Bayes factors

A summary of

Robert E. Kass, Adrian E. Raftery (1995) ***Journal of the American Statistical Association.*** **90 (430): 791**

Jonas S. Juul

Imagine that you have some data, and several competing theories related to these, and want to quantify the evidence for each theory. In addition, imagine that you consider some of these theories to be more likely to be true than the others, before you collect your data. Or maybe imagine that you are mostly interested in the value of some variable, that can be estimated from each of the theories, but are unsure about how to estimate this using your normal approaches, because you do not want to discard either of the theories for good. In each case, *Bayes factors* turn out to be a very useful statistical tool to apply. In this summary of the 1995 review by Kass & Raftery, I will introduce the basic mindset and mathematics behind hypothesis testing using Bayes factors. After this, I will comment on some of the advantages and disadvantages the method has. Finally I will mention 2 examples of the application of Bayes factors, and illustrate what the advantage of the method was in each case.

Mindset and Math. Traditionally, hypothesis testing is the process of attempting to reject a null hypothesis. This usually involves formulating a null-hypothesis, choosing a level of significance, calculating a p-value, and rejecting the null-hypothesis if the p-value is smaller than the chosen level of significance. If a null-hypothesis fails to be rejected, it is accepted for the time being. The mindset of using Bayes factors is different. Instead of attempting to reject a theory as being true, one attempts to quantify the evidence *for* the theory. The method uses Bayes Theorem, and hence allows for prior probabilities to be incorporated. Kass & Raftery introduce the Bayes factor in the following way. Assume that we collect some data D, and assume that these occured as a result of either of the hypotheses H_1 or H_2. Assume that the prior probability of H_1 is $p(H_1) = 1 - p(H_2)$. Bayes theorem then tells us that the probability of H_i given data is

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{p(D|H_1)p(H_1) + p(D|H_2)p(H_2)}. \quad (1)$$

Taking the ratio $p(H_1|D)/p(H_2|D)$ gives

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)}\frac{p(H_1)}{p(H_2)} = B_{12}\frac{p(H_1)}{p(H_2)}. \quad (2)$$

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

TABLE I. Interpretation of Bayes factor $B_{10} = 1/B_{01}$

In words, this reads: Posterior odds $=$ Bayes factor $\times$ Prior odds. So the posterior odds can be obtained by multiplying the prior odds with the Bayes factor, and hence the Bayes factor indicates whether the data mostly support H_0 or H_1. The value of the Bayes factor $B_{10} = 1/B_{01}$ can be interpreted using Table I. Note that the table is adopted from the paper, but since the Bayes factor compares two theories, the second column could equally well be titled "Evidence for H_1, when compared to H_0".

One important difference between hypothesis testing using Bayes factors, and using maximum likelihood-methods occurs when theories include some unknown parameters. If H_i has parameters θ_i (a vector), $p(D|H_i)$ is obtained by the following integral

$$p(D|H_i) = \int p(D|\theta_i, H_i)p(\theta_i|H_i)d\theta_i. \quad (3)$$

So compared to maximum-likelihood methods, where the optimal parameters are found for observed data, Bayes Factors calculates the probability of the observed data given a theory $p(D|H_i)$, as it could be calculated before the data was taken. Importantly, notice that integrating over the parameters like this, effectively introduces a punishment for including extra parameters, similar to what is normally referred to as Occam's razor.

Advantages and Disadvantages. That models with less parameters are favored might be appealing to some people. Another thing that is important to consider when contemplating using Bayes factors, is the choice of priors. That the method requires prior distributions both for the hypotheses and parameters might both be a strength and a weakness. It is a strength, if one has, or can derive, meaningful priors. This is especially difficult for the parameter priors $p(\theta_i|H_i)$. Several methods can help with this, and are described in detail in the paper. If these prove too difficult, and if the sample size is large, one might avoid the integrals by using the "Schwarz criterion". In this case, one computes a value, S, which converges to the logarithm of the Bayes factor as the number of data points goes to infinity. Generally, the Bayes factor is sensitive to the choice of parameter priors, and the authors suggest using several different choices to gain knowledge on this sensitivity.

One important advantage of the Bayes factors is that they do not require one to choose between models. As mentioned above, the Bayes factors describe the evidence in favor of a given model. This can be exploited if one has competing theories, but is interested in knowing a value of an entity r, which can be found from each of these (e.g. a half-time, when one is not sure about the exact form a decay takes). If one has $n+1$ hypotheses,$H_1, \ldots, H_n$ that are all compared to a hypothesis H_0, and the corresponding Bayes factors are computed B_{i0}, then the posterior probability of H_i can be expressed

$$p(H_i|D) = B_{i0}\frac{P(H_i)}{P(H_0)} / \sum_{m=0}^{n} B_{m0}\frac{P(H_m)}{P(H_0)} \quad (4)$$

These can be used as weights in estimating the entity r

$$p(r|D) = \sum_{i=0}^{n} p(r|D, H_i)p(H_i|D) \quad (5)$$

From this we can obtain the expectation value or variance of r *without choosing a model definitely*! That we can work with multiple models, each weighted by its credibility, and thereby taking model uncertainty into account in computing e.g. the expectation value of r, is an advantage of the method of Bayes factors.

Two examples. As a final part of this summary, I will mention two examples that illustrate how the Bayes product may be applied, as presented by Kass & Raftery.

Ozone exceedances in Houston, TX. A high level of ozone near the ground indicates air pollution, and must be decreased. Ozone levels often exceeding a threshold level, had previously been reported in Houston, and now the measures taken to decrease ozone levels had to be evaluated. Given time-series data on observations of exceedances, three hypotheses were formulated: 1) No decrease; 2) Gradual decrease; 3) Abrupt decrease. Since 3) has a discontinuous likelihood function, normal frequentist methods fail to compare the hypotheses (or are very involved), while the Bayes factors turned out to be simple to calculate. The Bayes factors were $B_{10} = 0.02$, suggesting strong evidence against gradual decrease, $B_{20} = 2.75$, evidence against no decrease was "worth no more than a bare mention", and $B_{21} = 135$, very strong evidence for abrupt rather than gradual decrease. So *if* there was a decrease, it was probably abrupt. This was hypothesized to be due to improvements in measurement devices, that lead to higher accuracy and less extreme values measured. This was found to be in agreement with recent changes in technology. So Bayes factors proved much simpler to compute, and yielded a possibility to evaluate three different hypotheses against each other, and conclude that Houston, TX, had not implemented measures that lead to a decrease in ozone exceedances.

The second example concerns *E. coli*-bacteria showing *Acetat utilization deficiency* (AUD). This example is biologically involved, so I will exclude some detail, and describe the important statistical observations. A group of researchers hypothesized that AUD would occur due to a specific DNA repair mechanism. They observed that the consequence of this would be that two lines of cells (one selected for trait A, the other unselected for) would, surprisingly, have equal proportions of cells with AUD ($p_1 = p_2$). In the experiment, p_1 turned out to be roughly equal to p_2, but because their null-hypothesis was $p_1 = p_2$, the chi-square test that they used had no way to quantify the certainty with which $p_1 = p_2$. In addition, they had more data, that could provide statistical priors for another hypothesis. They formulated the hypotheses $H_0 : p_1 = p_2$, and H_1, an alternative hypothesis based on the additional cell lines. The Bayes factor turned out to be $B_{10} = 0.065$ corresponding to "positive evidence for H_0".

In summary, Bayes factors estimate the evidence *for* competing hypotheses. They include prior information (which can be good or bad), act as an Occam's razor, and allows one to compute quantities given several competing hypotheses.