AIC SCORES AS EVIDENCE: A BAYESIAN INTERPRETATION

Malcolm Forster and Elliott Sober

1 INTRODUCTION

Akaike [1973] helped to launch the field in statistics now known as model selection theory by describing a goal, proposing a criterion, and proving a theorem. The goal is to figure out how accurately models will predict new data when fitted to old. The criterion came to be called AIC, the Akaike Information Criterion:

$$AIC(M) = \textit{log-likelihood of } L(M) - k,^{1}$$

where the model M contains k adjustable parameters and $L(M)$ is the member of M obtained by assigning to the adjustable parameters in M their maximum likelihood values. Akaike's theorem is that AIC is an unbiased estimator of predictive accuracy, given some assumptions that are widely applicable.[2] The three parts of this achievement should be kept separate. Not conflating the goal and the criterion is important, since criteria other than AIC might do a better job in some circumstances in achieving the goal. The criterion and the theorem also need to be distinguished, since, in general, the fact that an estimator is unbiased does not suffice to show that it should be used.

The theorem that Akaike proved made it natural to understand AIC as a frequentist construct. AIC is a device for estimating the predictive accuracy of models just as a kitchen scale is a device for estimating the weight of objects. Bayesians assess an estimator by determining whether the estimates it generates are probably true or probably close to the truth. Evaluating how probable it is that a melon weighs two pounds, given that the scale says that it weighs two pounds, requires that one have a prior probability for the melon's weighing two pounds. Akaike's theorem says nothing about prior or posterior probabilities, so there was no reason to think of AIC in Bayesian terms. Rather, what Akaike did was what

[1]Notice that AIC scores are negative numbers, with smaller likelihoods being "more negative" (i.e., farther from zero). Thus, here we adopt the convention used in [Forster and Sober, 1994] wherein higher AIC scores (those closer to zero) are better — they indicate a higher degree of predictive accuracy. In the statistics literature the AIC score is usually represented as -2 times the AIC formula we use here, so the opposite convention is followed – higher AIC scores are worse, because these scores indicate that the fitted model is more distant from the truth.

[2]See [Sakamoto *et al.*, 1986] for a thorough explanation and proof of Akaike's theorem.

Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.
Volume editors: Prasanta S. Bandyopadhyay and Malcolm R. Forster. General Editors: Dov M. Gabbay, Paul Thagard and John Woods.

frequentists generally do when they assess an estimator. They establish one or more of its "long-run operating characteristics." The "long-run" average of an unbiased estimator of a quantity is, by definition, centered on the quantity's true value. If you repeatedly weigh an object on an unbiased scale, the average of all the readings of the scale converges to the object's true weight. On any given occasion, the scale's reading may be too high or too low. To say that an estimator is unbiased leaves open what its variance is.

This fact about Akaike's theorem is sometimes taken to cast doubt on AIC. But here we must be careful — there may be more to AIC than Akaike's theorem established.

In fact, AIC isn't just an unbiased estimator. The expected squared error of an unbiased estimator is strictly less than that of any other estimator that differs from it by a constant.[3] AIC is in this respect a better estimator of predictive accuracy than BIC, the Bayesian Information Criterion first derived by Schwarz [1977]. BIC is both a biased estimator of predictive accuracy and has a higher expected squared error. To be sure, BIC was not developed as a method for estimating predictive accuracy, but rather was formulated as a method for estimating a model's average likelihood. Nonetheless, the point is that there is more to be said on behalf of AIC than that it is unbiased. This does not prove that AIC is the best of all possible estimators of predictive accuracy. However, we suggest that estimators should be evaluated in the same way that empirical scientific theories are. Rather than asking whether general relativity is the best of all possible theories, we should ask whether it is better than the alternative theories that have been identified. The same goes for AIC.

The goal of this paper is not to provide a full-scale assessment of AIC, but to show that it is an estimator whose estimates should be taken seriously by Bayesians, its frequentist pedigree notwithstanding.[4] Frequentists often maintain that the question of how an individual *estimate* should be interpreted is meaningless — that the only legitimate question concerns the long-term behavior of *estimators*. Bayesians reply that both questions are important and that the interpretation of individual estimates is pressing in view of the fact that a given estimate might be produced by any number of different estimation procedures (see, for example, [Howson and Urbach, 1993]). We agree with Bayesians on this point and so we will pose a question about individual estimates that frequentists typically disdain — do AIC scores provide evidence concerning how predictively

[3]**Proof**: Consider an estimator that differs from AIC by a constant c, and denote the true predictive accuracy a model has by AIC*. AIC is an unbiased estimate of AIC* if and only if $E[\text{AIC}] = \text{AIC}^*$, where E is "expected value." We may now calculate the expected squared error of the new estimator using well-known properties of expected values: $E\left[(\text{AIC} + c - \text{AIC}^*)^2\right] = E\left[(\text{AIC} - \text{AIC}^*)^2\right] + c^2$. The "cross-term" is zero because AIC is unbiased. Therefore, the expected squared error of AIC is strictly less than the expected squared error of any estimator that differs from it by a non-zero constant. This includes BIC and any other estimator that is equal to the log-likelihood plus some penalty for complexity.

