BAYES FACTORS*

by

James O. Berger
Purdue University

Technical Report # 96-9C

Department of Statistics
Purdue University
West Lafayette, IN USA

March 1996

* Prepared for the Encyclopedia of Statistical Sciences, and supported by the National Science Foundation under grant DMS - 9303556.

*

BAYES FACTORS

Bayes factors are the primary tool used by Bayesians for hypothesis testing and model selection. They also are used by non- Bayesians in the construction of test statistics. For instance, in the testing of simple hypotheses, the Bayes factor equals the ordinary likelihood ratio. BAYESIAN MODEL SELECTION discusses the use of Bayes factors in model selection. STATISTICAL EVIDENCE discusses the general use of Bayes factors in quantifying statistical evidence. To avoid overlap with these articles, we concentrate here on the motivation for using Bayes factors, particularly in hypothesis testing.

Suppose we are interested in testing two hypotheses:

$$H_1 : X \text{ has density } f_1(x|\theta_1) \quad \text{versus} \quad H_2 : X \text{ has density } f_2(x|\theta_2).$$

If the parameters θ_1 and θ_2 are unknown, a Bayesian generally specifies prior densities $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$ for these parameters, and then computes the Bayes factor of H_1 to H_2 as the ratio of the marginal densities of x under H_1 and H_2,

$$B = m_1(x)/m_2(x), \tag{1}$$

where

$$m_i(x) = \int f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i. \tag{2}$$

The Bayes factor is typically interpreted as the "odds provided by the data for H_1 to H_2," although this interpretation is strictly valid only if the hypotheses are simple; otherwise, B will also depend on the prior distributions. Note, however, that B does not depend on the prior probabilities of the hypotheses. If one wants a full Bayesian analysis, these prior probabilities, $P(H_1)$ and $P(H_2)$ (which, of course must sum to one), must also be specified. Then the posterior probability of H_1 is given by

$$P(H_1|x) = B/[B + (P(H_2)/P(H_1))] \tag{3}$$

(and, of course, the posterior probability of H_2 is just $1 - P(H_1|x)$.)

The attraction of using a Bayes factor to communicate evidence about hypotheses is precisely that it does not depend on $P(H_1)$ and $P(H_2)$, which can vary considerably among consumers of a study. Any such consumer can, if they wish, determine their own $P(H_1)$

and convert the reported Bayes factor to a posterior probability using (3). Although B will still depend on $\pi_1(\theta_1)$ and/or $\pi_2(\theta_2)$, the influence of these priors is typically less significant than the influence of $P(H_1)$ and $P(H_2)$. Also, default choices of $\pi_1(\theta_1)$ and $\pi_2(\theta_2)$ are sometimes available (see section 5), in which case B can be used as a default measure of evidence. General discussion of Bayes factors can be found in Jeffreys [29], Berger [2], Kass and Raftery [30], and Berger and Pericchi [11].

Example 1. Suppose we observe an i.i.d. normal sample, $X_1, X_2, \ldots, X_n$, from a $N(\theta, \sigma^2)$ distribution, with σ^2 known. It is desired to test $H_1 : \theta = \theta_0$ versus $H_2 : \theta \neq \theta_0$. For the prior distribution of θ under H_2, it is common to choose a $N(\theta_0, \tau^2)$ distribution, where the standard deviation τ is chosen to reflect the believed plausible range of θ if H_2 were true. (One could, of course, also choose prior means other than θ_0.)

