BAYES FACTORS BASED ON TEST STATISTICS

VALEN JOHNSON

Abstract. Traditionally, the use of Bayes factors has required the specification of proper prior distributions on model parameters implicit to both null and alternative hypotheses. In this paper, I describe an approach to defining Bayes factors based on modeling test statistics. Because the distributions of test statistics do not depend on unknown model parameters, this approach eliminates the subjectivity 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 can be simply expressed in closed form.

1. Introduction

Bayes factors are the cornerstone of Bayesian hypothesis testing (e.g., Jeffreys 1961). In contrast to 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 assigned to the model parameters 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 such Bayes factors makes it impractical

to review here, but readers interested in a comprehensive review of this methodology can consult Kass and Raftery (1995). Readers more interested in controversies and comparisons of p values to Bayes factors might consult, among many other references, Edwards et al (1963), Berger and Sellke (1989), or Sellke, Bayarri, and Berger (2001).

In this article, I propose a new approach towards defining Bayes factors. This approach eliminates much of the subjectivity associated with their definition and drastically simplifies their computation. This simplification is achieved by modeling the sampling distributions of test statistics instead of the sampling distribution of individual observations. Because the distribution of a test statistic under the null hypothesis is completely specified–that is, it does not depend on unknown parameters–no prior specification on model parameters is required. In cases for which the alternative hypothesis is only vaguely specified, this approach often leads to a convenient and parsimonious parameterization of the distribution of the test statistic under a reasonably broad class of alternative models. In such cases, I show that minimum bounds on the Bayes factor in favor of the null hypotheses can be determined by maximizing over the marginal likelihood of the data under the alternative hypothesis (see also, Good 1986, who explores maximization over Bayes factors in more traditional settings). For standard test statistics, including χ^2, F, t and z statistics, maximization of the marginal likelihood under the alternative hypothesis can often be achieved analytically, leading to simple, closed form expressions for the resulting Bayes factors.

2. χ^2 Tests Associated with Multinomial Data

2.1. **Simple Null Hypotheses versus Vague Alternatives.** 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. Let $\mathbf{p}' = (p_1, \ldots, p_K)$ denote the probabilities of these cells under the null hypothesis, and let $\mathbf{q}' = (q_1, \ldots, q_K)$ denote the multinomial probability vector under the alternative hypothesis. Define $\boldsymbol{\mu} = \{p_i - q_i\}$ and assume that the elements of $\boldsymbol{\mu}$, $\{\mu_i\}$, are $O_p(1/\sqrt{n})$, $n = \sum n_i$. From a practical perspective, this is the case of primary interest, as it is neither feasible to detect smaller deviations from the null as the sample size becomes large, nor is it difficult to detect larger deviations. Let $\boldsymbol{\kappa}$ denote the 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 standard results on the distribution of quadratic forms. Here and for the remainder of the article, I adopt notation similar to that used in Rao (1973). Proofs of lemmas follow directly from theorems and results provided there.

Lemma 1. *Under the alternative hypothesis, the asymptotic distribution of* $x \equiv \mathbf{V}'\mathbf{V}$ *is* $\chi^2_{K-1}(n\boldsymbol{\kappa}'\boldsymbol{\kappa})$*, a* χ^2 *distribution on* $K-1$ *degrees of freedom and noncentrality 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.

Because the distribution of x under the null hypothesis is completely specified, we need only specify a prior distribution on the non-centrality parameter that appears in the χ^2 distribution under the alternative hypothesis in order to calculate a Bayes factor between the two models.

To motivate a model for the non-centrality parameter $n\boldsymbol{\kappa}'\boldsymbol{\kappa}$, assume that under the alternative hypothesis the probability vector $\mathbf{q}$ is drawn from a Dirichlet distribution with parameter $c\,\mathbf{p}$. That is, the prior mean of $\mathbf{q}$ is $\mathbf{p}$ and the variance of the components of $\mathbf{q}$ is inversely proportional to $c+1$. To maintain the constraint that $\boldsymbol{\mu} = O_p(1/\sqrt{n})$, assume also that $c = O(n)$. This assumption follows the general philosophy espoused by Jeffreys (1961) and subsequently used by many others, including, in this context, Albert (1990). According to it, the value of a model 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. *For large c, the distribution of $(1+c)\boldsymbol{\kappa}'\boldsymbol{\kappa}$ is χ^2_{K-1}, a central χ^2 distribution on $K-1$ degrees of freedom.*

This result does not rely heavily on the assumption that the true probability vector $\mathbf{q}$ is drawn from a Dirichlet distribution; that assumption is made only to facilitate the conceptual modeling of $\mathbf{q}$ in what follows. Other distributions that

approach a multivariate normal distribution for large values of their parameter and having the same first and second order moments lead to the same result.

