Supplementary materials for this article are available at http://pubs.amstat.org/toc/jasa/104/487.

Assessing Robustness of Intrinsic Tests of Independence in Two-Way Contingency Tables

George CASELLA and Elías MORENO

For testing nested hypotheses from a Bayesian standpoint, a desirable condition is that the prior for the alternative model concentrates mass around the smaller, or null, model. For testing independence in contingency tables, the intrinsic priors satisfy this requirement. Furthermore, the degree of concentration of the priors is controlled by a discrete parameter, t, the training sample size, which plays an important role in the resulting answer. In this article we report on the robustness of the tests of independence for small or moderate sample sizes in contingency tables with respect to intrinsic priors with different degrees of concentration around the null. We compare these tests to frequentist tests and other robust Bayes tests. For large sample sizes, robustness is achieved because the intrinsic Bayesian tests are consistent. Examples using real and simulated data are given. Supplemental materials (technical details and data sets) are available online.

KEY WORDS: Bayes factor; Bayesian inference; Chi-squared test; Monte Carlo integration; Monte Carlo method; Volume test.

1. INTRODUCTION

The problem of testing independence in contingency tables has, to say the least, a long history (mainly from a frequentist viewpoint). Many controversies have arisen, concerning the question of whether to condition on marginal totals, whether inference should be asymptotic or exact, and what statistics should be used for testing. A good introduction to this topic is the review article by Agresti (1992); see also the textbook by Agresti (1996).

1.1 Frequentist and Bayesian Approaches

The standard test of independence is the classical Pearson chi-squared test. The justification of the test is asymptotic, based on the assumption of having a table with fixed margins, with a multinomial distribution conditional on the margins. But Pearson's chi-squared test is unreliable with small samples or sparse tables; in this situation, an alternative is an exact test.

Exact frequentist inference in contingency tables can be done by applying the same test statistic to all tables with the same marginals, and then assessing where the observed table has fallen in this set of reference tables. The first such test was Fisher's exact test, and this idea is the basis of the statistical package StatXact (*www.cytel.com*) (see Mehta, Patel, and Senchaudhuri 2000). These p-values have the form

$$\frac{1}{\sum_{k=1}^{N} w_k} \sum_{k=1}^{N} w_k 1_{(\chi_k^2 \geq \chi_{obs}^2)}, \tag{1}$$

where 1_A is the indicator function of the set A, N is the total number of tables under consideration, χ_{obs}^2 is the chi-squared statistic of the observed table, and χ_k^2 is the chi-squared statistic of the kth table. Standard "exact" weighted or unweighted p-values are computed, taking N to be the number of tables with the same margin as the observed table and

$$w_k = 1 \qquad \text{or} \qquad w_k = \frac{\prod r_i(\mathbf{x})! \prod c_j(\mathbf{x})!}{n! \prod_{ij} x_{ij}!},$$

depending on whether an unweighted test or multinomial weighting is used. Some interesting computational problems are associated with calculating the statistic (1); for example, one may need to calculate N, and also enumerate (or sample from) the set of tables.

An alternative, which conditions only on the table total, is obtained from the *volume test* of Diaconis and Efron (1985). The p-value of this test is given by (1) with $w_k = 1$ and N the total number of unrestricted tables (i.e., all tables with the same table total as the observed table) and unrestricted marginal totals. Although the volume test provides an inference based on a more realistic sampling model, it tends to be more conservative than the restricted tests, typically favoring the null hypothesis.

Most of the Bayesian literature on contingency tables focuses on the estimation of cell probabilities and smoothing parameters that appear by decomposing the log of the original parameters θ_{ij} into a sum of row effect parameters α_i, column effect parameters β_j, and interaction effect parameters λ_{ij} (e.g., Leonard 1975; Laird 1978; Albert and Gupta 1982, 1983; Albert 1987; Nazaret 1987; Epstein and Fienberg 1992; for reviews, see Albert 2004 or Agresti and Hitchcock 2005). A relatively smaller number of articles are devoted to testing one-sided hypotheses. There the use of conjugate priors is common (as in Novick and Grizzle 1965 and Altham 1969); some exceptions are the work of Howard (1998) and Kadane et al. (2002). Howard (1998) considered two conditionally independent binomial sampling distributions with parameters p_1 and p_2, and the one-sided null $H_0 : p_1 > p_2$. He argued in favor of considering a dependent prior for p_1 and p_2. there the null plays a role (albeit a somewhat weak one) in the construction of the Bayesian model for computing the posterior probability of H_0.