A simple computation then shows that the Bayes factor of H_1 to H_2 is

$$B = (1 + \frac{n\tau^2}{\sigma^2})^{1/2} \exp\{-\frac{1}{2} z^2/(1 + \frac{\sigma^2}{n\tau^2})\}, \tag{4}$$

where $z = \sqrt{n}(\bar{x} - \theta_0)/\sigma$ is the usual standardized test statistic. A frequently used "default" choice of τ^2 is $\tau^2 = 2\sigma^2$ (the quartiles of the prior on θ are then roughly $\pm\sigma$), in which case

$$B = \sqrt{1 + 2n} \exp\{-z^2/(2 + \frac{1}{n})\}. \tag{5}$$

For instance, if $n = 20$ and $|z| = 1.96$, then $B = 0.983$. As this very nearly equals 1, which would correspond to equal odds for H_1 and H_2, the conclusion would be that the data provides essentially equal evidence for H_1 and H_2.

Note that $|z| = 1.96$ corresponds to a P-value of 0.05, which is typically considered to be significant evidence against H_1, in contradiction to the message conveyed by $B = 0.983$. This conflict between P-values and Bayes Factors is discussed further below. For now, it is interesting to note that the conflict magnifies in severity as $n \to \infty$ (or $\tau^2 \to \infty$) in (4). Indeed, for any fixed z, (4) then converges to ∞, so that B would indicate overwhelming evidence for H_1 even though z was *any* (fixed) large value (and the P-value was, correspondingly, *any* fixed small value). Various versions of this phenomenon have become known as Jeffreys's Paradox (Jeffreys [29]), Lindley's Paradox (Lindley [33]), and Bartlett's Paradox (Bartlett [1]). The "paradox" depends crucially on H_1 being a believable exact point null

hypothesis. While this may sometimes be true, a point null hypothesis is more typically an approximation to a small interval null, and the validity of the approximation disappears as $n \to \infty$ (but not as $\tau^2 \to \infty$); see Berger and Delampady [7] for discussion of this and its impact on the "paradox."

MOTIVATION FOR USING BAYES FACTORS

Since posterior probabilities are an integral component of Bayesian hypothesis testing and model selection, and since Bayes factors are directly related to posterior probabilities, their role in Bayesian analysis is indisputable. We concentrate here, therefore, on reasons why their use should be seriously considered by all statisticians.

Reason 1. Classical P-values can be highly misleading when testing precise hypotheses. This has been extensively discussed in Edwards, Lindeman, and Savage [20], Berger and Sellke [13], Berger and Delampady [7], and Berger and Mortera [8]. In Example 1, for instance, we saw that the P-value and the posterior probability of the null hypothesis could differ very substantially. To understand that the problem here is with the P-value, imagine that one faces a long series of tests of new drugs for AIDS. To fix our thinking, let us suppose that 50% of the drugs that will be tested have an effect on AIDS, and that 50% are ineffective. (One could make essentially the same point with any particular fraction of effective drugs.) Each drug is tested in an independent experiment, corresponding to a normal test of no effect, as in Example 1. (The experiments could all have different sample sizes and variances, however.) For each drug, the P-value is computed, and those with P-values smaller than 0.05 are deemed to be effective. (This is perhaps an unfair caricature of standard practice, but that is not relevant to the point we are trying to make about P-values.)

Suppose a doctor reads the results of the published studies, but feels confused about the meaning of P-values. (Let us even assume here that all studies are published, whether they obtain statistical significance or not; the real situation of publication selection bias only worsens the situation.) So the doctor asks the resident statistician to answer a simple question: "A number of these published studies have P-values that are between 0.04 and 0.05; of these, what fraction of the corresponding drugs are ineffective"?

The statistician cannot provide a firm answer to this question, but can provide useful bounds if the doctor is willing to postulate a prior opinion that a certain percentage of the drugs being originally tested (say, 50% as above) were ineffective. In particular, it is then the case that at least 23% of the drugs having P-values between 0.04 and 0.05 are ineffective, and in practice typically 50% or more will be ineffective (see Berger and Sellke [13]). Relating to this last number, the doctor concludes: "So if I start out believing that a certain percentage of the drugs will be ineffective, say 50%, then a P-value near 0.05 does not change my opinion much at all; I should still think that about 50% are ineffective".

