear. Before doing so I want to preface my remarks with two comments. First, I think that we Bayesians should be grateful to Stein, Stone, Fraser and Monette, for their interesting examples, all of which have some real substance to them. Theories require good criticism in order to grow, and the lack of such criticism has been detrimental to the Bayesian theory. Secondly, I think it is essential that we keep in mind the distinction between the likelihood principle (by which I mean the formal likelihood principle of BW) and various implementations or interpretations of the likelihood principle. I shall try to demonstrate that none of the examples speak against the likelihood principle as such, but rather that they constitute frequentistic arguments against the use of specific improper (or diffuse finitely additive) prior distributions. I shall then explain why I think such arguments have no real teeth to them.

Let us begin with the example of Stein. Although it was originally presented by Stein as an argument against the likelihood principle, with an argument against lazy Bayesians tacked on at the end, I regard it as primarily an argument against Bayesians (such as myself) who use improper or finitely additive prior distributions to obtain approximate posterior distributions, and also against the theory of de Finetti (which I follow) which in principle does not rule out any finitely additive prior distribution. To begin with, the likelihood principle does not justify either (5.3.4) or (5.3.5) of BW, since it does not suggest a way of attaching probability to sets. It is true of course that some individuals who support the likelihood principle (perhaps with qualifications), such as George Barnard and A. W. F. Edwards, also sometime recommend the use of such probabilistic interpretations of the likelihood function, but that is by virtue of additional assumptions, whether explicit or implicit, and is not really part of the likelihood principle. BW apparently accept that (5.3.6) is a strong argument against the use of a uniform improper prior distribution for theta, but suggest that there is no difficulty for Bayesians because on the one hand theta is a scale parameter and so it is the logarithm of theta (if anything) that should be given a uniform prior distribution, as in Barnard's reply to Stein; and on the other hand, and more importantly, they argue that with proper prior distributions the type of interval that Stein shows has bad frequentistic properties can occur only rarely since "Y is almost certain to be enormous." Although I agree with both arguments of BW, it seems to me that the issue being raised by Stein is not whether a sensible Bayesian can avoid the intervals (5.3.5), but rather whether by virtue of carelessness or because his theory permits such intervals (as for example is true of the de Finetti theory) the unwary or even wary Bayesian will become frequentistic prey. If there really were a trap with teeth to it then Stein's example would suggest either that one stick with proper prior distributions, or else be quite careful in the choice of improper, or merely finitely additive, prior distributions, and as Stein says, this would make the "prior distribution used depend on

accidental features of the decision problem." Now to me this seems to be a real and important issue. Suppose we are discussing a real-world parameter, whose existence, definition, and meaning, in no way depend upon the experiment to be performed. (The existence of such parameters may be far less common than is usually assumed, but presumably we could all agree that at least some such parameters exists, or at any rate are worth discussing, and confine attention to these. When the "parameter" depends upon the experiment for its existence and meaning, then, of course, the likelihood principle does not apply.) In the subjective Bayesian theory of de Finetti and L. J. Savage, the prior distribution for such a parameter would be chosen to represent one's opinions about that parameter, and whether the measurement is to be made according to the normal model or according to (5.3.3) should not in any way affect the prior distribution. If for some reason I thought that a uniform improper prior for theta was appropriate as an approximation under the normal model, and then learned that in fact the measurement error was distributed according to (5.3.3), but with the nearly identical likelihood functions that Stein produces, then it seems to me that the uniform prior should still provide a satisfactory approximation in obtaining my posterior distribution. Furthermore, a Bayesian who would use the uniform prior for theta when the measurement error is normally distributed, but would use a uniform prior distribution for the logarithm of theta when the measurement error is distributed according to (5.3.3), is coming very close to violating the likelihood principle in this example, since he is making very different inference about theta in the two cases even though the likelihood functions are in a certain sense very "close," and theta is the same fixed quantity. See Savage (1970) for a related argument. (Some Bayesians, for example Box and Tiao, actually recommend that prior distributions be made to depend upon the sampling scheme, and so would use a different prior distribution for the parameter of a Bernoulli sequence if the experiment were of binomial form than if the experiment were of negative binomial form, even when the choice of the experiment is