Key references for Bayesian tests of independence in contingency tables are the series of articles by Good (1967, 1976), Crook and Good (1980), and Good and Crook (1987), which

George Casella is Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611 (E-mail: *casella@stat.ufl.edu*). Elías Moreno is Professor, Department of Statistics, University of Granada, 18071 Granada, Spain (E-mail: *emoreno@ugr.es*). Casella was supported by National Science Foundation grants DMS-04-05543, DMS-0631632, and SES-0631588. Moreno was supported by Ministerio de Ciencia y Tecnología, grant SEJ-65200 and Junta de Andalucía grant SEJ-02814. This work was started while Casella was on sabbatical at the University of Granada. The authors thank the editor, associate editor, and referees for their careful reading and thoughtful comments.

Journal of the American Statistical Association
September 2009, Vol. 104, No. 487, Theory and Methods
DOI: 10.1198/jasa.2009.tm08106

Table 1. Posterior probabilities of H_0 for the procedures of Good and Crook (1987), Albert (1990), and the intrinsic procedure, with $t = 1$, described in Sections 1.2 and 3

		Table		
		2 2 2 2 2 2 2 2 2	6 6 6 6 6 6 6 6 6	5 0 0 5 0 0 5 0 0
Posterior probability of H_0	Good	0.450	0.327	0.520
	Albert	0.5	0.5	0.375
	Intrinsic	0.648	0.891	0.964

are based on Bayes factors, that is, on the statistic $B_{10}(y) = m_1(y)/m_0(y)$, where $m_0(y)$ is the marginal of the data y under the null model and $m_1(y)$ under the unrestricted model. (See also Jeffreys 1961; Gunel and Dickey 1974; Albert 1990; and Kadane et al. 2002.) A general reference on Bayes factors is Kass and Raftery (1995).

Good (1976) calculated Bayes factors using a row-column independence prior for the parameters of the multinomial distribution under H_0 and mixtures of symmetric Dirichlet distributions under the alternative. A hyperparameter α, the common Dirichlet parameter (called the flattening constant), was assumed to be log-Cauchy distributed. This hyperprior distribution was extended by Crook and Good (1980), and Good and Crook (1987), where the hyperprior was extended to a log-Student distribution with degrees of freedom υ. For $\upsilon > 15$, the log-Student approximates a lognormal distribution, and $\upsilon = 1$ approximates a log-Cauchy distribution. Finally, these authors concluded that the Bayes factors were robust with respect to variations of the hyperparameters, and recommended the use of the log-Cauchy distribution.

With all of the attention that has been paid to Bayesian tests of independence in contingency tables, why do we introduce yet another test? We do so because some deficiencies remain in some key references of Bayesian tests of independence; consider, for example, the contingency tables in Table 1. These three tables are clearly in support of H_0, and we give the posterior probabilities of the procedures of Good and Crook (1987) and Albert (1990), along with those of the intrinsic procedure developed here. In cases where the null should be favored, the Good/Crook and Albert procedures do not do so; the intrinsic procedure gives more reasonable results.

1.2 Testing Nested Hypothesis: The Intrinsic Prior Class

In general, many priors that might be appropriate for estimation purposes cannot be recommended as priors for Bayesian tests of nested hypothesis. This is because the null hypothesis is not taken into account in the formulation of the prior. If this is not done, then it is impossible to guarantee that the prior distribution on the unrestricted parameters of the model will concentrate around the null hypothesis, a condition that is widely accepted and should be required of a prior for testing a hypothesis. (See, e.g., Jeffreys 1961, chap. 5; Gunel and Dickey 1974, who noted that this is the "Savage continuity condition"; Berger and Sellke 1987; Casella and Berger 1987; Morris 1987; Berger 1994; Robert 2001, who also discussed the Jeffreys–Lindley paradox; and Consonni and La Rocca 2008.)