This is essentially right, and this is essentially what the Bayes factor conveys. In Example 1 we saw that the Bayes factor is approximately one when the P-value is near 0.05 (for moderate sample sizes). And a Bayes factor of one roughly means that the data are equally supportive of the null and alternative hypotheses, so that posterior beliefs about the hypotheses will essentially equal the prior beliefs.

We cast the above discussion in a frequentist framework to emphasize that this is a fundamental fact about P-values; in situations such as that above, a P-value of 0.05 essentially does not provide any evidence against the null hypothesis. (Note, however, that the situation is quite different in situations where there is not a precise null hypothesis; then P-values and posterior probabilities often happen to be reasonably similar — see Casella and Berger [16].) That the meaning of P-values is commonly misinterpreted is hardly the fault of consumers of statistics. It is the fault of statisticians for providing a concept so ambiguous in meaning. The real point here is that the Bayes factor essentially conveys the right message easily and immediately.

Reason 2: Bayes factors are consistent for hypothesis testing and model selection. Consistency is a very basic property. It's meaning is that, if one of the entertained hypotheses (or entertained models) is actually true, then a statistical procedure should guarantee selection of the true hypothesis (or true model) if enough data is observed. Use of Bayes factors guarantees consistency (under very mild conditions), while use of most classical selection tools, such as P-values, C_p, and AIC, does not guarantee consistency (cf. Gelfand and Dey [25]).

In model selection it is sometimes argued that consistency is not a very relevant concept because no models being considered are likely to be exactly true. There are several possible counterarguments. The first is that, even though it is indeed typically important to recognize that entertained models are merely approximations, one should not use a procedure that fails the most basic property of consistency when you do happen to have the correct model under consideration. A second counterargument is based on the results of Berk [14] and Dmochowski [18]; they show that asymptotically (under mild conditions) use of the Bayes factor for model selection will choose the model that is closest to the true model in terms of Kullback-Leibler divergence. This is a rather amazing and compelling property of use of Bayes factors. It should be noted, however, that not all criteria support Bayes factors as optimal when the true model is not among those being considered; see Shibata [37] and Findley [23], for examples.

Reason 3. Bayes factors behave as automatic Ockham's razors, favoring simple models over more complex models, if the the data provides roughly comparable fits for the models. Overfitting is a continual problem in model selection, since more complex models will always provide a somewhat better fit to the data than will simple models. In classical statistics overfitting is avoided by introduction of an ad hoc penalty term (as in AIC), which increases as the complexity (i.e., the number of unknown parameters) of the model increases. Bayes factors act naturally to penalize model complexity, and hence need no ad hoc penalty terms. For an interesting historical example and general discussion and references, see Jefferys and Berger [28].

Reason 4. The Bayesian approach can easily be used for multiple hypotheses or models. Whereas classical testing has difficulty with more than two hypotheses, consideration of such poses no additional difficulty in the Bayesian approach. For instance, one can easily extend the Bayesian argument in Example 1 to test between $H_0 : \theta = 0, H_1 : \theta < 0$, and $H_2 : \theta > 0$.

Reason 5. The Bayesian approach does not require nested hypotheses or models, standard distributions, or regular asymptotics. Classical hypothesis testing has difficulty if the hypotheses are not nested or if the distributions are not standard. There are general classical approaches based on asymptotics, but the Bayesian approach does not require any of the assumptions under which an asymptotic analysis can be justified. Consider the following

example as an illustration of some of these notions.

Example 2. Suppose we observe an i.i.d. sample, $X_1, X_2, \ldots, X_n$, from either a normal or a Cauchy distribution f, and wish to test the hypotheses

$$H_1 : f \text{ is } N(\mu, \sigma^2) \text{ versus } H_2 : f \text{ is } C(\mu, \sigma^2).$$

This is awkward to do classically, as there is no natural test statistic and even no natural null hypothesis. (One can obtain very different answers depending on which test statistic is used and which hypothesis is considered to be the null hypothesis.) Also, computations of error probabilities are difficult, essentially requiring expensive simulations.

