Bayes Factors: What they are and what they are not

Michael Lavine and Mark J. Schervish

Bayes factors have been offered by Bayesians as alternatives to P-values (or significance probabilities) for testing hypotheses and for quantifying the degree to which observed data support or conflict with a hypothesis. Schervish (1996) showed how the interpretation of P-values as measures of support suffers a certain logical flaw. In this paper, we show how Bayes factors suffer that same flaw. We investigate the source of that problem and consider what are the appropriate interpretations of Bayes factors.

KEY WORDS: Measures of support; P-values.

1. INTRODUCTION

Consider tosses of a coin known to be either fair, two-headed or two-tailed. There are six nontrivial hypotheses about θ, the probability of heads:

$$\begin{array}{lll} H_1 : \theta = 1 & H_2 : \theta = 1/2 & H_3 : \theta = 0 \\ H_4 : \theta \neq 1 & H_5 : \theta \neq 1/2 & H_6 : \theta \neq 0 \end{array}$$

Jeffreys (1960) introduced a class of statistics for testing hypotheses that are now commonly called Bayes factors. The *Bayes factor* for comparing a hypothesis H to its complement, the alternative A, is the ratio of the posterior odds in favor of H to the prior odds in favor of H.

To make this more precise, let Ω be the parameter space and let $\Omega_H \subset \Omega$ be a proper subset. Let μ be a probability measure over Ω and for each $\theta \in \Omega$, let $f_{X|\Theta}(\cdot|\theta)$ be the density function (or probability mass function) for some observable X given $\Theta = \theta$. The predictive density of X given $H : \Theta \in \Omega_H$ is $f_H(x)$ equal to the average of $f_{X|\Theta}(x|\theta)$ with respect to μ restricted to Ω_H. Similarly, the predictive density of X given $A : \Theta \notin \Omega_H$ is $f_A(x)$ equal to the average of $f_{X|\Theta}(x|\theta)$ with respect to μ restricted to Ω_A. If p is the prior probability that H is true, then the posterior odds in favor of H is the ratio $pf_H(x)/[(1-p)f_A(x)]$. The Bayes factor is the ratio $f_H(x)/f_A(x)$.

Michael Lavine is Associate Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251. Mark J. Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.

Example 1. Consider four tosses of the coin mentioned earlier, and suppose they all land Heads. If the hypothesis of interest is $H_2 : \Theta = 1/2$, then

$$f_{H_2}(x) = \frac{1}{16} \qquad \text{and} \qquad f_{H_5}(x) = \frac{\mu(1)}{\mu(1) + \mu(0)},$$

where μ is the prior distribution. The Bayes factor in favor of H_2 is

$$\frac{f_{H_2}(x)}{f_{H_5}(x)} = \frac{\mu(1) + \mu(0)}{16\mu(1)}.$$

When a Bayesian tests a hypothesis H using a loss function that says the cost of type II error is constant over the alternative and the cost of type I error is constant over the hypothesis and is c times the cost of type II error, then the formal Bayes rule is to reject H if its posterior probability is less than $1/[1 + c]$. This is equivalent to rejecting H if the posterior odds in its favor are less than $1/c$. This in turn is equivalent to rejecting H if the Bayes factor in favor of H is less than some constant k implicitly determined by c and the prior odds.

It would seem then that a Bayesian could decline to specify prior odds, interpret the Bayes factor as "the weight of evidence from the data in favour of the ...model" (O'Hagan 1994, p. 191) and test a hypothesis "objectively" by rejecting H if the Bayes factor is less than some constant k. In fact, Schervish (1995, p. 221) says "The advantage of calculating a Bayes factor over the posterior odds ...is that one need not state a prior odds..." and then (p. 283) that Bayes factors are "ways to quantify the degree of support for a hypothesis in a data set." Robert (1994, p. 183), discussing hypothesis testing, says "the Bayes factor ...is still proposed as an 'objective' Bayesian answer ...."

What we show in this paper is that such a use of the Bayes factor has dangerous consequences that may not be readily apparent and that do not occur when using the posterior odds ratio to test H.

