Why Likelihood?

Malcolm Forster and Elliott Sober*
Philosophy Department
University of Wisconsin, Madison

March 8, 2001

ABSTRACT: The Likelihood Principle has been defended on Bayesian grounds, on the grounds that it coincides with and systematizes intuitive judgments about example problems, and by appeal to the fact that it generalizes what is true when hypotheses have deductive consequences about observations. Here we divide the Principle into two parts -- one qualitative, the other quantitative -- and evaluate each in the light of the Akaike information criterion. Both turn out to be correct in a special case (when the competing hypotheses have the same number of adjustable parameters), but not otherwise.

KEYWORDS: Akaike, AIC, Bayesian, evidence, hypothesis testing, likelihood, model selection, statistical inference.

---

* We would like to thank Ken Burnham, Ellery Eells, Branden Fitelson, Ilkka Kieseppä, Richard Royall, and the editors of this volume for helpful comments on an earlier draft.

Mark Anthony said that he came to bury Caesar, not to praise him. In contrast, our goal in connection with the likelihood concept is neither to bury likelihood nor to praise it. Instead of praising the concept, we will present what we think is an important criticism. However, the upshot of this criticism is not the conclusion that likelihood should be buried, but paradoxically, a justification of likelihood, properly understood.

Before we get to our criticism of likelihood, we should say that we agree with the criticisms that likelihoodists have made of Neyman-Pearson-Fisher statistics and of Bayesianism (Edwards 1987, Royall 1997). In our opinion, likelihood looks very good indeed when it is compared with these alternatives. However, the question remains of what can be said in defense of likelihood that is positive. Royall begins his excellent book with three lines of justification.

He points out, first, that the likelihood principle makes intuitive sense when probabilities are all 1's and 0's. If the hypothesis H_1 says that what we observe *must* occur, and the hypothesis H_2 says that it *cannot*, then surely it is clear that the observations strongly favor H_1 over H_2. The likelihood principle seems to be an entirely natural generalization from this special case; if O strongly favors H_1 over H_2 when $\Pr(O \mid H_1) = 1$ and $\Pr(O \mid H_2) = 0$, then surely it is reasonable to say that O favors H_1 over H_2 if $\Pr(O \mid H_1) > \Pr(O \mid H_2)$.

Royall's second argument is that the likelihood ratio is precisely the factor that transforms a ratio of prior probabilities into a ratio of posteriors. This is because Bayes's theorem (and the definition of conditional probability from which it derives) entails that

$$\frac{\Pr(H_1 \mid O)}{\Pr(H_2 \mid O)} = \frac{\Pr(O \mid H_1)}{\Pr(O \mid H_2)} \; \frac{\Pr(H_1)}{\Pr(H_2)} .$$

It therefore makes sense to view the likelihood ratio as a valid measure of the evidential meaning that the observations possess.

Royall's third line of defense of the likelihood principle is to show that it coincides with intuitive judgments about evidence when the principle is applied to specific cases. Here Royall is claiming for likelihood what philosophers typically claim in defense of the explications they recommend. For example, Rawls (1971) said of his theory of justice that his account is to be assessed by seeing whether the principles he describes have implications that coincide with our intuitive judgments about what is just and what is not in specific situations.

While we think there is value in each of these three lines of defense, none is as strong as one might wish. Even if the likelihood principle has plausible consequences in the limit case when all probabilities are 1's and 0's, the possibility exists that other measures of evidential meaning might do the same. Though these alternatives agree with likelihood in the limit case, they may disagree elsewhere. If so, the question remains as to why likelihood should be preferred over these other measures. As for Royall's second argument, it does show why likelihood makes sense if you are a Bayesian; but for anti-Bayesians such as Royall himself (and us), a very important question remains—when the hypotheses considered cannot be assigned objective probabilities, why think that likelihood describes what the evidence tells you about them? As for Royall's third argument, our hesitation here is much like the reservations we have about the first. Perhaps there are measures other than likelihood that coincide with what likelihood says in the cases Royall considers, but disagree elsewhere.[1]

