Bayesianism – Its Scope and Limits

Elliott Sober*

1. The Math and the Philosophy (Bayes' Theorem Bayesianism)

Bayes' theorem is a consequence of the definition of conditional probability. However, this way of putting things tends to conceal a proviso that needs to be recognized. What is true is that

$$Pr(H * O) = Pr(O * H)Pr(H)/Pr(O),$$

if each quantity mentioned in the theorem is well-defined.

It is not inevitable that all propositions should have probabilities. That depends on what one means by probability, a point to which I'll return. The claim that all propositions have probabilities is a philosophical doctrine, not a theorem of mathematics. This is where Bayesianism begins and Bayes' theorem leaves off. But there is more to Bayesianism that this. Bayesianism, in its strongest formulation, maintains not just that propositions have probabilities, but that all epistemological concepts that bear on empirical inquiry can be understood in terms of the probabilistic relationships described by Bayes's theorem. Of course, more modest Bayesianisms also can be contemplated.

As an illustration of what Bayesianism amounts to, consider the continuing philosophical puzzlement over the epistemic significance of simplicity. Scientists and philosophers often maintain that simplicity or parsimony is relevant to evaluating the plausibility of hypotheses. The challenge to Bayesianism is to map this informal talk of plausibility onto formal talk of probability. More specifically, a double application of Bayes' theorem yields the following comparative principle:

$$Pr(H1 * O) > Pr(H2 * O) \text{ if and only if } Pr(O * H1)Pr(H1) > Pr(O * H2)Pr(H2).$$

If "more plausible" is interpreted to mean *higher posterior probability*, then there are just two ingredients that Bayesianism gets to use in explaining what makes one hypothesis more plausible than another. This means that if simplicity *does* influence plausibility, it must do so via the prior probabilities or via the likelihoods.[1] If the relevance of simplicity cannot be accommodated in one of these two ways, then either simplicity is epistemically irrelevant or (strong) Bayesianism is mistaken.[2]

2. The Usual Objection – Priors

The standard objection to Bayesianism is to my mind correct. It often does not make sense to talk about propositions' having objective prior probabilities. This is especially clear in the case of hypotheses that attempt to specify laws of nature. Newton's universal law of gravitation, when suitably supplemented with plausible background assumptions, can be said to confer probabilities on observations. But what does it mean to say that the law has a probability in the light of those observations? More puzzling still is the idea that it has a probability before any observations are taken into account. If God chose the laws of nature by drawing slips of paper from an urn, it would make sense to say that Newton's law has an objective prior. But no one believes this process model, and nothing similar seems remotely plausible.

Bayesians used to reply to this challenge by trying to specify a sensible version of the Principle of Indifference. This has turned out to be a dead end. The problem is that there is no unique way to translate ignorance into an assignment of priors. Consider my garden, which is a square plot of land that is between 10 and 20 feet on a side. Based on this information, what is the probability that the garden is between 10 and 15 feet on each side? It might seem natural to say that every length between 10 and 20 feet has the same probability (density), in which case the probability is ½ that each side is between 10 and 15 feet . However, the information I gave you is equivalent to saying that the garden has an area that is between 100 and 400 square feet. This description makes it sound natural to say that every area between 100 and 400 square feet has the same probability, in which case the probability is ½ that the area is between 100 and 250 square feet. However, this entails that the probability is ½ that the square is between 10 and $\sqrt{250} = 15.8$ feet on a side. Applying the principle simultaneously to length and to area leads to contradiction. If the principle is to apply to just one of these, which should it be? No satisfactory answer has ever been provided.

Most contemporary Bayesians have given up on objective Bayesianism, and have gone the subjective route. If Newton's law of gravitation does not have an objective prior probability, perhaps each agent has a subjective degree of belief in that proposition before the evidence starts to roll in. If point values cannot be assigned to these degrees of belief, perhaps they can be said to fall in reasonably well-defined intervals. There are interesting questions to be addressed here concerning this psychological hypothesis. But even supposing that agents have subjective degrees of belief, my problem with subjective Bayesianism is that subjective prior probabilities do not have probative force.

To explain what I mean by this, I want to examine the fairly standard evolutionary idea that the (near) universality of the genetic code is evidence that all organisms now alive trace back to a single common ancestor. Crick (1968) argued that the code now in use is a "frozen accident" – which nucleotide triplet codes for which amino acid is functionally arbitrary. If this is right, it is clear why the universality of the code favors the hypothesis of one common ancestor over the hypothesis that current life traces back to twenty-seven separate start-ups. The evidence discriminates between the two hypotheses in this way because there is a likelihood inequality:

Pr(the code is universal *one common ancestor) >

Pr(the code is universal * 27 mutually unrelated groups).

