What’s wrong with Bayesian methods?

E. T. Jaynes

St. John’s College and Cavendish Laboratory, Cambridge, CB2 1TP, UK

April 1984

Our title is the title of a talk given in Cambridge on Feb. 3, 1984, by Professor G. A. Barnard. For him it was a question to be taken seriously, and answered seriously. But for us it is only a rhetorical question; for while we see many things in Bayesian methods that are still incomplete and in need of further technical development, we are unable to see anything basically wrong with them.

However, Barnard’s argument proved to be valuable, because it framed our differences in such a clear and sharp way. Thirty years ago Jimmie Savage remarked of statistics that “there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel.” Pondering Barnard’s remarks we were able to see where our communication has failed, more clearly than before.

For decades Bayesians have been accused of “supposing that an unknown parameter is a random variable”; and we have denied, hundreds of times and with increasing vehemence, that we are making any such assumption. We have been unable to comprehend why our denials have no effect, and that charge continued to be made.

Sometimes, in our perplexity it has seemed to us that there are two fundamentally different kinds of mentality in statistics; those who see the point of Bayesian inference at once; and need no explanation; and those who never see it: however much explanation is given. But Bernard’s remarks provided a clue to what has been causing the genuine Tower of Babel situation.

Barnard defined the term “statistics” as follows:

(A) It is not concerned with decision making; and it is not concerned with scientific inference in general.

(B) Rather, statistics is “that part of inference where experiments are repeatable, their results only partially so.”

(C) Any models we use must be verifiable — or at least criticizable — and the probabilities we use must be frequencies.

He then explained that Bayesian methods of parameter estimation, which express the result of a parameter estimation, are illogical; for how could the distribution of a parameter possibly become known from data which were taken with only one value of the parameter actually present? Assignment of prior probabilities was dismissed as an "uncriticizable assumption about a chance distribution."

1 The Bayesian reaction

Let us note the instinctive first reaction that a Bayesian has to Bernard's arguments. His definition of "Statistics" seems to cut it off from most of the problems where we had been led to believe that "statistics" was the appropriate tool. For it is simply a fact of life that in most of the real problems faced by scientists, engineers, economists, and administrators:

(A) we are concerned with decision making, and with inference in general.

(B) There is no repeatable experiment involved; the reason why inference is needed is not "random errors" but incomplete information.

(C) Officially, Bayesians do not know what it means to "verify" a model or "criticise" a prior probability, because our models and prior probabilities are only summaries of what we knew about the phenomenon being observed and about the possible values of the parameters. The evidence in support at them has already been taken into account when we propose them.

Unofficially, of course, we may wish like anybody else to make a "trial run" experiment with some model or prior that does not represent actual knowledge, but only a whimsy, to see what happens. This might, for example, help us to decide whether it would be worth the effort to get a certain kind of prior information.

Bernard's subsequent complaint appears to us as an example of a semantic trap caused by habitual use of the phrase "distribution of the parameter" when one should have said "distribution of the probability." Our communication problems arise in large part from the difficulty that orthodox terminology is not adapted to expressing Bayesian ideas (in this respect it reminds one of the Orwellian NEWSPEAK, a language within whose vocabulary and grammar it was not possible to express dissenting views).

In Bayesian inference, both the prior and posterior distributions represent, not any measurable property of the parameter, but only our own state of knowledge about it. The width of the distribution indicates not the range of variability of the true values of the parameter, but rather the range of values that are consistent with (i.e. not ruled out by) our prior information and data. Honesty therefore compels us to admit them as possible values.

What is "distributed" is not the parameter, but the probability. A terminology which always makes this clear is much needed. The phraseology "probability distribution function (pdf) for a parameter" is a step in the right direction, but perhaps we might find something more brief and explicit.

Bernard’s argument is then absolutely mind-boggling to a Bayesian: for to try to “verify” a distribution which expresses only a state of knowledge about a parameter by performing random experiments on the parameter, is the logical equivalent of trying to verify a boy’s love for his dog by performing experiments on the dog. But just to have our differences appear in such acute form suggests a plausible — and at least to the writer, new and startling — hypothesis about where our communication has failed.