In any event, friends of likelihood might want a stronger justification than the three-point defense that Royall provides. Whether there *is* anything more that can be said remains to be seen. In this regard, Edwards (1987, p. 100) endorsed Fisher's (1938) claim that likelihood should be regarded as a "primitive postulate;" it coincides with and systematizes our intuitions about examples, but nothing more can be said in its behalf. This, we should note, is often the fate of philosophical explications. If likelihood is epistemologically fundamental, then we should not be surprised to find that it cannot be justified in terms of anything that is more fundamental. It would not be an objection to the likelihood concept if it turned out to be an item of rock-bottom epistemology.

Royall follows Hacking (1965) in construing the likelihood principle as a two-part doctrine. There is first of all the idea, noted above, which we will call the qualitative Likelihood Principle:

(QUAL) O favors H_1 over H_2 if and only if $\Pr(O \mid H_1) > \Pr(O \mid H_2)$.

Notice that this principle could be expressed, equivalently, by saying that the likelihood ratio must be greater than unity or that the difference in likelihoods must be greater than 0.

Hacking adds to this a second and logically stronger claim—that the likelihood ratio measures the degree to which the observations favor one hypothesis over the other:

(DEGREE) O favors H_1 over H_2 to degree x if and only if O favors H_1 over H_2 and $\Pr(O \mid H_1)/ \Pr(O \mid H_2) = x$.

Obviously, (DEGREE) includes (QUAL) as a special case, but not conversely. However, even if (QUAL) were correct, a further argument would be needed to accept (DEGREE). Why choose the likelihood ratio, rather than the difference, or some other function of likelihoods, as one's measure of strength of evidence? There are many alternatives to consider, and the choice makes a difference because pairs of measures frequently fail to be ordinally equivalent (see Fitelson 1999). Consider, for example, the following four likelihoods:

[1] Here we should mention Royall's demonstration (which he further elaborates in Royall, forthcoming) that in a certain class of problems, a bound can be placed on the probability that the evidence will be misleading when one uses a likelihood comparison to interpret it. That is, for certain types of competing hypotheses H_1 and H_2, if H_1 is true, a bound can be placed on the probability that one's data will be such that $\Pr(\text{Data} \mid H_1) < \Pr(\text{Data} \mid H_2)$. We agree with Royall that this is welcome news, but we do not think that it shows that a likelihood comparison is the uniquely correct way to interpret data. The question remains as to whether other methods of interpreting data have the same property. Furthermore, we think it is important to recognize a point that Royall (forthcoming) makes in passing, that it is no criticism of likelihood that it says that a false hypothesis is better supported than a true one when the data are misleading. In fact, this is precisely what likelihood *should* say, if likelihood faithfully interprets what the data indicate; see Sober (1988, pp. 172-183) for discussion.

$\Pr(O_1 \mid H_1) = .09 \qquad \Pr(O_1 \mid H_2) = .02$
$\Pr(O_2 \mid H_3) = .8 \qquad \Pr(O_2 \mid H_4) = .3$

If we measure strength of evidence by the ratio measure, we have to say that O_1 favors H_1 over H_2 more strongly than O_2 favors H_3 over H_4. However, if we choose the difference measure, we get the opposite conclusion.[2] Royall does not take up the task of defending (DEGREE). Our discussion in what follows will focus mainly on the principle (QUAL), but we will have some comments on the principle (DEGREE) as well.

There is a third element in Royall's discussion that we should mention, one that differs from both (QUAL) and (DEGREE). This is his criterion for distinguishing O's *weakly* favoring H_1 over H_2 from O's *strongly* favoring H_1 over H_2:

(R) O strongly favors H_1 over H_2 if and only if O favors H_1 over H_2 and
$\Pr(O \mid H_1)/ \Pr(O \mid H_2) > 8$.

O weakly favors H_1 over H_2 if and only if O favors H_1 over H_2 and
$\Pr(O \mid H_1)/ \Pr(O \mid H_2) \leq 8$.