This reasoning is grounded in objective (if not incontrovertible) considerations about the evolutionary process. What would be added to this if one specified one's subjective degrees of prior belief in the two hypotheses? People may have different feelings here. And even if people have the same feeling, I don't see why that common feeling is epistemologically relevant. If science is about the objective and public evaluation of hypotheses, these subjective feelings do not have scientific standing.[3] When scientists read research papers, they want information about the phenomena under study, not autobiographical remarks about the authors of the study. A report of the author's subjective posterior probabilities blends these two inputs together. This is why it would be better to expunge the subjective element and let the objective likelihoods speak for themselves.

I am not suggesting that we should avoid Bayesian thinking even in the privacy of our own homes. If you have a subjective degree of belief in a hypothesis, by all means use Bayes' theorem to update that degree of belief as you obtain new evidence. For those of us who feel at a loss to say anything about the plausibility that many hypotheses have in the absence of evidence, this is an invitation we will want to decline. However, the most important point is that when opinions clash, the disagreement cannot be resolved by pointing to the fact that different agents have different subjective priors. If the disagreement boils down to this, the agents have simply agreed to disagree.

Consistent with this objection to (strong) Bayesianism, there remains an important domain of scientific problems in which Bayesianism is entirely legitimate. When the hypotheses under consideration describe possible outcomes of a chance process (Hacking 1965, Edwards 1972), it can make perfect sense to talk about objective prior and posterior probabilities. If you draw at random from a standard deck of cards, the probability that you'll draw the six of spades is 1/52. This is a "valid prior", but not because it is obtained *a priori* from some version of the Principle of Indifference, and not because it reports your subjective degree of belief. The prior is legitimate because it is based on empirical information about the process at work.[4] There are many contexts in which Bayesianism has important applications -- medical diagnosis and legal proceedings provide plenty of examples. My point is just that Bayesianism can't be the whole story about scientific inference.

3. The Retreat to Likelihoodism

As the example about the universality of the genetic code suggests, likelihoods are often more objective than prior probabilities. This makes it attractive to regard likelihood as an epistemology unto itself (Edwards 1972, Royall 1997). In doing so, one is changing the question one expects one's epistemology to answer. As Royall points out, likelihoods don't tell you what to believe or how to act or which hypotheses are probably true; they merely tell you how to compare the degree to which the evidence supports the various hypotheses you wish to consider.

Likelihoodism[5] is sometimes criticized for entailing that perfectly absurd hypotheses often have likelihoods that cannot be bettered. If you draw the six of spades from a deck of cards, the hypothesis that this was due to the intervention of an evil demon bent on having you draw that very card has a likelihood of unity, but few of us would regard this hypothesis as very plausible. Doesn't it sound strange to say that your drawing the six of spades supports the demon hypothesis more than it supports the hypothesis that the card was drawn at random from a normal deck? Yet this is precisely what likelihoodism asserts.

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. This is the familiar point that when a hypothesis entails an observation, and the observational outcome was not certain to occur, and the hypothesis's prior probability is neither zero nor one,[6] the observation confirms. It is entirely consistent with this point that the probability of the demon hypothesis remains very low and the normal hypothesis' probability remains very high. But if confirmation concerns the diachronic question of how probabilities *change* rather than the synchronic question of what a probability's *absolute value* is, then Bayesians have to concede that the observation of the six of spades confirms the demon hypothesis. If so, they should not cast a jaundiced eye on the likelihoodist's claim about differential support.

Likelihoodists can and should admit that the demon hypothesis is implausible or absurd, notwithstanding the fact that it has a likelihood of unity relative to the single observation under consideration. It's just that likelihoodists decline to represent this epistemic judgment by assigning the hypothesis a probability. Likelihoodist epistemology is modest in its ambitions; support gets represented formally, but plausibility does not. It thereby contrasts with (strong) Bayesianism, which, as I've explained, aims to characterize *all* genuine epistemic concepts.

There's another Bayesian criticism of likelihoodism that I want to consider. This is the idea that likelihoods have epistemic significance only when they are used in the context of Bayes' theorem. Bayesians sometimes present this claim as if it were obvious, but it isn't. Why is Edwards (1972, p. 100) wrong when he endorses Fisher's (1938) view that likelihood is a "primitive postulate" -- it stands on its own and requires no more ultimate justification? Furthermore, it should be clear that the likelihood principle can be evaluated in the same way that any philosophical explication can be – by seeing if it accords with and systematizes intuitions about examples. Finally, I should add that there are nonBayesian inferential frameworks in which likelihood plays a prominent role (Forster and Sober 2002); I'll mention one of these at the end of this paper.

4. The Trouble with Likelihoods

Likelihoodism's objection to Bayesianism comes back to haunt it when one considers composite hypotheses. A simple statistical hypothesis confers a sharp probability on each possible

observation. Composite hypotheses do not. When composite hypotheses have objective likelihoods, their values are often unknown; often they cannot be said to have objective likelihoods at all.