[4]By "frequentists" we are referring to the classical school of statisticians led primarily by Neyman and Pearson, and Fisher.

accurate a model will be?

2 ESTIMATES AS EVIDENCE

When Bayesians seek to explain why a positive pregnancy test is evidence that the person taking the test is pregnant, they inevitably refer to the Law of Likelihood [Hacking, 1965]:

> Observation O favors hypothesis H_1 over hypothesis H_2 if and only if $P(O|H_1) > P(O|H_2)$.

This concept of "favoring" ignores the values of prior probabilities and focuses exclusively on the evidential meaning of the present observations, which is said to be captured by the likelihoods of the competing hypotheses. The odds version of Bayes's theorem makes it clear that, for Bayesians, likelihoods are the sole vehicle by which the present observations can change one's degree of belief in the hypotheses:

$$\frac{P(H_1|O)}{P(H_2|O)} = \frac{P(O|H_1)}{P(O|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

The ratio of posterior probabilities differs from the ratio of priors precisely when the likelihood ratio differs from unity. And the more the likelihood ratio differs from unity, the greater the transformation that the observations achieve.

The Law of Likelihood applies to point estimates of a continuous parameter just as much as it does to the presence or absence of a dichotomous trait like pregnancy. For example, if a thermometer, applied to object o, yields a reading $R(o) = x$, this is evidence that o's temperature, $T(o)$, is y rather than x precisely when

$$(T_1) \qquad P[R(o) = y|T(o) = y] > P[R(o) = y|T(o) = x].$$

If AIC scores are like thermometer readings in this respect, an AIC score of x is evidence that the model has a predictive accuracy of x rather than y precisely when

$$(A_1) \qquad P[\mathrm{AIC}(M) = y|PA(M) = y] > P[\mathrm{AIC}(M) = y|PA(M) = x],$$

where $PA(M)$ denotes the predictive accuracy of model M. Akaike's theorem, that AIC is an unbiased estimator of predictive accuracy, does not entail (A_1) any more than a thermometer's being an unbiased estimator of temperature entails (T_1). For suppose a thermometer works like this: when an object has a temperature of y, the thermometer has 90% chance of reading $(y - a)$ and a 10% chance of reading $(y + b)$. The thermometer is an unbiased estimator of temperature precisely when

$$y = (0.9)(y - a) + (0.1)(y + b),$$

which is true if and only if

$$a/b = 1/9.$$

Suppose, for example, that $a = 10$ and $b = 90$. Then an object whose true temperature is +3 has a 90% chance of having the thermometer say its temperature is -7 and a 10% chance of having the thermometer say its temperature is +93. Since the average (expected) reading is $(0.9)(-7)+(0.1)(93) = 3$, the thermometer is unbiased. Yet, if an object's temperature is 3, the probability is zero that the thermometer will say that its temperature is 3. If the thermometer produces a reading of x, the maximum likelihood estimate of the object's temperature is $(x + 10)$. This unbiased thermometer has an asymmetric error distribution; its readings have a higher probability of being too low than too high. This is why (T_1) is false. The following proposition is true instead:

(T_2) $$P[R(o) = y|T(o) = y + 10] = 90\% > P[R(o) = y|T(o) = x], \text{ for all } x, y \text{ such that } y + 10 \neq x$$

The thermometer's readings provide evidence about temperature because (T_2) is true; the notion of evidence used here is the one sanctioned by the Law of Likelihood.

3 DIFFERENCES IN AIC SCORES

The reason we have labored the point that an unbiased estimator can have an asymmetric error distribution, and that this does not prevent its estimates from providing evidence, is that AIC is both an unbiased estimator and has an asymmetric error distribution. The proposition (A_1) is false, but AIC scores still provide evidence about the predictive accuracies of models.

Consider two models, M_1 and M_2, where the first is nested in the second, meaning that M_1 is a special case of M_2. Suppose that the AIC score of the second model is y units larger than the AIC score of the first. It turns out that this difference in AIC scores does not obey the following principle:

(A_2)

$$P[AIC(M_2) - \text{AIC}(M_1) = y|PA(M_2) - PA(M_1) = y] > P[\text{AIC}(M_2) - \text{AIC}(M_1) = y|PA(M_2) - PA(M_1) = x], \text{ for all } x, y \text{ such that } x \neq y.$$

The falsity of (A_2) is shown by the different curves in Figure 1, which represents two nested models that differ by one in the number of adjustable parameters they have. Each curve in Figure 1 pertains to a difference in AIC scores; differences in AIC scores are observations, akin to the observation that one object has a higher thermometer reading than another. For any one observation, different hypotheses about the difference in the two models' predictive accuracies confer different probabilities on that observation. Each curve in Figure 1 is in this sense a likelihood function. The derivation of these curves for the general case of a pair of nested

models that differ by k in their number of adjustable parameters is given in the Appendix.