Is it possible that, for all these years, those who have seemed immune to all Bayesian explanation have just been misunderstanding our purpose? All this time, we had thought it clear from our subject matter context that we are trying to estimate the value that the parameter had *when the data were taken.* Put differently, we are trying to draw inferences about what actually did happen; not about what might have happened but did not.

Nothing could be further from our purpose than to make statements about how our parameter might be “distributed” in other situations that we are not reasoning about. Indeed, our posterior distribution for a parameter is not necessarily a predictive distribution for values that it might have in future experiments; this depends on further details of our prior knowledge, that were not relevant in the problem we had addressed.

But now it appears that our critics may have been trying to interpret our work in a different way, imposed on them by their habits of terminology as an attempt to solve a very different problem. If so, our past communication difficulties would become understandable: the problem they impute to us has — as they correctly see — no solution from the information at hand. The fact that we nevertheless get a solution then seems miraculous to them, and we are accused of trying to get something for nothing.

“Statistics” as defined by Barnard and “Scientific Inference” as defined by Jeffreys are concerned with different problems; and so, far from having reason to argue the merits of our different methods, we have no reason to compare them at all. We can restore peace in both fields simply by going our separate ways.

Yet it is clear that neither George Barnard nor I wants to do this; for we both see that, in spite of the above, our different problems are closely related mathematically. Rising above past criticisms — which now appear to have been only misunderstandings of our purpose — Bayesian are in a position to help orthodox statistics in some of its most serious current difficulties. For the Bayesian procedure is flexible enough to apply to many different problems, including both those by Barnard and Jeffreys.

2 Let’s not confuse two different problems

In the following it is essential that we understand clearly what the two problems are and which problem we are talking about.

In the Jeffreys scenario we are *estimating*, from our prior information and data, the unknown constant value that the parameter had when the data were taken.

In Barnard’s we are *deducing*, from prior knowledge of the frequency distribution of the parameter of some large class C of repetitions of the whole experiment, the frequency distribution that it has in the subclass $C(D)$ of cases that yield the same data D. The problems are so different that one would expect them to be solved by different procedures.

But Bayesian inference need not adhere constantly to the Jeffreys problem, for nothing prohibits us from estimating a frequency distribution instead of a fixed value, if that happens to be the thing of interest. But instead of saying that the probability *is* the frequency, we would calculate the probability that the frequency lies in various intervals, enabling us to make statements about the accuracy of the estimate.

Likewise, orthodox statistics could in principle switch back and forth between the problems of deducing conditional frequency distributions of a parameter, and inferring fixed values of a parameter, depending on whether the parameter is or is not considered "random." However, we are not sure that this switching has ever occurred, for we know of no real problem in which anyone has actually used Bayes' theorem for the purpose that Barnard supposed.

In the case where the parameter is considered to be a fixed constant, application of Bayes' theorem for Barnard's purpose would indeed be illogical; or rather idle, for it presupposes that we already know, in the singular prior, the frequency distribution of the parameter in every subclass $C(D)$, and so there is nothing more to be learned from the data.

But we never know that singular frequency distribution in advance (if we did know it, we would not be considering the problem). Orthodox statistics then reverts necessarily to inference concerning a fixed value of the parameter rather than a distribution of values. Then the orthordoxian and Bayesian are trying to solve the same problem if neither has any prior information about the parameter; and it makes sense to argue our different philosophies and compare our different methods.

3 What happens if we consider the same problem?

Since orthodoxy sees no meaning in a probability which is not also a frequency, it is obliged to seek other tools than probability theory. Lacking guiding theoretical principles, Neymannian orthodoxy is reduced to inventing *ad hoc* procedures like confidence intervals or significance tests based on some statistic chosen by intuition.

Fisherian orthodoxy is in a better position because it recognizes that such inference is valid only when we are using sufficient statistics or conditioning on ancillary statistics. Thus it avoids the wild anomalies that can arise in Neymannian inference, some of which were noted by Barnard.

For the Bayesian, who does see meaning in a probability that is not a frequency, all the needed theoretical principles are contained in the product and sum rules of probability theory. He views them, not merely as roles for calculating frequencies (which they are, but trivially); but also rules for conducting inference — a nontrivial property requiring mathematical demonstration. But that demonstration is a long since accomplished fact, as noted below, and the result is those rules tell us to use Bayes' theorem in the manner of Laplace and Jeffreys. So how do the pragmatic results compare?