As was true of (DEGREE), principle (R) includes (QUAL) as a special case, but not conversely. Furthermore, (R) and (DEGREE) are logically independent. Royall recognizes that the choice of cut-off specified in (R) is conventional. In what follows, our discussion of (QUAL) will allow us to make some critical comments on (R) as well.

One of Royall's simple but extremely important points is that it is essential to distinguish carefully among three questions that you might want to address when evaluating the testimony of the observations:

(1) What should you do?
(2) What should you believe?
(3) What do the observations tell you about the hypotheses you're considering?

Royall argues that Neyman-Pearson statistics addresses question (1). However, if this *is* the question that the Neyman-Pearson approach tries to answer, then the question falls in the domain of decision theory, in which case utilities as well as probabilities need to be considered. Question (2) is the one that Bayesians address. Question (3) is different from (2) and also from (1); (3) is the proper province for the likelihood concept. Royall considers the possibility that Fisher's idea of statistical significance might be used in Neyman-Pearson statistics to address question (3). However, a good part of Royall's book is devoted to showing that the likelihoodist's answer to question (3) is better.

---

[2] Even if we restrict (DEGREE) to comparisons of hypotheses relative to the same data, it remains true that (DEGREE) is logically stronger than (QUAL). In our example, the ratio and difference measures disagree even when $O_1 = O_2$.

We will argue in what follows that (3) needs to be subdivided. There are at least two different questions you might ask about the bearing of evidence on hypotheses:

(3a) What do the observations tell you about the *truth* of the hypotheses you're considering?
(3b) What do the observations tell you about the *predictive accuracy* of the hypotheses you're considering?

Question (3a) is what we think Royall has in mind in his question (3). The second concept, of predictive accuracy, is something we'll discuss later. It is an important feature of this concept that a false model is sometimes more predictively accurate than a true one. The search for truth and the search for predictive accuracy are different; this is why we separate questions (3a) and (3b).

To develop these ideas, we now turn to a type of inference problem that is, we think, the Achilles heel of the likelihood approach. This is the problem of "model selection." By a "model," we mean a statement that contains at least one adjustable parameter. Models are composite hypotheses. Consider, for example, the problem of deciding whether the dependent variable y and the independent variable *x* are related linearly or parabolically:

(LIN) $y = a + bx + u$
(PAR) $y = a + bx + cx^2 + u$.

In these models, a, b, and c are adjustable parameters, while u is an error term with a normal distribution with zero mean and a fixed variance. Fixing the parameter values picks out a specific straight line or a specific parabola.

How can likelihood be used to choose between these models? That is, how are we to compare Pr(Data | LIN) and Pr(Data | PAR)? As Royall says, there are no general and entirely valid solutions of the problem of assessing the likelihoods of composite hypotheses. Let us consider some alternative proposals and see how they fare. The model LIN is a family, composed of the infinite set of straight lines in the *x*-*y* plane. Strictly speaking, the likelihood of LIN is an average of the likelihoods of all these straight lines:

$$\Pr(\text{Data} \mid \text{LIN}) = \Sigma_i \Pr(\text{Data} \mid \text{L}i)\Pr(\text{L}i \mid \text{LIN}).$$

If we have no way to figure out how probable different straight lines are, conditional on LIN, then we cannot evaluate the likelihood. Suppose, however, that every straight line has the same probability as every other, conditional on LIN. In this case Pr(Data | LIN) = 0 and the same is true of Pr(Data | PAR) (Forster and Sober 1994). This is an unsatisfactory conclusion, since scientists often believe that the data discriminate between LIN and PAR.[3]

[3] Rosenkrantz (1977) proposed that average likelihoods (based on a uniform prior) would favor simpler models, while Schwarz (1978) provided a mathematical argument for that conclusion, together with an approximate formula (BIC) for how the average likelihood depends on the maximum likelihood and the number of adjustable parameters. However, Schwarz's derivation is suspect in examples like the one we discuss. Uniform priors are improper in the

