J. R. Statist. Soc. B (2005)
67, *Part* 5, *pp.* 689–701

Bayes factors based on test statistics

Valen E. Johnson

University of Texas, Houston, USA

[Received May 2004. Final revision May 2005]

Summary. Traditionally, the use of Bayes factors has required the specification of proper prior distributions on model parameters that are implicit to both null and alternative hypotheses. I describe an approach to defining Bayes factors based on modelling test statistics. Because the distributions of test statistics do not depend on unknown model parameters, this approach eliminates much of the subjectivity that is normally associated with the definition of Bayes factors. For standard test statistics, including the χ^2-, F-, t- and z-statistics, the values of Bayes factors that result from this approach have simple, closed form expressions.

Keywords: Bayesian hypothesis tests; Model selection; Posterior model probabilities; Posterior odds; Strength of evidence

1. Introduction

Bayes factors are the corner-stone of Bayesian hypothesis testing (e.g. Jeffreys (1961)). In contrast with classical p-values, the value of a Bayes factor has a direct interpretation in terms of whether or not a hypothesis is true: it represents the factor by which data modify the prior odds of two hypotheses to give the posterior odds. Unfortunately, the values of Bayes factors often depend critically on the prior densities that are assigned to the model parameters that are inherent to null and alternative hypotheses. In addition, the calculation of Bayes factors usually involves the evaluation of high dimensional integrals. For this reason, Bayes factors are employed less frequently than they otherwise would be, although progress in developing methodology to reduce both the computational burden and the subjectivity of Bayes factors is proceeding rapidly.

The volume of research on Bayes factors makes it impractical to review here. Readers who are interested in a recent overview of this topic can consult Kass and Raftery (1995). Controversies surrounding the use of Bayes factors and comparisons with p-values are described by, among others, Edwards *et al.* (1963), Berger and Sellke (1987), Casella and Berger (1987), Efron and Gous (2001) and Sellke *et al.* (2001). Recent developments on the use of Bayes factors for model selection and objective interpretations are summarized in Rao and Wu (2001), Chipman *et al.* (2001), Berger and Pericchi (2001) and in accompanying discussion in Lahiri (2001).

In this paper, I propose a new approach towards defining Bayes factors. My approach eliminates much of the subjectivity that is associated with the definition of Bayes factors and drastically simplifies their computation. This innovation is achieved by modelling the sampling distributions of test statistics directly, rather than modelling the sampling distributions of

Address for correspondence: Valen E. Johnson, Department of Biostatistics and Applied Mathematics, M. D. Anderson Cancer Center, University of Texas, 1515 Holcombe Boulevard, Houston, TX 77030-4009, USA.
E-mail: vejohnson@mdanderson.org

 1369–7412/05/67689

individual observations. Because the distribution of a test statistic under the null hypothesis is completely specified—i.e. it does not depend on unknown parameters—no prior specification on model parameters is required under the null hypothesis. When the alternative hypothesis represents the negation of the null hypothesis, it is often possible to obtain a parsimonious parameterization of the distribution of the test statistic under a reasonably broad class of alternative models. In such cases, I show that bounds on the Bayes factor can be obtained by maximizing over the marginal likelihood of the data under the alternative hypothesis. (Upper bounds on Bayes factors against point null hypotheses have been discussed by Edwards *et al.* (1963), Dickey (1977), Good (1950, 1958, 1967, 1986), Berger (1985), Berger and Sellke (1987), Casella and Berger (1987), Berger and Delampady (1987) and Delampady and Berger (1990), among others.) For standard test statistics, including χ^2-, F-, t- and z-statistics, the marginal maximum likelihood estimate (MMLE) of parameters that are implicit to the alternative hypothesis can be determined analytically, leading to simple, closed form expressions for the associated Bayes factors.

The primary objection that might be raised to this formulation involves the manner in which statistical models for the raw data are circumvented. However, the practice of modelling transformations of data is not uncommon in statistics: analyses of principal components, binning data to intervals, modelling reconstructed images rather than raw image data and performing cluster analysis and statistical tests on processed probe cell intensities measured from gene chips are but a few of the many examples in which raw data are discarded to simplify subsequent analyses. Still, modelling test statistics rather than raw data is cause for concern and becomes an important issue if the test statistic that is selected to test a hypothesis does not capture most of the information that is contained in the data for that purpose. Of course, similar comments apply also to p-values. Whether this methodology is more useful than traditional Bayes factors in a particular application ultimately depends on whether the loss of information that is incurred by modelling the distribution of the test statistic is offset by the elimination of the requirement to specify prior distributions on model parameters when default or subjective choices for these priors are not available or are difficult to obtain.

There is also an issue of whether it is appropriate to estimate model parameters that are implicit to the alternative hypothesis by using MMLE. On the one hand, Bayes factors that are based on the MMLE result in the least conservative report of evidence against the null hypothesis (i.e. they provide the largest value of the Bayes factor in favour of the alternative). Bayes factors based on the MMLE are thus more favourable to the alternative hypothesis than the usual Bayes factors. On the other hand, use of the MMLE may be warranted if the parameter space over which the distribution of the test statistic has been marginalized under the alternative model is large—in that case the prior weight that is assigned to any particular data value by the alternative model will be relatively small. The MMLE also has the advantage of providing an objective report of the Bayes factor. In general, however, I prefer to set the value of the alternative model's dispersion parameter subjectively. Both approaches are explored in the examples that follow.

A more serious concern that stems from modelling the distribution of a test statistic rather than raw data involves the potential loss of coherence, i.e. the Bayes factor between, say, models A and B does not necessarily equal the Bayes factor between models A and C multiplied by the Bayes factor between models C and B. This is so because the test statistics that are used to compute these Bayes factors may represent different transformations of the data.

