

BOOK REVIEWS

Bayes's Theorem (Proceedings of the British Academy, Vol. 113) edited by Richard Swinburne. Oxford: Oxford University Press, 2002. 149 pp. \$24.95 (hardback). ISBN 0-19-726249-X.

This book is a collection of papers presented at a symposium held at the British Academy in March of 2001. It contains contributions by the editor and five other authors, and Thomas Bayes's original essay on what we now know as Bayes's Theorem, published in 1763, with an historical introduction by G. A. Barnard. As with many collections of this type, I found the quality of the chapters uneven. Most of the contributors are philosophers, the sole exception being Philip Dawid, who is a well-known Bayesian statistician. In consequence, the papers tend to concentrate on the philosophical aspects of Bayesianism rather than its practical applications.

The introductory first chapter, written by the editor, consists of a competent if uninspired discussion of probability theory and Bayes's Theorem, together with a recap of some of the main points made by the various other contributors. The subjective and objective schools of Bayesianism are highlighted along with a discussion of some apparent paradoxes on priors, which show that the issue of priors is still alive and kicking (more about this when I get to Elliott Sober's article). There is also a discussion of simplicity and Ockham's Razor as it may arise in various contexts.

The second chapter, by Elliott Sober, is intended as a rigorous criticism of the underlying basis of Bayesianism. Sober is himself a likelihoodist: that is, he accepts that inference must obey the Likelihood Principle, which says that all of the information in the data is contained in the likelihood function for the data actually observed. The Likelihood Principle is violated by many commonly used statistical procedures, in particular many of those labeled 'frequentist', but is respected by both likelihood-based procedures and by Bayesian methods. As a card-carrying Bayesian, I found much to disagree with in Sober's essay, so I will spend much of my time discussing it. Many of my disagreements were also considered by Colin Howson, in the next chapter, who does a creditable job of answering many of Sober's objections. The two chapters need to be read together.

The first of Sober's examples is flawed. He asks whether one can have an objective notion of selecting priors. His example is of a square garden which is between 10 and 20 feet on a side (prior information). He points out that if one puts a uniform prior on the length of one of the sides, then this means that the probability is $1/2$ that the length of the garden is between 10 and 15 feet, so that the probability that the area of the garden is between 100 and 225 square feet is $1/2$. But, he notes, if we were to put a uniform prior on the

area itself, we would find that the probability that the area is between 100 and 250 square feet is $1/2$. He asks, rhetorically, which of these is correct, and states that "No satisfactory solution [by objective Bayesian methods] has ever been provided."

This isn't right, for a satisfactory objective Bayesian solution has been known at least since the time of Harold Jeffreys in the 1930s, and perhaps even earlier. In the light of this, I don't know of any experienced Bayesian who would want to put a uniform prior on either the length or the area of the garden if nothing else were known. The reason is that if one is considering quantities like length and area, the natural prior is the Haar prior that is invariant under the natural invariance group of scale changes (i.e., changes in the units one uses to measure length or area). Thus, it should not matter whether we measure length in meters or feet, or area in hectares or acres, we should get the same results. When one takes this constraint into account, one arrives at a prior that is inversely proportional to the length or the area, respectively. An easy calculation shows that such a prior gives identical results, regardless of whether one decides to look at length or area. Thus, I conclude that whatever the merits of Sober's position against objective Bayesian methods, this example fails.

It is astonishing that Sober seems unaware of the extensive Bayesian literature on the selection of objective and "reference" priors, since this is just an elementary application of these well-known ideas.

Sober then considers an alternative: Likelihoodism. He says that sometimes likelihoodism is criticized because it produces absurd results, such as that if you draw the six of spades from a deck of cards, this would support the hypothesis that an evil demon was intent on making us draw the six of spades. The likelihood is unity that we would draw the six of spades if that evil demon hypothesis is correct, but only $1/52$ if the hypothesis that this happened by chance is correct. He then says, "Whatever the merits of this objection, it is not something that a Bayesian should embrace. The reason is that Bayes's theorem tells us that the observation of the six of spades *confirms* the demon hypothesis, in the sense that it raises its probability." But this objection doesn't stand up under scrutiny. It is an example of what Jaynes called the "sure thing" hypothesis, picking one's hypothesis after looking at the data. One of the things that a Bayesian is supposed to do is to consider *all* competing hypotheses *prior* to looking at the data, and unless someone tells me *a priori* that the evil demon likes the six of spades over all other possibilities, I must find myself entertaining alternative hypotheses (e.g., the evil demon prefers the three of clubs or the ace of diamonds, etc.) and before I pick the card I don't know which hypothesis to prefer. This requires me to spread the prior probability under the demon hypothesis over all of the 52 possible outcomes, so that (assuming we did this uniformly) we would find that observing the six of spades, while supporting the "demon prefers six of spades" hypothesis *by itself*, neither supports nor undermines the hypothesis that "a demon made me pick some outcome (which turned out to be the six of spades)." Frankly, be-

tween likelihoodism and Bayesianism as applied to this problem, I prefer the Bayesian approach.