This example, we should mention, also illustrates a problem for Bayesianism, one that Popper (1959) noted. Because LIN is nested inside of PAR, it is impossible that Pr(LIN | Data) > Pr(PAR | Data), no matter what the data say. When scientists interpret their data as favoring the simpler model, it is impossible to make sense of this judgment within the framework of Bayesianism.[4]

An alternative is to shift focus from the likelihoods of LIN and PAR to the likelihoods of L(LIN) and L(PAR). Here L(LIN) is the likeliest straight line, given the data, and L(PAR) is the likeliest parabola. The suggestion is to compare the families by comparing their likeliest special cases. The problem with this solution is that it is impossible that Pr[Data | L(LIN)] > Pr[Data | L(PAR)]. This illustrates a general point—when models are nested, it is almost certain that more complex models will fit the data better than models that are simpler. However, scientists don't take this as a reason to conclude that the data always favor PAR over LIN. Instead, they often observe that simplicity, and not just likelihood, matters in model selection. If L(LIN) and L(PAR) fit the data about equally well, it is widely agreed that one should prefer L(LIN). And if LIN is compared with a polynomial that has 100 terms, scientists will say that even if L(POLY-100) fits the data *much* better than L(LIN) does, that *still* one might want to interpret the data as favoring L(LIN). Likelihood is an incomplete device for interpreting what the data say. It needs to be supplemented by due attention to simplicity. But how can simplicity be represented in a likelihood framework?

---

sense that the probability density cannot integrate to 1. If the density is zero, then it integrates to zero. If it is non-zero, then it integrates to infinity, no matter how small. That is, the density is improper because it cannot be normalized. For the purpose of calculating the posterior probability, this does not matter because if the density is everywhere equal to some arbitrary non-zero constant, then the posterior density of any curve is proportional to its likelihood, and the arbitrary constant drops out when one normalizes the posterior (which one should do whenever possible). So, Bayesians got used to the idea that improper priors are acceptable. However, they are not acceptable for the purpose of calculating average likelihoods because there is no such thing as the normalization of likelihoods (they are not probabilities). Of course, one could instead assume that the arbitrary constant is the same for all models. Then each average likelihood is proportional to the sum (integral) of all the likelihoods. But to compare integrals of different dimensions is like comparing the length of a line with the area of a rectangle—it makes little sense when there is no principled way of specifying the units of measurement (Forster and Sober 1994). For example, if one compares the very simple model $y = a$ with $y = a + bx$, and then with $y = a + 2bx$, the added b-dimension is scaled differently. So, when one integrates the likelihood function over parameter space, the result is different (unless one can justify the formula $y = a + bx$ uniquely as the principled choice).

An uneasiness about Schwarz's derivation has since led some Bayesians to invent other ways to compute average likelihoods (see Wasserman, 2000, for an easy technical introduction to the Bayesian literature). One proposal, which also gets around our objection, is the theory of "intrinsic Bayes factors" due to Berger and Pericchi (1996). The idea is to "pre-conditionalize" on each datum in the data set and average the results to obtain a well-conditioned and "approximately" uniform "prior." Then the average likelihoods are non-zero and can be compared straightforwardly. First, we note that this is not a vindication of Schwarz's argument because it yields a different criterion. Second, this solution appears to us to be unacceptably ad hoc as well.

[4] Bayesians sometimes address this problem by changing the subject. If we define PAR* to be the family of parabolas that does not include straight lines (i.e., c is constrained to be nonzero), then the axioms of probability do not rule out the possibility that LIN might have a higher probability than PAR*. However, it remains unclear what reason a Bayesian could have for thinking that Pr($c = 0$) > Pr($c \neq 0$).

We think that it cannot, at least not when simplicity is a consideration in model selection.[5] This is our criticism of likelihood. However, there is another inferential framework in which the role of simplicity in model selection makes perfect sense. This framework was proposed by the statistician H. Akaike (one of his earliest articles is Akaike, 1973; one of his latest is Akaike, 1985; see Sakamoto *et al.* 1986 for a thorough introduction to his method, and Burnham and Anderson 1998 for some scientific applications of Akaike's approach). The Akaike framework assumes that inference has a specific goal; the goal is not to decide which hypothesis is most probably true, or most likely, but to decide which will be most predictively accurate.