With these facts in hand, the strategy for defining a Bayes factor in this context can be summarized as follows. The null hypothesis that the multinomial probability is equal to $\mathbf{p}$ has been operationalized by recasting the null hypothesis as the statement that x is distributed as a χ^2 random variable on $K-1$ degrees of freedom. The alternative hypothesis that the multinomial probability vector is not equal to $\mathbf{p}$ has been recast as the statement that x is distributed as a non-central χ^2 random variable on $K-1$ degrees of freedom. Finally, by assuming that the distribution of the multinomial probability vector under the alternative hypothesis is distributed around $\mathbf{p}$ with a Dirichlet distribution, the asymptotic distribution of the non-centrality parameter of the alternative's non-central χ^2 distribution is found to be distributed as a scaled version of a central χ^2 distribution.

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

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

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 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

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

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 chi-squared probability density function–we can use (1) to express the Bayes factor between the null and alternative hypothesis as

$$
\begin{aligned}
\text{Bayes factor} &= \frac{g\left(x \;\middle|\; \frac{K-1}{2}, \frac{1}{2}\right)}{g\left[x \;\middle|\; \frac{K-1}{2}, \frac{1+c}{2(1+c+n)}\right]} \\
&= \left(\frac{1+c+n}{1+c}\right)^{\frac{K-1}{2}} \exp\left[\frac{-nx}{2(1+c+n)}\right] \qquad (2)
\end{aligned}
$$

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

$$
\text{Bayes factor} = \left(\frac{\alpha+1}{\alpha}\right)^{\frac{K-1}{2}} \exp\left[\frac{-x}{2(\alpha+1)}\right] \qquad (3)
$$

When the chi-squared statistic x exceeds its expectation under the null hypothesis (i.e., when $x > K-1$), the value of α that maximizes the marginal density of the data under the alternative hypothesis (or equivalently, the value of α that minimizes of the Bayes factor of M_0 to M_a) is

$$
\alpha = \frac{K-1}{x-(K-1)}. \qquad (4)
$$

At this value of α, the Bayes factor equals

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

This value represents a lower bound on the weight of evidence in favor of the null hypotheses and is explored further in Section 2.2. When $x < K-1$, the minimum value of the Bayes factor is 1, and this value is achieved by letting α become large (i.e., when the alternative hypothesis concentrates its mass near $\mathbf{p}$).

2.2. **Composite Hypotheses.** Next, consider a null hypothesis in which the multinomial cell probabilities represent functions of a s-dimensional parameter vector $\boldsymbol{\theta}$, where $s < K-1$. That is, assume that the multinomial cell probabilities $p_1(\boldsymbol{\theta}), \ldots, p_K(\boldsymbol{\theta})$ are specified functions of a parameter vector $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$ (or another efficient estimator of $\boldsymbol{\theta}$ in the sense specified in Cramér (1946)). Suppose also that each $p_k(\boldsymbol{\theta})$ possesses continuous first partial derivatives with respect to each of the components of $\boldsymbol{\theta}$ and define $\mathbf{M}$ to be the $(K \times s)$ matrix of rank s having elements $\{p_i^{-1/2}\partial p_i/\partial\theta_j\}$. Let $\boldsymbol{\theta}_0$ denote the point in the s-dimensional space of $\boldsymbol{\theta}$ for which the Kullback-Leibler information between $\mathbf{p}(\boldsymbol{\theta})$ and $\mathbf{q}$, the true value of the multinomial probability vector under the alternative hypothesis, is maximized. The Kullback-Leibler information is defined at any value of $\boldsymbol{\theta}$ by