It is important to realize that if a prior on the unrestricted hypothesis H_1 concentrates probability near H_0, this does not necessarily favor H_0, but rather focuses the test on model alternatives that are close to H_0. If H_0 is reasonable, then it is important to be able to distinguish H_0 from reasonable, close alternatives. Putting high prior probability on extreme models, far from H_0, is wasteful. If such models are truly generating the data, then this will be easy to discover with any procedure. If they are not generating the data (which is more likely), giving them high probability will distort the resulting test and discount the more reasonable alternatives.

The foregoing arguments motivate the consideration of the intrinsic priors for testing independence in contingency tables. Starting from a default prior for the parameters of a contingency table $\{\theta_{ij}\}$, which typically will not concentrate probability near the null hypothesis H_0 of independence but instead will spread it out in H_1 giving high probability to models far from H_0, the intrinsic prior construction creates a family of new priors, $\{\pi^I(\theta_{ij}|H_0, t), t \geq 1\}$. There priors (a) concentrate probability near H_0; (b) control the degree of concentration of the prior around the null with the integer parameter t, the "training sample size"; and (c) maintain consistency of the tests; that is, as the sample size n becomes infinite, the test will always make the correct decision for any arbitrary fixed value of t. This means that when using the intrinsic prior class and an observed sample $\{y_{ij}\}$, we will obtain a set of posterior answers, one answer for each t, associated with different degrees of concentration of the prior around the null. When these answers essentially convey the same message, we say that the test is robust with respect to the concentration of the prior around the null.

Intrinsic priors were introduced in hypothesis testing to convert improper priors into proper ones (Berger and Pericchi 1996; Moreno 1997; Moreno, Bertolino, and Racugno 1998), but there is no inherent limitation in using them when the default prior is proper. The intrinsic prior construction for a testing problem is as follows. Consider a Bayesian testing problem

$$H_0 : \{f_0(x|\theta_0), \pi_0(\theta_0)\} \quad \text{vs} \quad H_1 : \{f_1(x|\theta_1), \pi_1(\theta_1)\}, \tag{2}$$

where $f_0(x|\theta_0)$ is nested in $f_1(x|\theta_1)$ and $\pi_0(\theta_0)$ and $\pi_1(\theta_1)$ are default estimation priors that might be improper. The intrinsic prior for θ_1 conditional on H_0 and t is given by

$$\pi^I(\theta_1|H_0, t) = \pi_1(\theta_1) E_{\mathbf{x}|\theta_1} \frac{m_0(\mathbf{x})}{m_1(\mathbf{x})}, \tag{3}$$

where $\mathbf{x} = (x_1, \ldots, x_t)$, $m_i(\mathbf{x})$, $i = 0, 1$, are the respective marginals and the expectation is taken with respect to $f_1(\mathbf{x}|\theta_1) = \prod_{j=1}^{t} f_1(x_j|\theta_1)$.

It is important to note that here we are using a *theoretical* $\mathbf{x}$, in that no actual data are used in constructing the intrinsic prior. In our calculations, the training sample $\mathbf{x}$ is distributed according to either $f_0(\mathbf{x}|\theta_0)$ or $f_1(\mathbf{x}|\theta_1)$, with sample size t.

To illustrate the role of t in the shape of the intrinsic prior, consider the simple case of sampling from a binomial distribution $B(y|n,p)$ with n known. A default prior for estimating the parameter p is usually chosen from one of the following distributions: the uniform (Bayes 1783; Laplace 1812), the Jeffreys prior (Jeffreys 1961; Bernardo 1979), Zellner's prior (Zellner 1977), or that of Novik and Hall (1965). The first two of these, the most popular, are proper. The third is proper, and the fourth is improper. Any of these distributions can be used as a reasonable default prior for estimating p in the absence of subjective prior information (see, e.g., Berger 1985, p. 89).