The extent to which coherency is violated by using 'similar' test statistics is not examined in this paper. Instead, I focus attention on the specific problem of testing null hypotheses against their negation.

2. χ^2-tests associated with multinomial data

2.1. Simple null hypotheses

To illustrate the essential ideas behind the use of test statistics to compute Bayes factors, consider Pearson's χ^2 goodness-of-fit statistic for testing a simple null hypothesis *versus* the negation of that hypothesis. Under the assumption of multinomial sampling, suppose that data have been binned into K predefined cells, and let $\mathbf{n}' = (n_1, \ldots, n_K)$ denote the observed frequencies in the K cells, with $n = \Sigma n_i$. Let $\mathbf{p}' = (p_1, \ldots, p_K)$, $p_i > c^2 > 0$ for $i = 1, \ldots, K$, denote the fixed multinomial probability vector of these cells under the null hypothesis. Under the alternative hypothesis, let $\mathbf{q}' = (q_1, \ldots, q_K)$ denote the multinomial probability for the cell counts and assume, *a priori*, that this vector is drawn from a distribution that concentrates its mass around $\mathbf{p}$. More specifically, letting $\boldsymbol{\mu} = \mathbf{q} - \mathbf{p}$, it is assumed that the random vector $\boldsymbol{\mu}$ is $O_p(1/\sqrt{n})$ under the alternative hypothesis, and that $\mathbf{p} = \mathbf{q}$ under the null hypothesis.

Also, let $\boldsymbol{\kappa}$ denote the random vector with components $\mu_i/\sqrt{p_i}$, and define

$$\mathbf{V}' = \left(\frac{n_1 - np_1}{\sqrt{(np_1)}}, \ldots, \frac{n_K - np_K}{\sqrt{(np_K)}} \right).$$

Under these assumptions, lemma 1 follows from the standard results on the distribution of quadratic forms. Here and for the remainder of the paper, I adopt notation that is similar to that used in Rao (1973). In addition, if $\mathbf{w}_n$ is a sequence of s-dimensional random variables and f_n is a sequence of positive numbers, then $\mathbf{w}_n = O_p(f_n)$ requires that for every $\varepsilon > 0$ there be an M_ε such that $\Pr(\|\mathbf{w}_n\| > M_\varepsilon f_n) < 1 - \varepsilon$, where $\|\cdot\|$ denotes the Euclidean norm. Furthermore, indexing of $\mathbf{q}$, $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ with sample size is suppressed to simplify the exposition. Proofs of lemmas 1 and 2 follow directly from theorems and results provided in, for example, Rao (1973).

Lemma 1. Suppose that $\mathbf{q}$ is a probability vector drawn from a distribution for which $\boldsymbol{\mu} = O_p(1/\sqrt{n})$ and that, given $\mathbf{q}$, $\mathbf{n}$ represents a draw from a multinomial distribution with probability $\mathbf{q}$ and denominator n. Then, under the conditions that were stated above, the asymptotic distribution of $x \equiv \mathbf{V}'\mathbf{V}$ is $\chi^2_{K-1}(n\boldsymbol{\kappa}'\boldsymbol{\kappa})$, that of a χ^2-distribution on $K - 1$ degrees of freedom and non-centrality parameter $n\boldsymbol{\kappa}'\boldsymbol{\kappa}$.

Of course, under the null hypothesis, the asymptotic distribution of x is χ^2_{K-1}, a central χ^2-distribution on $K - 1$ degrees of freedom.

At this point, several comments regarding the assumption that $\boldsymbol{\mu}$ is $O_p(1/\sqrt{n})$ are warranted. From a mathematical perspective, this assumption guarantees that x converges to a non-central χ^2 random variable under the alternative hypothesis as $n \to \infty$. For a given multinomial observation, this approximation is accurate when the variances of the elements of $\mathbf{V}$ are not too far from $1 - p_i$ and the covariance between elements of $\mathbf{V}$ is approximately $-\sqrt{(p_i p_j)}$. More precise approximations to the sampling distribution of x under a specified alternative hypothesis could be estimated by using numerical methods (and then used to compute a Bayes factor), but for the remainder of this section the asymptotic χ^2-approximation to the distribution of x under the alternative hypothesis is assumed to be adequate.

Substantively, the assumption that $\boldsymbol{\mu} = O_p(1/\sqrt{n})$ implies that the non-centrality parameter appearing in lemma 1 is $O_p(1)$. For moderate to moderately large sample sizes, this is the case of practical interest: larger deviations from the null hypothesis would make differentiation between the null and alternative hypotheses trivial, whereas differentiating between the null and alternative hypotheses under smaller order deviations is not practical. As Efron and Gous (2001) and Andrews (1994) pointed out, this is also the case in which the evidence that is contained in Bayes factors and p-values is most easily reconciled. Implications of this assumption for (very) large sample sizes are discussed in Section 5.

With these results in hand, the definition of the Bayes factor between the alternative and null hypotheses requires only the specification of a prior distribution on the non-centrality parameter of the χ^2-distribution under the alternative hypothesis. Suppose then that $\mathbf{q}$ is drawn from a Dirichlet distribution with parameter $c\mathbf{p}$. The constraint that $\boldsymbol{\mu} = O_p(1/\sqrt{n})$ requires that $c = O(n)$. Centring the distribution of $\mathbf{q}$ on $\mathbf{p}$ follows the general philosophy that was espoused by Jeffreys (1961) and subsequently used by many others, including in this context Albert (1990) and Delampady and Berger (1990). Accordingly, the value of a parameter in a vaguely specified alternative model is assumed to be distributed near its value under the null hypothesis for the simple reason that the null hypothesis would not be subjected to testing if it was not at least considered plausible.