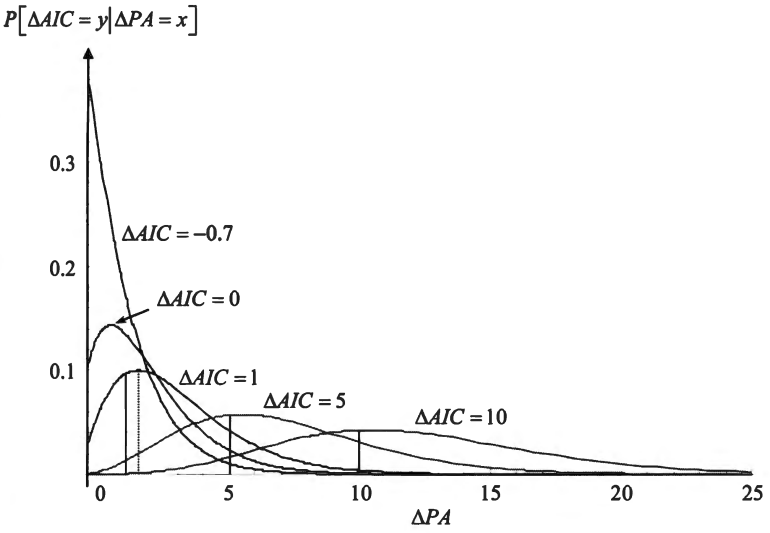

Figure 1. The graphs show the likelihood function for various values of the statistic ΔAIC. The meta-model describes how the difference in predictive accuracy of two nested models (call them object models) will affect the difference in their AIC scores. The meta-model has one adjustable parameter ΔPA, and each curve shows the likelihood of the hypothesis corresponding to the value of ΔPA read from the horizontal axis. In this case, ΔPA is the difference in predictive accuracy of two object models that differ in the number of adjustable parameters by only 1.

Notice in Figure 1 that the peak of the likelihood function for the observation that $\mathrm{AIC}(M_2) - \mathrm{AIC}(M_1) = y$ does not correspond to the hypothesis that $PA(M_2) - PA(M_1) = y$. Rather, the maximum likelihood hypothesis is that $PA(M_2) - PA(M_1) = x$, for some number x that is greater than y. This is easiest to see in Figure 1 by looking at the case where $\Delta\mathrm{AIC} = 1$. This value is made more probable by $\Delta PA = 1.5$ than it is by $\Delta PA = 1$. The same point holds for the curves depicted in Figure 2, which again describes a pair of nested models, but this time one model has 10 more adjustable parameters than the other.

The following generalization holds. If $\mathrm{AIC}(M_2) - \mathrm{AIC}(M_1) > 0$, then there exists a positive number x such that

(A_3)

$$P[\mathrm{AIC}(M_2) - \mathrm{AIC}(M_1) = y | PA(M_2) - PA(M_1) = x] > P[\mathrm{AIC}(M_2) - \mathrm{AIC}(M_1) = y | PA(M_2) - PA(M_1) = z], \text{ for all } z \neq x.$$

AIC differences therefore favor some hypotheses about predictive accuracy over

others, assuming that the concept of evidential favoring is given by the Law of Likelihood.

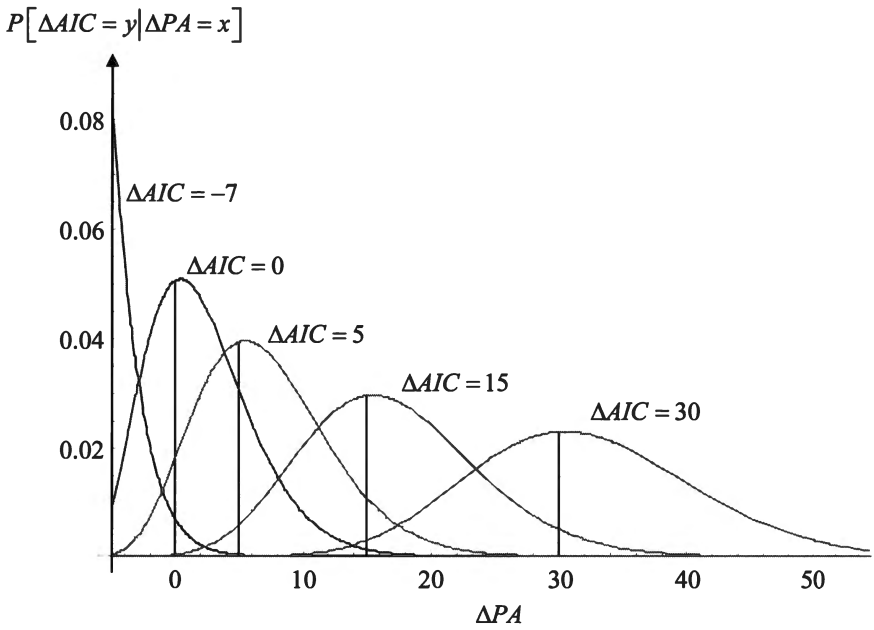

Figure 2. The graphs show the likelihood function for various values of the statistic ΔAIC. As in Figure 1, the meta-model has one adjustable parameter ΔPA, and each curve shows the probability of the statistic for various values of the parameter. In this case, ΔPA is the difference in predictive accuracy of two object models that differ in the number of adjustable parameters by 10.