made by randomization and is thus uninformative, and this clearly violates the likelihood principle. Whatever else may be said of such an approach, it is certainly not a part of the ordinary subjective Bayesian theory, in which the prior distribution for a parameter of the type we are discussing does not depend upon what experiments may or may not be performed at some future time. Furthermore, if BW choose to use improper prior distributions, but only when these do not lead to bad frequentist properties, then they too are perilously close to a violation of the likelihood principle, since their choice will turn out to depend upon the sampling distribution, just as with Box and Tiao.) What the Stein example actually demonstrates is that if a Bayesian uses the uniform prior distribution for theta, then his posterior probability for the interval (5.3.5), given any y, is at least .95, while given any theta, the frequentist probability for the interval is very tiny according to (5.3.6). This is the phenomenon of nonconglomerability. Conglomerability is a property of a probability distribution, and was defined by de Finetti (1972, p. 99) as follows: if the conditional probability of an event, given each element of a partition, lies between p and q, then also the probability of the event lies between p and q. Conglomerability always holds for countably additive probability distributions and countable partitions, but need not hold for merely finitely additive distributions, and in fact, as shown recently in Hill and Lane (1983) using only elementary mathematics and verifying a conjecture of de Finetti, conglomerability and countable additivity are equivalent for countable spaces. The uniform improper prior distribution can be given a finitely additive interpretation, which is why the nonconglomerability exhibited by Stein can occur. Thus for the partition based upon the value of theta, we have (5.3.6), while for the partition based upon the value of Y, the intervals (5.3.5) have posterior probability at least .95 for all possible such Y values, when the uniform prior distribution for theta is used. The unconditional probability of the interval has not been defined, but whatever value it is given must exhibit a nonconglomerability with respect to one or the other of the two partitions. In Hill (1981) I

gave some general arguments as to why nonconglomerability cannot be avoided in the subjective Bayesian framework, and as to why I believe there is really very little that is operationally meaningful in the type of superficially frightening calculation exhibited in (5.3.6). After discussing the other examples, I will return to this issue, and suggest a new argument as to why there is no way to demonstrate any undesirable consequences if one uses an improper prior distribution. Thus although I agree with BW that the Bayesian can always avoid the trap by using proper distributions, I also like to use improper prior distributions or merely finitely additive prior distributions when I think they yield a simple and satisfactory approximation to my posterior distribution, and do not accept (as BW seem to do) that there are any operationally meaningful ill consequences to so using such distributions (even for all possible values of Y in the Stein example). A general theory pertaining to the type of consequences that arise in nonconglomerable situations has been formulated and elegantly presented by Heath and Sudderth (HS) in Heath and Sudderth (1978) and by Lane and Sudderth (1984), and as we shall see later all of the examples purportedly against the likelihood principle, are in fact merely more examples of the type of incoherence discussed by HS. (See HS example 5.2 for a very simple example similar to that of Stein.) It is my opinion, however, that the HS requirement for coherence, to the extent it goes beyond the de Finetti form of coherence (which only requires avoidance of sure loss with a finite number of gambles), is too restrictive, and at least in the special case of the Monette-Fraser example, I will argue that the apparent ill consequences of violating the HS condition for coherence cannot really be made operational.

The Stone example does not directly pertain to the likelihood principle, and has been analysed by myself in Hill (1981) from a finitely additive point of view. In addition to observing that a finitely additive diffuse uniform distribution on the "length" of path yields the standard confidence result, it was also pointed out that in order to obtain the uniform distribution on the location of the treasure that Stone criticizes it

is necessary to employ a diffuse finitely additive prior distribution which gives odds of nine to one in favor of paths of length j+1 versus paths of length j-1, for all j > 1, and such a prior distribution seems rather silly in this example. Nonetheless, just as in the Stein example the de Finetti theory does not rule out such prior distributions, and the question is once again whether a serious case can be made against their use. It may be noted that the posterior obtained with this prior is also incoherent in the sense of Heath and Sudderth.