2. BAYES FACTORS ARE NOT MONOTONE IN THE HYPOTHESIS

Some Bayesians would like to use Bayes factors to measure the amount of support the data offer to the hypothesis without having to introduce the prior odds. Bertolino, Piccinato, and Racugno (1995, p. 213) say "Bayes factors are currently used as measures of evidence in favor of a given statistical hypothesis H_0 versus a general alternative H_A ..." and Kass and Raftery (1995, p. 777) say "The Bayes factor is a summary of the evidence provided by the data in favor of one scientific theory, ..., as opposed to another." We show, by example, that such informal use of Bayes factors suffers a certain logical flaw that is not suffered by using the posterior odds to measure support. The removal of the prior odds from the posterior odds to produce the Bayes factor has consequences that affect the interpretation of the resulting ratio.

Example 2. Consider the four coin tosses once again and let the prior distribution be

$$\mu(1) = .01, \qquad \mu(1/2) = .98 \quad \text{and} \quad \mu(0) = .01.$$

The six predictive probabilities are

$$\begin{array}{ccc} f_{H_1}(x) = 1, & f_{H_2}(x) = 1/16, & f_{H_3}(x) = 0, \\ f_{H_4}(x) \approx .0619, & f_{H_5}(x) = 1/2, & f_{H_6}(x) \approx .072. \end{array}$$

and the six nontrivial Bayes factors are

$$f_{H_1}(x)/f_{H_4}(x) \approx 16.16, \qquad f_{H_2}(x)/f_{H_5}(x) = 1/8, \qquad f_{H_3}(x)/f_{H_6}(x) = 0$$

and their inverses. Suppose that we use the Bayes factors to test the corresponding hypothesis. That is, we reject a hypothesis if the Bayes factor in its favor is less than some fixed number k. If we choose $k \in (.0619, 1/8)$ then we reject H_4 because $1/16.16 < k$ but accept H_2 because $1/8 > k$. That is, we face the apparent contradiction of accepting $\theta = 1/2$ but rejecting $\theta \in \{0, 1/2\}$. This problem does not arise if we choose to test the hypotheses by rejecting when the posterior odds is less than some number k'. The posterior odds in favor of H_2 is never more than the posterior odds in favor of H_4.

In Example 2, we were testing two hypotheses H_2 and H_4 such that H_2 implies H_4. Gabriel (1969) introduced a criterion for simultaneous tests of nested hypotheses. The tests of H_2 and H_4 are *coherent* if rejecting H_4 entails rejecting H_2. One typical use of a measure of support for hypotheses is to reject those hypotheses (that we want to test) that have small measures of support. We can translate the coherence condition into a requirement for any measure of support for hypotheses. Since any support for H_2 must a fortiori be support for H_4, the support for H_2 must be no greater than the support for H_4. Using the Bayes factor as a measure of support violates the coherence condition. Schervish (1996) showed that using P-values as measures of support also violates the coherence condition. Examples of coherent measures are the posterior probability, the posterior odds ratio, and various forms of the likelihood ratio test statistic

$$LR(H) = \frac{\sup_{\theta \in \Omega_H} f_{X|\Theta}(x|\theta)}{\sup_{\theta \in \Omega} f_{X|\Theta}(x|\theta)}, \qquad \text{and} \qquad LR'(H) = \frac{\sup_{\theta \in \Omega_H} f_{X|\Theta}(x|\theta)}{\sup_{\theta \in \Omega_A} f_{X|\Theta}(x|\theta)}. \tag{1}$$

The nonmonotonicity of Bayes factors is actually very general. Suppose that there are three nonempty, disjoint, and exhaustive hypotheses H_1, H_2, and H_3 as in Examples 1 and 2. Let H_4 be the complement of H_1 (the union of H_2 and H_3) as in the examples, so that H_2 implies H_4. Straightforward algebra shows that if $f_{H_3}(x) < \min\{f_{H_2}(x), f_{H_1}(x)\}$, then the Bayes factor in favor of H_4 will be smaller than the Bayes factor in favor of H_2 regardless of the prior probabilities of the three hypotheses H_1, H_2, and H_3. For instance, the nonmonotonicity will occur in Example 2 no matter what one chooses for the (strictly positive) prior distribution μ. What happens is that the Bayes factor penalizes H_4 for containing additional parameter values (those in H_3) that make the observed data less likely than all of the other hypotheses under consideration. An applied example of this phenomenon was encountered by Olson (1997) who was comparing three modes of inheritance in the species *Astilbe biternata*. All three modes are represented by simple hypotheses concerning the distribution of the observable data. One hypothesis H_1 is called tetrasomic inheritance, while the other two hypotheses H_2 and H_3 (those which happen to have the smallest and largest likelihoods) together form a meaningful category, disomic inheritance. The Bayes factor in favor of H_2 will be larger than the Bayes factor in favor of $H_2 \cup H_3$ no matter what strictly postive prior one places over the three hypotheses because H_3 has the smallest likelihood.