Indeed, I see no way that the likelihoodist can handle the restated problem. Because the likelihoodist refuses to put a prior on the various possible outcomes in the hypothesis "a demon determined the outcome of the experiment," there is no way that he can announce, after the six of spades has been chosen, that the demon hypothesis has been supported.

The demon hypothesis is an example of a compound hypothesis. One of the real drawbacks of the pure likelihoodist approach, as Sober admits, is the difficulty it has in dealing with simple versus compound hypotheses. Oddly, he doesn't seem to recognize that the demon hypothesis is actually one such! To make it a simple hypothesis, one would have to predict *in advance* that the six of spades would be the only outcome if the demon hypothesis is correct. Then, if the six of spades is selected, it would indeed support that hypothesis. But, what is so strange about that? Another hypothesis might be that the six of spades would be forced on you by an expert magician. That is equally in accord with the data, and its prior probability, given what we *know* about expert magicians and *may believe* about demons, is surely far greater than that of the demon hypothesis who forces sixes of spades on unsuspecting victims. So, if one calculates the posterior probabilities of forcing by "magician" versus "demon", given that six of spades was predicted, it is true that observing a six of spades supports the "demon" hypothesis against random chance, but it equally supports "magician" against random chance, so that the posterior odds of "magician" against "demon" are unchanged and remain high.

Sober discusses the important practical problem of evaluating phylogenetic trees based on DNA data. The likelihoodist approach has been popular in solving problems of this sort; however, in recent years the Bayesian approach has been found to have distinct advantages. One is that there are so many possible phylogenetic trees, even under modest amounts of data, that it is impossible in principle to evaluate all the possible trees to find the "maximum likelihood" tree that is best supported by the data. Another is that the maximum likelihood tree, even if it could be found, is probably not the actual tree that resulted in the data, and that there are many other trees that are about as good as the absolute maximum at predicting the data; indeed, there is probably no single "actual" tree, since real biological populations contain a diversity of alleles, many of which will contribute to the following generations. The likelihood landscape for such problems tends to have quite a few isolated peaks. The maximum likelihood approach will incorrectly ignore this population of alternative alleles and peaks as well as nearby "good" but not optimal trees. A better approach would be to use Bayesian methods to explore the entire posterior distribution of tree space, obtaining results of interest by Bayesian model averaging, with models weighted by their posterior probabilities. Computer programs like John Huelsenbeck and Fredrik Ronquist's "MrBayes" have made this approach very popular. This is not to say that

there are not serious issues with priors: there are. But despite these problems, the approach is very promising and has had significant successes.

Sober's final complaint is about simplicity, which he regards as "the Achilles Heel of (Strong) Bayesianism." Frankly, I found his argument strange. He seems to think that Bayesians have some obligation *a priori* to assign higher priors to simpler models (measured by the number of parameters they contain) than to more complex models. I do not see this point. Indeed, a fully Bayesian analysis of simple versus complex models incorporates an automatic "Bayesian Ockham's Razor" that balances simplicity against goodness of fit in a very appealing way.¹ These methods are used daily in practical statistical inference, and incorporate another nice feature: The ability to do *model averaging* when all of the models are mere approximations, and one does not wish either to overfit or underfit the data (e.g., when approximating a physical process by a finite polynomial given a finite amount of data). In such a case, an average over several models, weighted by their posterior probabilities, may be better than settling on a specific "best" model (as with the phylogenetic tree problem).

Sober's solution to the problem of comparing simple and complex models is the Akaike Information Criterion (AIC). But AIC has significant drawbacks: It is only valid asymptotically, and in cases in which there is a large amount of data, it tends to pick excessively large models when the true model is fixed. Given that this is an essay about alleged deficiencies of Bayesian methods, it is odd that Sober does not mention other criteria, such as Schwarz's Bayesian Information Criterion (BIC), a modest modification of the AIC that avoids the problem of picking excessively large models by penalizing them in a way that depends on the amount of data. Surely Sober has reason to prefer AIC over BIC (or similar criteria of note like George's Risk Inflation Criterion), but we don't hear of it.

I found the following article by Colin Howson much more satisfactory. After a brief discussion of early Bayesian theory and its eclipse, he recounts the revival of Bayesian methods in the latter part of the twentieth century. He gives a good account of the Lindley paradox, which demonstrates the deficiencies of the frequentist p-value methods. He then discusses the problems of pure likelihoodism and defends the need for prior information. Howson briefly describes various notions of 'informationless' priors that have been proposed, including a discussion of analogs to the problems that Sober mentioned with "objective" priors. (I would have preferred a term like "reference" prior, since, in fact, no prior comes completely without information. For example, my earlier example for the garden problem used information in group invariance under scale transformations.) He gives a cogent discussion of what simplicity means in selection of priors, and answers one of the comments that Swinburne made regarding simplicity in his introductory essay.