Under these assumptions, the asymptotic distribution of $\boldsymbol{\kappa}'\boldsymbol{\kappa}$ is specified in lemma 2.

Lemma 2. Suppose that $\mathbf{q}$ is drawn from a Dirichlet distribution with parameter $c\mathbf{p}$, $c = O(n)$, and that, for the given value of $\mathbf{q}$, $\mathbf{n}$ represents a multinomial random variable with probability $\mathbf{q}$ and denominator n. Then the asymptotic distribution of $(1+c)\boldsymbol{\kappa}'\boldsymbol{\kappa}$ is χ^2_{K-1}, which is a central χ^2-distribution on $K-1$ degrees of freedom.

The probability density function of a non-central $\chi^2_s(\lambda)$ random variable y can be expressed as

$$f(y|s,\lambda) = \exp\left(-\frac{\lambda}{2}\right) \sum_{r=0}^{\infty} \frac{1}{r!\,\Gamma(r+s/2)} \left(\frac{\lambda}{2}\right)^r \left(\frac{1}{2}\right)^{r+s/2} y^{r+s/2-1} \exp\left(-\frac{y}{2}\right).$$

It follows that the conjugate prior density for the non-centrality parameter is a gamma distribution. If $z \equiv n\boldsymbol{\kappa}'\boldsymbol{\kappa}$, then, according to the prior model that is assumed for the non-centrality parameter under the alternative hypothesis, the marginal density of the χ^2-statistic x, say $m_a(x)$, under the alternative hypothesis can be expressed in closed form as

$$\begin{aligned} m_a(x) &= \int_0^{\infty} f(x|K-1,z)\, g\left(z \,\middle|\, \frac{K-1}{2}, \frac{1+c}{2n}\right) \mathrm{d}z \\ &= g\left\{x \,\middle|\, \frac{K-1}{2}, \frac{1+c}{2(1+c+n)}\right\}. \end{aligned} \tag{1}$$

Here, the function $g(\cdot|a,b)$ represents a gamma density with shape parameter a and scale parameter b.

Coupled with the simple form of the marginal density of x under the null hypothesis—a χ^2 probability density function—we can use equation (1) to express the Bayes factor in favour of the alternative hypothesis as

$$\begin{aligned} \text{Bayes factor} &= \frac{g\{x|(K-1)/2, (1+c)/2(1+c+n)\}}{g(x|(K-1)/2, \frac{1}{2})} \\ &= \left(\frac{1+c}{1+c+n}\right)^{(K-1)/2} \exp\left\{\frac{nx}{2(1+c+n)}\right\}. \end{aligned} \tag{2}$$

Recalling that $c = O(n)$ and letting $c = \alpha n - 1$, $\alpha > 1/n$, equation (2) can be rewritten as

$$\text{Bayes factor} = \left(\frac{\alpha}{\alpha+1}\right)^{(K-1)/2} \exp\left\{\frac{x}{2(\alpha+1)}\right\}. \tag{3}$$

Thus, the Bayes factor reduces to a function of a single nuisance parameter α.

Several approaches might be taken for setting the value of α. The most objective approach is to set α as its MMLE under the alternative hypothesis. Alternatively, we might consider a constrained marginal maximum likelihood estimation of α, in which a constraint on α is imposed to prevent too much prior mass from concentrating in the neighbourhood of **p**. Or a subjective view can be adopted and the value of α (or a prior distribution on α) can be specified on the basis of scientific considerations and available prior knowledge. For example, in most scientific investigations this specification can be based on equating the standard deviation of the components of **q** under the alternative hypothesis to what is considered to be a substantively important deviation from **p**. Such judgments are routinely made in designing clinical and other statistical trials and are, of course, required in statistical power calculations.

The value of α that maximizes the marginal density of the data under the alternative hypothesis is

$$\alpha = \frac{K-1}{x-(K-1)}, \tag{4}$$

provided that the χ^2-statistic x exceeds its expectation under the null hypothesis (i.e. $x > K-1$). At this value of α, the Bayes factor equals

$$\left(\frac{K-1}{x}\right)^{(K-1)/2} \exp\left\{\frac{x-(K-1)}{2}\right\}. \tag{5}$$

This value represents an upper bound on the weight of evidence against the null hypotheses and is explored further in Section 2.3. When $x < K-1$, the Bayes factor is 1, as it should be since there is then no evidence against the null hypothesis. As stated previously, this value is achieved by letting $\alpha \to \infty$, or when the alternative hypothesis concentrates its mass on **p**.

2.2. Composite hypotheses

Now consider a null hypothesis in which the multinomial cell probabilities represent functions of an s-dimensional parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $s < K-1$, i.e. assume that the multinomial cell probabilities $p_1(\boldsymbol{\theta}), \ldots, p_K(\boldsymbol{\theta})$ are specified functions of a parameter vector $\boldsymbol{\theta}$ with $\Sigma_i\, p_i(\boldsymbol{\theta}) = 1$, $p_i(\boldsymbol{\theta}) > c^2 > 0$ for all i, and let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$ (or another efficient estimator of $\boldsymbol{\theta}$ in the sense that was specified in Cramér (1946)). Suppose also that each $p_k(\boldsymbol{\theta})$ has continuous first and second partial derivatives with respect to each of the components of $\boldsymbol{\theta}$ and define **M** to be the $K \times s$ matrix of rank s having elements $\{p_i^{-1/2} \partial p_i / \partial \theta_j\}$.