In contrast, there is a natural and standard automatic Bayesian test for such hypotheses. In fact, for comparison of any location-scale distributions, it is shown in Berger, Pericchi, and Varshavsky [12] that one can legitimately compute Bayes factors using the standard noninformative prior density $\pi(\mu, \sigma^2) = 1/\sigma^2$. For testing H_1 versus H_2 above, the resulting Bayes factor is available in closed form (see Franck [23] and Spiegelhalter [38]).

Reason 6. The Bayesian approach can account for model uncertainty and is often predictively optimal. Selecting a hypothesis or model on the basis of data, and then using the same data to estimate model parameters or make predictions based upon the model, is well known to yield (often severely) overoptimistic estimates of accuracy. In the classical approach it is often thus recommended to use part of the data to select a model and the remaining part of the data for estimation and prediction. When only limited data is available, this can be difficult.

The Bayesian approach takes a different tack: ideally, all models are left in the analysis with, say, prediction being done using a weighted average of the predictive distributions from each model, the weights being determined from the posterior probabilities (or Bayes factors) of each model. See Geisser [24] and Draper [19] for discussion and references.

Although keeping all models in the analysis is an ideal, this can be cumbersome for communication and descriptive purposes. If only one or two models receives substantial posterior probability, it would not be an egregious sin to eliminate the other models from consideration. Even if one must report only one model, the fact mentioned above, that Bayes factors act as a strong Ockham's razor, means that at least the selected model will

not be an overly complex model, and so estimates and predictions based on this model will not be quite so overly optimistic. Indeed, one can even establish certain formal optimality properties of selecting models on the basis of Bayes factors. Here is one such:

Result 1. Suppose it is of interest to predict a future observation Y under, say, a symmetric loss $L(|Y - \hat{Y}|)$, where L is nondecreasing. Assume that two models, M_1 and M_2, are under consideration for the data (present and future), that any unknown parameters are assigned proper prior distributions, and that the prior probabilities of M_1 and M_2 are both equal to $1/2$. Then the optimal model to use for predicting Y is M_1, or M_2, as the Bayes factor exceeds, or is less than, one.

Reason 7. Bayes factors seem to yield optimal conditional frequentist tests. The standard frequentist testing procedure, Neyman-Pearson testing, has the disadvantage of requiring the report of fixed error probabilities, no matter what the data. (The data-adaptive versions of such testing, namely P-values, are not true frequentist procedures and suffer from the rather severe interpretational problems discussed earlier.) In a recent surprising development (based on ideas of Kiefer [32]), Berger, Brown, and Wolpert [5] and Berger, Boukai, and Wang [6] show for simple versus simple testing and for testing a precise hypothesis, respectively, that tests based on Bayes factors (with, say, equal prior probabilities of the hypotheses) yield posterior probabilities which have direct interpretations as conditional frequentist error probabilities. Indeed, the posterior probability of H_1 is the conditional Type I frequentist error probability, and the posterior probability of H_2 is a type of average conditional Type II error probability (when the alternative is a composite hypothesis). Note that the reported error probabilities thus vary with the data, exactly as do the posterior probabilities. Another benefit that accrues is the fact that one can accept H_1 with a specified error probability (again data dependent).

The necessary technical detail to make this work is the defining of suitable conditioning sets upon which to compute the conditional error probabilities. These sets necessarily include data in both the acceptance and the rejection regions, and can roughly be described as the sets which include data points providing equivalent strength of evidence (in terms of Bayes factors) for and against H_1. Computation of these sets is, however, irrelevant to practical implementation of the procedure.