There is a second interpretation of the results described in Figures 1 and 2 that Bayesians can extract. It is obtained by using Royall's [1997] concept of a likelihood *interval*. Royall supplements the Law of Likelihood by proposing a definition of what it means for an observation to provide *strong* evidence favoring H_1 over H_2; he suggests that this is true precisely when the likelihood ratio is at least 8. Royall uses this convention to draw an interval around the maximum likelihood point on a likelihood function that has the following property: the maximum likelihood value has a likelihood that is at least 8 times as large as that possessed by any point that is outside the interval. Interpreted in this way, some of the observations described in Figures 1 and 2 provide *strong* evidence favoring the maximum likelihood estimate (which is positive) over *any* negative estimate.[5]

[5]In Figure 2, ΔAIC = −7 provides strong evidence that ΔPA is negative, although this does not happen in Figure 1.

Again, this fails to be true when the difference in AIC scores is very small.

A further assimilation of AIC scores to the Bayesian framework can be achieved by adding a prior distribution over differences in predictive accuracies. This, plus the likelihood function associated with the observed difference in AIC scores, allows one to calculate the posterior probability (density) that one model's predictive accuracy exceeds another's by x. One also can calculate the posterior probability that the difference in predictive accuracies is positive.

There is a common puzzle about Bayesianism that has been clarified here. Bayesians often talk about comparing the probabilities of nested models. But, in practice, they compare nested models in terms of the ratio of their likelihoods, the Bayes factor, as in [Schwarz, 1978], which may be viewed as an application of LL. This makes sense because the Bayes factor can favor either model. Unfortunately, this comparison contradicts a more fundamental Bayesian comparison in terms of posterior probabilities if the first model entails the second, for then the probability of the first model can never be greater than the second. We have shown that the problem disappears as soon as Bayesians focus on the predictive accuracy of the models, rather than their truth. For now it is possible for the predictive accuracy of the first model to be higher than the second model, even if it is impossible for the first model to be true while the second model is false. Bayes factors are no longer involved in the comparison; now we are comparing meta-hypotheses about predictive accuracy. This shows why Bayesians need not be threatened, or puzzled, when they see scientists comparing models that are related by strict logical entailment.[6]

4 CONCLUSION

AIC began life with Akaike's [1973] theorem, which established that AIC is an unbiased estimator of predictive accuracy. Because this proof described a long-run operating characteristic of the estimator, and not the evidential meaning of the particular estimates that the estimator might provide, and also because prior and posterior probabilities were not part of the framework used by Akaike and his school (see, for example, [Sakamoto *et al.*, 1986]), AIC came to be viewed as a frequentist construct. However, these facts about earlier defenses of AIC do not establish that AIC scores are meaningless from a Bayesian point of view. We have argued that AIC scores provide evidence concerning the predictive accuracies of models in the sense of "evidence" sanctioned by the Law of Likelihood, which is a central Bayesian principle.[7] AIC scores are no more essentially tied to frequentism than thermometer readings are.

[6]We are grateful to Prasanta Bandyopadhyay for suggesting this point.

[7]The result reported here pertains to any pair of nested models. The case of non-nested models remains to be investigated.

APPENDIX

Suppose that we are interested in corn crop yields in a field that has been divided into numerous plots. The plots are subjected to different conditions, they have different drainage properties, different exposures to the wind, slightly different soils, and perhaps they are treated and irrigated differently as well. The question is whether these different conditions affect the height of the corn plants. Assume that the height of each plant i is represented by a random variable X_i whose value is denoted by x_i. Also to simplify the mathematics, we assume that the random variables for plants in the same plot are independent and identically distributed with a normal distribution and unit variance; $X_i \sim N(\mu_{j(i)}*, 1)$, where $j(i)$ denotes the plot number in which plant i is found. The * indicates that this is the true value of the mean, not something hypothetical. Sometimes we will just write μ_j*.

The various null hypotheses one might consider here are *false* because they falsely assert that two or more plots have the same mean, whereas in fact all the μ_j* are different (although only slightly in some cases). A typical hypothesis is false, but it's not false that it has some predictive accuracy. The model under test assigns a degree of predictive accuracy to some model about corn yields. This meta-model attributes a property to the object model. The object model is false, but the meta-model may be true.