$$\mathbf{E}\left[\log\left(\frac{\mathbf{p}(\boldsymbol{\theta})}{\mathbf{q}}\right)\right] = \int \log\left(\frac{\mathbf{p}(\boldsymbol{\theta})}{\mathbf{q}}\right)\mathbf{q}\, d\mathbf{q},$$

where the dependence on data has been suppressed in both densities. If $\mathbf{V}$ is now redefined to represent the vector

$$\mathbf{V}' = \left(\frac{n_1 - p_1(\hat{\boldsymbol{\theta}})}{\sqrt{np_1(\hat{\boldsymbol{\theta}})}}, \ldots, \frac{n_K - p_K(\hat{\boldsymbol{\theta}})}{\sqrt{np_K(\hat{\boldsymbol{\theta}})}} \right),$$

and $\boldsymbol{\mu}$ is redefined to be the vector with components $\{p_i(\boldsymbol{\theta}_0) - q_i\}$, then the following lemma applies.

Lemma 3. *Under the alternative hypothesis, the asymptotic distribution of* $\mathbf{V}'\mathbf{V}$ *is* $\chi^2_{K-s-1}(n\boldsymbol{\kappa}'\boldsymbol{\kappa})$*, where* $\boldsymbol{\kappa}$ *is the vector having components* $\mu_i/\sqrt{p_i(\boldsymbol{\theta}_0)}$*.*

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

Specifying an appropriate alternative model for the deviation of $\mathbf{q}$ from $\mathbf{p}(\boldsymbol{\theta}_0)$ 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 both $\mathbf{q}$ and $\mathbf{p}(\boldsymbol{\theta}_0)$ are generated jointly from the following sampling procedure. First, a point $\mathbf{p}^*(\boldsymbol{\theta})$ satisfying the constraints 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, the true multinomial probability $\mathbf{q}$ 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

$$\boldsymbol{\mu} \stackrel{a}{=} \left(\mathbf{I} - \mathbf{M}\mathbf{J}^{-1}\mathbf{M}'\right)(\mathbf{q} - \mathbf{p}^*) = \mathbf{q} - \mathbf{p}(\boldsymbol{\theta}_0)$$

where

$$\mathbf{p}(\boldsymbol{\theta}_0) \stackrel{a}{=} \mathbf{p}^* + \mathbf{M}\mathbf{J}^{-1}\mathbf{M}'(\mathbf{q} - \mathbf{p}^*)$$

and $\mathbf{J} = \mathbf{M}'\mathbf{M}$. Here, $\stackrel{a}{=}$ denotes asymptotic equivalence. Given this alternative model for the generation of $\mathbf{q}$, we obtain the following result.

Lemma 4. *Under the assumptions stated above, if $\boldsymbol{\kappa}$ denotes the vector with components $\mu_i/\sqrt{p_i(\boldsymbol{\theta}_0)}$, the asymptotic distribution of $(1+c)\boldsymbol{\kappa}'\boldsymbol{\kappa}$ is χ^2_{K-s-1}.*

Noting that $\mathbf{p}(\boldsymbol{\theta}_0)$ maximizes the Kullback-Leibler information to $\mathbf{q}$ among probability vectors satisfying the given constraints, the proofs of these lemmas follow directly from results given in Rao (1973).

The similarity of Lemmas 3 and 4 to Lemmas 1 and 2 implies that the results of Section 2.1 can be applied to composite hypotheses by simply substituting $(K-s-1)$ for $(K-1)$ in (1-5) (when $x > K-s-1$).

