Bayesian Statistics in Radiocarbon Calibration

Daniel Steel†‡
University of Pittsburgh

Critics of Bayesianism often assert that scientists are not Bayesians. The widespread use of Bayesian statistics in the field of radiocarbon calibration is discussed in relation to this charge. This case study illustrates the willingness of scientists to use Bayesian statistics when the approach offers some advantage, while continuing to use orthodox methods in other contexts. The case of radiocarbon calibration, therefore, suggests a picture of statistical practice in science as eclectic and pragmatic rather than rigidly adhering to any one theoretical position.

1. Introduction. According to Deborah Mayo:

Scientists do not proceed to appraise claims by explicit application of Bayesian methods. They do not, for example, report results by reporting their posterior probability assignments to one hypothesis compared with others—even dyed-in-the-wool Bayesians apparently grant this. (1996, 89)

Rather, scientists use what is often dubbed “classical statistics,” which are the sorts of things that one learns about in any introductory statistics course: significance tests, confidence intervals, confidence coefficients, and so on.

The assertion that scientists are not Bayesians is sometimes presented as an objection to the Bayesian point of view. For example, consider these statements taken from a published exchange between Mayo and Colin

†Send requests for reprints to the author, 1017 Cathedral of Learning, University of Pittsburgh, Pittsburgh, PA 15260; email: dpsst19+@pitt.edu.
‡I would like to thank Clark Glymour, Deborah Mayo, John Norton, and John Earman for helpful comments on earlier drafts of this essay.
1. Mayo refers to such methods as “error statistics.”

Copyright 2001 by the Philosophy of Science Association. All rights reserved.
Howson regarding the subjective Bayesian "solution" to the Duhem problem. Mayo asks what answer Howson can give to

the charge that, in practice, it is error statistical methods that are appealed to in checking auxiliaries, distinguishing real effects from artifacts, estimating backgrounds, discriminating different errors and so on in all the tasks called for in reliably pinpointing blame. (Mayo 1997, 323–324; italics added)

That scientists use these non-Bayesian methods, Mayo goes on to say, is an indication that they find them useful. She then asserts:

Faced with this indictment of the Bayesian solution, Howson promotes an alternative hypothesis that saves the Bayesian Way—scientists are just confused by the ambiguities of ordinary language. (1997, 324; italics added)

The words "charge" and "indictment" in these passages leave no reasonable interpretation except that Mayo is objecting to Bayesianism on the grounds that the statistical methods employed in science are not Bayesian.

This objection, as we will shortly see, depends on an appeal to a naturalistic approach to philosophy of science, according to which philosophers should attempt to understand, rather than debunk, standard scientific practices. In what follows, I discuss the scientists-are-not-Bayesians objection in relation to a field, radiocarbon calibration, in which Bayesian methods are not only explicitly advocated by some researchers but are also employed in the standard statistical software package, called CALIB. I argue that radiocarbon calibration presents statistical practice in science as eclectic and pragmatic: eclectic because of a willingness to use, often in closely related contexts, methods based on fundamentally conflicting principles; pragmatic because the choice of method is often motivated by practical concerns (such as computational tractability) rather than by deep theoretical convictions about the aim of statistical inference. I conclude, therefore, that one who accepts a naturalistic view of philosophy of science should adopt a pluralistic theory of scientific inference, rather than rigidly adhering to a single perspective.

2. Naturalism and the Scientists-Are-Not-Bayesians Argument. The assertion that scientists do not use Bayesian methods can be transformed into an objection when linked with the naturalistic premise that philosophies of inductive inference should primarily aim to explicate rather than legislate actual scientific practice. For example, Mayo asserts:

It may seem surprising, given the current climate in philosophy of science, to find philosophers (still) declaring invalid a standard set of
experimental methods rather than trying to understand or explain why scientists evidently (still) find them so useful. I think it is surprising. Is there something special about the philosophy of experimental inference that places it outside of the newer naturalistic attitudes? (1996, 70–71)

Mayo then takes Bayesians to task—particularly Howson and Peter Urbach (1993)—for proceeding as if they were free to ignore the naturalistic view of philosophy of science. She concludes with the statement, “I do not believe that an adequate philosophy of experiment can afford to be at odds with statistical practice in science” (1996, 71).