To generalize the results of the previous section to composite hypotheses, it is necessary to assume that $\mathbf{p}(\hat{\boldsymbol{\theta}})$ is not too far from **q** when the alternative hypothesis is actually true. For this, let $\boldsymbol{\theta}_q$ denote the point in the s-dimensional space of $\boldsymbol{\theta}$ for which the Kullback–Leibler information between $\mathbf{p}(\boldsymbol{\theta})$ and the true (but unknown) value of **q** is maximized, and let $\boldsymbol{\mu} = \mathbf{q} - \mathbf{p}(\boldsymbol{\theta}_q)$. With this definition of $\boldsymbol{\theta}_q$, I assume that the prior distribution for **q** under the alternative hypothesis is specified so that $\boldsymbol{\mu} = O_p(1/\sqrt{n})$. If **V** is now redefined to represent the vector

$$\mathbf{V}' = \left(\frac{n_1 - n\, p_1(\hat{\boldsymbol{\theta}})}{\sqrt{\{n\, p_1(\hat{\boldsymbol{\theta}})\}}}, \ldots, \frac{n_K - n\, p_K(\hat{\boldsymbol{\theta}})}{\sqrt{\{n\, p_K(\hat{\boldsymbol{\theta}})\}}}\right),$$

then lemma 3 follows from theorem 1 of Mitra (1958).

Lemma 3. Suppose that **q** is a probability vector drawn from a distribution for which $\boldsymbol{\mu} = O_p(1/\sqrt{n})$ and that, given **q**, **n** denotes a draw from a multinomial distribution with probability **q** and denominator n. Then, under the conditions that were stated above, as $n \to \infty$, $\mathbf{V}'\mathbf{V}$ converges to a $\chi^2_{K-s-1}(n\boldsymbol{\kappa}'\boldsymbol{\kappa})$ distribution, where $\boldsymbol{\kappa}$ is the vector with components $\mu_i / \sqrt{p_i(\boldsymbol{\theta}_q)}$.

The asymptotic distribution of $\mathbf{V}'\mathbf{V}$ under the null hypothesis is χ^2_{K-s-1}.

Specifying an appropriate prior model for $\mathbf{q}$ under the alternative hypothesis is somewhat more complicated here than it was in the case of a simple null hypothesis. The difficulty arises from the constraint that $\mathbf{q}$ be 'close' to a probability vector satisfying the functional constraints $\mathbf{p}(\boldsymbol{\theta})$. However, a natural way to view this problem is to assume that $\mathbf{q}$ is generated from the following sampling procedure. First, a point $\mathbf{p}(\boldsymbol{\theta}^*)$ satisfying the constraints that are imposed by the null model is selected at random. (The prior distribution from which the given value of $\boldsymbol{\theta}^*$ is drawn is arbitrary and does not affect the asymptotic results that follow.) Under the alternative hypothesis, $\mathbf{q}$ is then drawn from a Dirichlet distribution with parameter $c\,\mathbf{p}(\boldsymbol{\theta}^*)$. For large c, the error term $\boldsymbol{\mu}$ can be written as

$$\boldsymbol{\mu} = \mathbf{q} - \mathbf{p}(\boldsymbol{\theta}_q) \stackrel{\text{a}}{=} (\mathbf{I} - \mathbf{M}\mathbf{J}^{-1}\mathbf{M}')(\mathbf{q} - \mathbf{p}(\boldsymbol{\theta}^*))$$

where $\mathbf{J} = \mathbf{M}'\mathbf{M}$. Here, $=^{\text{a}}$ denotes asymptotic equivalence. Under these assumptions, we obtain the following result.

Lemma 4. Suppose that $\mathbf{q}$ is drawn from a Dirichlet distribution with parameter $c\,\mathbf{p}(\boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ and $c = O(n)$. Given $\mathbf{q}$, let $\mathbf{n}$ denote a sample drawn from a multinomial distribution with probability $\mathbf{q}$ and denominator n. Then, under the assumptions that were stated above, if $\boldsymbol{\kappa}$ denotes the vector with components $\mu_i/\sqrt{p_i(\boldsymbol{\theta}_q)}$, the asymptotic distribution of $(1+c)\boldsymbol{\kappa}'\boldsymbol{\kappa}$ is χ^2_{K-s-1}.

The similarity of lemmas 3 and 4 to lemmas 1 and 2 implies that the results of Section 1 can be applied to composite hypotheses by simply substituting $K-s-1$ for $K-1$ in expressions (1)–(5) (when $x > K-s-1$).

2.3. Comparison of p-values and Bayes factors based on the same test statistic

In current standard statistical practice, the value of Pearson's χ^2-statistic is used to calculate a p-value against a null hypothesis. Usually, the null hypothesis is rejected when a p-value of less than 0.05 is observed. It is therefore of some interest to examine the probability that the null hypothesis is true (as calculated from expression (5) or (3)) when the p-value of the test just achieves its critical value of 0.05 and equal probability is assigned to each hypothesis *a priori*. Fig. 1 displays this probability as a function of the degrees of freedom of the χ^2 test statistic when α is replaced by its MMLE. Because the marginal density of the data under the alternative hypothesis has been maximized with respect to the parameter α, the probabilities that are displayed in Fig. 1 represent the smallest possible probability that could be assigned to the null hypothesis when the alternative hypothesis takes the form that is specified above. For 1 degree of freedom, the minimum probability that the null hypothesis is true is 0.32; at 100 degrees of freedom, the minimum probability that the null hypothesis is true is 0.22. It is interesting that these results are quite comparable with those which were obtained by Delampady and Berger (1990) for simple hypotheses. In that setting, they modelled the multinomial probability directly and considered two classes of priors on it. With equal prior probability assigned to both null and alternative hypotheses, their minimum posterior probability on the null hypothesis when the χ^2-test on 1 degree of freedom just achieved significance at the 0.05-level was 0.32 under a conjugate class of priors on $\mathbf{p}$ and was 0.29 under a unimodal symmetric class of prior densities.