Composite hypotheses are frequently disjunctions – whether finite or infinite – of simple hypotheses. When this is so, their likelihoods are *averages*. Consider, for example, how the standard Mendelian model of reproduction may be used to compute the likelihoods of different hypotheses about the genotypes of an organism's parents, given an observation of the offspring's genotype:

Pr(Offspring is Aa *Parents are AA and AA) = 0
Pr(Offspring is Aa *Parents are AA and Aa) = ½
Pr(Offspring is Aa *Parents are AA and aa) = 1.0
Pr(Offspring is Aa *Parents are Aa and Aa) = ½
Pr(Offspring is Aa *Parents are Aa and aa) = ½
Pr(Offspring is Aa *Parents are aa and aa) = 0.

Hypotheses about the genotype of the parental pair are simple. However, the hypothesis (H1) that the offspring's mother is AA is composite; its likelihood is an average:

Pr(Offspring is Aa *Mother is AA) =

Pr(Offspring is Aa *Mother is AA & Father is AA)Pr(Father is AA *Mother is AA) +
Pr(Offspring is Aa *Mother is AA & Father is Aa)Pr(Father is Aa *Mother is AA) +
Pr(Offspring is Aa *Mother is AA & Father is aa)Pr(Father is aa *Mother is AA) =

$(0)w_1 + (½)w_2 + (1)w_3$ (where $w_1 + w_2 + w_3 = 1.0$).

The hypothesis H1 that the mother was AA is a disjunction of simple hypotheses – either she was AA and the father was AA, *or* she was AA and the father was Aa, *or* she was AA and the father was aa. The same point holds with respect to the hypothesis (H2) that the mother was a heterozygote. However, in this case the relevant simple hypotheses about the parental pair confer the same probability on the observation, namely ½.. This means that H2's likelihood is ½. The likelihoods of the two hypotheses, as a function of the father's genotype, are depicted in Figure 1.

Figure 1

Which hypothesis, H1or H2, has the higher likelihood? To answer this question, we must evaluate the likelihood of H1, but to do this we have to know the values of the weighting terms w_1, w_2, and w_3, which represent the properties of the mating scheme by which males and females in the parental generation came together to reproduce. If one had empirical information about whether mating was random or assortative (and to what degree), there would be no problem. But in the absence of such information, it is hard to see how values for these weighting terms can be specified unless one regards

them as reflecting one's subjective degrees of belief.

In this example the weighting terms are "nuisance parameters." Our interest is in inferring the mother's genotype, but the father's genotype gets in the way. One solution that is sometimes employed in science is to estimate the values of nuisance parameters by finding the values for those parameters that maximize the likelihood of the composite hypothesis in question. In the case at hand, it is easy to see that $\omega_1 = 0$, $\omega_2 = 0$, and $\omega_3 = 1.0$ are the values that maximize the likelihood of H1. If we assume that AA mothers always pair with aa fathers, we can conclude that H1 has a likelihood of unity and so is more likely than H2.

Although this procedure for handling nuisance parameters may seem to solve our problem, it does not. Rather, we have merely changed the subject. We have not compared H1 and H2; instead, we have compared L(H1) and H2, where L(H1) specifies a specific set of values for the nuisance parameters in H1. The likelihood of L(H1) is greater than the likelihood of H1, not equal to it. H1 is composite, but L(H1) is simple. Instead of assessing the *average* likelihood of the composite hypothesis with which we began, we have evaluated the likelihood of its likeliest special case.[7]

This example is artificial, but it illustrates a genuine issue that arises in scientific practice. As a more realistic example, consider the idea that the likelihood concept can be used to discriminate among competing phylogenetic hypotheses (first proposed by Edwards and Cavalli-Sforza 1964; recently reviewed by Lewis 1998). Suppose we are trying to ascertain the phylogenetic relationships that connect species X, Y, and Z and that our data consist of the characteristics these species are observed to exhibit. There are three possible phylogenetic trees
-- (XY)Z, X(YZ), and (XZ)Y. The first of these possibilities is depicted in the accompanying figure. The tips of the tree represent species that exist now; interior nodes represent common ancestors. The (XY)Z tree says that X and Y have a common ancestor that is not an ancestor of Z.

Figure 2

How can data on the similarities and differences these species exhibit be used to decide which genealogical hypothesis is best supported? The likelihood approach is to find the tree that confers the highest probability on the data. The problem, however, is that a phylogeny, by itself, does not tell us how probable it is that the three species should have the characteristics we observe. What is needed in addition is a quantitative model of how different traits evolve in different lineages. For example, the value of Pr[Data * (XY)Z] depends on the rules of evolution that each trait obeys in the four branches depicted in the accompanying figure and on the characteristics that the root species possesses. The likelihood of the tree topology is an average over the different possible values that these nuisance parameters can take:

$$\Pr[\text{Data} * (XY)Z] = \sum_i \Pr[\text{Data} * (XY)Z \,\&\, N_i]\Pr(N_i * (XY)Z].$$

The hypothesis (XY)Z is composite – it is an infinite disjunction in which each disjunct consists of the topology (XY)Z with the nuisance parameters fixed at a particular set of values. Biologists who use maximum likelihood typically estimate the values of nuisance parameters from the data. Thus, instead of comparing the likelihoods of (XY)Z, X(YZ), and (XZ)Y, they compare the likelihoods of L[(XY)Z], L[X(YZ)], and L[(XZ)Y]. The problem has been changed from one that is intractable to one that can be solved.