An early indication of things to come was the demonstration by Jeffreys (1939) that the orthodox *t*-test follows exactly from a few lines of Bayesian analysis, using the Jeffreys uninformative priors for the location and scale parameters. The same is easily shown to be true for the *F*-test and the orthodox test for the parameter of a Poisson distribution.

Lindley (1958) then proved that if any problem is equivalent (to within a change of variables) to a location/scale parameter problem and has sufficient statistics so that fiducial inference is possible, then that fiducial distribution is identical with a Bayesian posterior distribution. Unfortunately, location/scale parameter problems do not in general have sufficient statistics; but they do have a complete set of ancillary statistics. The writer has shown (Jaynes, 1976) that the "best" confidence interval for any location or scale parameter (i.e. the shortest one that meets Fisher's requirement of conditioning on all ancillary statistics) is identical with the Bayesian posterior probability interval at the same level, based on the Jeffreys uninformative priors. The proof does not even require independent sampling.

It appears to the writer that all of the procedures which the "orthodox" statistician himself considers fully satisfactory. follow trivially from the Bayesian approach with noninformative priors. If there are exceptions to this conjecture; it would be interesting to learn about them and study them.

But many problems encounter technical difficulties (nuisance parameters, nonexistence of sufficient or ancillary statistics, flat-topped likelihood function, inability to use prior information) which have not been overcome by any satisfactory orthodox procedure. We find, when we apply Bayesian methods to such problems, that the difficulties are overcome effortlessly yielding substantial improvements over orthodox results (Jaynes, 1976).

Finally, Bayesian methods with the adjunct of Maximum Entropy — which can be thought of either as a factor in the prior or as a utility function — apply also to a mass of new problems that cannot be formulated at all in orthodox terms; and computers are now busy grinding out the useful solutions. They are performing very non-trivial data analysis in such diverse fields as spectrum estimation, medical instrumentation, underwater acoustics, radio astronomy, geophysics, optical image reconstruction, physical chemistry, crystallography, and what will probably become the largest area of application, biological macromolecular structure determination,

Here in Cambridge, computers are now routinely locating constrained entropy maxima in spaces of over a million dimensions. The numerical results are so impressive that the methods are moving steadily into new areas, and major efforts are under-way in many places, to develop still more powerful programs.

In view of this, we are not surprised to find that criticisms of Bayesian and/or Maximum Entropy methods deplore only our philosophy, and stop short of examining our actual numerical results in real problems (which seems a pity, because we are proud of those results and think they would be easy to defend, although we could hardly compare them to non-existent orthodox results).

But our philosophy is very easy to defend also as soon as one recognises our purpose as explained above; for it is not an opinion, but a theorem (Cox, 1946), that any set of rules for inference, in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to the Laplace-Jeffreys rules, or inconsistent (in the sense that one could find two methods of calculation, each permitted by the rules, which yield different results).

In the simpler problems of this type, orthodox intuition was powerful enough to invent *ad hockeries* that proved to be equivalent to Bayesian methods with uninformative priors.

In technically more complicated problems where we obtain different results, orthodoxy does not usually formulate its rules completely enough, or apply them far enough, for the aforementioned kind of inconsistency to appear. But it is always lurking just beneath the surface, and sometimes does come into view. The statistical literature — not all Bayesian — contains many examples of the anomalous results that orthodox methods can give in particular cases.

For example, confidence intervals not based on sufficient statistics and not conditioned on ancillary statistics lead to different conclusions from different choices of the statistic; even to grotesquely impossible conclusions, because they ignore cogent information in the sample, that Bayesian methods take into account automatically.

On the other hand, whenever someone has claimed to exhibit an anomaly in Bayesian results, it has turned out that there was an error in the calculation or the Bayesian method was misapplied. Typically, the user has extra information, highly relevant for the inference, that he failed to take into account in the calculation. Venn's polemical attack on Laplace's Rule of Succession, answered by Fisher (1956), is perhaps the classic example.

For Jeffreys' problem of inferring a fixed value, then, it seems to us a long since demonstrated fact — on both the theoretical and pragmatic level — that the orthodox statistician has a great deal to gain in useful results (and as far as we can see, nothing to lose but his ideological chains) by joining the Bayesian camp and finally taking advantage of the powerful tool that Harold Jeffreys created and offered to him 45 years ago. Failing to do this, he faces rapid obsolescence as the new applications of Jeffreys' principles pass far beyond his domain.