In current 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 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 (5)) when the p value of the test just achieves its critical value of 0.05. Figure 1 displays this probability as a function of the degrees of freedom of the χ^2 test statistic. Because the marginal density of the data under the alternative hypothesis has been *maximized* with respect to the parameter α, the probabilities displayed in Figure 1 represent the *minimum* probability that the null hypothesis is true when the alternative hypothesis takes the form specified above. For one degree of freedom, the probability

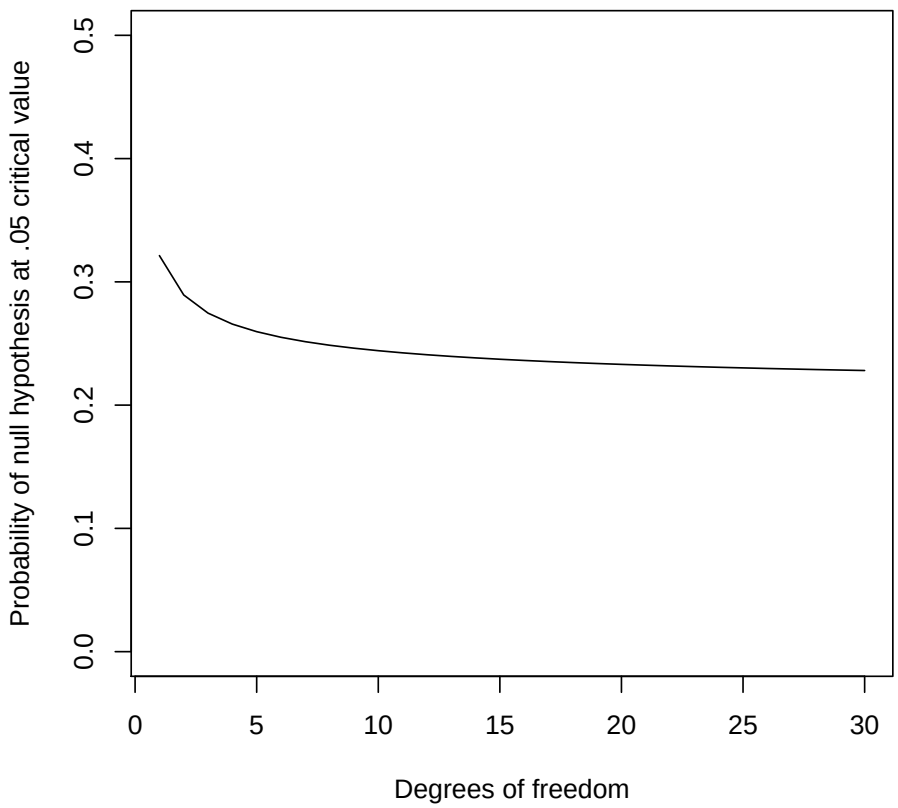

FIGURE 1. The posterior probability that the null hypothesis is true when Pearson's χ^2 statistic is observed to equal its .95 quantile under the null and equal prior probability is assigned to the null and alternative hypotheses.

that the null is true is 0.32; at 30 degrees of freedom, the probability that the null is true is 0.23.

The compliment to Figure 1 is provide in Figure 2. In Figure 2, p values of the χ^2 statistics that lead to a 5% probability that the null is true are displayed. Perhaps not surprisingly, these p-values are substantially smaller than 0.05.