2.3.1. A contingency table example

I now compare the Bayes factor that is based on the χ^2-statistic with more traditional Bayes factors in the context of testing independence of row and column classifications in contingency

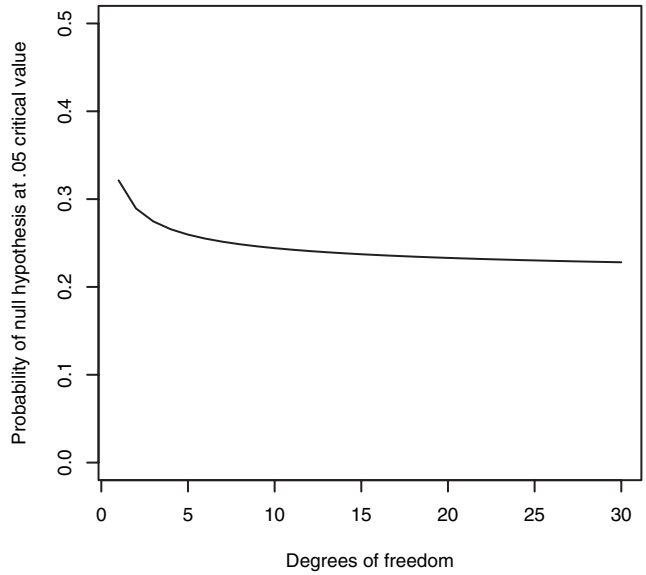

Fig. 1. Posterior probability that the null hypothesis is true when Pearson's χ^2-statistic is observed to equal its 0.95-quantile under the null hypothesis, equal prior probability is assigned to the null and alternative hypotheses, and α is set to its MMLE

Table 1. White and Eisenberg's (1959) classification of cancer patients

Site	*Results for the following blood groups:*		
	O	*A*	*B or AB*
Pylorus and antrum	104	140	52
Body and fundus	116	117	52
Cardia	28	39	11
Extensive	28	12	8

tables. Of course, the values of traditional Bayes factors depend on the prior densities that are assumed for the multinomial probability vector under the null and alternative models. I consider two prior specifications here. Both are based on priors that are approximately equivalent to the implicit assumption that is made on the alternative hypothesis assumed in the derivation of the Bayes factors above. The first, based on Albert (1990), uses a prior density for the multinomial probability under the alternative model that is 'concentrated about the "independence surface"'. The second, based on methodology that was described in Good and Crook (1987), employs a mixed Dirichlet prior with hyperparameter values that are determined by using empirical Bayes methodology.

The particular contingency table that is used for this illustration (Table 1) is taken from White and Eisenberg (1959) and is also discussed in Albert (1990). The data represent a cross-classification on cancer site and blood type for 707 patients with stomach cancer.

Pearson's χ^2-statistic for the test of independence for this table is 12.65 on 6 degrees of freedom. On the basis of expression (5), the Bayes factor against the independence model is 2.97.

The prior models underlying the computation of the Bayes factors that were proposed by Albert (1990) and Good and Crook (1987) are rather intricate, as are the methods for numerically evaluating them. For this reason, a detailed description of these methodologies is not presented here. Instead, only the details that are required for the replication of results are presented; interested readers should consult Albert (1990) and Good and Crook (1987) for more complete accounts.

The computation of the Bayes factor for independence under Albert's (1990) model requires the specification of a hyperparameter w, which (accepting Albert's recommendation) I take to be equal to 1. A second parameter K is used to control the dispersion of the multinomial probability vector around the independence surface under the alternative model. The maximum Bayes factor against independence in this formulation can be obtained by maximizing an approximation to Albert's Bayes factor with respect to K. Doing so leads to a Bayes factor against independence equal to 3.02.

To compute the Bayes factor under Good and Crook's (1987) model assumptions, a prior density is required on a hyperparameter k_0, and a second hyperparameter κ must be estimated. Following the estimation procedure that was suggested by Good and Crook leads to a maximum Bayes factor against the independence model of 3.06. This figure agrees with the Bayes factor that was obtained by using Albert's prior assumptions, suggesting some degree of robustness of Bayes factors obtained when this general approach towards specifying vague alternative models is adopted.

Both of these Bayes factors also agree with the Bayes factor that is based on the χ^2-statistic, suggesting that little information has been lost by modelling the distribution of the test statistic directly.

3. *F*-, *t*- and *z*-tests

Consider now the problem of testing the validity of a linear constraint on a regression parameter. Suppose that

$$\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

where $\mathbf{y}$ is an $n \times 1$ observation vector, $\boldsymbol{\beta}$ is an $r \times 1$ regression parameter, $\mathbf{X}$ is an $n \times r$ matrix of rank r and σ^2 is a scalar variance parameter. Assume further that, under the null hypothesis, $\mathbf{H}'\boldsymbol{\beta} = \xi$, where $\mathbf{H}$ is an $r \times k$ matrix of rank k whose range space is contained in the range space of $\mathbf{X}'$. As Rao (1973), page 191, noted, there then is a matrix $\mathbf{C}$ such that $\mathbf{H} = \mathbf{X}'\mathbf{X}\mathbf{C}$ where the rank of $\mathbf{X}\mathbf{C}$ is k.