A solution to this problem that is truer to the tenets of likelihoodism is to identify regions of parameter space where the likelihoods of the phylogenetic hypotheses have one ordering, and other regions where they have another. Perhaps when the nuisance parameters fall in one region, (XY)Z is the hypothesis that makes the data most probable, whereas when the nuisance parameters fall in a different region, X(YZ) is the hypothesis of maximum likelihood. This conditional assessment does not have the finality of the unqualified conclusion that (XY)Z is the maximum likelihood hypothesis. Rather, it points to further biological questions that must be answered if we wish to say more (Sober 1988).

The problem I have just surveyed – that composite hypotheses often do not have known objective likelihoods – is not a problem for likelihoodism, if that philosophy is sufficiently modest, but it *is* a problem for Bayesianism, if that philosophy is sufficiently *im*modest. Modest likelihoodists can and do admit that the likelihoods of composite hypotheses often cannot be evaluated. However, Bayesians who want their epistemology to provide a complete account of scientific inference here confront a second problem. The difficulty with priors also attaches to likelihoods.[8]

5. A Further Problem for the Best-Case Strategy of Dealing with Nuisance Parameters

Although Bayesians and likelihoodists should not confuse the likelihood of a composite hypothesis with the likelihood of its likeliest special case, it is worth exploring a further difficulty that arises if one changes the subject in this way. Let's return to the topology depicted in the figure and consider what a fully realistic model of evolution in that tree will look like. We will want to have nine nuisance parameters for each dichotomous character. Eight of these are branch transition probabilities; the ninth assigns a probability to the root species' occupying character state 0. This model acknowledges that there can be *between-trait and within-trait heterogeneity*;
different traits can evolve according to different rules, and the rules that a trait follows on one branch may differ from the rules that it follows on another. The model says that heterogeneities are possible, but it does not demand that they be actual. Two different parameters may have different values, but they also can have the same value. In addition to these nine parameters for each trait, a fully realistic model will also need parameters that represents the degree of independence with which each pair of traits evolves in each branch. The presence of these parameters in our model does not commit us to saying that traits are correlated in their evolution,
but merely says that this is possible.

The complex model I have just described is realistic, but it has an embarrassing consequence – if we use this model and deploy the best-case solution to the problem of nuisance parameters, the result is that all phylogenies have a likelihood of unity. For example, consider a characteristic in which X is in state 1, Y is in state 1, and Z is in state 0; it is easy to find values for the nine nuisance parameters that pertain to this character's evolution that entail that this distribution of characters has a probability of unity. When we move to another character that has a different distribution, we can do the same thing. Other topologies are no different. What this means is that it is impossible to discriminate among phylogenetic hypotheses if we use the best case strategy in the context of a fully realistic model of the evolutionary process.

This has not stopped evolutionary biology in its tracks. Rather, biologists assign likelihoods to tree topologies by using constrained models. These constraints fall into two categories:

> *Within trait constraints*: The strongest version of this idea is that a trait's probability of changing in a unit of time is the same everywhere in the tree. A weaker constraint is that a trait's probability of changing in a unit of time is the same at any two simultaneous temporal intervals. In terms of the topology depicted in Figure 2, the latter constraint entails that $e_1 = e_2$, but says nothing about the relationship of e_1 and e_3; the former says that $e_1 = e_2$ *and* that $e_1 = e_3$ if lineages 1 and 3 have the same durations.
>
> *Between trait constraints*: Traits evolve independently of each other, and different traits in a lineage follow the same rules of evolution.

This second category of constraints has been applied both globally, to all traits in the data, and locally, within classes of traits.

Although these constraints save the problem of phylogenetic inference from collapsing under its own weight, they are manifestly unrealistic. Do we really believe that a trait's rules of evolution are exactly the same in different lineages? Do we really believe that different traits follow exactly the same rules of evolution? These are idealizations.[9] What is clearly true is the unconstrained model, which says that different traits may or may not evolve independently, that they may or may not follow the same rules of evolution, and that a given trait may or may not follow the same rules in different lineages. We thus have arrived at a dilemma: we can use a realistic model and give up on the idea of inferring phylogenies by best-case maximum likelihood, or we can use a constrained model to infer phylogenies, but leave our inference vulnerable to the charge that the model used is not realistic.

As noted earlier, the best-case strategy for dealing with nuisance parameters does not conform to the dictates of likelihoodism, which is perfectly clear on the difference between the average likelihood of a composite hypothesis H and the likelihood of the simple hypothesis L(H). However, I doubt that this point will give pause to biologists who use this strategy to reconstruct phylogenies. This is because these biologists are frequentists, not likelihoodists; they use the frequentist

technique known as the likelihood ratio test.[10] This test prevents the collapse I have just described. Instead of automatically opting for hypotheses of high likelihood, they ask whether more complex hypotheses have likelihoods that are *significantly* greater than the likelihoods of simpler hypotheses.