What does it mean to talk about the predictive accuracy of a model, like LIN? Imagine that we sample a set of data points from the true underlying distribution and use that data to find the best fitting straight line, namely L(LIN). We then use L(LIN) to predict the location of a new set of data. We draw these new data and see how close L(LIN) comes to predicting their values. Imagine repeating this process many times, using an old data set to find L(LIN) and then using that fitted model to predict new data. The average closeness to new data (as measured by the per-datum log-likelihood) is LIN's predictive accuracy. If the true underlying distribution is in fact linear, LIN may do poorly on some of these trials, but, on average, it will do well. On the other hand, if the true underlying distribution is highly nonlinear, LIN may do fairly well occasionally, but, on average, it will do a poor job of predicting new data. Obviously, the predictive accuracy of a model depends on what the true underlying distribution is. However, in making an inference, we of course don't know in advance what the truth is. Maximizing predictive accuracy might be a sensible goal, but, so far, it appears to be epistemologically inaccessible. Is it possible to figure out, given the single data set before us, how predictively accurate a model is apt to be?

Akaike proved a surprising theorem,[6] one that shows that predictive accuracy is epistemologically accessible. He showed that an unbiased estimate of a model's predictive accuracy can be obtained by taking the log-likelihood of its likeliest case, relative to the data at hand, and correcting that best-case likelihood with a penalty for complexity:

An unbiased estimate of the predictive accuracy of model M = Log Pr[Data | L(M)] − k.

Here k is the number of adjustable parameters in the model (see Forster 1999 for a more exact description of the meaning of k). LIN contains two adjustable parameters, whereas PAR contains

[5] In some inference problems, simplicity or parsimony can be shown to be relevant because simplicity influences likelihood. Phylogenetic inference is a case in point; see Sober (1988) for discussion.

[6] Akaike's theorem rests on some assumptions: a Humean 'uniformity of nature assumption' (that the old and new data sets are drawn from the same underlying distribution), and a surprisingly weak "regularity" assumption that implies (amongst other things) that the true distribution of the parameter estimates, when the number of data n is sufficiently large, is a multivariate normal distribution with a covariance matrix whose terms are inversely proportional to n, and that this covariance structure is mirrored approximately in the likelihood function. The central limit theorems in their various forms (Cramér 1946), entail a similar result for the distributions of the sums of random variables. The full details of the "normality assumption" are complex, and we refer the interested reader to Sakamoto et al. (1986) as providing the simplest technical introduction to these details. Akaike's result can also hold exactly for small sample sizes when additional conditions are met (*e.g.*, Kieseppä 1997).

3, and POLY-100 contains 101.[7] Akaike's theorem says that likelihood provides information about the predictive accuracy of a model, but the information is always distorted. Likelihood is like a bathroom scale that always tells you that you are lighter than you are. Its outputs are evidentially relevant, but they need to be corrected.

We now can explain our earlier remark that a true model can be less predictively accurate than a false one. Suppose you know that the relationship of x and y is nonlinear and parabolic. It *still* can make sense to use LIN to predict new data from old, if L(LIN) fits the data about as well as L(PAR). The truth can be a misleading predictor. It is a familiar fact that idealizations are valuable in science when a fully realistic model is either unavailable or is mathematically intractable. The Akaike framework reveals an additional virtue that idealizations can have—even when we possess a fully realistic (true) model, a (false) idealization can be a better predictor (Forster and Sober 1994, Sober 1999, Forster 2000a).[8]

As we mentioned earlier, model selection is the Achilles heel of likelihood. Yet Akaike's theorem describes a general circumstance in which likelihood provides an unbiased estimate of a model's predictive accuracy—*when two models have the same number of parameters, the likelihoods of their likeliest cases provide an unbiased indication of which can be expected to be more predictively accurate.* Likelihood needn't be viewed as a primitive postulate. We needn't resign ourselves to the idea that we value likelihood for its own sake. If predictive accuracy is your goal, likelihood is one relevant consideration because it helps you estimate a model's predictive accuracy. And when the models under consideration are *equally* complex, likelihood is the *only* thing you need to consider. Likelihood is a means to an end and is justified relative to that end.[9]