Naturalism need not be understood to mean that philosophers are never at liberty to judge certain scientific procedures to be misguided. Rather, the idea is that philosophers should refrain from condemning standard scientific procedures that are regarded as acceptable by the relevant experts. Moreover, the naturalistic view should not be confused with the proposal that the solution to all problems of philosophy of science is to “just ask the scientists.” Justifications of standard methods and procedures often are not transparent, even to the scientists themselves. According to the naturalist, it is the job of philosophers of science to identify the rationale underlying such practices.

The naturalistic view of philosophy as understanding science but not legislating it is also taken up by Ronald Giere:

The role of the philosopher, I now think, should be to understand both statistical theory and practice from the perspective of a knowledgeable outsider who is concerned with more general issues regarding the nature of modern science, such as representation, empirical warrant, and the growth of scientific knowledge. (1997, S182; italics in original)

Giere concludes, “It seems to me that Mayo, more than H&U [i.e., Howson and Urbach], embodies this latter understanding of the role of the philosopher of statistics” (1997, S182). Since the scientists-are-not-Bayesians argument has the capacity to sway an informed commentator like Giere, it is worthwhile to examine its cogency.

3. Conflicts of Principle. It will be useful to note briefly the conflicts of principle between Bayesian and classical statistics that are relevant to the present case. Bayesianism is founded on the claim that (ideally) rational people have degrees of belief that can be represented as probabilities, and that scientific inference is a matter of changing these degrees of belief in accordance with Bayes’ theorem as new information is received. Suppose that \( \{H_1, H_2, \ldots \} \) is a set of mutually exclusive and collectively exhaustive
hypotheses, and let $H$ be some member of this set. Then Bayes’ theorem can be stated as follows:

$$P(H \mid E) = \frac{P(H)P(E \mid H)}{\Sigma P(H_i)P(E \mid H_i)}.$$ 

$P(H \mid E)$ is called the posterior probability, $P(H)$ the prior probability, and $P(E \mid H)$ the likelihood. Typically, it is assumed that after the data $E$ has been learned, the posterior probability becomes the new prior probability of $H$. Hence, the probabilities represent the beliefs of (some ideally rational) person at a given time: as new information is acquired, yesterday’s posterior probability becomes today’s prior.

Some points about terminology will be helpful. The prior probability distribution is a function that tells us the value of $P(H_i)$, for any $i$. Likewise, the posterior probability distribution is a function specifying the value of $P(H_i \mid E)$, for any $i$. The likelihood function tells us the value of $P(E \mid H_i)$, for any $i$. Notice that given the prior probability distribution and likelihood function, the posterior distribution can be computed by Bayes’ theorem. One type of prior probability distribution that will be frequently mentioned below is the uniform prior probability distribution. The prior probability distribution is said to be uniform just in case there is some constant $k$ such that $P(H_i) = k$, for all $i$.

The principles underlying classical statistics conflict with the Bayesian approach in a variety of ways. The most obvious difference is that probabilities associated with conclusions (e.g., $p$ values, confidence coefficients) do not represent degrees of belief in hypotheses, but as estimates of the relative frequencies with which a method produces certain types of results. For example, estimates in classical statistics are often expressed via confidence intervals and associated confidence coefficients. In the case of a radiocarbon date, the confidence interval might be $5000 \pm 100$ years before present and the confidence coefficient $95\%$. The confidence coefficient is not, as might be thought, the probability that the age of the item is between $5100$ and $4900$ years before present. Rather, the confidence coefficient tells us something about the method by which the estimate was generated (cf. Lindgren 1993, 282–283). That is, in a large number of repetitions, the method would very likely produce an estimate within $100$ years of the correct date with a relative frequency of approximately $.95$ (cf. Lupton 1993, 57).