Suppose that there are only three plots, with mean values $\mu_1^*, \mu_2^*, \mu_3^*$. An arbitrary hypothesis asserts that the mean values are μ_1, μ_2, μ_3 for some particular triple of numbers. A model will leave these as adjustable parameters, but possibly add constraints such as $\mu_1 = \mu_2$. Each such constraint will reduce the number of adjustable parameters by one. For example, the model that (falsely) asserts that the three plots have the same mean yield has one adjustable parameter. It asserts for all plants, $X_i \sim N(\mu, 1)$, for some unknown parameter μ.

Let's say that there are n_1 plants sampled from plot 1, n_2 plants from plot 2, n_3 plants from plot 3, with a total of $n = n_1 + n_2 + n_3$ plants. The log-likelihood of any particular hypothesis in the object model (picked out by a particular triple of numbers μ_1, μ_2, μ_3) is

$$\begin{aligned} l(\mu_1, \mu_2, \mu_3) &= -\tfrac{n}{2}\log 2\pi - \tfrac{1}{2}\sum_{i=1}^{n_1}(x_i - \mu_1)^2 - \tfrac{1}{2}\sum_{i=n_1+1}^{n_1+n_2}(x_i - \mu_2)^2 - \tfrac{1}{2} \\ &\quad \sum_{i=n_1+n_2+1}^{n}(x_i - \mu_3)^2. \end{aligned}$$

Since μ_1, μ_2, μ_3 are constants, $-2l(\mu_1, \mu_2, \mu_3) - n\log 2\pi$ is the sum of the squares of n normal random variables of unit variance and mean $\mu_1^* - \mu_1, \mu_2^* - \mu_2$, or $\mu_3^* - \mu_3$, depending on the plot in which the plant is sampled. That means that $-2l(\mu_1, \mu_2, \mu_3) - n\log 2\pi$ is a chi-square random variable with n degrees of freedom, and a non-centrality parameter equal to the sum of the squares obtained by substituting the mean value for each random variable in the quadratic. That is,

$$-2l(\mu_1, \mu_2, \mu_3) - n\log 2\pi \sim \chi^2(n, \lambda),$$

where

$$\lambda = n_1 (\mu_1^* - \mu_1)^2 + n_2 (\mu_2^* - \mu_2)^2 + n_3 (\mu_3^* - \mu_3)^2 .$$

Since the mean of a chi-square random variable is equal to the number of degrees of freedom (n in this case) plus the non-centrality parameter, we can calculate the predictive accuracy of any hypothesis, where the predictive accuracy of such a hypothesis is, by definition, its expected log-likelihood.

$$\begin{aligned} PA(\mu_1, \mu_2, \mu_3) &= -\tfrac{n}{2} (\log 2\pi + 1) - \\ &\quad \tfrac{1}{2} \left(n_1 (\mu_1^* - \mu_1)^2 + n_2 (\mu_2^* - \mu_2)^2 + n_3 (\mu_3^* - \mu_3)^2 \right) . \end{aligned}$$

Note that this is the predictive accuracy of a "point" hypothesis or non-composite hypothesis.

Now consider the hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu$, which has one adjustable parameter. This is a composite hypothesis, not a point hypothesis. Since this distinction is important, we will mark it by always referring to a composite hypothesis as a *model*, and never referring to point hypotheses as models. This is because the equality above does not specify any particular numerical value for μ. The question at hand is: How should we define the predictive accuracy of a model? If we fit a model to the actual set of data, we get a point hypothesis, namely the maximum likelihood hypothesis that belongs to the model. Its predictive accuracy is already well defined, but we do not want to define the predictive accuracy of the model directly in terms of this number because the actual data set may not be typical. So, we imagine that we repeatedly generate data sets of the same size as the actual data set, and define the predictive accuracy of the model as the average predictive accuracy of the maximum likelihood hypotheses generated in this random way. The predictive accuracy of a model is therefore the predictive accuracy of a "typical" maximum likelihood hypothesis.

From a mathematical point of view, things just got complicated because there is now a double expectation involved in calculating the predictive accuracy of a model. First, we take the general formula for $PA(\mu, \mu, \mu)$, which is itself defined as an expected value of a particular point hypothesis, $\mu_1 = \mu_2 = \mu_3 = \mu$, where we are thinking of μ as a fixed number. Recall that the predictive accuracy, PA, is defined here as the expected log-likelihood of the point hypothesis relative to some newly generated data set, which we may call test data. How well the hypothesis fits this new data is naturally thought of as a measure of its predictive accuracy. The next step is to put μ equal to the maximum likelihood estimate of μ determined by the actual data set, which is denoted by $\hat{\mu}$. Thus, we get$PA(\hat{\mu}, \hat{\mu}, \hat{\mu})$. But $\hat{\mu}$ may not be typical, so we want to average $PA(\hat{\mu}, \hat{\mu}, \hat{\mu})$ over all possible values of $\hat{\mu}$ that we would obtain if we were to draw data from the same (unknown) probability distribution. This average is defined as the expected value of $PA(\hat{\mu}, \hat{\mu}, \hat{\mu})$ as determined by the value of $\hat{\mu}$ that would be obtained from any data set that could be used to initially fit the model. We think of this data set as a *calibration data set* because it is used to fix the values of adjustable parameters. The calibration data is conceptually quite different from a test data set. To define the notion of

prediction, these data sets must be kept separate, and it is therefore essential that we define the predictive accuracy of a model as a double expectation.