Howson follows this with an introduction to the logical theory that makes Bayesianism, if not inevitable, at least plausible as a form of multivalued logic. This uses the well-known "Dutch Book" argument, which says that if

one is to have a consistent scheme for evaluating the truth of propositions, and is not to be caught out in an inconsistency, then one must use ordinary probability theory, updating one's probabilities *via* Bayes's theorem as new evidence comes in. Howson does not mention the equally persuasive, if mathematically more sophisticated argument by Ed Jaynes (and independently Jack Good) which assumes that ordinary logic applies when propositions are either true or false, that logic on indeterminate propositions (plausibility) must be consistent, and that the plausibility of propositions A and B both being true has to be some (unknown but to be determined) function of the plausibility that proposition B is true, and of the plausibility of A being true, given that B is true; and similarly if one reverses the roles of A and B. Both routes lead to the same probability calculus and to Bayes's theorem.

The next article, by Philip Dawid, is perhaps the most interesting of the lot. Dawid is a practicing statistician, and his article is devoted to the question of how juries should weigh evidence using Bayes's theorem. I believe that even Professor Sober would take relatively little exception to this article, since there are reasonably objective ways to assess the required prior probabilities (given known data on DNA, human populations, and other relevant factors). Dawid notes the problem of the "prosecutor's fallacy," wherein the probability of evidence given guilt is confused with the probability of guilt given evidence. A simple example of this is that if the probability of evidence given guilt is one in a million, but the population under consideration consists of one million individuals, then it is highly likely that there is a second individual in the population to whom the same evidence would apply. Thus, the probability of innocence, given just this evidence, isn't one in a million but more like one in two. The "prosecutor's fallacy" is correctly accounted for by Bayes's theorem. Dawid demonstrates with an actual case how various independent pieces of evidence ought ideally to be combined by a jury to determine innocence or guilt. Unfortunately, the final outcome of this case on appeal demonstrates how much work needs to be done to explain to the courts how Bayesian reasoning can lead to more reliable verdicts in the judicial system. It isn't surprising that many judges and lawyers don't understand basic probabilistic reasoning; neither do many physicians and other professionals.²

John Earman attacks the more daunting problem of the plausibility of actual miracles. He goes back to the examples of David Hume and Richard Price, who were rough contemporaries of Thomas Bayes. There is no question that Price corresponded with both Hume and Bayes, but it is only conjectural that Hume and Bayes were in contact. Earman analyzes Hume's anti-miracle argument from a Bayesian point of view and claims to have shown that from this point of view it is "a shambles." At the same time he "hopes to have given no comfort to the pro-miracle forces." I found Earman's article to be confused and unconvincing. Better articles on Hume and Price have been published.³

The last article, by David Miller, attempts to connect propensities with Bayes's Theorem. I found it thin and of little interest.

In summary, this book has three articles of real interest, by Sober, Howson and Dawid. Sober's article is, in my opinion, deeply flawed, but it is a challenging summary of anti-Bayesian arguments and worth reading in conjunction with Howson's. Dawid's article discusses some important real-world applications of Bayesian reasoning, even if the judicial system is so far unready to make effective use of them.

WILLIAM H. JEFFERYS
Department of Astronomy
University of Texas at Austin
Austin TX 78712
bill@astro.as.utexas.edu

Notes

¹ Jefferys W., & Berger J. (1992) *American Scientist*, 80, 64–72.

² Gigerenzer G. (2002) *Calculated Risks*. New York: Simon and Schuster.

³ Dawid P., & Gillies D. (1989) *The Philosophical Quarterly*, 39, 5765.

The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century by David Salsburg. New York: W. H. Freeman, 2001. xi + 340 pp. \$23.95 (cloth). ISBN 0-7167-4106-7; Owl [Imprint] Henry Holt, 2002. \$16.00 (paper). ISBN 0-8050-7134-2.

The development of statistics as a distinct discipline is described without resort to mathematics, by telling about some of the main actors and what their contributions were. I have been assured by I. J. "Jack" Good that the book is largely sound as to technical matters. (However, some chemists and physicists in particular may balk at the claim that "By the end of the twentieth century, almost all of science had shifted to using statistical models" [p. viii].)

The book makes plain how statistical inference is a matter of exercising judgment rather than a cut-and-dried application of mathematical formulas to convert facts into authoritative answers: "Cause and effect are not so simple to prove, after all" (p. 194). This is something that the conventional wisdom and the popular media seem to not understand. Again, "Logic and probabilistic arguments are incompatible" (p. 300), because logic shows how to make deductions *with certainty*. "The real-life meaning of probability is well established for sample surveys. . . . It is not well established when statistical methods are used for observational studies in astronomy, sociology, epidemiology, law, or weather forecasting. . . . different mathematical models will give rise to different conclusions. If we cannot identify the space of events that generate the probabilities being calculated, then one model is no more valid than another. . . . two expert statisticians working with the same data can disagree