Since the case study to be discussed here involves the estimation of ages of items excavated from archeological sites, the classical theory of estimation is the most relevant for our concerns. The classical theory of estimation asserts that that an estimator should be unbiased and efficient. An estimation method is unbiased just in case the expected value of its estimate equals the parameter of interest. Efficiency is a measure of how probable it is that the estimate will be within a given range of its expected
value: the more probable that the estimate will be close to its expected value, the greater the efficiency. The appeal of unbiased and efficient estimators is easy to see, since an estimator that is unbiased and efficient is highly likely to produce an estimate close to the truth. That is, an unbiased and efficient estimator will, if repeated many times, produce an accurate estimate with a high relative frequency, no matter the value of the parameter in question.

Bayesian estimation methods may easily fail to be unbiased in the sense defined above. For example, suppose we have some information suggesting that some values of the parameter are more probable than others. Then according to the Bayesian approach, our prior probability distribution must reflect this information. However, given such a prior distribution, Bayes’ theorem may be a biased estimator in the classical sense and, typically, can be assured of being an unbiased estimator only given a uniform prior probability distribution.

4. Some Basics of Radiocarbon Calibration. Radiocarbon dating is based on the decay of the radioactive isotope carbon 14 ($^{14}$C) that occurs with a half-life of 5,730 years (Campbell and Loy 1996, 26–27). $^{14}$C is continually created in the upper atmosphere as a result of collisions among neutrons and nitrogen atoms, and some of the $^{14}$C atoms join with oxygen to form $^{14}$CO$_2$ molecules (ibid.). Such $^{14}$CO$_2$ molecules are absorbed by plants through the process of photosynthesis, which breaks down the molecule into $^{14}$C, which is used in the construction of cells, and O$_2$, which is released into the atmosphere. Thus, the plant comes to contain a ratio of $^{14}$C to the stable carbon 12 ($^{12}$C) that corresponds to the ratio of $^{14}$CO$_2$ to $^{12}$CO$_2$ in the atmosphere, and this ratio is then passed along through the food chain. Once the plant or animal dies, however, it no longer absorbs additional $^{14}$C, and the proportion of $^{14}$C to $^{12}$C gradually decreases over time as $^{14}$C atoms decay. Given estimates of the half-life of $^{14}$C and the ratio of $^{14}$CO$_2$ to $^{12}$CO$_2$ in the atmosphere, it is then possible to make an estimate of the date at which the plant or animal died.

However, this simple picture is made enormously more complex by the fact that the proportion of atmospheric $^{14}$CO$_2$ to $^{12}$CO$_2$ is known to have fluctuated throughout history and to be sensitive to certain local conditions, such as volcanic eruptions. Consequently, within the last thirty years a great deal of work in the field of radiocarbon dating has been consecrated to an activity known as “calibration,” by which is meant the following process. First, estimate the age of the sample on the assumption that the current atmospheric $^{14}$CO$_2$ to $^{12}$CO$_2$ ratio has remained constant throughout the earth’s history, and call this estimate the “radiocarbon age” of the sample (Stuiver and Pearson 1992, 19). Next, adjust this estimate by taking into account the historical fluctuations of atmospheric $^{14}$CO$_2$ to $^{12}$CO$_2$, and call
the adjusted estimate the "calendar age" of the sample. "Calibration" is the name for the process of deriving the calendar age from the radiocarbon age.

The strategy pursued for calibrating radiocarbon dates has been to find organic materials, which can be radiocarbon dated, and which can be independently dated by some other method. By correlating radiocarbon dates to independently obtained calendar dates, researchers have created a "calibration curve" that allows radiocarbon dates to be properly adjusted. Initially, the calibration curve was constructed from timber samples that could be reliably dated by dendrochronology (i.e., tree-ring dating) (cf. Stuiver and Pearson 1993). More recently, the time range of the calibration curve has been greatly extended through the use of corals, which can be dated on the basis of uranium/thorium decay as well as 14C dating (Stuiver et al. 1998). Currently, the calibration curve stretches from about zero to 24,000 years before present (Stuiver et al. 1998, 1041).