The primary limitation of this Bayesian - frequentist equivalence is that there will typically be a region, which is called the "no-decision region," in which frequentist and Bayesian interpretations are incompatible. Hence this region is excluded from the decision space. In Example 1, for instance, if the default $N(0, 2\sigma^2)$ prior is used and $n = 20$, then the no-decision region is the set of all points where the usual z-statistic is between 1.18 and 1.95. In all examples we have studied, the no-decision region is a region where both frequentists and Bayesian would feel indecisive, and hence its presence in the procedure is not detrimental from a practical perspective.

There are more surprises arising from this equivalence of Bayesian and conditional frequentist testing. One is that, in sequential testing using these tests, the stopping rule is largely irrelevant to the stated error probabilities. In contrast, with classical sequential testing the error probabilities depend very much on the stopping rule. For instance, consider a sequential clinical trial which is to involve up to 1000 patients. If one allows interim looks at the data (after, say, each 100 patients), with the possibility of stopping the experiment if the evidence appears to be conclusive at an interim stage, then the classical error probability will be substantially larger than if one had not allowed such interim analysis. Furthermore, computations of classical sequential error probabilities can be very formidable. It is thus a considerable surprise that the conditional frequentist tests mentioned above not only provide the freedom to perform interim analysis without penalty, but also are much simpler than the classical tests. The final surprise is that these conditional frequentist tests provide frequentist support for the stopping rule principle (see Berger and Berry [4]).

CARE IN SPECIFICATION OF HYPOTHESES

In classical statistics, whether one formulates a test as a one-sided or a two-sided test makes little practical difference; the alpha level or P-value changes by at most a factor of two. In Bayesian hypothesis testing, however, the difference between the formulations can lead to strikingly different answers, and so considerable care must be taken in formulation of the hypotheses. Let us begin with two examples.

Example 3. Suppose one is comparing a standard chemotherapy treatment with a new radiation treatment for cancer. There is little reason to suspect that the two treatments

could have the same effect, so that the correct test would be a one-sided test comparing the two treatments.

Example 4. Suppose two completely new treatments for AIDS are being compared. One should now be concerned with equality of treatment effects, because both treatments could easily have no (and hence equal) effect. Hence one should test the null hypothesis of no treatment difference against the alternative that there is a difference. (One might well actually formulate three hypotheses here, the null hypothesis of no difference, and the two one-sided hypotheses of each treatment being better; this is perfectly permissible and adds no real complications to the Bayesian analysis.)

The difference in Example 3 is that the standard chemotherapy treatment is presumably known to have a nonzero effect, and there is no reason to think that a radiation treatment would have (nearly) the same nonzero effect. Hence possible equality of treatment effects is not a real concern in Example 3. (In Bayesian terms, this event would be assigned a prior probability of essentially zero.) Note, however, that if the second treatment had, instead, been the same chemotherapy treatment, but now with (say) steroids added, then equality of treatments would have been a real possibility, since the steroids might well have no effect on the cancer.

Deciding whether or not to formulate the test as testing a precise hypothesis or as a one-sided test thus centers on the issue of deciding if there is a believable precise hypothesis. Sometimes this is easy, as in testing for the presence of extrasensory perception, or testing that a proposed law of physics holds. Often it is less clear; for instance, in medical testing scenarios it is often argued that any treatment will have some effect, even if only a very small effect, and so that exact equality of treatment effects will never occur. While perhaps true, it will still typically be reasonable to formulate the test as a test of no treatment difference, since such a test can be shown to be a good approximation to the "optimal" test unless the sample size is very large (cf. Berger and Delampady [7]).

Another aspect of this issue is that Bayesians cannot test precise hypotheses using confidence intervals. In classical statistics one frequently sees testing done by forming a confidence region for the parameter, and then rejecting a null value of the parameter if it does not lie in the confidence region. This is simply wrong if done in a Bayesian formulation (and if the

null value of the parameter is believable as an hypothesis).

USING BAYES FACTORS

Several important issues that arise in using Bayes factors will be discussed here.