3. BAYES FACTORS ARE MEASURES OF CHANGE IN SUPPORT

The fact that Bayes factors are not coherent as measures of support does not mean that they are not useful summaries. It only means that one must be careful how one interprets them. What the Bayes factor actually measures is the *change* in the odds in favor of the hypothesis when going from the prior to the posterior. In fact, Bernardo and Smith (1994, p. 390) say "Intuitively, the Bayes factor provides a measure of whether the data $\boldsymbol{x}$ have increased or decreased the odds on H_i relative to H_j." In terms of log-odds, the posterior log-odds equals the prior log-odds plus the logarithm of the Bayes factor. So, for example, if one were to use log-odds to measure support (a coherent measure) then the logarithm of the Bayes factor would measure how much the data changes the support for the hypothesis.

Testing hypotheses by comparing Bayes factors to standard levels is similar to confusing $\Pr(A|B)$ with $\Pr(B|A)$. In Example 2, even though the Bayes factor $f_{H_4}(x)/f_{H_1}(x) = .0619$ is small the posterior odds $\Pr[H_4|x]/\Pr[H_1|x] = .99/.01 \times .0619 \approx 6.13$ is large and implies $\Pr[H_4|x] \approx .86$. The small Bayes factor says that the data will lower the probability of H_4 a large amount relative to where it starts (.99), but it does not imply that H_4 is unlikely.

4. WHY COHERENCE?

Is coherence a compelling criterion to require of a measure of support? Aside from the heuristic justification given earlier, there is a decision theoretic justification. As before, we assume that a typical application of a measure of support will be to reject hypotheses that have low support. Hence, we will justify coherence as a criterion for simultaneous tests. Consider the most general loss function L that is conducive to hypothesis testing. That is, let the action space have two points 0 and 1, where 0 means accept H and 1 means reject H, and let $H : \Theta \in \Omega_H$ be the hypothesis. We assume that $L(\theta, 0) > L(\theta, 1)$ for all $\theta \notin \Omega_H$ and $L(\theta, 0) < L(\theta, 1)$ for all $\theta \in \Omega_H$. This says that error is more costly than correct decision but otherwise places no restrictions on the loss. Now, suppose that we have two hypotheses $H_1 : \Theta \in \Omega_1$ and $H_2 : \Theta \in \Omega_2$ with corresponding loss functions L_1 and L_2 of the above form. For the simultaneous testing problem, we use loss $L(\theta, (a_1, a_2)) = L_1(\theta, a_1) + L_2(\theta, a_2)$, where a_i is the action for testing H_i for $i = 1, 2$. We impose one other condition, namely that for those parameters that are in both hypotheses or in both alternatives, the costs of error be the same in both testing problems. In symbols, this means that for all $\theta \in (\Omega_1 \cap \Omega_2) \cup (\Omega_1^C \cap \Omega_2^C)$, we have $L_1(\theta, 0) - L_1(\theta, 1) = L_2(\theta, 0) - L_2(\theta, 1)$. Under these conditions, we can prove two simple results. For non-Bayesians, incoherent tests are inadmissible. For Bayesians, incoherent tests are not formal Bayes rules.