In the simplest case, radiocarbon calibration involves transferring a single radiocarbon date into an estimate of a calendar date, but calibration often involves more complex problems (Buck et al. 1991, 812; Pazdur and Michczynska 1989, 824). For example, it is often desirable to combine several radiocarbon dates into one calendar date estimate. Moreover, prior knowledge sometimes makes it possible to place absolute limits on the range of possible ages of the sample or samples being dated. Another situation involves what is known in archeology as "phasing." Archeological sites often exhibit distinct strata corresponding to a series of occupations of the same location, and the term "phasing" refers to the process of reconstructing these stages. The term "phase" is used to refer to such occupational stages within a site. Naturally, estimating the beginning and ending dates of the phases of sites is a common concern of archeologists. Moreover, archeological excavation typically provides important information on this score, since we may reasonably assume that strata further down in the excavation are older than higher strata.

5. The Calibration Curve and CALIB. The most important feature of the calibration curve, for our purposes, is that it is very "wiggly," which has the result that one radiocarbon date may correspond to several calendar dates. Thus, a probability distribution associated with a calibrated radiocarbon date is usually multi-peaked (cf. Pazdur and Michczynska 1989). This feature creates immediate difficulties for applying the classical theory of estimation, since an estimator would often produce several point estimates of the (unique) calendar date. In such a context, the classical notion of an unbiased estimator makes little sense. The first statistical program designed for radiocarbon calibration that I know of utilized a type of orthodox statistical method known as a maximum likelihood estimator (Orton 1983). However, this program failed to become widely used, pri-
arily because of certain computational difficulties involved in its implementation that apparently arose from the irregular shape of the calibration curve (Naylor and Smith 1988, 591; Buck et al. 1991, 812–813). Mathematical complexities raised by the form of the calibration curve, then, appear to have played an important role in the emergence of Bayesian methods in radiocarbon calibration.

The first version of CALIB was published in 1986 in Radiocarbon, the main journal in the field (Stuiver and Reimer 1986). Since then CALIB has become "[t]he most well-known and commonly used computer program . . . for the calibration of radiocarbon determinations" (Buck et al. 1996, 215). A 1993 special issue of Radiocarbon dedicated to calibration came equipped with a diskette bearing CALIB. A 1998 issue of Radiocarbon (40 (3)), whose centerpiece is the presentation of an updated calibration curve, provides the address to a web-site from which the latest version of CALIB can be downloaded.

Let us consider how CALIB deals with the simplest type of case, in which we wish to infer a calendar date from a single radiocarbon estimate (cf. Buck et al. 1996, 212–215). Here the data is a particular radiocarbon estimate (say, 5000 years BP). The hypotheses assert different possible calendar ages for the item being dated. Hence, we want to compute a posterior probability distribution for these hypotheses given the radiocarbon estimate.

CALIB assumes that the prior distribution is uniform over the calendar dates. The likelihood function is generated in the following way. Suppose that the calendar date of the item in question is \( t \), and let \( \mu(t) \) represent the corresponding radiocarbon date given by the calibration curve. However, the process of generating radiocarbon estimates involves a degree of random error, so the particular estimate produced on a given occasion is likely to differ from the correct value. Let \( X \) be a random variable that ranges over possible values of radiocarbon estimates. It is then assumed that the distribution of \( X \) conditional on \( t \) is normally distributed with a mean of \( \mu(t) \) with a standard deviation of \( \sigma \).

Given the prior distribution and a procedure for generating likelihoods, one might suppose that the posterior probability distribution is directly computed via Bayes' theorem. However, such a direct approach is generally not feasible for practical purposes, so CALIB employs a shortcut computational algorithm that produces an approximation of the posterior distribution.\(^2\) Non-Bayesian methods typically use shortcut computational procedures of their own, and as was noted above, the main obstacle for a

---

2. The computational algorithm of this sort most commonly used for Bayesian applications is the Gibbs sampler. See Casella and George 1992, and Buck et al. 1996, Chapter 8.
maximum likelihood estimator approach to radiocarbon calibration lay in difficulties facing their computational algorithms. Thus, an important cause of the popularity of CALIB was the serendipitous fact that Bayesian computational algorithms could more easily accommodate the complexities raised by the irregular form of the calibration curve.