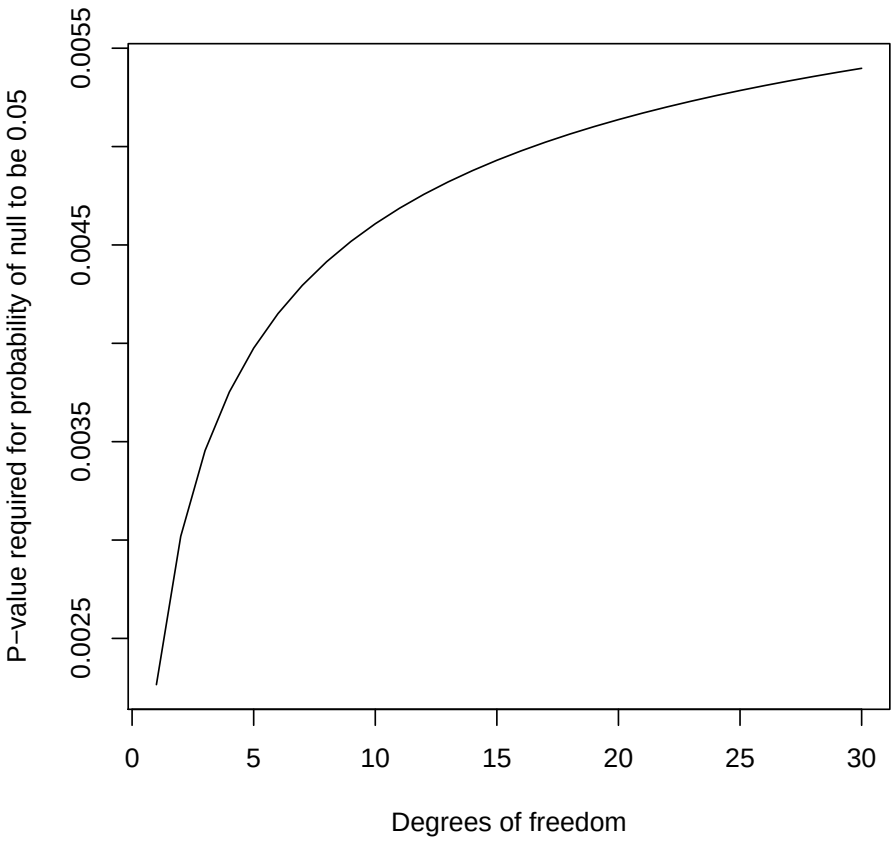

FIGURE 2. The p values required of the χ^2 test statistic for the null hypothesis to be true with posterior probability 0.05 when the prior odds are 1.

2.2.1. *A contingency table example.* It is interesting to compare the Bayes factor based on the χ^2 statistic, as proposed above, to more traditionally-computed Bayes factors for the purpose of testing independence of row and column classifications in contingency tables. Of course, the values of the traditional Bayes factors depend on the prior densities assumed for the multinomial probability vector under the null and alternative models. For that reason, we consider Bayes factors based on only two prior specifications here. Both are based on priors that are approximately

	Blood Group		
Site	O	A	B or AB
Pylorus and antrum	104	140	52
Body and fondus	116	117	52
Cardia	28	39	11
Extensive	28	12	8

TABLE 1. White and Eisenberg's classification of cancer patients

equivalent to the implicit assumption 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 described in Good and Crook (1987), employs a mixed Dirichlet prior with hyperparameter values determined from an empirical Bayes approach.

The particular contingency table considered here is taken from White and Eisenberg (1959) and was also considered in Albert (1990). The data represent a cross-classification on cancer site and blood type for 707 stomach cancer patients. The data appear in Table 1.

Pearson's χ^2 statistic for the test of independence for White and Eisenberg's data is 12.65 on 6 degrees of freedom. Based on (5), the Bayes factor on the odds for the independence model against a general alternative is 0.337.

The prior models underlying the computation of the Bayes factors proposed in 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 those details required for the replication of results are presented; interested readers should consult the original articles for more complete accounts.

The computation of the Bayes factor for independence under Albert's model requires the specification of a hyperparameter w. Albert recommends a value of 1 for this hyperparameter; this value corresponds to placing a uniform prior on second stage Dirichlet distributions for the marginal multinomial probabilities under the null. Accepting that recommendation, we take $w = 1$. A second parameter, K, is used to control the dispersion of the multinomial probabilty vector around the independence surface under the alternative model. The minimum Bayes factor against independence in this formulation can be obtained by minimizing an approximation to the Bayes factor given in Albert with respect to K. Doing so leads to a Bayes factor in favor of independence equal to 0.331.

To compute the Bayes factor under Good and Crook's model assumptions, a prior density is required on a hyperparameter k_0 that determines the degree of smoothing applied in an empirical Bayes prior density on the row and column probabilities under the null model. To estimate this probability, Good and Crook suggest mixing over a log-Cauchy density with lower and upper quartiles given by 10 and 50 divided by the number of rows or columns. Accepting this recommendation, if the Bayes factor is minimized over the value of a second hyperparameter κ, and if Good and Crook's suggestion to assume that the mixing density on the Dirichlet priors represents a point mass at $h(\kappa)$, then a mininum Bayes factor in favor of independence of 0.327 is obtained. This figure agrees well with Bayes factor obtained using Albert's prior assumptions, and suggests 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 well with the Bayes factor based on the χ^2 statistic, suggesting that little information has been lost by modeling the distribution of the test statistic directly.