If we define R_1^2 by

$$R_1^2 = \min\{(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\},$$

minimized over all $\boldsymbol{\beta}$ subject to the constraint $\mathbf{H}'\boldsymbol{\beta} = \xi$, and R_0^2 to be the corresponding minimum when $\boldsymbol{\beta}$ is unconstrained, then under the null hypothesis the quantity

$$f = \frac{(R_1^2 - R_0^2)/k}{R_0^2/(n-r)}$$

is distributed as $F_{k,n-r}$, a central F-distribution on $(k, n-r)$ degrees of freedom.

Now suppose that under the alternative hypothesis $\boldsymbol{\beta}$ is generated from the following mechanism. First, a value of the regression parameter satisfying the null hypothesis is selected. Denote

this value by $\boldsymbol{\beta}^*$. Next, $\boldsymbol{\beta}$ is drawn from an r-variate normal distribution centred on $\boldsymbol{\beta}^*$ and having covariance matrix $\tau\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Under this scheme for generating $\boldsymbol{\beta}$, the distribution of $\mathbf{H}'\boldsymbol{\beta}$ is normal with mean ξ and covariance matrix equal to $\tau\sigma^2\mathbf{H}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}$. Under both the null and the alternative hypotheses, the distribution of $R_1^2 - R_0^2$ is a scaled $\chi_k^2(\lambda)$ distribution, where the non-centrality parameter λ can be expressed as

$$\lambda = \sigma^{-2}(\mathbf{H}'\boldsymbol{\beta} - \xi)'(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}(\mathbf{H}'\boldsymbol{\beta} - \xi).$$

Under the alternative, it follows that λ/τ is distributed as a χ_k^2 random variable, and that the distribution of f given λ has a non-central F-distribution with density function

$$p(f|\lambda) = \left(\frac{k}{m}\right)^{k/2} \exp\left(-\frac{\lambda}{2}\right) \sum_{r=0}^{\infty} \left(\frac{k\lambda}{2m}\right)^r \frac{1}{r!} B\left(\frac{k}{2} + r, \frac{m}{2}\right) \frac{f^{r-1+k/2}}{\{1 + (k/m)f\}^{r+(k+m)/2}}.$$

In this equation, $m = n - r$ and

$$B(s,t) = \frac{\Gamma(s+t)}{\Gamma(s)\,\Gamma(t)}.$$

Marginalizing over λ, it can be shown that the distribution of $f/(1+\tau)$ under the alternative hypothesis has a central $F_{k,m}$-distribution.

For $f > 1$, the MMLE of τ based on the observed value of f under the alternative hypothesis is $\tau = f - 1$. At this value of τ, the marginal density of f is

$$p(f|\tau = f-1) = B\left(\frac{k}{2}, \frac{m}{2}\right)\left(\frac{k}{m}\right)^{k/2} \frac{1}{(1+k/m)^{(k+m)/2}} \frac{1}{f}. \tag{6}$$

It follows that the Bayes factor that is based on the MMLE of τ in favour of the alternative hypothesis for $f > 1$ is

$$\text{Bayes factor} = \left(\frac{m/k + f}{m/k + 1}\right)^{(k+m)/2} f^{-k/2}. \tag{7}$$

For large f, this Bayes factor is approximately $f^{m/2}$.

The case $k = 1$ is of particular interest as it corresponds to the t-test for a normal mean when the variance is unknown. In this case, the Bayes factor that is based on the MMLE of τ against the null reduces to

$$\left(\frac{m+f}{m+1}\right)^{(m+1)/2} f^{-1/2} \tag{8}$$

where $f = t^2$.

The one-sample z-statistic can be obtained from expression (8) by taking the limit as $m \to \infty$. Taking this limit, we find that the Bayes factor for testing the value of a normal mean is

$$\text{Bayes factor} = f^{-1/2} \exp\left(\frac{f-1}{2}\right). \tag{9}$$

This matches the result that was derived in Section 2 based on a χ_1^2-distribution.

3.1. Hald's data

The performance of Bayes factors based on F-test statistics can be illustrated by using Hald's

Table 2. Bayes factors for Hald's data†

Model	$BF_{\max}$	*AIBF1*	*AIBF2*	*ZS*	BF_9
1,2,3,*c*	1.0	0.29	0.29	0.3	0.32
1,2,4,*c*	1.0	0.26	0.26	0.3	0.32
1,3,4,*c*	1.0	0.31	0.32	0.36	0.40
2,3,4,*c*	1.99	1.2	1.2	1.11	1.75
1,2,*c*	1.0	0.18	0.19	0.26	0.23
1,3,*c*	36823	8242	15873	2439	2221
1,4,*c*	1.36	0.46	0.45	0.56	0.71
2,3,*c*	526	216	361	90.9	285
2,4,*c*	9415	2774	5071	833	1335
3,4,*c*	20.5	13.1	13.8	7.14	20.42
1,*c*	20643	4159	8531	3125	1997
2,*c*	5557	1910	3564	1176	1178
3,*c*	111508	22842	52084	11494	3318
4,*c*	5037	851	1705	1087	1126
c	235712	19722	37830	11236	4134

†The second column provides the Bayes factor that was obtained from the F-statistic for the submodel regression against the full model when τ is determined by its MMLE (7). The third and fourth columns provide the arithmetic intrinsic Bayes factors that are based on reference and modified Jeffreys priors respectively. The fifth column lists the Bayes factors that were proposed in Zellner and Siow (1980), and the sixth column the Bayes factor that was obtained from the F-statistic with $\tau = 9$. Values of AIBF1, AIBF2 and ZS are taken from Berger and Pericchi (1996).

regression data. This data set was discussed by Zellner (1984) and was used by Berger and Pericchi (1996) to compare intrinsic Bayes factors with Bayes factors calculated under model assumptions that were described by Zellner and Siow (1980).