Fraser in his discussion of my Valencia article Hill (1981) maintained that the Stone example also has implications with regard to the likelihood principle, and gave the example reported by BW. The example as initially presented did not seem appropriate to me, since it required that first theta, the true path to the treasure, be selected as in Stone's example, next that the observed path of the Stone experiment be given, and finally that a randomization be performed that leaves one with the same likelihood function as before. In this situation, where the second experiment consists of the first experiment together with an irrelevant randomization, the likelihood principle follows from just the sufficiency principle, and is barely worth commenting on. However, the Fraser example can be modified so that this is no longer the case, for example, one can imagine that a new experiment E* is performed as follows: first a path z from the origin is selected according to a probability distribution that depends upon theta, in such a way that z is equally likely to be any of the four paths for which theta is a one block extension or retraction of z, and we observe this z. Next, someone else who somehow or other happens to know the true theta, unobservedly retracts or extends z back (or forth) to the true theta (which therefore remains the true parameter of the experiment), and from there does the experiment with the Fraser likelihood function as presented by BW, with z = x(0). One then observes in this last experiment a path X. The likelihood function for theta based upon the data Z = z, X = z, in the experiment E*, is then identical with the likelihood function derived from the Stone experiment with the same

observed path z, and so with this modification the Fraser example does meet the conditions for the likelihood principle to apply. If one adopts a Bayesian point of view, then as BW argue, one has precisely the same apriori information about theta no matter which experiment is performed, and it is certainly reasonable to draw the same inference in each experiment. Suppose, however, one imagines that it is meaningful to consider the case of no prior information (whatever this means), so that Bayesian inference is not possible. It would be interesting to know what the appropriate non-Bayesian inference about theta would be under E* as opposed to the Stone experiment. Would, for example, a non-Bayesian now treat theta as though it were equally likely to be any of the four possible paths? Rather than calling into question the likelihood principle it seems to me that this example may raise some serious problems for non-Bayesians.

Now let us turn to the new example by Monette and Fraser (MF). This example does not seem to pertain directly to the likelihood principle, since there is only one experiment under discussion. It does, like the other examples, suggest that according to frequentist standards a certain improper, or diffuse finitely additive, prior distribution is unsatisfactory, and BW, as in the Stein example, argue that for proper prior distributions, and even for the conventional improper prior distribution for something akin to a scale parameter, there is no difficulty. Although again I agree with BW that ordinarily one need only consider quite proper prior distributions, and also that the particular improper or finitely additive distributions that are being castigated may be of no special interest, I would nonetheless like to argue that as yet very little has been demonstrated against the use of such prior distributions. My argument would be much the same in all examples, but will be presented here in connection with the MF example, which is the simplest. What has been shown is that choice of an improper uniform prior distribution (or a finitely additive diffuse prior distribution) for theta would lead to a posterior distribution, such that if I were to bet in accord with it, I would be a loser in the Heath-Sudderth sense (this is closely related to a lack of

extended admissibility). Since I regard the finitely additive uniform distribution as useful for approximations, and as having as much justification as any other distribution (to be given full rights, as de Finetti says), and in any case I don't think that it matters whether theta is akin to a scale parameter, so that I cannot take refuge in the BW argument unless I dispense entirely with both merely finitely additive distributions and improper priors, I am loathe to give it up so easily. So suppose I fall into the trap and agree to post odds in accord with the posterior distribution that is uniform over the three possible values for theta, given x. Let us see to what extent MF can take advantage of such foolishness as I am willing to exhibit. In order to do so they must construct a real world version of their mathematical model. So first of all they must somehow or other pick a theta, and then pick an x in accord with their model. The Heath-Sudderth gambling scenario seems to be a convenient and appropriate way of describing the operational consequences of my potential incoherency (even for those who think that they don't gamble), and if desired, can easily be translated into non-gambling terms. Thus suppose that theta is picked from amongst the positive integers by the master of ceremonies in any way he likes, and then X is selected according to the MF distribution for X, given theta. After we are all given the value x that X takes on, I then use the posterior distribution based upon the uniform prior distribution for theta to determine the odds that I, as bookie, will give for the various values of theta. Also, after observing x, MF are entitled to place any finite number of bets concerning theta they wish, and finally theta is revealed by the master of ceremonies and all bets are paid off. Suppose that MF bet a dollar on the event that theta takes on the value $\delta_1(X)$, and let G denote the final payoff from me to them. Given theta, there is at least probability 2/3 that $\delta_1(X)$ will equal theta, and so the expectation of G, given theta, is at least \$1, for all possible theta, and I am incoherent in the sense of Heath-Sudderth. However, to make the transaction operationally meaningful it is necessary to specify precisely how X will be revealed, for