But these priors are not appropriate for testing a null hypothesis, because they do not concentrate mass around the null hypothesis as the intrinsic priors do. For example, for testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, starting from the proper uniform prior $\pi(p) = 1_{[0,1]}(p)$, the intrinsic prior for p, conditional on the null value p_0 and training sample size t, is

$$\begin{aligned}\pi^I(p|p_0,t) &= E_{x|p}\frac{B(x|t,p_0)}{\int_0^1 B(x|t,p)\,dp}\\ &= \frac{1}{t+1}\sum_{i=0}^{t}\mathrm{Be}(p|i+1,t-i+1)\\ &\quad\times \mathrm{Be}(p_0|i+1,t-i+1),\end{aligned}$$

where the expectation is taken with respect to the binomial $B(x|t,p)$ and $\mathrm{Be}(p|a,b)$ represents the beta distribution for p with parameters a and b.

Figure 1 shows intrinsic priors $\pi^I(p|p_0,t)$ for two values of p_0 and $t = 1, 2, \ldots, 25$. The prior is always unimodal, and for $t = 1$ it is a linear function of p, but as t increases, it concentrates more probability mass in the neighborhood of p_0. When we start with the Jeffreys prior and $t \geq 2$, the resulting intrinsic priors are very close to those obtained when starting from the uniform.

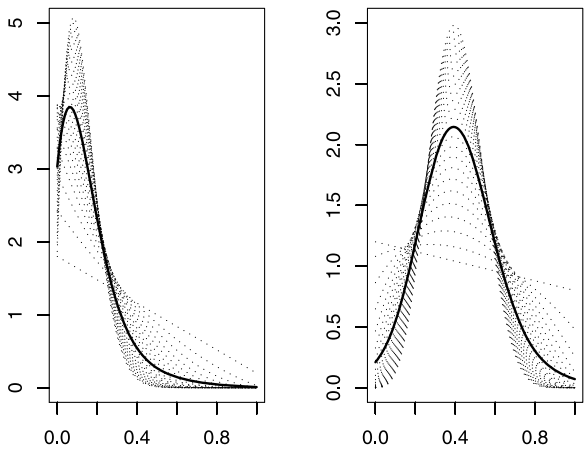

Figure 1. Intrinsic priors from the uniform prior for $p_0 = 0.1$ (left) and $p_0 = 0.4$ (right) for $t = 1, 2, \ldots, 25$. As t increases, the prior concentrates more probability mass in the neighborhood of p_0. The solid curve represents the average intrinsic prior.

Thus, in this simple case, the default prior provides intrinsic priors concentrated around the null hypothesis. We show that this is the case in more realistic examples. Finally, we note that the construction of the intrinsic prior class is fully automatic.

1.3 Summary

The rest of the article is organized as follows. In Section 2 we recall the sampling models used for tests of independence, and in Section 3 we develop the intrinsic priors and the resulting posterior probabilities in the 2×2 case. Details of the derivations for general $a \times b$ tables are provided in the Supplementary Materials. In Section 3.3 we explore consistency, with technical details relegated to the Appendix. In Section 3.4 we also discuss calculation of the intrinsic priors, which requires summing over all possible tables with table total t or row totals t_i. In Section 4 we evaluate the performance of intrinsic posterior probabilities with a number of examples, both real and artificial. We conclude with a discussion in Section 5.

2. SAMPLING MODELS AND INFERENCE

There are several possible sampling models for contingency tables. Good and Crook (1987) described three sampling procedures:

P_1 : Condition only on the table total,

P_2 : Condition only on the totals of one margin,

P_3 : Condition on the totals of both margins.

They noted that P_3 is not a very common sampling model, and that P_1 and P_2 are more useful. They derived Bayesian tests under P_1 and P_2 and illustrated the performance of their method on some example tables, both real and artificial. In general, their answers were reasonable, indicating that calibration of the set of all tables is possible. There are some disturbing anomalies, however. For example, for the 3×3 table in which every cell has a 6 (and of course has a p-value $= 1$), Good and Crook reported an average Bayes factor of 2.1, which would lead to a posterior probability of the null hypothesis of $1/(1 + 2.1) = 0.327$. In contrast, our intrinsic procedure yields posterior probabilities of the null varying between 0.839 and 0.891 as the concentration parameter t varies.