Not only does the qualitative likelihood principle receive a circumscribed justification from the Akaike framework; in addition, we can use Akaike's theorem to evaluate the (DEGREE) principle and principle (R) as well. The theorem provides an unbiased criterion for when one model will be more predictively accurate than another:

M_1 is estimated to be more predictively accurate than M_2 if and only if

$$\text{log-Pr}[\text{DATA} \mid \text{L}(M_1)] - k_1 > \text{log-Pr}[\text{DATA} \mid \text{L}(M_2)] - k_2.$$

This can be rewritten as

M_1 is estimated to be more predictively accurate than M_2 if and only if

$$\Pr[\text{DATA} \mid \text{L}(M_1)]/\Pr[\text{DATA} \mid \text{L}(M_2)] > \exp(k_1 - k_2).$$

[7] The number of adjustable parameters should also include the variance of the (assumed) error distribution, and any other free parameters used to define it. However, we have chosen to ignore this complication because it is not relevant to the main point of this essay.

[8] Recall that we defined 'model' to mean a statement containing at least one adjustable parameter. Our point about idealizations would not be correct for statements containing no adjustable parameter.

[9] Our point is in accord with Akaike's (1973) observation that AIC is an "extension of the maximum likelihood principle."

If $k_1 = k_2$, this can be stated, equivalently, by saying that the likelihood ratio must be greater than 1 or by saying that the difference in likelihoods must be greater than 0.[10] However, if $k_1 \neq k_2$, a *ratio* criterion can be formulated, but there is no equivalent criterion that can be stated purely in terms of likelihood *differences*.[11] This helps distinguish some measures of strength of evidence from others, as (DEGREE) requires.

Similar remarks apply to Royall's principle (R). If $k_1 = k_2$, Royall's stipulation in (R) of the number 8 as the cut-off separating strong from weak evidence favoring H_1 over H_2 is a possibility (though other cut-offs are possible as well). However, when $\exp(k_1 - k_2) > 8$, the difference between strong and weak evidence cannot be defined by the proposed cut-off of 8. Akaike's theorem does not determine how the distinction between weak and strong evidence should be drawn, but it does restrict the search to criteria defined in terms of likelihood ratios.

Our defense of (DEGREE) is not a defense of everything it implies. Remember that (DEGREE) equates the degree to which O favors H_1 over H_2 with the likelihood ratio, and this implies two things: (a) that the likelihood ratio is the correct way of capturing how the degree of fit between O and H_1 and between O and H_2 influences the degree to which O favors H_1 over H_2, and (b) nothing else influences the degree to which O favors H_1 over H_2. Part (a) rules out the possibility that any other measures of fit, such as the sum of the absolute values of the residuals between the data and the mean curve of the hypothesis, affect the relative degree of support. For AIC, H_1 and H_2 are the likeliest hypotheses $L(M_1)$ and $L(M_2)$, and the degree to which O favors $L(M_1)$ over $L(M2)$ depends on the likelihood ratio and $\exp(k_1 - k_2)$. The latter term corrects for an expected overfitting bias, which would otherwise provide the more complex model with an unfair

[10] The most important special case here is when $k_1 = 0 = k_2$, which is just the non-model selection problem of comparing two specific point hypotheses. In this case, there are many independent reasons in favor of a likelihood ratio law of this form, including the classical Neyman-Pearson theorems that prove that a decision rule based on such a rule is the most powerful test of its size (see Hogg and Craig 1978, p. 246). The limitations of these theorems is that they presuppose a very simple 0 or 1 measure of discrepancy between a hypothesis and the truth. Lele (this volume) develops an alternative analysis based on more interesting discrepancy measures, which also speaks in favor of the likelihood ratio criterion.