2.3. **Bayes factors between specific hypotheses.** The discussion above assumes that the alternative hypothesis has been vaguely specified in the sense that it represents only the negation of the null. Computing the Bayes factor from comparable test statistics obtained from two well defined hypotheses is also straightforward. If x_0 and x_1 represent the values of the test statistic under each model, and $f_0(\cdot)$ and $f_1(\cdot)$ represent their sampling densities, then the Bayes factor between the two hypotheses based on the test statistics is simply

$$\text{Bayes factor} = \frac{f_0(x_0)}{f_1(x_1)}.$$

If the test statistics are nominally χ^2 on ν_0 and ν_1 degrees of freedom, then

$$\begin{aligned} 2\log(\text{Bayes factor}) &= -x_0 + x_1 + (\nu_0 - 2)\log(x_0) - (\nu_1 - 2)\log(x_1) - \\ &\qquad (\nu_0 - \nu_1)\log(2) + 2\log\left[\Gamma\left(\frac{\nu_0}{2}\right)\right] - 2\log\left[\Gamma\left(\frac{\nu_1}{2}\right)\right] \\ &= \text{constant} - x_0 + x_1 + (\nu_0 - 2)\log(x_0) - (\nu_1 - 2)\log(x_1). \end{aligned} \tag{6}$$

There is an interesting connection between (6) and the BIC criterion. If we consider the asymptotic case in which both the number of observations and the degrees of freedom for each model is large, then the term $2\log(x_0/x_1)$ is of smaller order than the remaining terms and so can be ignored. Noting that the expected value of a χ^2 statistic is equal to its degrees of freedom, and assuming that $\nu_i \approx n$—so that $\log(x_i) \approx \log(n)$—for large values of x_0 and x_1, we see that (6) takes

the approximate form of the (scaled) difference between BIC values for comparing models M_0 and M_1. This similarity is even more pronounced when the test statistics x_0 and x_a included in this equation represent realizations of deviance statistics rather than realizations of Pearson's χ^2 statistic.

3. F, t AND z TESTS

Consider now the problem of testing the validity of a linear constraint on a regression parameter. Adopting notation similar to that used in Rao (1973, page 191), suppose that

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

where $\mathbf{y}$ is an $n \times 1$ observation vector, $\boldsymbol{\beta}$ is an $r \times 1$ regression parameter, $\mathbf{X}$ is a $n \times r$ matrix of rank r, and σ^2 is a scalar variance parameter. Suppose that under the null hypothesis, $\mathbf{H}'\boldsymbol{\beta} = \xi$ where $\mathbf{H}$ is an $m \times k$ matrix of rank k whose range space is contained in the range space of $\mathbf{X}'$. As Rao notes, there then exists 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 condition $\mathbf{H}'\boldsymbol{\beta} = \xi$, 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 by 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 a r-variate normal distribution centered on $\boldsymbol{\beta}^*$ and having covariance matrix $\tau\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Again, this is the case of practical interest because values of $\boldsymbol{\beta}$ not drawn from a distribution similar to this will either be accepted or rejected with probability close to 0 or 1 as the number of observations becomes large. Note also that the marginal variances of the components of $\boldsymbol{\beta}$ around the point $\boldsymbol{\beta}^*$ are typically $O(1/n)$, making the deviation of the components of $\boldsymbol{\beta}$ away from the null hypothesis $O_p(1/\sqrt{n})$ under the alternative.

Under this scheme for generating $\boldsymbol{\beta}$ under the alternative hypothesis, the distribution of $\mathbf{H}'\boldsymbol{\beta}$ is normally distributed 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 alternative hypotheses, the distribution of $R_1^2 - R_0^2$ is $\chi_k^2(\lambda)$ where the non-centrality parameter λ is given by

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

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

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

In this equation, $m = n - r$ and $B(s,t) = \Gamma(s+t)/[\Gamma(s)\Gamma(t)]$. 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.