There are four regressors in this data set. Following Berger and Pericchi (1996) we denote them by 1, 2, 3 and 4, and let c denote the constant term corresponding to the intercept. As Berger and Pericchi pointed out,

> 'this data set is somewhat extreme because of the very small sample size ($n = 13$) and because the design matrix is nearly singular'.

Berger and Pericchi calculated several versions of their intrinsic Bayes factor for the submodels that were obtained by including subsets of regressors in the normal regression model. They also provided a table in which these values were displayed next to Bayes factors obtained under Zellner and Siow's (1980) model assumptions. On the basis of these comparisons, Berger and Pericchi argued in favour of the use of arithmetic intrinsic Bayes factors based on either improper reference priors or modified Jeffreys priors.

Table 2 displays the intrinsic Bayes factors that were recommended by Berger and Pericchi (1996) and the Bayes factors that were produced under Zellner and Siow's (1980) model. Also displayed are Bayes factors that are based on the F-statistic. Two such Bayes factors are provided. The first represents the Bayes factor that is based on the MMLE of τ ($\mathrm{BF}_{\max}$); the second by fixing $\tau = 9$ (BF_9). This value of τ represents an assumption that the standard deviation of β under the alternative hypothesis is three times greater than the standard error of the least squares estimate of the regression parameter under the null hypothesis.

As expected, the values of BF_{max} that are displayed in Table 2 provide an upper bound for the remaining Bayes factors in the table. Loosely speaking, BF_{max} tends to be 2–3 times larger than the values of the arithmetic intrinsic Bayes factors that are based on the modified Jeffreys priors, and 4–6 times greater than the values of the arithmetic intrinsic Bayes factors that are based on the reference priors. This is not entirely unexpected since the values of BF_{max} represent the maximum Bayes factor against the null hypothesis based on the F-statistic, whereas the remaining Bayes factors have not been maximized in the same way (in contrast, all Bayes factors applied to White and Eisenberg's (1959) data represented the maxima). Interestingly, the multiplicative inverses of the values of BF_{max} are closest to the classical p-values.

There is relatively close agreement between Zellner and Siow's (1980) Bayes factors and the Bayes factor that is based on the F-statistic with $\tau = 9$.

4. Extensions to other test statistics

Conclusions from Section 2 can be extended to other χ^2-statistics, like the score test, likelihood ratio test and Wald's test, although the motivation for the probability models underlying the alternative hypotheses is less natural for those statistics than it is for Pearson's statistic. To see why, consider the score test. If the efficient score is denoted by $\mathbf{V}$ and the information matrix by $\mathbf{J}$, then the score statistic is $\mathbf{V}'\mathbf{J}^{-1}\mathbf{V}$. The most direct line of reasoning leading to a 'conjugate hypothesis', under which the score statistic has a non-central χ^2-distribution, is an assumption that the distribution of $\mathbf{V}$ under the alternative hypothesis is Gaussian with a non-zero mean, say λ, and covariance matrix $\mathbf{J}$. If λ is assumed to follow a Gaussian distribution, then the results of Section 2 can also be extended to the score statistic. However, the specification of an alternative probability model on the score vector itself, rather than on a parameter in a data model, seems less intuitive than the specification of a Dirichlet prior on a multinomial probability vector. Still, the specification of a scaled χ^2-distribution on the non-centrality parameter, with degrees of freedom equal to that of the test statistic, appears to work well for other χ^2-statistics and makes subsequent analyses tractable. As a 'conjugate' alternative, this approach seems to offer many advantages.

Bayes factors can be defined from test statistics in many small sample settings as well. Fisher's exact test provides a case in point. By conditioning on row and column totals in a 2×2 table, the counts in a contingency table are known to follow a (central) hypergeometric distribution. When the null hypothesis is false, the natural alternative model is that the counts follow a non-central hypergeometric distribution with a non-centrality parameter, say, ϕ. If ϕ is parameterized to represent the odds ratio, then it is natural to define a class of alternative models by assuming that $\log(\phi)$ is drawn from a symmetric distribution centred on 0 with scale parameter, say, σ. With such a definition of the alternative model, it is simple to maximize numerically the marginal likelihood of the data with respect to the scale parameter σ to obtain the Bayes factor of the test. And, of course, the use of Bayes factors in this context eliminates the necessity of determining which of several possible tail probabilities are relevant to the calculation of the p-value.

Fisher's tea-tasting experiment (Fisher, 1935) is perhaps the most famous example of the exact test for independence in contingency tables. In this experiment, a colleague of Fisher claimed to be able to distinguish whether tea was added to milk or milk to tea. After being told that four cups of tea had been prepared each way, she could correctly identify three of four cups of each preparation after tasting them in a randomized order. The resulting 2×2 table contained entries (3,1,1,3). The probability of this table according to a central hypergeometric distribution is 0.229. The only table that is more extreme is the table (4,0,0,4), corresponding

to all correct identifications. That table has probability 0.014, leading to a one-sided p-value of 0.243.

The Bayes factor against the alternative hypothesis, when $\log(\phi)$ is assumed to be drawn from an $N(0, \sigma^2)$ distribution and σ is estimated by marginal maximum likelihood estimation, is 1.11. Thus, there is some evidence against the null hypothesis but its posterior probability (assuming equal prior odds) is relatively high, equalling 0.47.