Suppose that $\Omega_1 \subset \Omega_2$ and let ϕ_i be a test of H_i. That is $\phi_i(x) = 1$ means reject H_i and $\phi_i(x) = 0$ means accept H_i. Let $C = \{x : \phi_1(x) < \phi_2(x)\}$, that is C is the set where we make the incoherent decision to reject H_2 while accepting H_1. Suppose that there exists $\theta \in \Omega_2 \setminus \Omega_1$ with $P_\theta(C) > 0$. Let (ψ_1, ψ_2) be another pair of tests such that, for $i = 1, 2$, $\psi_i(x) = \phi_i(x)$ for all $x \notin C$ and $\psi_i(x) = \phi_{3-i}(x)$ for all $x \in C$. That is, ψ switches the two decisions when incoherence occurs in ϕ. We can now show that ψ dominates ϕ and that it has smaller posterior risk. To these ends, define

$$g(\theta) = L_1(\theta, 0) - L_1(\theta, 1) - L_2(\theta, 0) + L_2(\theta, 1),$$

so that $g(\theta) > 0$ for all $\theta \in \Omega_2 \setminus \Omega_1$. The risk function of ϕ minus the risk function of ψ is $g(\theta)$ times $P_\theta(C)$ for all $\theta \in \Omega_2 \setminus \Omega_1$. The two risk functions are equal elsewhere. Hence, ϕ is inadmissible. Similarly, the difference between the posterior risk of ϕ and the posterior risk of ψ is the integral of $g(\theta)$ over $\Omega_2 \setminus \Omega_1$ with respect to the posterior distribution. So long as the posterior risks are finite and $\Omega_2 \setminus \Omega_1$ has positive posterior probability, ϕ cannot be the formal Bayes rule.

5. DISCUSSION

Coherence is a property of two or more hypothesis tests considered jointly, but we can gain some insight into it by considering a single test on its own. When comparing two hypotheses it is useful to rephrase the question as *How well, relative to each other, do the hypotheses explain the data?* In the case of comparing two simple hypotheses, there is wide agreement on how this should be done. As Berger (1985, p. 146) points out, the Bayes factor is the same as the likelihood ratio LR' from (1) in this case. Also, in the case of two simple hypotheses, the P-value is just the probability in the tail of one of the distributions beyond the observed likelihood ratio, hence it is a monotone function of the Bayes factor. So, the Bayes factor and the P-value really can measure the support that the data offers for one simple hypothesis relative to another, and in a way that is acceptable to Bayesians and non-Bayesians alike. One should also note that coherence is not an issue in the case of two simple hypotheses because there do not exist two nonempty distinct nested hypotheses with nonempty complements. On the other hand, as we noted at the end of Section 3, just because the data lends a lot of support to a hypothesis H

relative to its complement does not necessarily make H more likely than its complement, it only makes H more likely than it was a priori.

When at least one of the hypotheses is composite, interpretations are not so simple. One might choose either to maximize, to sum, or to average over composite hypotheses. Users of the likelihood ratio statistic maximize: they find the value of θ within each hypothesis that best explains the data. Users of posterior probabilities sum: the posterior probability of a hypothesis is the sum (or integral) of the posterior probabilities of all the θ's within it. Users of Bayes factors average: the Bayes factor is the ratio of $f_{X|\Theta}(x|\theta)$ averaged with respect to the conditional prior given each hypothesis. But averaging has at least two potential drawbacks. First, it requires a prior to average with respect to, and second, it penalizes a hypothesis for containing values with small likelihood. As we noted at the end of Section 2, interpreting the Bayes factor as a measure of support is incoherent because of the second drawback.

REFERENCES

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Bernardo, J. and Smith, A. F. M. (1994), *Bayesian Theory*, New York: John Wiley.

Bertolino, F., Piccinato, L., and Racugno, W. (1995), "Multiple Bayes Factors for Testing Hypotheses," *Journal of the American Statistical Association*, 90, 213–219.

Gabriel, K. R. (1969), "Simultaneous Test Procedures—Some Theory of Multiple Comparisons," *The Annals of Mathematical Statistics*, 40, 224–250.

Jeffreys, H. (1960), *Theory of Probability* (3 ed.), Oxford: Carendon Press.

Kass, R. and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*, Cambridge: University Press.

Olson, M. (1997), "Application of Bayesian Analyses to discriminate between disomic and tetrasomic inheritance in *Astilbe biternata*," Technical report, Department of Botany, Duke University.

Robert, C. (1994), *The Bayesian Choice*, New York: Springer-Verlag.

Schervish, M. J. (1995), *Theory of Statistics*, New York: Springer-Verlag.

——— (1996), "P-values: What they are and what they are not," *The American Statistician*, 50, 203–206.