Not that all is well if one embraces frequentism to solve this problem. In addition to the conceptual objections that Bayesians and likelihoodists have developed against frequentism, frequentism has a practical limitation in the problem at hand – the frequentist's likelihood ratio test applies only to nested hypotheses. This can be illustrated by considering our three competing tree topologies and some of the process models already described. Consider a data set that describes the character states of each species for ten dichotomous characters. The accompanying table represents the different best case hypotheses that are generated by bringing a tree topology together with a process model and then finding the maximum likelihood estimate of the nuisance parameters.[11] Models in the same column are nested, but entries in different columns are not.[12] Frequentism has no way to implement diagonal comparisons.

##########		tree topologies		
		(XY)Z	X(YZ)	(XZ)Y
process models	All characters evolve independently (N90).	L[(XY)Z & N90]	L[X(YZ) & N90]	L[(XZ)Y & N90]
	All characters evolve independently and follow the same rules (N9).	L[(XY)Z & N9]	L[X(YZ) & N9]	L[(XZ)Y & N9]
	All characters evolve independently, follow the same rules, and evolve at a constant rate (N5).	L[(XY)Z & N5]	L[X(YZ) & N5]	L[(XZ)Y & N5]

6. Simplicity – the Achilles Heel of (Strong) Bayesianism

In the list of process models displayed in the table, N5 is simpler than N9, and N9 is simpler than N90. Don't be misled by the lengths of the verbal descriptions; the relevant consideration is the number of adjustable parameters. Model N90 has 90 nuisance parameters – nine for each of the ten character distributions in the data set. Because it has so many parameters, this model is able to leave open whether different traits evolve according to the same or different rules, and also whether a given

trait follows the same rules in different branches. In the table's list of models, simpler models entail models that are more complex. This is the kind of situation that Popper (1959) was thinking about when he equated simplicity with falsifiability.[13] As Popper observed, the epistemic relevance of simplicity in this instance cannot be captured by stipulating that simpler theories are more probable. If N5 entails N90, N5 cannot be more probable than N90.

Howson (1988) correctly notes that there is no logical prohibition against assigning simpler models higher priors when models are *not* nested. However, Popper's point remains true for nested models. What are we to conclude? It may seem that the question that needs to be addressed is whether scientists should compare nested models. As noted before, they in fact do so; scientists are often frequentists and the likelihood ratio test *requires* that models be nested. Bayesians may reply that this is a confusion from which scientists need to emancipate themselves. Since nested models are not incompatible, why should we regard them as competitors? I'll return to this question in a while; the point I want to emphasize here is that the insistence on non-nested models does not pluck the Bayesian fat from the fire. This is because it is obscure what justification the Bayesian can offer for assigning simpler models higher priors. For example, suppose we reformulate N90 so that N9 is no longer nested in it – let N90* use different parameters for different traits with the stipulation that different traits cannot have exactly the same branch transition probabilities. What justification could there be for assigning N9 a higher prior than N90*? If a gun were put to my head, I'd allow dimensionality considerations to lead me to bet on just the opposite judgment -- I'd say that it is more probable that two traits have different probabilities of evolving than that they have exactly the same probabilities of evolving.

If Bayesianism can't capture the epistemic relevance of simplicity by defending an objective ordering of prior probabilities, the other possibility is that it might be able to explain the relevance of simplicity via the vehicle of likelihoods. However, we have just seen that serious difficulties stand in the way of this undertaking. The likelihoods of composite hypotheses often cannot be evaluated objectively. And if we change the subject by using the best case strategy, the problem is that simpler models inevitably come out with lower likelihoods, not higher ones.[14]

I mentioned at the outset that Bayesianism has just two resources for explaining the epistemic relevance of simplicity – priors and likelihoods. Neither of these appears to be at all promising. Does it follow that Bayesianism is mistaken? There is a way out to consider – perhaps one should deny that simplicity has any epistemic relevance at all. Perhaps simplicity is merely an aesthetic frill. Scientists *like* simpler theories for various reasons, but that does not mean that simplicity is epistemically significant.

I think this way out is blocked for two reasons. First, the practice of science makes it very hard to believe that simplicity always counts for nothing. It does no good for the Bayesian to point to examples of scientific inference in which simplicity plays no role. Granted, there are such cases. However, in biology and the social sciences, scientists frequently compare models that contain different

numbers of adjustable parameters. Simplicity is central to science because model selection is a pervasive problem.