example, that X will be expressed to the base 10 (or in any other specified form whatsoever), and that a certain finite time limit is prescribed during which the game is to be played. Now I think that all of us could come to agreement that given the constraints of the world we live in, there is an upper bound, say N, to the value of X that can be reported to us as data in the prescribed form and in the prescribed time, for example, an N such that in the present state of technology even the fastest computer could not display an integer greater than N in the time allotted for the experiment. (To be even more realistic, the same is true with regard to theta, but for the purpose of the present argument we need not assume any constraint on the magnitude of theta, and shall follow MF in assuming that the master of ceremonies can choose any value whatsoever, and then can and does select an X in the way that MF specify. Of course theta, like X, cannot actually be reported it if exceeds N, but one might wish to consider cases where the master of ceremonies has extraordinary powers, and is entrusted to announce who wins the gamble in situations where X does not exceed N but 2 X does. This points out that there are in fact a variety of ways to make the Heath-Sudderth scenario operationally meaningful, and that our assumption that X cannot be reported if it exceeds some known N, is merely the minimal real-world constraint. This gives the present argument greater generality in that it may apply even when theta is a real-world physical parameter for which there would be no known bounds. If a bound on theta were available then of course the argument would apply all the more. However, the point is that whether or not there is such a bound on theta, there is necessarily a bound on the possible value of X that can be reported. If we do take into account known bounds on possible theta, or on possible reported values of theta, then this would lead us to proper prior distributions as in BW. However, it is not necessary to introduce such considerations in the present example since, as we shall soon see, the boundedness of the X that can be reported already destroys the frequentistic argument.) Suppose then that theta and X are selected by the master of ceremonies in accord with the MF model, without any constraint upon the magnitude of either, and that N is a

known upper bound for any X that can possibly be reported as data. We do not assume that N is the least possible upper bound for a reportable X, but merely that it is an upper bound. (It is, of course, desirable that N be not too much larger than the least upper bound, but the argument does not depend upon this.) Thus our experiment now consists in precisely the MF experiment, together with the modest real world constraint that if $X > N$, then no value of X will be reported (since it would be impossible to do so), and hence that any bets that depend upon the value of X will be called off. In this situation the actual gamble as to whether theta is $\delta_1(X)$ is called off whenever $X > N$, and we are dealing with a conditional gamble in the sense of de Finetti (1974, Ch. 4). Consequently the payoff from me to MF is now as before if X is actually reported, but all gambles are called off if $X > N$. (There is nothing underhanded here with regard to the reduction to conditional gambles: in order that transactions can occur, so that the scenario has operational meaning, it is necessary that the bets are conditional bets, given that a value of X is reported, and hence conditional upon the event that the X selected in the MF experiment does not exceed N. If $X > N$ then no X is reported and no gambles can be made concerning whether theta = $\delta_1(X)$. Note also that it is not necessary to assume that X must be reported if $X < N$, but merely that X cannot be reported if $X > N$, and that X must be reported if it is possible to do so in the fashion prescribed.) It is interesting now to see what becomes of the frequentist argument that showed that the conditional expectation of my loss, given theta, is at least \$1, for all possible theta. I am still using the same prior distribution as before, so that if I am actually given a value of X (necessarily $\leq N$) then I post the same odds as before against the event that theta = $\delta_1(x)$. If theta is sufficiently small so that X both can and must be reported (hence necessarily theta $< N/2$), then the expectation of G, given such a theta, is the same as before, at least \$1. On the other hand, if theta $> N/2$, then the only values of X that can possibly be reported are X = theta/2 or (theta-1)/2, depending on whether theta is even or odd. Hence given a value of theta $> N/2$, and given that the gamble is not called off, it is certain that

theta is not equal to $\delta_1(X)$, and so the conditional expectation of G, given such a theta and that the gamble is not called off, is -$1, while the unconditional expectation of G, given such a theta, lies between -$.33 and $0. Whether in gambling terms or in coverage probability terms, it is thus seen that when a real-world constraint as to the value of X that can be reported is incorporated into the MF example, then the example breaks down, and in fact if a value of X is actually reported, then the very same $\delta_1(X)$ that appeared so desirable from the MF point of view, becomes impossible as the value of theta when theta > N/2. (A variation of this scenario would require me also to post odds on theta, given the information that X exceeds N. This would require care in obtaining the posterior distribution for a finitely additive prior distribution, but in any event the $\delta_i(X)$ are still not available, and the frequentistic argument still breaks down.)