The marginal maximum likelihood estimate of τ based on the observed value of f under the alternative hypothesis is $\tau = f - 1$ when $f > 1$. At this value of τ, the

marginal density of f is

$$p(f \mid \tau = f-1) = B\left(\frac{k}{2}, \frac{m}{2}\right)\left(\frac{k}{m}\right)^{k/2} \frac{1}{\left(1+\frac{k}{m}\right)^{(k+m)/2}} \frac{1}{f}. \tag{7}$$

Finally, the minimum Bayes factor in favor of the null hypothesis for $f > 1$ is

$$\text{Bayes factor} = \left[\frac{\frac{m}{k}+1}{\frac{m}{k}+f}\right]^{\frac{k+m}{2}} f^{\frac{k}{2}}. \tag{8}$$

For large f, the minimum 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 minimum Bayes factor against the null reduces to

$$\left(\frac{m+1}{m+f}\right)^{\frac{m+1}{2}} \sqrt{f} \tag{9}$$

where $f = t^2$.

Figure 3 depicts the minimum posterior probability that the null is true for t-tests as a function of the degrees of freedom m, assuming prior odds of 1 between the null and alternative. As $m \to \infty$, the limiting value of this probability at $f = 1.96^2$ is 0.32 , which is consistent with the corresponding χ^2 test reported in the previous section.

The one-sample z statistic can be obtained from (9) by taking the limit as m becomes large. Taking this limit, we find that the Bayes factor for testing the value of a normal mean against the alternative that the mean has different value is

$$\text{Bayes factor} = \sqrt{f}\exp\left(-\frac{f-1}{2}\right) \tag{10}$$

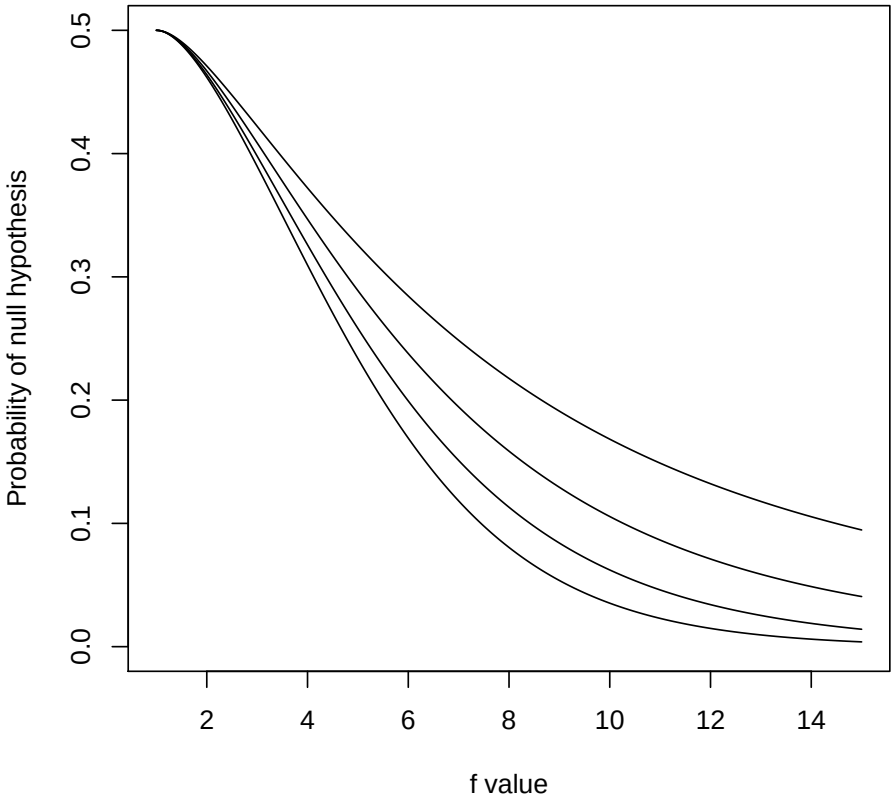

FIGURE 3. The posterior probability that the null hypothesis is true as a function of the observed f statistic when the numerator degrees of freedom is 1 (assuming prior odds equal to 1). From top to bottom, the curves represent the null's posterior probability when the degrees of freedom in the denominator are 5, 10, 25, and 500.

This is equivalent to the result given in Section 2 based on a χ^2_1 distribution.