Computation: Computation of Bayes factors can be difficult. A useful simple approximation is the Laplace approximation; see BAYESIAN MODEL SELECTION for its application to Bayes factors. The standard method of numerically computing Bayes factors has long been Monte-Carlo importance sampling. There have recently been a large variety of other proposals for computing Bayes factors; see BAYESIAN MODEL SELECTION and Kass and Raftery [30] for discussion and references.

Multiple hypotheses or models: When considering k (greater than two) hypotheses or models, Bayes factors are rather cumbersome as a communication device, since they involve only pairwise comparisons. There are two obvious solutions. One is just to report the marginal densities, $m_i(x)$, for all hypotheses and models. But since the scaling of these is arbitrary, it is typically more reasonable to report a scaled version, such as

$$P_i^* = m_i(x)/(\sum_{j=1}^{k} m_j(x)).$$

The P_i^* have the additional benefit of being interpretable as the posterior probabilities of the hypotheses or models, if one were to assume equal prior probabilities. Note that a consumer of such a report who has differing prior probabilities, $P(H_i)$, can compute his or her posterior probabilities as

$$P(H_i|x) = P_i^* P(H_i)/[\sum_{j=1}^{k} P_j^* P(H_j)].$$

Minimal statistical reports and decision theory: While Bayes factors do summarize the evidence for various hypotheses or models, they are obviously not complete summaries of the information from an experiment. In Example 1, for instance, along with the Bayes factor one would typically want to know the location of θ given that H_2 were true. Providing the posterior distribution of θ, conditional on the data and H_2 being true, would clearly suffice in this regard. A 'minimal' report would, perhaps, be a credible set for θ based on this posterior, along with the Bayes factor of course. (It is worth emphasizing that the credible set alone would not suffice as a report; the Bayes factor is needed to measure the strength of evidence against H_1.)

We should also emphasize that, often, it is best to approach hypothesis testing and model selection from the perspective of decision analysis (see Bernardo and Smith [15] for discussion of a variety of decision and utility based approaches to testing and model selection.) It should be noted that Bayes factors do not necessarily arise as components of such analyses. Frequently, however, the statistician's goal is not to perform a formal decision analysis, but to summarize information from a study in such a way that others can perform decision analyses (perhaps informally) based on this information. In this regard, the 'minimal' type of report discussed above will often suffice as the statistical summary needed for the decision-maker.

Updating Bayes factors: Full posterior distributions have the pleasant property of summarizing all available information, in the sense that if new, independent information becomes available, one can simply update the posterior with the new information through Bayes rule. The same is not generally true with Bayes factors. Updating Bayes factors in the presence of new information typically also requires knowledge of the full posterior distributions (or, at least, the original likelihoods). This should be kept in mind when reporting results.

DEFAULT BAYES FACTORS

Ideally, $\pi_1(\theta_1)$ and/or $\pi_2(\theta_2)$ are derived as subjective prior distributions. In hypothesis testing, especially nested hypothesis testing, there are strong reasons to do this. In Example 1, for instance, the Bayes factor clearly depends strongly on the prior variance, τ^2. Thus, at a minimum, one should typically specify this prior quantity (roughly the square of the prior guess as to the possible spread of θ if H_2 is true) to compute the Bayes factor. An attractive alternative for statistical communication is to present, say, a graph of the Bayes factor as a function of such key prior inputs, allowing consumers of a study to easily determine the Bayes factor corresponding to their personal prior beliefs (cf. Dickey [17] and Fan and Berger [21]).