6. CALIB and the Scientists-Are-Not-Bayesians Argument. The above example shows that summary generalizations to the effect that real scientists are not Bayesians are misleading at best, if not outright false. Nor is the present case an isolated occurrence, since the number of articles published in scientific journals that discuss applications of Bayesian methods has steadily increased within the last decade (Malakoff 1999, 1461). Moreover, CALIB raises a special problem for the scientists-are-not-Bayesians argument, since it is not a procedure recommended by just a few renegade radiocarbon researchers or archeologists. On the contrary, it is the standard statistical tool used for radiocarbon calibration, and it is endorsed by the most prestigious journal in the field of radiocarbon dating. Given the naturalistic perspective of philosophy of science described in Section 2, therefore, it would seem that philosophers ought to attempt to understand, rather than condemn, the use of Bayesian procedures in CALIB.

Let us consider how CALIB conflicts with classical principles. Most obviously, CALIB conflicts with the classical position that probabilities associated with conclusions drawn from statistical data do not represent the degrees of confidence that one may reasonably have in hypotheses. CALIB produces a posterior probability distribution over a set of hypotheses concerning the calendar age of the item being dated. Suppose we generated an interval of .95 probability from such a posterior distribution. The interval and associated probability would assert that a reasonable person could have a degree of confidence of .95 that the true date lies within the interval in question. The probabilities computed by CALIB in no way represent the relative frequency with which the program would produce estimates within a certain range of the correct result if repeated a large number of times. In sum, a posterior probability distribution is not an estimate of efficiency, in the sense of the classical theory of estimation.

I found no mention of the difference between the use of probabilities in CALIB and the classical view of the role of probabilities in statistical analysis. Thus, there is no evidence that the relevant experts in the fields of radiocarbon dating and archeology regard this violation of classical strictures as unacceptable. In defense of the scientists-are-not-Bayesians argument, it could be observed the CALIB manual makes no mention of Bayesian statistics (Buck et al. 1996, 215). Nevertheless, it is not rare for the implicitly Bayesian character of CALIB, and similar programs, to be noted in the literature (cf. Bowman and Leese 1995, 99; Buck et al. 1991,
812; Buck et al. 1996, 215). One review of developments in radiocarbon calibration, for example, refers to the “emergent use of Bayesian statistics” (Bowman and Leese 1995, 102). Anyone who kept up to date in this literature, then, would be aware of the implicitly Bayesian character of CALIB.

CALIB is a program employing Bayesian methods that can be justifiably regarded as an accepted, standard procedure within the field of radiocarbon calibration. Given this example and the growing interest in Bayesian statistical methods in science noted above, a naturalistically inclined philosopher of science would be led to a pluralistic view of scientific inference. According to such a pluralistic view, we should not expect to find a single unified theory capable of illuminating all aspects of inductive inference. Instead, we should accustom ourselves to the persistence of distinct theoretical approaches with sometimes overlapping and sometimes complementary applications. Since some have already embraced a pluralistic view of this kind (cf. Hellman 1997, 320–321), its connection with a naturalistic approach to philosophy of science will not be entirely unwelcome.

7. More Thoroughly Bayesian Approaches. Although CALIB violates classical strictures through assigning probabilities to hypotheses, it is very conservative from a Bayesian perspective owing to its uniform prior probability distribution. However, several researchers have advocated a more thoroughgoing Bayesian approach to calibration in which archeological and historical information is allowed to influence the prior probability distribution (cf. Buck et al. 1991; Buck et al. 1996; Christen et al. 1995; Goslar and Madry 1998). Moreover, some have created explicitly Bayesian calibration programs that can accommodate such information (cf. Ramsey 1995, 1998).