5. Summarizing remarks

By modelling the distribution of test statistics directly, Bayes factors can be computed in many standard problems without the full specification of subjective prior densities. Because the distribution of the test statistic does not involve unknown parameters, no prior densities are involved in the calculation of the marginal density of the data under the null hypothesis. Alternative models can often be defined in a natural way as the 'non-central' version of the test statistic's distribution under the null hypothesis; and doing so introduces a non-centrality parameter that must be modelled. For standard test statistics, however, a conjugate prior density or other convenient prior density for the non-centrality parameter is often apparent and typically involves only a single scale parameter. Marginalizing over the non-centrality parameter and maximizing with respect to this scale parameter leads to the MMLE of the density of the data under the alternative hypothesis, which in turn leads to what might be considered a default Bayes factor.

In this paper, I have focused on the calculation of Bayes factors that are based on 'moderately large' sample sizes. For truly large sample sizes, this approach can be modified to account for a 'non-point' null hypothesis if, in fact, the null hypothesis is not required to concentrate exactly on a single value of the parameter or to lie precisely in a specified subspace. Under such circumstances, the distribution of the test statistic under the null hypothesis might be more accurately modelled by using methodology that has previously been employed to obtain the distribution of the test statistic under the alternative hypothesis. For example, in the context of the F-tests that were cited in the previous section this modification could be achieved by assigning a relatively small value to τ to obtain the distribution of the test statistic under the null hypothesis, and assigning a larger value of τ to obtain the distribution of the F-statistic under the alternative hypothesis. Adopting such a strategy provides an escape from Lindley's paradox, though it would require careful elicitation of the dispersion of non-centrality parameters under both the null and the alternative hypotheses.

Acknowledgements

I thank Jianhua Hu, Peter Müller, two referees and the Associate Editor for numerous comments and suggestions that greatly improved the presentation of material in this paper.

References

Albert, J. H. (1990) A Bayesian test for a two-way contingency table using independence priors. *Can. J. Statist.*, **18**, 347–363.

Andrews, D. (1994) The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests. *Econometrica*, **62**, 1207–1232.

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

Berger, J. O. and Delampady, M. (1987) Testing precise hypotheses. *Statist. Sci.*, **2**, 317–335.

Berger, J. O. and Pericchi, L. R. (1996) The intrinsic Bayes factor for linear models. In *Bayesian Statistics 5*, pp. 25–44. Oxford: Clarendon.

Berger, J. O. and Pericchi, L. R. (2001) Objective Bayesian methods for model selection (with discussion). In *Model Selection* (ed. P. Lahiri), pp. 135–137. Beachwood: Institute of Mathematical Statistics.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. *J. Am. Statist. Ass.*, **82**, 112–122.
Casella, G. and Berger, R. L. (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *J. Am. Statist. Ass.*, **82**, 106–111.
Chipman, H., George, E. I. and McCulloch, R. (2001) The practical implementation of Bayesian model selection (with discussion). In *Model Selection* (ed. P. Lahiri), pp. 65–134. Beachwood: Institute of Mathematical Statistics.
Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton: Princeton University Press.
Delampady, M. and Berger, J. O. (1990) Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *Ann. Statist.*, **18**, 1295–1316.
Dickey, J. M. (1977) Is the tail area useful as an approximate Bayes factor? *J. Am. Statist. Ass.*, **72**, 138–142.
Edwards, W., Lindman, H. and Savage, L. J. (1963) Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**, 193–242.
Efron, B. and Gous, A. (2001) Scales of evidence for model selection: Fisher versus Bayes (with discussion). In *Model Selection* (ed. P. Lahiri), pp. 208–256. Beachwood: Institute of Mathematical Statistics.
Fisher, R. A. (1935) *Design of Experiments*. Edinburgh: Oliver and Boyd.
Good, I. J. (1950) *Probability and the Weighting of Evidence*. London: Griffin.
Good, I. J. (1958) Significance tests in parallel and in series. *J. Am. Statist. Ass.*, **53**, 799–813.
Good, I. J. (1967) A Bayesian significance test for multinomial distributions (with discussion). *J. R. Statist. Soc.* B, **29**, 399–431.
Good, I. J. (1986) The maximum of a Bayes factor against 'independence' in a contingency table, and generalizations to higher dimensions. *J. Statist. Computn Simuln*, **26**, 312–316.
Good, I. J. and Crook, J. F. (1987) The robustness and sensitivity of the mixed-Dirichlet Bayesian test for 'independence' in contingency tables. *Ann. Statist.*, **15**, 670–693.
Jeffreys, H. (1961) *Theory of Probability*. Oxford: Oxford University Press.
Kass, R. E. and Raftery, A. E. (1995) Bayes Factors. *J. Am. Statist. Ass.*, **90**, 773–795.
Lahiri, P. (ed.) (2001) *Model Selection*. Beachwood: Institute of Mathematical Statistics.
Mitra, S. K. (1958) On the limiting power function of the frequency chi-squared test. *Ann. Math. Statist.*, **29**, 1221–1233.
Sellke, T., Bayarri, M. J. and Berger, J. O. (2001) Calibration of p values for testing precise null hypotheses. *Am. Statistn*, **55**, 62–71.
Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*. New York: Wiley.
Rao, C. R. and Wu, Y. (2001) On model selection (with discussion). In *Model Selection* (ed. P. Lahiri), pp. 1–64. Beachwood: Institute of Mathematical Statistics.
White, C. and Eisenberg, H. (1959) ABO blood groups and cancer of the stomach. *Yale J. Biol. Med.*, **32**, 58–61.
Zellner, A. (1984) *Basic Issues in Econometrics*. Chicago: University of Chicago Press.
Zellner, A. and Siow, A. (1980) Posterior odds ratio for selected regression hypothesis. In *Bayesian Statistics*, pp. 585–603. Valencia: Valencia University Press.