The second reason not to deny the epistemic import of simplicity is that there exists an inferential framework that is neither Bayesian, nor likelihoodist, nor frequentist, which entails that simplicity is epistemically relevant and explains why this is so. This is the model selection framework and criterion developed by H. Akaike (1973) and his school (see Sakamoto *et al.* 1986). It turns out that the simplicity of a model, when measured in terms of the number of adjustable parameters it contains, is relevant to estimating how predictively accurate the model will be. Akaike's framework and criterion are nonBayesian, in that no prior probabilities are invoked. However, the criterion for model selection that Akaike derives does say that the likelihood of a model's likeliest special case is relevant to estimating the model's predictive accuracy. Akaike (1973) describes his proposal as an "extension" of the method of maximum likelihood. Likelihood is relevant to estimating predictive accuracy, but it is not the only thing that is relevant; simplicity is relevant too.[15]

Since this paper is about Bayesianism, not the work of Akaike, I won't try to explain in any detail how these ideas work. However, I will make a few brief comments, which I hope will whet the reader's appetite. Akaike suggested that the problem of model selection be conceived in terms of a certain goal; the goal is not to find models that are true, but models that will be predictively accurate. This conception of the goal of model selection is what I mean by Akaike's "framework." Akaike also proposed a means for achieving that goal; he proved a theorem that describes how one can obtain an unbiased estimate of a model's predictive accuracy. This theorem is the basis for what has come to be called the Akaike information criterion (AIC). This separation of Akaike's framework from his criterion is important; there may be circumstances in which AIC is *not* the best criterion to use in model selection, even granting the goal of maximizing predictive accuracy. The model selection literature contains a good deal of discussion of this point. My own view is not that AIC is the be-all and end-all; what I find philosophically interesting is the Akaike framework and a certain feature that many model selection criteria share – that simplicity is relevant because it helps one estimate predictive accuracy.

Akaike's idea of predictive accuracy has to be understood in terms of a two-step process. Models that contain adjustable parameters make predictions in the following sense: first one draws a set of data from the underlying distribution and uses that data to estimate the values of the model's parameters (by maximum likelihood estimation). One then uses that fitted model to predict a new data set drawn from the same distribution. In terms of our previous notation, we use a model M to make a prediction about new data by using the old data to find L(M) -- it is L(M) that makes a definite prediction. The predicted values may be close to the new data, or far away (as measured by the Kulback-Leibler distance measure). Imagine using the model repeatedly in this two-step process; there will doubtless be some variation among these repetitions in terms of how well the fitted model predicts new data. The *average* performance of the model is what defines its predictive accuracy. The predictive accuracy of M is the *expected* likelihood of L(M).

Having models that are predictively accurate may be a desirable goal, but how can one tell how predictively accurate a model is apt to be? That is, is predictive accuracy epistemically accessible? Akaike's (1973) remarkable theorem provides an answer:

An unbiased estimate of the predictive accuracy of model M . Log-likelihood[L(M)] - k.

One takes the logarithm of the likelihood of the fitted model and subtracts k, the number of adjustable parameters.[16] Complex models, when fitted to the data, tend to have higher likelihoods than simpler ones, but they also incur a larger penalty because of their complexity. For a complex model to have a higher AIC value than a simpler one, it isn't enough that the complex model fit the data better; it must fit the data better by a sufficient margin to overcome the fact that it is more complex.

There is more to the Akaike framework than Akaike's theorem. For example, even though AIC provides an unbiased estimate of a model's predictive accuracy, one may want to know how much error there is in this estimate. Sakamoto *et al.* (1986) describe a theorem that addresses this question (see Forster and Sober 1994). The model selection literature explores this and other properties of AIC and other model selection criteria. In addition, Akaike's concept of predictive accuracy needs to be supplemented. Akaike described what Forster (2002) calls *interpolative* predictive accuracy; the concept of *extrapolative* predictive accuracy has interestingly different properties.

Notice that it doesn't matter to the Akaike framework or to AIC whether the models one considers are nested or non-nested. Comparing nested models makes sense because nested models can make different predictions when fitted to the data. It *does* seem strange to compare nested models if the goal is to discover which model is true. Since nested models are not in conflict, why does one have to choose? It is here that the Akaike framework is fundamental. Bayesians typically see truth as the goal of inference – the point of evaluating data is to say which of the competing theories one has formulated has the highest probability of being true. When predictive accuracy is substituted for truth as the goal of inference, the epistemological landscape undergoes a fundamental change.

7. Conclusion

Objective Bayesianism has its place and so does subjective Bayesianism. By "objective Bayesianism," I don't mean a Bayesianism based on the principle of indifference (how could that be objective?), but one in which priors are objectively justified by a plausible account of a chance process. When I say that subjective Bayesianism has its place, I mean that agents who have degrees of belief in a proposition should use Bayes's Theorem to update. However, these two arenas for Bayesianism leave a large void in the theory of scientific inference. Many of the hypotheses of interest to science do not have objective prior probabilities. In addition, there are many composite hypotheses for which objective likelihoods cannot be provided. The reaction to these exigencies should not be a retreat to

subjective Bayesianism. This is because it is doubtful that people always have subjective degrees of belief in hypotheses before they have any evidence in hand. But more importantly, the scientific enterprise aims to separate objective evidence from subjective preconception.