The primary motivation for the development of such procedures has been the traditional Bayesian view that the prior probability distribution should accommodate available information. As Buck et al. (1996, 3) put it: “Bayesian methods permit, indeed demand, that just such information (if it is available) be included.” Christopher Ramsey argues that a uniform prior does not accurately represent the archeologist’s beliefs even in cases in which little is known about the age of the item being dated, since “any traces of living matter are much more likely to be recent than they are to be extremely ancient, all other things being equal” (1998, 470–471). Thus, he recommends as a default prior one that is inversely proportional to the

3. This is an incomplete list. Further examples of thoroughly Bayesian approaches to radiocarbon calibration can be found by doing a keyword search under “Bayesian AND Radiocarbon” in the Social Sciences Citation Index.
age. That is, the prior probability of a hypothesis that asserts the calendar age of the item to be \( t \) is proportional to \( 1/t^2 \) (ibid.).

It will be helpful to briefly consider an example of such a thoroughly Bayesian approach to radiocarbon calibration. In their 1991 essay, "Combining archaeological and radiocarbon information: a Bayesian approach to calibration", Buck et al. generate estimates of the beginning and end dates of two occupational phases of an archeological site by Bayesian means. The items to be dated were taken from either the bottom or top of one of these two phases. Thus, the idea was that by learning the dates of these items, the dates of the beginning and end of the two phases could also be ascertained. Buck et al. construct a prior distribution that takes into account the information that items originating from lower strata are older than those from higher ones. An additional motivation given for combining archeological information in the process of calibrating a radiocarbon date is that it often produces a posterior distribution more densely clustered over a narrow range of dates (Buck et al. 1996, 3–5).

The response to such Bayesian inroads into radiocarbon calibration and archeology in general so far seems mostly positive. For example, I located two reviews (Batt 1997; Scott 1997) of Buck et al.'s (1996) *Bayesian Approach to Interpreting Archaeological Data*, a textbook which recommends a thoroughly Bayesian approach to a broad array of archeological inference problems (including radiocarbon calibration). Both of these reviews were positive. One states that "Bayesian analysis is becoming more and more popular and this book will encourage that interest" (Scott 1997, 219). I found only one negative published reaction to thoroughgoing Bayesian approaches to radiocarbon calibration (Reece 1994), which was specifically directed at the Buck et al. (1991) essay just described. In his critique, Richard Reece argues against combining archeological information with radiocarbon dates in producing calendar date estimates. Primarily, Reece is concerned with the feasibility of representing qualitative archeological information via quantitative probabilities (1994, 849). In sum, though there is a thriving interest in thoroughly Bayesian approaches to radiocarbon calibration, it is not at present a standard method.

Despite the widespread use of CALIB and the increasing interest in more thoroughly Bayesian calibration procedures, classical statistics are still commonly employed in radiocarbon research. To take just one case, it is not rare to see classical methods employed in the construction of the calibration curve. For example, Knox and McFadgen (1997) describe a procedure, based on the method of least squares, for smoothing out some of the curve's wiggles. What is striking in all of this is the readiness of radiocarbon researchers to use classical statistics in one context and Bayesian methods in another, with little or no regard for the conflicting principles underlying the two approaches. The sense of a holy war between
the Bayesians and their classical enemies, so prevalent in the philosophy of science literature, is almost entirely absent.

8. Conclusion. The case described here shows that the charge that scientists-are-not-Bayesians rests on an overly simplistic portrait of statistical practice. Bayesian methods are one of the statistical approaches available to the modern scientist, and the present case demonstrates the willingness of scientists to use them when they perceive some advantage in doing so. Moreover, the case illustrates that scientists can be perfectly comfortable using Bayesian and classical methods side-by-side without paying much attention to conflicts of statistical principle. So the following would be a better a slogan: scientists are eclectic and pragmatic. Given a naturalistic approach to philosophy of science, this conclusion leads to a pluralistic theory of scientific inference. Of course, it is debatable to what extent naturalism is a tenable position. The burden of this essay, however, has been to show that, whatever its merits, naturalism does not generate a clear argument for any single approach to inductive inference.

REFERENCES