The maximum log-likelihood of the model is found by finding the maximum likelihood estimate of the adjustable parameter μ, which we denote by $\hat{\mu}$, and then writing down the log-likelihood of the particular hypothesis corresponding to this value of μ. The answer is, clearly,

$$\text{maximum log-likelihood} = l(\hat{\mu}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Now think of $\hat{\mu}$ as a quantity that varies randomly as we consider different sets of calibration data. It is well known (e.g., see [Hogg and Craig, 1978, p. 279]) that

$$\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \sim \chi^2(n-1, \lambda^*)$$

is a non-central chi-square distribution with $n-$ 1 degrees of freedom, whose non-centrality parameter is calculated by substituting the mean values for each random variable in the quadratic. The non-centrality parameter is therefore

$$\lambda^* = n_1(\mu_1^* - \mu^*)^2 + n_2(\mu_2^* - \mu^*)^2 + n_3(\mu_3^* - \mu^*)^2,$$

where

$$\mu^* \equiv \frac{n_1}{n}\mu_1^* + \frac{n_2}{n}\mu_2^* + \frac{n_3}{n}\mu_3^*.$$

Notice what went on here. We began with an object model, and considered its log-likelihood relative to the actual data. Then we introduced a meta-hypothesis about how the log-likelihood would behave statistically if it were determined by a new data set of the same size. In other words, we have treated $\hat{\mu}$ as well as $l(\hat{\mu})$ as random variables, and we have constructed a hypothesis about its statistical behavior. The meta-hypothesis and the object model are different (one may be true, while the other is false).

Now let's consider the general case in which there are n plants sampled from m plots ($n > m$) in numbers $n_1, n_2, \ldots, n_m$, where the hypothesis under consideration asserts that the means of two or more plots are equal. We can think of the plots being partitioned into clusters; sometimes there may be one plot in a cluster, or two plots in a cluster, and so on. Obviously, there are fewer clusters than plots, except in the special case in which there is only one plot in any cluster. In that case, the model has m adjustable parameters. In the other extreme case, in which all the plots are in one cluster, there is one adjustable parameter. In general, there are k adjustable parameters, where k is the number of clusters introduced by the hypothesis; clearly $\{1 \leq k \leq r\} 1 \leq k \leq m$. Let's denote the maximum log-likelihood of the model with k adjustable parameters by $l(\hat{\mu}_1, \hat{\mu}_1, \ldots, \hat{\mu}_k)$. Then

$$-2l(\hat{\mu}_1, \hat{\mu}_1, \ldots, \hat{\mu}_k) - n\log 2\pi \sim \chi^2(n-k, \lambda^*),$$

where the non-centrality parameter λ^* is calculated in the same way as before. The $*$ reminds us that this is a meta-hypothesis about how the log-likelihood of the object model behaves.

According to Akaike's theorem, AIC $= l(\hat{\mu}_1, \hat{\mu}_1, \ldots, \hat{\mu}_k) - k$ is an unbiased estimate of the predictive accuracy of the model M. That is, $E^*[\text{AIC}(M)] = \text{AIC}^*(M)$, where the expectation is taken with respect to the true distribution. But we have just learned that

$$\text{AIC}(M) \sim -\tfrac{1}{2}\chi^2(n-k, \lambda^*) - \tfrac{1}{2}n\log 2\pi - k,$$

where the mean value of $\chi^2(n-k, \lambda^*)$ is $n - k + \lambda^*$. Therefore,

$$PA(M) = -\tfrac{1}{2}(n + k + \lambda^*) - \tfrac{1}{2}n\log 2\pi.$$

Observe that $PA(M)$ becomes infinitely large as the number of data, n, tends to infinity. It is not standard in statistics to estimate a quantity that grows infinitely large as the sample size increases, and this has led to considerable confusion about the properties of AIC (see [Forster, 2000; 2001], for detailed discussions). But, since this issue is not being discussed here, we shall refer to the *per datum* predictive accuracy when the need arises, and reserve PA to refer to the per datum predictive accuracy times the number of data n.

Therefore, our result is written

$$PA(M) = -C - \tfrac{1}{2}\lambda^* - \tfrac{1}{2}k,$$

where $C = \frac{1}{2}(\log 2\pi + 1)n \simeq 1.41894n$. So defined, predictive accuracy is the expected log-likelihood for newly sampled data that are generated by the true hypothesis, since the true hypothesis has no adjustable parameters and zero non-centrality parameter (the postulated mean values are the true values).[8] Thus, $-C$ is the highest predictive accuracy that could be achieved by any *point* hypothesis in the model.