These problems come vividly into focus when they are brought to bear on the question of why simplicity matters in scientific inference. Bayesians have two resources to use in framing an answer. They can argue that simpler theories have higher prior probabilities or that they have higher likelihoods. When models are nested, it is impossible for the simpler model to have the higher prior (or posterior). For non-nested models, there is no logical contradiction in assigning simpler models higher priors, but what could justify that assignment? It is not enough that one *has* various prior degrees of belief. The question is why those assignments are right and others are wrong. Hopes for a likelihood account of the role of simplicity are likewise dim.[17] Models containing adjustable parameters are composite, and it often is obscure how the likelihoods of composite hypotheses can be compared objectively. One might be tempted to solve this problem by using the best-case strategy. However, this renders simpler hypotheses less likely, not more so.

Is it plausible to think that these problems for Bayesianism will be solved with more time and hard work? I tend to regard them as permanent and intractable. In contemplating the prospects for progress in this research program, it is worth considering the fact that simplicity is not a puzzlement in the Akaike framework; rather, its justification is patent. There are no prior probabilities here, and the problem of evaluating the likelihoods of composite hypotheses does not arise. I don't want to suggest that this newer framework is a paradise free of conceptual puzzles. But it is a framework well worth exploring, in view of Bayesianism's scope and limits.

Figure 1 Caption: The likelihoods of two hypotheses about the mother's genotype, relative to the observation that the offspring's genotype is Aa. H2 say that the mother was Aa; this hypothesis confers the same probability on the observation, regardless of what the father's genotype was. H1 says that the mother was AA; what probability this hypothesis confers on the observation depends on the father's (unknown) genotype.

Figure 2 Caption: In this phylogenetic tree, there are nine nuisance parameters for each dichotomous character (whose two possible states are 0 and 1). For each lineage i (i=1,2,3,4), e_i = Pr(lineage i ends in state 1 * lineage i begins in state 0) and r_i = Pr(lineage i ends in state 0 * lineage i begins in state 1). In addition, there is one parameter that describes the state of the root – Pr(R is in state 0).

References

Akaike, H. (1973): "Information Theory as an Extension of the Maximum Likelihood Principle." In B. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*. Budapest:

Akademiai Kiado, pp. 267-281.

Burnham, K. and Anderson, D. (1998): *Model Selection and Inference – a Practical Information-Theoretic Approach*. New York: Springer.

Crick, F. (1968): "The Origin of the Genetic Code." *Journal of Molecular Biology* 38: 367-379.

Edwards, A. (1972): *Likelihood*. Cambridge: Cambridge University Press.

Edwards, A. and Cavalli-Sforza, L. (1964): "Reconstruction of Evolutionary Trees." In V. Heywood and J. McNeill (eds.), *Phenetic and Phylogenetic Classification*. New York Systematics Association Publication No. 6, pp. 67-76.

Fisher, R. (1938): "A Comment on H. Jeffreys's 'Maximum Likelihood, Inverse Probability, and the Method of Moments.'" *Annals of Eugenics* 8: 146-151.

Forster, M. (1995): "Bayes and Bust – Simplicity as a Problem for a Probabilist's Approach to Confirmation." *British Journal for the Philosophy of Science* 46: 399-429.

Forster, M. (2000): "Hard Problems in the Philosophy of Science – Idealisation and Commensurability." In R. Nola and H. Sankey (eds.), *After Popper, Kuhn, and Feyerabend.* London: Kluwer, pp. 231-250.

Forster, M. (2002): "In Defense of the Predictive Accuracy Framework." *PSA 2000 – Proceedings of the Philosophy of Science Association*.

Forster, M. and Sober, E. (1994): "How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions." *British Journal for the Philosophy of Science* 45: 1-36

Forster, M. and Sober, E. (2002): "Why Likelihood?" In M. Taper and S. Lee (eds.), *The Nature of Scientific Evidence*, Chicago: University of Chicago Press.

Howson, C. (1988): "On the Consistency of Jeffreys's Simplicity Postulate and its Role in Bayesian Inference." *Philosophical Quarterly* 38: 68-83.

Kalbfleisch, J. and Sprott, D. (1970): "Application of Likelihood Methods to Models Involving Large Numbers of Parameters (with discussion)." *Journal of the Royal Statistical Society B* 32: 175-208.

Lewis, P. (1998): "Maximum Likelihood as an Alternative to Parsimony for Inferring Phylogeny Using Nucleotide Sequence Data." In D. Soltis, P. Soltis, and J. Doyle (eds.), *Molecular Systematics of*

Plants II. Boston: Kluwer, pp. 132-163.

McQuarrie, A. and Tsai, C. (1998): *Regression and Time Series Model Selection.* Singapore: World Scientific.

Popper, K. (1959): *Logic of Scientific Discovery.* London: Hutchinson.

