Bayes Factors: What They Are and What They Are Not

Author(s): Michael Lavine and Mark J. Schervish

Source: *The American Statistician*, May, 1999, Vol. 53, No. 2 (May, 1999), pp. 119-122

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2685729

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. In an earlier article, Schervish showed how the interpretation of P values as measures of support suffers a certain logical flaw. In this article, 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: Measure 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:

$$H_1 : \theta = 1 \quad H_2 : \theta = 1/2 \quad H_3 : \theta = 0$$
$$H_4 : \theta \neq 1 \quad H_5 : \theta \neq 1/2 \quad H_6 : \theta \neq 0.$$

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 (the complement of Ω_H). That is,

$$f_H(x) = \frac{\int_{\Omega_H} f_{X|\Theta}(x|\theta) d\mu(\theta)}{\mu(\Omega_H)}$$

and

$$f_A(x) = \frac{\int_{\Omega_A} f_{X|\Theta}(x|\theta) d\mu(\theta)}{\mu(\Omega_A)}.$$

If p is the prior probability that H is true—that is, $p = \mu(\Omega_H)$—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 $f_H(x)/f_A(x)$.

Example 1. Consider four tosses of the coin mentioned earlier, and suppose they all land heads. Let μ be the prior over the parameter space $\Omega = \{0, 1/2, 1\}$, where a point in Ω gives the probability of heads. If the hypothesis of interest is $H_2 : \Theta = 1/2$, then

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

The Bayes factor in favor of H_2 is

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

Suppose that a Bayesian observes data $X = x$ and tests a hypothesis H using a loss function that says the cost of type II error is some constant b over the alternative and the cost of type I error is constant over the hypothesis and is $c \times b$. The posterior expected cost of rejecting H is then $cb\Pr(H \text{ is true}|X = x)$, while the posterior expected cost of accepting H is $b(1 - \Pr(H \text{ is true}|X = x))$. The formal Bayes rule is to reject H if the cost of rejecting is smaller than the cost of accepting. This simplifies to rejecting H if its posterior probability is less than $1/[1 + c]$, which 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); "a summary of the evidence provided by the data in favor of one scientific theory . . . as opposed to another" (Kass and Raftery 1995, p. 777); or the "'odds for H_0 to H_1 that are *given by the data*' " (Berger 1985, p. 146) and test a hypothesis "objectively" by rejecting H if the Bayes factor is less than some constant k. In fact, Schervish (1995, p. 221) said "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." Of course, as these authors clarified, such an interpretation is not strictly justified. While the Bayes factor does not depend on the prior odds, it does depend on "how the prior mass is spread out over the two hypotheses" (Berger 1985, p. 146). Nonetheless, it sometimes happens that the Bayes factor "will be relatively insensitive to reasonable choices" (Berger 1985, p. 146), and then a common opinion would be that "such an interpretation is reasonable" (Berger 1985, p. 147).

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

Michael Lavine is Associate Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251 (Email: michael@stat.duke.edu). Mark J. Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.

of the prior odds from the posterior odds to produce the Bayes factor has consequences that affect the interpretation of the resulting ratio.

2. BAYES FACTORS ARE NOT MONOTONE IN THE HYPOTHESIS

Example 2. Consider once again the four coin tosses that all came up heads, let the parameter space be $\Omega = \{0, 1/2, 1\}$ (as in Example 1) and define a prior distribution μ by

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

The six predictive probabilities are

$$f_{H_1}(x) = 1, \quad f_{H_2}(x) = .0625, \quad f_{H_3}(x) = 0,$$
$$f_{H_4}(x) \approx .0619, \quad f_{H_5}(x) = .5, \quad f_{H_6}(x) \approx .072,$$

and the six nontrivial Bayes factors are

$$f_{H_1}(x)/f_{H_4}(x) \approx 16.16, \quad f_{H_2}(x)/f_{H_5}(x) = .125,$$
$$f_{H_3}(x)/f_{H_6}(x) = 0,$$

and their inverses. Suppose that we use the Bayes factors to test the corresponding hypotheses. That is, we reject a hypothesis if the Bayes factor in its favor is less than some fixed number k. If we choose $k \in (.0619, .125)$, then we reject H_4 because $1/16.16 < k$ but accept H_2 because $.125 > k$. That is, we face the apparent contradiction of accepting $\theta = .5$ but rejecting $\theta \in \{0, .5\}$. 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, and various forms of the likelihood ratio test statistic

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

The nonmonotonicity (incoherence) 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 largest and smallest likelihoods, respectively), 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 positive 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) said "Intuitively, the Bayes factor provides a measure of whether the data 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 change the support for the hypothesis.