The above form of argument suggests why there need not be anything wrong with using the finitely additive uniform distribution in connection with experiments conducted by human beings, i.e., where the reportable observation X, if not theta itself, must be bounded, and one can with a little thought always choose a generous upper bound. More generally, when theta is not chosen by any human, but is a parameter of the real world, then one may not be able to argue for any upper bound for theta, but in my opinion neither will there by any operationally meaningful scenario in which one who chooses a finitely additive distribution can be shown to be in trouble by virtue of frequentist properties. BW suggest using proper prior distributions for theta as a way of avoiding the apparent frequentistic difficulties in the above examples. However, if BW or Heath-Sudderth wish to use improper or merely finitely additive prior distributions, and if they choose to avoid nonconglomerability and its frequentistic consequences, as in the various examples, then it seems to me that they are in fact going to violate the likelihood principle, since the particular improper or finitely additive distributions that they must rule out in order to avoid nonconglomerability will depend upon the form of

the experiment, just as the prior distributions that Box and Tiao recommend depend upon the form of the experiment. (BW can avoid violating the likelihood principle by either restricting themselves to proper prior distributions, or by using improper prior distributions only when they provide an "adequate" approximation to the posterior distribution based upon some proper prior distribution. But I think it is too restrictive always to restrict oneself to proper prior distributions, and although, as mentioned earlier, I too ordinarily take the approximation point of view, I don't think the notion of what is an adequate approximation should depend upon frequentistic properties.) In the Stein example and in the Heath-Sudderth example (5.2), where according to the model (taken literally) the parameter and data are not discrete and the set of theta compatible with the data is not finite, the argument I have given above must be modified, but I think that here too, when real-world constraints are allowed for, the frequentistic argument will again break down, and I hope that my discussion of the MF example at least suggests some of the difficulties involved in trying to make the frequentistic argument operational. In my Valencia article I also suggested that as yet no serious argument for conglomerability had ever been given (since that time Lane and Sudderth (1984) have given such an argument, but I do not agree with their views concerning the appropriate gambles with which to define coherency), and suggested also that Stone's example had an implicit assumption of conglomerability for its castigation of the uniform prior. (Stone (1979) replied by asserting that Hill is like a prisoner condemned to death by guillotine who rejoices that the guillotine will be chosen from an infinite collection. I replied "Yes, Mervyn, but all your guillotines are made of butter." At a deeper level this concerns the appropriate interpretation of conditional probability, whether in terms of gambles that are called off if the conditioning event does not occur, as in de Finetti (1972, p. 81), or in the more usual way, but there is not space to go into this here.) Sir Harold Jeffreys once criticized conventional tests of significance because they reject hypotheses that may be true on the basis of data that have not occurred. Apparently some would also have us reject the

use of improper prior distributions because of experiments that cannot be performed.

Finally, let me mention an important real-world problem where exactly such considerations as I have been discussing arise. Consider a balanced one-way random effects analysis of variance model, with I rows and J columns. In Hill (1980) I examined the consequences of drawing inference about the ratio of the between to the within variance, τ, using as data only the ratio of the mean square between to the mean square within. It was shown that this can in fact be justified by a fully Bayesian analysis, and is appropriate when the prior distribution of the two variance components is such that their ratio is independent of the within variance, and the overall mean is given a diffuse prior distribution. The problem then reduces to one of inference about a simple location parameter, $\gamma = \ln(1+J\tau)$, based upon data $\hat{\gamma} = \ln(MSB/MSW)$, and with the distribution of $\hat{\gamma}-\gamma$, given γ, being that of the logarithm of a random variable having the F distribution with I-1 and I(J-1) degrees of freedom. The likelihood function for γ based upon the data $\hat{\gamma}$ is then the density of this ln(F) distribution, translated so that the mode is at $\hat{\gamma}$ (and with degrees of freedom reversed), except that the density must be truncated from below at 0 because γ is nonnegative (it is convenient and harmless to think of the likelihood function as being defined for all γ, so that even if $\hat{\gamma}$ is negative, the mode is at $\hat{\gamma}$, and then to make the truncation from below at 0 stem from the prior distribution.) If one uses the uniform prior distribution for γ, with $\gamma > 0$, then one is in precisely the type of situation that the Stein, HS (5.2), and MF examples, deal with. Although there is nothing magical or mandatory about use of this particular prior distribution, and in fact there is usually a great deal of prior information about the ratio of variance components in such problems, so that I would recommend use of a proper prior distribution for γ, at the same time, I think a great deal of insight can be obtained from the improper uniform prior on γ, and do not think it should be automatically ruled out merely because it may lead to bad frequentistic risk properties. As I argue in Hill (1980), the posterior expectation based upon this improper prior