The procedures P_1 and P_2 arise under two different sampling schemes and lead to two different distributions. In an $a \times b$ table, if sampling procedure P_1 is used, then the distribution of the frequencies is multinomial with ab cells and total equal to the table total n. There are $ab - 1$ free parameters, the cell probabilities. If sampling procedure P_2 is used, where the a row totals are fixed, then the distribution of the frequencies is that of a independent multinomials, each with $b - 1$ free parameters and a total equal to the row total. These are obviously different models.

It is possible for a statistic to have the same distribution under either P_1 or P_2; the asymptotics of the chi-squared statistic are the same. The question that we examine is whether or not such an equality is desirable. Good and Crook (1987) stated the following assumption:

Assumption 1 (Ancillarity of the row totals). Under P_1, the row totals alone (or the column totals alone) convey no evidence for or against H_0.

They then argued that this assumption should be reflected in the chosen statistic, and they chose their prior to force this to be the case. Nonetheless, we view P_1 and P_2 as distinct procedures with distinct structures, which thus should have distinct statistics. To defend our position, we look at the 2×2 case (although we could argue in the general case as well) and consider the joint density of (y_{11}, r, c), the $(1, 1)$ observation, the total of the first row, and the total of the first column. Using θ to denote the parameter, direct factorization yields

$$f(y_{11}, r, c|\theta) = f(y_{11}, c|r, \theta) f(r|\theta).$$

Assumption 1 requires that $f(r|\theta) \propto f(r|\theta_0)$, where θ_0 is a null parameter value. This occurs if r corresponds to the fixed n_i with the rows of the table being independent binomials, or in the 2×2 table with cell probabilities θ_{ij} and table total n, $r \sim \text{binomial}(r|n, \theta_{11} + \theta_{12})$, where the parameter is a marginal probability. Although this "approximate ancillarity" of r is well known, the distributions are different, and formally, there can never be equality of the sampling procedures P_1 and P_2.

3. INTRINSIC PRIORS FOR 2 × 2 TABLES

In this section we give a detailed derivation of the intrinsic posterior probabilities for the 2×2 table under sampling procedures P_1 and P_2. We present this simple case to provide insight into the workings of the priors and the resulting probabilities; the general case is treated in the Supplementary Materials.

3.1 Margins Unrestricted

We start with a 2×2 contingency table with n individuals classified into four cells each with an unknown probability θ_{ij}, $i, j = 1, 2$, and $\sum_{ij} \theta_{ij} = 1$. Under this sampling scheme (in which only n is fixed), the distribution of the possible tables, $\mathbf{y} = \{y_{11}, y_{12}, y_{21}, y_{22}\}$, is a three-parameter multinomial distribution, $M(\mathbf{y}|n, \theta_{ij})$. A default prior for θ_{ij} can be taken as either a three-dimensional Dirichlet with all parameters equal to $1/2$ (the Jeffreys prior) or the Dirichlet with all parameters equal to 1 (the uniform prior).

Under the independence assumption $\theta_{ij} = p_i q_j$, where $\sum_{i=1}^2 p_i = \sum_{j=1}^2 q_j = 1$, the two-parameter distribution of the table $\mathbf{y} = \{y_{11}, y_{12}, y_{21}, y_{22}\}$ is

$$f_0(\mathbf{y}|n, p_1, q_1) = \binom{n}{\mathbf{y}} p_1^{(y_{11}+y_{12})}(1-p_1)^{(y_{21}+y_{22})} \times q_1^{(y_{11}+y_{21})}(1-q_1)^{(y_{12}+y_{22})},$$

where $\binom{n}{\mathbf{y}} = \binom{n}{y_{11}y_{21}y_{12}y_{22}}$, the multinomial coefficient. This density is nested in the multinomial $M(\mathbf{y}|n, \theta_{ij})$. The prior $\pi(p_1, q_1) = \text{Uniform}(p_1|0, 1) \times \text{Uniform}(q_1|0, 1)$ is a default prior for the parameters (p_1, q_1).