Testing hypotheses by comparing Bayes factors to prespecified standard levels (like 3 or 1/3 to stand for 3-to-1 for or 1-to-3 against) 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 be-

Table 1. Partitions of the Set of Possible x Values by Two Pairs of Tests

	ϕ_2			ψ_2	
ϕ_1	0	1	ψ_1	0	1
0	F	C	0	F	$\emptyset$
1	E	D	1	$C \cup E$	D

fore, 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. The sample space is divided into four parts C, D, E, and F according to whether $(\phi_1(x), \phi_2(x))$ is (0,1), (1,1), (1,0), or (0,0), respectively. See the left side of Table 1. In particular, $C = \{x : (\phi_1(x), \phi_2(x)) = (0, 1)\}$ is the set where we make the incoherent decision to reject H_2 while accepting H_1. Create another pair (ψ_1, ψ_2) 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, $\psi = (\psi_1, \psi_2)$ switches the two decisions when incoherence occurs in $\phi = (\phi_1, \phi_2)$. Then the right side of Table 1 gives the sets where ψ_1 and ψ_2 take various pairs of values. Suppose that there exists $\theta \in \Omega_2 \setminus \Omega_1$ with $P_\theta(C) > 0$. We can now show that ψ dominates ϕ and that it has smaller posterior risk. The risk functions of the two pairs of tests are

$$\begin{aligned} R(\theta, \phi) &= L_1(\theta, 0)P_\theta(C \cup F) + L_2(\theta, 0)P_\theta(E \cup F) \\ &\quad + L_1(\theta, 1)P_\theta(D \cup E) \\ &\quad + L_2(\theta, 1)P_\theta(C \cup D), \\ R(\theta, \psi) &= L_1(\theta, 0)P_\theta(F) + L_2(\theta, 0)P_\theta(C \cup E \cup F) \\ &\quad + L_1(\theta, 1)P_\theta(C \cup D \cup E) \\ &\quad + L_2(\theta, 1)P_\theta(D). \end{aligned}$$

If we subtract these two we get $R(\theta, \phi) - R(\theta, \psi) = P_\theta(C)g(\theta)$, where

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

Our assumptions imply that $g(\theta) > 0$ for all $\theta \in \Omega_2 \setminus \Omega_1$ and it is 0 for all other θ. Since $P_\theta(C) > 0$ for some $\theta \in \Omega_2 \setminus \Omega_1$, ϕ is inadmissible. From the Bayesian perspective, if $x \in C$, the posterior risk of ϕ is $\int [L_1(\theta, 0) + L_2(\theta, 1)] d\mu_{\Theta|X}(\theta|x)$, where $\mu_{\Theta|X}$ is the posterior distribution. The posterior risk of ψ is $\int [L_1(\theta, 1) + L_2(\theta, 0)] d\mu_{\Theta|X}(\theta|x)$. The difference between these two posterior risks is easily seen to equal the integral of $g(\theta)$ with respect to the posterior distribution. If $x \notin C$, then the two rules make the same decision; hence, they have the same posterior risk. So long as the posterior risks are finite and $\Omega_2 \setminus \Omega_1$ has positive posterior probability, ϕ cannot be a formal Bayes rule.

5. DISCUSSION

Coherence is a property of tests of two or more nested hypotheses 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) pointed 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 increase the support for 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.

[Received January 1997. Revised September 1997.]

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: Wiley.

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

Jeffreys, H. (1960), *Theory of Probability* (3 ed.), Oxford: Clarendon 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, Duke University, Department of Botany.

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.