The constant term is unimportant for our purposes because it is the same for every model. The other terms are negative because the predictive accuracy is expected log-likelihood, and the log-likelihood is negative whenever the likelihood is less than 1. Nevertheless, it is important to understand that higher predictive accuracies (less negative) are better than lower predictive accuracies (more negative).

Firstly, the higher the value of $\frac{1}{2}\lambda^*$, the lower the predictive accuracy. To interpret this fact, notice that $\frac{1}{2}\lambda^*/n$ is independent of the total number of data because λ^* grows proportionally to n. In our simple case, we could write

$$\lambda^*/n = \frac{n_1}{n}(\mu_1^* - \mu^*)^2 + \frac{n_2}{n}(\mu_2^* - \mu^*)^2 + \frac{n_3}{n}(\mu_3^* - \mu^*)^2,$$

so if the proportions n_1/n, n_2/n, and n_3/n are held fixed, then λ^*/n is fixed. Therefore $-\left(C + \frac{1}{2}\lambda^*\right)/n$ is the per datum predictive accuracy achieved by the

[8] Equivalently, if the non-centrality parameter is not zero, then the hypothesis is false.

model as the number of data tends to infinity. Alternatively, we could say that $-\left(C + \frac{1}{2}\lambda^*\right)$ is the predictive accuracy that could be obtained if the parameter estimation were free of error. It is convenient to refer to $\left(C + \frac{1}{2}\lambda^*\right)$ as the *bias of a model*. Bias is bad and predictive accuracy is good, so the smaller the bias of a model, the higher the predictive accuracy that it could potentially achieve.

For small data sets especially, the term $\frac{1}{2}k$ is important because it measures the effect of the sampling errors involved in estimating the parameters of the model. The fewer the adjustable parameters, the greater the amount of data that can be used to estimate the value of each parameter. The negative effect that estimation errors has on the predictive accuracy of the maximum likelihood hypothesis is referred to as the *variance* of the model. Variance is bad and predictive accuracy is good, so they have opposite signs.

In summary, the predictive accuracy of a model can be defined in terms of its bias and variance, thus

$$PA = -(\text{bias} + \text{variance}),$$

where bias and variance are positive quantities. The minus sign reflects the fact that both bias and variance are bad for predictive accuracy. Bias measures the predictive accuracy of the best hypothesis in the model (the hypothesis that would be selected if there were no variance) while variance measures the loss in predictive accuracy that comes about by estimation errors. Variance is reduced in simpler models, but the bias is greater. So, maximizing predictive accuracy (if that is the goal) involves a tradeoff between bias and variance.

The predictive accuracy of a model is unknown because the model bias is unknown, and that's because the non-centrality parameter of the generating distribution is unknown. But hypotheses about its value can be tested, its value can be estimated, and confidence intervals can be inferred.

It may seem that our work is done, but we also want to know how *differences* in AIC scores provide evidence about *differences* in predictive accuracy. This is achieved by showing that differences in the maximum log-likelihoods are also governed by a non-central chi-square distribution. Or at least, we can show this when the models are nested (i.e., when one model is a special case of the other). First, we need the following theorem (which is a special case of a well known theorem proved, for example, in Hogg and Craig, 1978, p. 279).

THEOREM 1. *Let $Q_1 = Q_2 + Q$, where Q_1, Q_2, and Q are 3 random variables that are quadratic forms in n mutually stochastically independent random variables that are each normally distributed with means $\mu_1^*, \mu_2^*, \ldots, \mu_n^*$ and unit variance. Let Q_1 and Q_2 have chi-square distributions with degrees of freedom r_1 and r_2, respectively, and let Q be non-negative. Then:*

(a) Q_2 and Q are mutually stochastically independent, and

(b) Q has a chi-square distribution with $r = r_1 - r_2$ degrees of freedom.

To apply the Theorem, note that

$$Q_2 = -2l(\hat{\mu}_1, \hat{\mu}_2) - n \log 2\pi = \sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (x_j - \hat{\mu}_2)^2$$

is a chi-squared random variable with $n-$ 2 degrees of freedom. Then note that

$$Q_1 = -2l(\hat{\mu}) - n \log 2\pi = \sum_{k=1}^{n} (x_k - \hat{\mu})^2$$

is a chi-squared random variable with $n-$ 1 degrees of freedom. A long, but straightforward, calculation shows that

$$\Delta l = l(\hat{\mu}_1, \hat{\mu}_2) - l(\hat{\mu}) = \frac{n_1}{2} (\hat{\mu} - \hat{\mu}_1)^2 + \frac{n_2}{2} (\hat{\mu} - \hat{\mu}_2)^2 .$$