4 Now let's consider Barnard's problem

But suppose we do want to infer a frequency distribution for a parameter according to Bernard's scenario; how do our methods compare? The orthodoxian will then allow the use of Bayes' theorem in principle, because he can interpret every probability in it as a frequency. The prior probability can stand only for a frequency in the large class C. But this is almost always unknown in the real problem, and so he can almost never use Bayes' theorem in practice. We do not know of any case where Barnard's scenario has actually been enacted.

The orthodoxian also has a difficulty of principle. For, even if we did know the frequency distribution of the parameter on the large class C, we might not want to use it as a prior. Suppose we also happened to know something more, that did not involve frequencies, pertaining to the present experiment (for example, that because of special circumstances extreme values cannot occur). It appears to us that neither the orthodoxian's ideology nor his procedures would permit him to use that additional information to improve his estimates. Indeed, we suspect that he would not wish to consider the problem at all, because in this particular experiment the parameter would be in his view, "not randomly selected."

The difficulty applies equally well to the sampling distribution. Even if we knew the frequency distribution of samples for all values of the parameter, we might not want to use it as a sampling distribution. We might have knowledge of special circumstances that affect the possible values of data that can be observed in the present experiment (for example, that because of rotation of our spacecraft every third datum is subject to additional error not in the others, but we do not know which ones are thus affected). Surely, common sense will tell us that it would be wrong to analyze the problem as if we did not know this; yet what orthodox; principles would determine, or even justify using, a procedure that takes this into account? To do so would be to admit, with Jeffreys, that in inference a probability stands for more than just a frequency.

We do not contend that such difficulties will arise very often, but only want to point out that attempts to uphold frequency definitions of probability can lead to difficulties of principle whenever we have relevant information that does not consist of frequencies. Bayesian methods can take such information into account easily; such cases make the interesting homework problems for our student.

In our view, orthodox principles could deal with Barnard's scenario satisfactorily only when we have perfect knowledge of frequencies and no other relevant information (i.e. a just the case where Bayesian probabilities are equal to frequencies). That is, when we have the frequency information that the orthodoxian must have but the Bayesian can get along without, but lack the extra information that the Bayesian can use but the orthodoxian cannot. So again, as in the case of confidence intervals, the orthodox procedure will be satisfactory only when our information is such that the orthodox results agree with the Bayesian ones.

In contrast, the Bayesian is prepared to consider this problem in far greater generality and depth, because he can take into account any prior information that can be expressed by a prior distribution, and for this he is prepared to go into deeper and deeper hypothesis spaces. If he lacks knowledge of the frequency distribution of the parameter β in class C he can still assign, perhaps β maximum entropy, a prior that represents whatever partial information he has. If he has additional information about β beyond frequencies, he can take this into account equally well.

Most important of all, the Bayesian has a technical flexibility in that he can solve Barnard's problem, not on the parameter space B consisting of the possible values of β, but on the extension space $B(n) = B \times B \times \cdots \times B$ comprising the jointly possible values $[\beta_1 \ldots \beta_n]$ of β in each of any number of repetitions of the experiment. On this space it is possible to express information about the variability and correlations of β in different experiments — out to and including the limit of perfect correlation where he knows that β is an unknown constant, and the results on $B(n)$ reduce to those of the previous elementary Bayesian analysis on B.

By this means, we can transcend the crudity of supposing that the probability is the frequency, and develop the probability functional giving the relative probabilities of different frequency distributions, in the light of whatever information we have. In image reconstruction, physical chemistry, and geophysics we are now beginning to attack real problems where we have important prior information that can be expressed only on such an extension space.

Failure to appreciate the work of Jeffreys has been very costly to the field of statistics for decades — nearly fatal as far as the ability to participate in new developments is concerned. The non-Bayesian area is being left far behind, as the new applications are taken over instead by younger scientists who would never dream of consulting a statistician for advice, because they understand and use Bayesian analysis as naturally as they use Fourier analysis.

Of course, the field is open-ended, and many more technical advances will be needed in the future as problems become more sophisticated. Yet the Bayesian remedy for the elementary shortcomings of orthodoxy is already highly developed, and is available to anyone.