[11] If it were granted that the degree of evidence depends only on the likelihoods in some way, then there would be an independent reason for not using the difference measure. For in the case of continuous variables, likelihoods are equal to the probability *density* of an observed quantity x times an arbitrary multiplicative constant (Edwards 1987). To understand the reason for this arbitrary factor, consider a transformation of the variable x, $x' = f(x)$, for some one-to-one function f. Probabilities are invariant under such transformations, so consider the probability that x is observed in a given interval around x. This probability is equal to the area under the density curve within this interval. If the interval is small, then the area is equal to the density at x times the width of the interval. In order for the area to be invariant under any transformation of x, the density must change whenever the length of the interval changes. So densities, or the differences in densities, fail the requirement of language invariance. On the other hand, the difference of the *probabilities* is invariant, but it is proportional to the length of the small interval around x, which is arbitrary (see Forster 1995 for further discussion). Therefore the difference measure is caught in a dilemma--it either fails the desideratum of language invariance or it contains an arbitrary multiplicative factor. Fortunately, the arbitrary constant drops out when we take the *ratio* of the likelihoods, or any function of the likelihood ratio, so it is both language invariant and non-arbitrary. As far as we can see, this class of measures is unique in this regard, at least amongst likelihood measures.

advantage.[12] AIC therefore agrees with part (a) of (DEGREE), but denies part (b). This leads us to following general principle:

> (DEGREE Prime) The likelihood ratio is the correct way of capturing how the degree of fit between O and H_1 and between O and H_2 influences the degree to which O favors H_1 over H_2.[13]

In the special case of comparing simple hypotheses, H_1 and H_2, for which there are no selection biases, (DEGREE Prime) reduces to (DEGREE).

The argument for this principle arises out of the *form* of AIC—the fact that AIC can be expressed as: O favors $L(M_1)$ over $L(M_2)$ if and only if $\Pr(O \mid L(M_1))/\Pr(O \mid L(M_2)) > K$. However, we have to admit that our argument is only as strong as its premises, and not every statistician will agree that AIC stands on firm foundations.[14] In fact, the argument really depends on the premise that model selection should have this form for *some K*. The exact value of K does not matter. As it turns out, almost all model selection criteria in the literature can be expressed in this basic form, including BIC (Schwarz 1978), variations on AIC (*e. g.*, Hurvich and Tsai 1989), posterior Bayes factors (Aitkin 1991), and an important class of Neyman-Pearson hypothesis tests.[15] Moreover, all of these criteria apply to a wide variety of statistical applications, including contingency table analysis, regression models, analysis of variance, and time series. There are many independent arguments for (DEGREE Prime).

But is that the end of the debate? One still might dream of a unified perspective from which everything else follows, including a corrected version of Royall's distinction between weak and strong evidence. Akaike's framework goes one step beyond Fisher's view that likelihood is fundamental, but one still might wish for more.

---

[12] The degree to which O favors H_1 over H_2 may also be a function of the number of data, n, even though this is a fact about the data and is not a function of the likelihood ratio. We regard this kind of information as akin to the difference $k_1 - k_2$ because neither of them measure the fit of H_1 and H_2 to the data O.

[13] Note that (DEGREE Prime) only applies when the likelihoods of H_1 and H_2 are well defined.

[14] For example, it is often charged that AIC is inconsistent (but see Forster, 2000b, for a defense of AIC against this charge). Or it might be maintained that the goal of predictive accuracy is not the primary consideration at hand.