From this, we prove that

$$Q_1 = Q_2 + 2\Delta l.$$

From the Theorem, it follows that $2\Delta l$ is a chi-square random variable with one degree of freedom, which has a non-centrality parameter, call it $\Delta\lambda$, found by substituting the mean values for each variable in the quadratic. That is,

$$\Delta\lambda = n_1 (\mu^* - \mu_1^*)^2 + n_2 (\mu^* - \mu_2^*)^2 ,$$

where, by definition,

$$\mu^* = \frac{n_1}{n} \mu_1^* + \frac{n_2}{n} \mu_2^*.$$

So, the evidential problem reduces to something simple. The statistic $2\Delta l = 2(l_2 - l_1)$ provides evidence about the predictive accuracy of M_2 compared to M_1. Recall that models with lower numbers of adjustable parameters have higher biases, which confirms that $\Delta\lambda$ has the same sign as Δbias. However, in order to keep the difference in the variances positive, we need to define

$$\Delta\text{variance} = \text{variance}\,(M_2) - \text{variance}\,(M_1) = k_2 - k_1 = \Delta k,$$

because the more complex model has a greater variance than the simpler model.

Now define the difference in predictive accuracy as the advantage that the complex model has over the simpler model. That is,

$$\Delta PA = PA(M_2) - PA(M_1).$$

Then the tradeoff between bias and variance is expressed by the formula

$$\Delta PA = \Delta\text{bias} \; - \; \Delta\text{variance} \; = \tfrac{1}{2}\Delta\lambda - \tfrac{1}{2}\Delta k.$$

Since the difference in AIC scores is straightforwardly calculated from Δl, it too provides evidence for hypotheses about the difference in predictive accuracy. All

we require is that any particular hypothesis about the value of ΔPA is associated with a particular chi-square distribution, which requires that we know the degrees of freedom and find the non-centrality parameter. The degree of freedom is Δk, which is the difference in the number of adjustable parameters, whereas the non-centrality parameter is given by

$$\Delta\lambda = \Delta k + 2\Delta PA.$$

From this, one can make statistical inferences about differences in predictive accuracy from differences in AIC scores, for we know that $2\Delta l \sim \chi^2(\Delta k, \Delta\lambda)$.

Note that, since $\Delta\lambda \geq 0$, $\Delta PA \geq -\Delta k/2$. That is, simpler models are limited in the advantage in predictive accuracy they can have over more complex models. More interestingly, the advantage is constrained to become closer and closer to zero as the number of data increases. Thus, in the large sample limit, there is no difference in predictive accuracy.

In summary, given any value of ΔPA, Δl is the value of a random variable with the distribution $\frac{1}{2}\chi^2(\Delta k, \Delta\lambda)$. So, ΔAIC is a random variable with the distribution $\frac{1}{2}\chi^2(\Delta k, \Delta k + 2\Delta PA) - \Delta k$, whose expected value is ΔPA. Or, in other words, ΔAIC is a random variable with a known distribution whose expected value is ΔPA, which is the content of Akaike's theorem. But this result goes far beyond Akaike's theorem by providing the probability distribution that one can use whatever method of statistical inference about ΔPA one may wish to deploy.

Here is a final remark about what it means for AIC to be an unbiased estimator of predictive accuracy. It means that if we fix ΔPA and resample the calibration data, we will get values of the statistic ΔAIC whose mean value is equal to ΔPA. It does not mean that for any value of ΔAIC, the mean value of ΔPA will be equal to ΔAIC. Not only would such a statement depend on the assignment of a Bayesian prior distribution over values of ΔPA, but in some cases there is no prior that could even make it true. To see this, suppose ΔPA has its lowest possible value of $-\frac{1}{2}\Delta k$. The statistic ΔAIC will be higher in value sometimes *and lower in value sometimes*, which means that sometimes it will be lower than the lowest possible value of ΔPA, even though statisticians will still say that ΔAIC is an unbiased estimator of ΔPA.

ACKNOWLEDGMENTS

We thank Jim Hawthorne and Prasanta Bandyopadhyay for useful suggestions.

BIBLIOGRAPHY

[Akaike, 1973] H. Akaike. 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, 1973.

[Forster, 2000] M. R. Forster. Key Concepts in Model Selection: Performance and Generalizability, *Journal of Mathematical Psychology* **44**: 205-231, 2000.

[Forster, 2001] M. R. Forster. The New Science of Simplicity. In A. Zellner, H. A. Keuzenkamp, and M. McAleer (eds.) *Simplicity, Inference and Modelling.* Cambridge University Press, pp. 83-119, 2001.

[Forster and Sober, 1994] M. R. Forster and E. Sober. 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, 1994.

[Hacking, 1965] I. Hacking. *The Logic of Statistical Inference.* Cambridge: Cambridge University Press, 1965.

[Howson and Urbach, 1993] C. Howson and P. Urbach. *Scientific Reasoning – the Bayesian Approach.* Peru, IL: Open Court, 1993.

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

[Sakamoto *et al.*, 1986] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics.* New York: Springer, 1986.

[Schwarz, 1978] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics* **6**: 461-465, 1978.