Royall, R. (1997): *Statistical Evidence – a Likelihood Paradigm*. Boca Raton: Chapman and Hall.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986): *Akaike Information Criterion Statistics*. New York: Springer.

Schwarz, G. (1978): "Estimating the Dimension of a Model." *Annals of Statistics* **6**: 461-465.

Sober, E. (1988): *Reconstructing the Past – Parsimony, Evolution, and Inference.* Cambridge: MIT Press.

Sober, E. (1990): "Let's Razor Ockham's Razor." In D. Knowles (ed.), *Explanation and Its Limits*, Royal Institute of Philosophy Supplementary Volume 27, Cambridge University Press, pp. 73-94.

Sober, E. (1998): "Instrumentalism Revisited." *Critica* 31: 3-38.

Sober, E. (1999): "Modus Darwin." *Biology and Philosophy* 14: 253-278.

Sober, E. (2002a): "Instrumentalism, Parsimony, and the Akaike Framework." *PSA 2000 – Proceedings of the Philosophy of Science Association*.

Sober, E. (2002b): "Reconstructing the Character States of Ancestors – A Likelihood Perspective on Cladistic Parsimony." *The Monist*

Notes

*. I thank Martin Barrett, Ellery Eells, Branden Fitelson, Richard Royall, and Michael Steel for useful discussion.

1. In this paper I use the terms "likelihood" and "likely" in the technical sense introduced by R.A. Fisher – the likelihood of a hypothesis H in the light of observations O is the probability that H confers on H, not the probability that O confers on H. H's likelihood is Pr(O * H), while its probability is Pr(H * O).

2.Bayesians need not argue that simplicity is always relevant just by way of influencing priors, or that it is always relevant just by way of influencing likelihoods. See Sober (1990) for discussion.

3. It would be a different matter if one had an empirically well-confirmed theory that allowed one to say how often life can be expected to emerge from nonlife in various environments, and how often whole phylogenetic trees can be expected to go extinct. A process theory of this kind would provide an objective basis for the prior probabilities. See Sober (1999) for discussion.

4. The prior probability is properly so-called, not because it is *a priori* (it is not), but because it is in place prior to one's taking the new evidence into account.

5. By likelihoodism, I mean the comparative principle that O supports H1 more than O supports H2 if and only if Pr(O * H1) > Pr(O * H2). It is a further claim that *degree* of differential support is measured by the likelihood *ratio*. Formulations of the Likelihood Principle often combine these two elements; see Forster and Sober (2002) for discussion.

6. Since Bayesians usually reserve priors of 0 and 1 for tautologies and contradictions, I take it that they will want to assign the demon hypothesis an intermediate prior probability.

7. This "best-case procedure" is discussed by Kalbfleisch and Sprott (1970), by Edwards (1972, pp. 109-119), and by Royall (1998, pp. 158-159).

8. There is a Bayesian proposal for evaluating the average likelihoods of composite hypotheses. This is Schwarz's (1978) Bayesian information criterion (BIC). This approach imposes a flat distribution on parameter values that are near the data and a probability of zero on values that are far away; in addition, it renders commensurable the average likelihoods of composite hypotheses containing different adjustable parameters by introducing stipulations that fail to be invariant under reparameterization. See Forster and Sober (1994, pp. 23-24) for discussion.

9. For discussion of the relationship between idealization and simplification, see Sober (1998, 2002) and Forster (2000, 2002).

10. Don't be misled by the terminology – the likelihood ratio test is not consistent with likelihoodism.

11. In this table, I've written, for example, "L[(XY)Z & N9]" and not "(XY)Z & L[N(9)]." The reason is that conjunctions in the same row often have their nuisance parameters set at different values, depending on the phylogeny to which they are attached.

12. Although the conjunctions in this table that include the same phylogeny are nested, it is perfectly possible for two such conjunctions to be non-nested (e.g., let one assume that there is between-trait homogeneity and leave open whether there is within-trait homogeneity and let the other do the opposite).

13. Popper, of course, realized that simplicity is sometimes not epistemically relevant; he introduced his equation to explain what makes simplicity epistemically relevant when it is so in fact.

14. For further discussion of Bayesianism's and likelihoodism's inadequate treatment of simplicity, see Forster and Sober (1994) and Forster (1995).

15. For further discussion of Akaike's framework and theorem, see Burnham and Anderson (1998), McQuarrie and Tsai (1998), Forster and Sober (1994, 2002), Forster (2002), and Sober (2002a).

16. More exactly, the formulation of Akaike's result that Forster and Sober (1994) and Forster (2002) recommend is that an unbiased estimate of the model's predictive accuracy *per datum* is (1/N){Log-likelihood[L(M)] - k}, where N is the number of data.

17. There is a special case in which I think this pessimism is misplaced. Cladistic parsimony is a method of inference used in phylogeny reconstruction. I suspect that this method makes sense to the extent that it reflects likelihood considerations; see Sober (1988, 2002b) for discussion.