yields, at least in some respects, a more plausible estimator for a multivariate mean (the realized random effects) than does the positive-part Stein estimator. For example, the posterior expectation cannot shrink all the way to the grand mean of the observations, since the weight given to the row means $\bar{y}_{i\cdot}$ decreases only to 2/(I+1) as the ratio of the between to within mean square goes to 0, whereas of course the positive-part Stein estimator can give zero weight to the row means, and this is not always sensible. The behavior of the posterior expectation stems partly from the particular form of the prior distribution for the variance components (especially the fact that the ratio of the variance components is apriori independent of the within variance component), and partly from the truncation of the posterior distribution of γ from below at 0, neither of which do non-Bayesians incorporate into their analysis. In my opinion due respect for the likelihood principle, and proper allowance for these aspects of the problem, are far more important than any frequentistic arguments against the use of improper prior distributions, while at the same time, as BW would presumably agree, a proper prior distribution for the variance components would ordinarily be reasonable, and give the best of both worlds.

DISCUSSION OF THE SECOND EDITION BY PROFESSOR HILL

Since the publication of the first edition of the monograph by Berger and Wolpert, I have written several articles pertaining to the validity of the likelihood principle, and to its role in Bayesian data-analysis. I believe that the example of Hill (1987a,b) clearly shows that the original statement of the likelihood principle by Birnbaum in terms of an abstract concept of evidence was faulty. The difficulty in the likelihood principle is easily remedied, however, and this was done in my statement of the restricted likelihood principle in those articles. In my formulation one speaks not of the evidence in some undefined abstract sense, but rather only of the evidence about the *value* of θ, and excludes from the discussion any assertion about how θ might relate to other unknowns, whether hypotheses or parameters. Thus my

example can be viewed as showing that two different experiments that yield proportional likelihood functions for θ do not necessarily provide the same evidence about θ, since we can learn, for example, that θ has a different 'color' in the two experiments. The color might be an important part of the overall evidence about θ. Of course the color can be included in the parameter, but the likelihood principle, as usually formulated, does not require one to do so. It is hoped that once this point is understood, others will, like myself, become even stronger supporters of the essential part of the likelihood principle.

The basic point of my example is related to fundamental questions that arise in theories of causality, for example, concerning determinism and the possibility of independence in the real world. Such questions arise in critical discussions of quantum mechanics and relativity theory, for example, in connection with Bell's inequality, as well as in philosophy.

In Hill (1985-86, p. 223) I have given an account of how the likelihood principle must be further modified to deal with Bayesian data-analysis, where through exploration of the data, one may modify the original model. The same article, p. 202f, argues that even apart from inadmissibility, incoherence, and the failure to utilize available information, the frequentist approach breaks down completely in connection with such data-analysis, since all frequentistic assertions must be conditional not only upon the diagnostics used, but their order, and even the thoughts that cross one's mind. Such conditional probabilities are plainly both unknown and unknowable. Finally, Hill (1988) gives a very short, and partly new, proof of the stopping rule principle, i.e., that the stopping rule is irrelevant for inferential and decision-making purposes, or that "sequential analysis is a hoax," as concluded by Anscombe (1963, p. 381). Here the proof does not depend upon the likelihood principle, or even the restricted likelihood principle. Instead, it is shown that on a post-data basis, i.e., given the *realized* data, sequential analysts purport to extract information over and above that following from the corresponding fixed sample size experiment, from a *logically certain event*. In this article the

important distinction between the pre-data and post-data considerations is emphasized. Once one is given the data, the primary aim must be to make intelligent and rational decisions, for which the Bayesian approach seems quite well suited. Of course sequential *design* need not necessarily be a hoax, but it appears that not very much is known about this potentially important subject, perhaps because of the confusion between pre-data and post-data considerations, as discussed in Hill (1988).