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 as an example 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 distribution of 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 noncentrality 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 an interesting 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 that counts follow a non-central hypergeometric distribution with, say, non-centrality parameter ϕ. If ϕ is parameterized so as 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 centered on 0 with scale parameter, say, σ. With this definition of the alternative model, it is a simple

matter to numerically maximize 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 probabilties are relevant to the calculation of the p value.

Fisher's tea-tasting experiment (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 was able to correctly identify three of four cups of each preparation after tasting them in 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 .229. The only table that is more extreme is the table (4,0,0,4), corresponding to all correct identifications. That table has probability .014, leading to a one-sided p value of .243.

The Bayes factor in favor of the null, when $\log(\phi)$ is assumed drawn from a $N(0, \sigma^2)$ distribution and the marginal density of the alternative is maximized with respect to σ, is .90. The maximum marginal likelihood of the data is achieved when $\sigma = 1.3$. Thus, there is some evidence against the null, but its posterior probability (assuming equal prior odds) is relatively high, equalling .47.

5. Summary

By modeling the distribution of test statistics directly, Bayes factors can be computed in many standard problems without the 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. 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. Doing so introduces a noncentrality parameter that must be modeled, but for standard test statistics a conjugate prior density or other convenient prior density for the noncentrality parameter is often apparent and typically involves a only single scale parameter. Marginalizing over the noncentrality parameter and maximizing with respect to the scale parameter leads to the maximum marginal likelihood estimate of the density of the data under the alternative, which in turn leads to what might be considered a default Bayes factor.

Bayes factors defined in this way are numerically easy to compute, and require neither the specification of prior densities on model parameters nor the explicit specification of alternative models. For normal-theory test statistics, they are actually easier to compute than p values, and so can be applied routinely to common testing problems.

The most important aspect of this framework is that it provides practitioners with an alternative to p values for summarizing evidence against null hypotheses. Because the value of a Bayes factors represents the modification of the probability that a hypothesis is true based on test data, the routine use of default Bayes factors would reduce the confusion that often occurs when p values are reported to the public.

References

[1] Albert, J.H. (1990), "A Bayesian test for a two-way contingency table using independence priors," *Canadian Journal of Statistics*, **18**, 347-363.

[2] Albert, J. H. , and Gupta, A. K. (1982), "Mixtures of Dirichlet distributions and estimation in contingency tables", *Annals of Statistics*, **10**, 1261-1268.

[3] Berger, J.O. 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.

[4] Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press: Princeton, N.J.

[5] Crook, J. F. , and Good, I. J. (1980), "On the application of symmetric Dirichlet distributions and their mixtures to contingency tables," *Annals of Statistics*, **8** , 1198-1218.

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

[7] Fisher, R.A. (1935) *Design of Experiments* (8th edition, 1966), Oliver & Boyd: Edinburgh.

[8] Good, I. J. (1967), "A Bayesian significance test for multinomial distributions (with discussion)", *Journal of the Royal Statistical Society, Series B*, **29**, 399-431.

[9] Good, I. J. (1976), "On the application of symmetric Dirichlet distributions and their mixtures to contingency tables", *Annals of Statistics*, **4**, 1159-1189.

[10] Good, I. J. (1986), "The maximum of a Bayes factor against 'independence' in a contingency table, and generalizations to higher dimensions", *Journal of Statistical Computation and Simulation*, **26** , 312-316

[11] Good, I. J. , and Crook, J. F. (1987), "The robustness and sensitivity of the mixed-Dirichlet Bayesian test for 'independence' in contingency tables", *Annals of Statistics*, **15** , 670-693.

[12] Gûnel, E. and Dickey, J. (1974), "Bayes factors for independence in contingency tables", *Biometrika*, **61**, 545-557.

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

[14] Jeffreys, H. (1961) *Theory of Probability*, Oxford University Press: Oxford.

[15] Sellke, T., Bayarri, M.J., Berger, J.O. (2001), "Calibration of p values for testing precise null hypotheses," *The American Statistician*, **55**, 62-71.

[16] Rao, C. R. (1973), *Linear statistical inference and its applications*, John Wiley & Sons: New York.

[17] White, C. and Eisenberg, H. (1959) "ABO blood groups and cancer of the stomach," *Yale Journal of Biology and Medicine*, **32**, 58-61.