Another possibility is to use robust Bayesian methods, presenting conclusions that are valid simultaneously for a large class of prior inputs. In Example 1, for instance, one can show that the lower bound on the Bayes factor over all possible τ^2, when $z = 1.96$, is 0.473. While this is a useful bound here, indicating that the evidence against H_1 is no more than 1 to 2, such bounds will not always answer the question. The problem is that the upper bounds

on the Bayes factor, in situations such as this, tend to be infinite, so that one may well be left with an indeterminate conclusion. (Of course, one might very reasonably apply robust Bayesian methods to a reduced class of possible prior inputs, and hope to obtain sensible upper and lower bounds on the Bayes factor.) Discussions of robust Bayesian methods in testing can be found in Edwards, Lindman, and Savage [20], Berger [2], Berger and Sellke [13], Berger and Delampady [7], and Berger [3].

Sometimes use of noninformative priors is reasonable for computing Bayes factors. One-sided testing provides one such example, where taking limits of symmetric proper priors as they become increasingly vague is reasonable and can be shown to give the same answer as use of noninformative priors (cf. Casella and Berger [16]). Another is non-nested testing, when the models are of essentially the same type and dimension. Example 2 above was of this type. See Berger, Pericchi, and Varshavsky [12] for discussion of other problems of this type.

In general, however, use of noninformative priors is not legitimate in hypothesis testing and model selection. In Example 1, for instance, the typical noninformative prior for θ (under H_2) is the constant density, but any constant could be used (since the prior is improper regardless) and the resulting Bayes factor would vary with the arbitrary choice of the constant. This is unfortunate, especially in model selection because, at the initial stages of model development and comparison, it is often not feasible to develop full subjective proper prior distributions. This has led to a variety of alternative proposals for the development of default Bayes factors. A few of these methods are briefly discussed below. For discussion of other methods and comparisons, see Berger and Pericchi [11] and Iwaki [27].

The most commonly used default procedure is the Bayes Information Criterion (BIC) of Schwarz [36], which arises from the Laplace approximation to Bayes factors. See BAYESIAN MODEL SELECTION for discussion. BIC is often a quite satisfactory approximation (cf. Kass and Wasserman [31]), but it avoids the problem of prior specification only by simply ignoring that term of the expansion.

Another common approach is to simply choose default proper prior distributions. Thus, in Example 1, we commented that the $N(0, 2\sigma^2)$ distribution is a standard default prior for this testing problem. Jeffreys [29] pioneered this approach (although he actually recommended a

$C(0, \sigma^2)$ default prior in Example 1); see also Zellner and Siow [40] and the many references to this approach in Berger and Pericchi [11].

An attempt to directly use noninformative priors, but with a plausible argument for choosing particular constant multiples of them when they are improper, was proposed for linear models in Spiegelhalter and Smith [39].

Two recent default approaches are the intrinsic Bayes factor approach of Berger and Pericchi [9, 10, 11] and the fractional Bayes factor approach of O'Hagan [34, 35]. These use, respectively, parts of the data ("training samples") or a fraction of the likelihood to, in a sense, create a default proper prior distribution. These approaches operate essentially automatically, and apply in great generality to hypothesis testing and model selection problems. And the better versions of these approaches can be shown to correspond to use of actual reasonable default proper prior distributions. They thus provide the best general default methods of testing and model selection that are currently available.

References

[1] Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, **44**, 533-534.

[2] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edition). Springer-Verlag, New York.

[3] Berger, J. (1994). An overview of robust Bayesian analysis. *Test*, **3**, 5–124.

[4] Berger, J. and Berry, D. (1988). The relevance of stopping rules in statistical inference. *Statistical Decision Theory and Related Topics IV*, S. S. Gupta and J. Berger (eds.), Springer-Verlag, New York.

[5] Berger, J., Brown, L. and Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *Annals of Statistics*, **22**, 1787–1807.

[6] Berger, J., Boukai, B., and Wang, Y. (1994). Unified frequentist and Bayesian testing of a precise hypothesis. Technical Report 94-25C, Purdue University, West Lafayette.

[7] Berger, J. and Delampady, M. (1987). Testing precise Hypotheses. *Statistical Science*, **3**, 317-352.