[15] There are two cases to consider. In the case of comparing simple hypotheses, H_0 and H_1, where H_0 is the null hypothesis, a *best test* of size α is by definition (Hogg and Craig 1978, p. 243) a test with a critical region C of size α such that for any other critical region A of size α, $\Pr(C \mid H_1) \geq \Pr(A \mid H_1)$. That is, a best test maximizes the probability of rejecting H_0 when H_1 is true. Hogg and Craig (1978, p. 246) show that for any best test with critical region C, there is a number K, such that a data set lies within C if and only if the likelihood ratio of H_1 to H_0 is greater than or equal to K. These are the Neyman-Pearson theorems mentioned in note 8. In the case of comparing a simple hypothesis H_0 against a composite alternative M, a *uniformly most powerful critical region* of size α is by definition (Hogg and Craig 1978, p. 252) a region C that provides a best test for comparing H_0 against any point hypothesis in M. So, if a uniformly most powerful test exists, the test between H_0 and any representative of M is a likelihood ratio test. In examples in which the assumptions of Akaike's theorem hold (see note 6), and the composite hypothesis has one adjustable parameter, a uniformly most powerful Neyman-Pearson test with $\alpha = .05$ effectively trades off the maximum log-likelihood against simplicity to a degree somewhere between AIC and BIC (Forster 2000b).

References:

Aitkin, M. (1991): "Posterior Bayes Factors," *Journal of the Royal Statistical Society* **B 1:** 110-128.

Akaike, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle." B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory:* 267-81. Budapest: Akademiai Kiado.

Akaike, H. (1985): "Prediction and Entropy." In A. C. Atkinson and S. E. Fienberg (eds.), *A Celebration of Statistics*. New York: Springer. 1-24.

Blackwell, David (1953): "Equivalent Comparisons of Experiments." *Annuls of Mathematical Statistics* **24:** 265-272.

Berger, James O. and L. Pericchi (1996): "The Intrinsic Bayes Factor for Model Selection and Prediction," *The Journal of the American Statistical Association*, **91**: 109-122.

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

Cramér H. (1946): *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Edwards, A. W. F. (1987): *Likelihood.* Expanded Edition. Baltimore and London: Johns Hopkins University Press.

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

Fitelson, Branden (1999): "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity." *Philosophy of Science* **66** (Proceedings): S362-S378.

Forster, Malcolm R. (1995): "Bayes and Bust: The Problem of Simplicity for a Probabilist's Approach to Confirmation." *The British Journal for the Philosophy of Science* **46:** 399-424.

Forster, Malcolm R. (1999): "Model Selection in Science: The Problem of Language Variance." *British Journal for the Philosophy of Science* **50**: 83-102.

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

Forster, Malcolm R. (2000b): "Key Concepts in Model Selection: Performance and Generalizability," *Journal of Mathematical Psychology* **44**: 205-231.

Forster, Malcolm R. and Elliott Sober (1994): "How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions." *The British Journal for the Philosophy of Science* **45**: 1-35.

Hacking, Ian (1965). *Logic of Statistical Inference.* Cambridge: Cambridge University Press.

Hogg, Robert V. and Allen T. Craig (1978): *Introduction to Mathematical Statistics*. Fourth Edition. New York: Macmillan Publishing Co.

Hurvich, C. M. and Tsai, C. (1989): "Regression and Time Series Model Selection in Small Samples." *Biometrika* **76**: 297-307.

Kieseppä, I. A. (1997): "Akaike Information Criterion, Curve-fitting, and the Philosophical Problem of Simplicity." *British Journal for the Philosophy of Science* **48**: 21-48.

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

Rawls, John (1971): *A Theory of Justice*. Cambridge: Harvard University Press.

Rosenkrantz, Roger D. (1977): *Inference, Method, and Decision*. Dordrecht: D. Reidel.

Royall, Richard M. (1997): *Statistical Evidence: A likelihood paradigm*. Boca Raton: Chapman & Hall/CRC.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa (1986): *Akaike Information Criterion Statistics.* Dordrecht: Kluwer Academic Publishers.

Schwarz, Gideon (1978): "Estimating the Dimension of a Model." *Annals of Statistics* **6:** 461-5.

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

Sober, Elliott (1999): "Instrumentalism Revisited." *Critica* **31**: 3-38.

Wasserman, Larry (2000): "Bayesian model selection and model averaging." *Journal of Mathematical Psychology* **44**: 92-107.

Zucchini, Walter (2000): "An Introduction to Model Selection." *Journal of Mathematical Psychology* **44**: 41-61.