The likelihood principle is often mistakenly assumed to be largely equivalent to the Bayesian approach. The likelihood principle, as proposed by Birnbaum, in terms of an abstract and empty concept of evidence, was in fact the last gasp (intellectually speaking) of the theory of classical statistics, with its naive pretence at objectivity. Indeed, Birnbaum (1962, p. 277) quotes Jimmie Savage as follows. "Rejecting both necessary and personalistic views of probability left statisticians no choice but to work as best they could with frequentist views... The frequentist is required, therefore, to seek a concept of evidence, and of reaction to evidence, different from that of the primitive, or natural, concept that is tantamount to application of Bayes' theorem."

"Statistical theory has been dominated by the problem thus created, and its most profound and ingenious efforts have gone into the search for new meanings for the concepts of inductive inference and inductive behavior. Other parts of this lecture will at least suggest concretely how these efforts have failed, or come to a stalemate. For the moment, suffice it to say that a problem which after so many years still resists solution is suspect of being ill formulated, especially since this is a problem of conceptualization, not a technical mathematical problem like Fermat's last theorem or the four-color problem."

Birnbaum then states that "The present paper is concerned primarily with approaches to informative inference which do not depend upon the Bayesian principle of inverse probability." It would therefore appear that Birnbaum regarded his approach to evidence as meeting the objections that Savage and others had raised. However, just as the Michelson-Morely experiment spelt the

death knell for classical physics (which was at least a highly successful and useful subject), one must wonder what is left of classical statistics, without even Birnbaum's likelihood principle to sustain it. All that appears to be left is the restricted likelihood principle, which is implied by the Bayesian approach, and is somewhat more general than the Bayesian approach, since it allows for versions of Bayesian data analysis such as in Hill (1988). I know of no way to demonstrate even the restricted likelihood principle, however, other than through the Bayesian approach.

I think that nowadays it will be readily understood that the pretence at objectivity in classical statistics was equivalent to taking a particular subjectivistic Bayesian view, that based upon diffuse prior distributions, and by fiat declaring that this constitutes objectivity. Such prior distributions play an important role in Bayesian statistics, via the stable estimation argument of Jimmie Savage, but do not acquire any magical status in the Bayesian theory.

The nature of "objectivity" was never seriously discussed in classical statistics, despite the fact that this was and is a notoriously difficult question in philosophy. Even in statistics, numerous examples exist showing that this pretence cannot be made, without leading to absurdities. There are many examples in which the *realized* likelihood function is nearly flat, no matter what the pre-data expected information may have been. This occurs, for example, in inference about variance components when the classical unbiased estimator of the between variance component is negative, as in Hill (1965, 1967). A more sophisticated example of the need for a subjective view occurs in deciding whether a particular observation is an "outlier," as in Hill (1974b, Section 4) and Hill (1988, Section 3). What the so-called objectivists do, as Jack Good says, is to SUTC (sweep the subjective aspects under the carpet). Probability and statistics, as related to the real world, are fundamentally subjective or personalistic. In certain situations, however, one may obtain practical objectivity by means of a consensus as to appropriate prior distributions and models. See Hill (1985-86, 1988). Also, sometimes certain

"objectivistic" methods, such as the fiducial approach, can be justified Bayesianly, as for example with A(n) in Bayesian nonparametric statistics, Hill (1987c). Finally, by a delicious irony, it also turns out that the few important objective *criteria* that frequentists have recommended, such as admissibility, extended admissibility, etc., lead inevitably back to the Bayesian approach.

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."

James's eloquent statement can serve as a preamble to the theory and practice of Bayesian data analysis and decision-making, which is a synthesis of the empiricism-pragmatism of John Locke, David Hume, Charles Peirce, and William James, with the rationalism of Plato, Descartes, Kant, and others, and to which I believe that the next century will be devoted.