[8] Berger, J. and Mortera, J. (1991). Interpreting the stars in precise hypothesis testing. *International Statistical Review*, **59**, 337–353.

[9] Berger, J., and Pericchi, L. R. (1995). The intrinsic Bayes factor for linear models. *Bayesian Statistics 5.* J. M. Bernardo, et. al. (eds.), pp. 23-42, Oxford University Press, London.

[10] Berger, J. and Pericchi, L. R. (1995). On the justification of default and intrinsic Bayes factors. Technical Report 95–18C, Purdue University, West Lafayette.

[11] Berger, J., and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, March.

[12] Berger, J., Pericchi, L., and Varshavsky, J. (1995). An identity for linear and invariant models, with application to non-Gaussian model selection. Technical Report 95-7C, Purdue University, West Lafayette.

[13] Berger, J., and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P-values and evidence. *Journal of the American Statistical Association,* **82**, 112-122.

[14] Berk, R. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, **37**, 51–58.

[15] Bernardo, J. M., and Smith, A. F. M. (1994). *Bayesian Theory.* Wiley, New York.

[16] Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, **82**, 106-111.

[17] Dickey, J. (1973). Scientific reporting. *Journal of the Royal Statistical Society, B*, **35**, 285–305.

[18] Dmochowski, J. (1995). *Properties of Intrinsic Bayes Factors.* Ph.D. Thesis, Purdue University, West Lafayette.

[19] Draper, D. (1995). Assessment and propagation for model uncertainty (with discussion). *Journal of the Royal Statistical Society B*, **57**, 45-98.

[20] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review,* **70**, 193-242.

[21] Fan, T. H. and Berger, J. (1995). Robust Bayesian displays for standard inferences concerning a normal mean. Technical Report, Purdue University, West Lafayette.

[22] Findley, D. (1991). Counterexamples to parsimony and BIC. *Annals of the Institute of Statistical Mathematics*, **43**, 505–514.

[23] Franck, W. E. (1981). The most powerful invariant test of normal versus Cauchy with applications to stable alternatives. *Journal of the American Statistical Association*, **76**, 1002–1005.

[24] Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall, London.

[25] Gelfand, A., and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society*, Ser. B, **56**, 501-514.

[26] George, E. I., and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881-889.

[27] Iwaki, K. (1995). Posterior expected marginal likelihood for comparison of hypotheses. Technical Report, Purdue University, West Lafayette.

[28] Jefferys, W. and Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist,* **80**, 64–72.

[29] Jeffreys, H. (1961). *Theory of Probability* (3rd. Ed.). Clarendon Press, Oxford.

[30] Kass, R. E., and Raftery, A. (1995). Bayes factors and model uncertainty. *Journal of the American Statistical Association*, **90**, 773-795.

[31] Kass, R. E., and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association,* **90**, 928-934.

[32] Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, **72**, 789–827.

[33] Lindley, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187-192.

[34] O'Hagan, A. (1994). *Bayesian Inference*. Edward Arnold, London.

[35] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *Journal of the Royal Statistical Society*, Ser. B, **57**, 99-138.

[36] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

[37] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

[38.] Spiegelhalter, D. (1985). Exact Bayesian inference on the parameters of a Cauchy distribution with vague prior information. *Bayesian Statistics 2*, J. M. Bernardo, et. al. (eds.), pp. 743–750, North-Holland, Amsterdam.

[39] Spiegelhalter, D. J., and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society*, Ser. B, **44**, 377-387.

[40] Zellner, A., and Siow, A. (1980). Posterior odds for selected regression hypotheses. *Bayesian Statistics* 1, J. M. Bernardo et al. (eds.), pp. 585-603, Valencia University Press, Valencia.

(BAYESIAN MODEL SELECTION
STATISTICAL EVIDENCE
SCIENTIFIC METHODS AND STATISTICS
SCHWARZ CRITERION
PRINCIPLE OF PARSIMONY)

James O. Berger