A default analysis of the testing problem $H_0 : \theta_{ij} = p_i q_j$ versus $H_1 : \theta_{ij}$ is to choose between M_0 and M_1, where

$$\begin{aligned} &M_0 : \{f_0(\mathbf{x}|n, p_1, q_1), \pi(p_1, q_1)\} \qquad \text{and} \\ &M_1 : \{M(\mathbf{x}|n, \theta_{ij}), \mathcal{D}_3(\theta_{ij}|1, 1, 1, 1)\}. \end{aligned} \tag{4}$$

Note that the default prior $\mathcal{D}_3(\theta_{ij}|1, 1, 1, 1)$ does not depend on the null. We use this prior to create an intrinsic prior for θ_{ij}, a prior that does depend on H_0. We then substitute $\mathcal{D}_3(\theta_{ij}|1, 1, 1, 1)$ for the intrinsic prior $\pi^I(\theta_{ij}|t)$ in (4), where t is the training sample size.

Applying (3), it is straightforward to see that, based on a training sample size t, the intrinsic prior for θ_{ij} is

$$\pi^I(\theta_{ij}|t) = \frac{(t+3)!}{[(t+1)!]^2} \sum_{\mathbf{x}:\sum_{ij} x_{ij} = t} \binom{t}{\mathbf{x}} \left(\prod_{i=1}^2 r_i(x)!\right) \times \left(\prod_{j=1}^2 c_j(x)!\right)\left(\prod_{i,j} \frac{\theta_{ij}^{x_{ij}}}{x_{ij}!}\right), \tag{5}$$

where $r_i(x) = \sum_{j=1}^2 x_{ij}$ and $c_j(x) = \sum_{i=1}^2 x_{ij}$ are the sum of the rows and columns. For a data set $y = \{y_{ij}\}$, the Bayes factor $B_{10,t}(y)$, for (6) versus a uniform prior for p_1 and q_1, is equal to $m_1^I(y|t)/m_0(y|t)$, where

$$m_1^I(y|t) = \binom{n}{\mathbf{y}} \frac{(t+3)!}{[(t+1)!]^2(2t+3)!} \sum_{\{\mathbf{x}:\sum_{ij} x_{ij} = t\}} \left(\prod_{i=1}^2 r_i(\mathbf{x})!\right) \times \left(\prod_{j=1}^2 c_j(\mathbf{x})!\right) \prod_{ij} \frac{(x_{ij}+y_{ij})!}{(x_{ij}!)^2}$$

and

$$m_0(y|t) = \binom{n}{\mathbf{y}} \frac{(\prod_{i=1}^2 r_i(\mathbf{y})!)(\prod_{j=1}^2 c_j(\mathbf{y})!)}{[(t+1)!]^2}.$$

If we assume that, a priori, $P(M_0) = P(M_1) = 1/2$, then for any training sample size t, the posterior probability of the null is given by $P(M_0|y, t) = 1/(1 + B_{10}(y, t))$.

3.2 One Margin Fixed

In this case the sampling scheme is that of sampling from two binomial distributions, $B(y_1|n_1, p_1)$ and $B(y_2|n_2, p_2)$, where n_1 and n_2 are fixed. The interest is in testing

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 \neq p_2,$$

which is the problem of choosing between the null model

$$M_0 : \{B(y_1|n_1, p_0)B(y_2|n_2, p_0), \pi^U(p_0) = 1_{(0,1)}(p_0)\},$$

and the alternative

$$M_1 : \{B(y_1|n_1, p_1)B(y_2|n_2, p_2), \pi^U(p_1, p_2) = 1_{(0,1)}(p_1)1_{(0,1)}(p_2)\}.$$

As in Section 3.1, the conventional uniform prior for (p_1, p_2) does not depend on the null model M_0, but the intrinsic prior for (p_1, p_2) concentrates probability mass around the null hypothesis (the line $p_1 = p_2$). With training sample sizes t_1 and t_2, the intrinsic prior is a convex combination of the product of beta distributions, that is,

$$\begin{aligned} &\pi^I(p_1, p_2|t_1, t_2) \\ &= \sum_{i=0}^{t_1} \sum_{j=0}^{t_2} \binom{t_1}{i}\binom{t_2}{j} \frac{\Gamma(i+j+1)\Gamma(t_1+t_2-i-j+1)}{\Gamma(t_1+t_2+2)} \\ &\quad \times \text{Be}(p_1|i+1, t_1-i+1)\,\text{Be}(p_2|j+1, t_2-j+1), \end{aligned}$$

where the parameters p_1 and p_2 are not *a priori* independent.

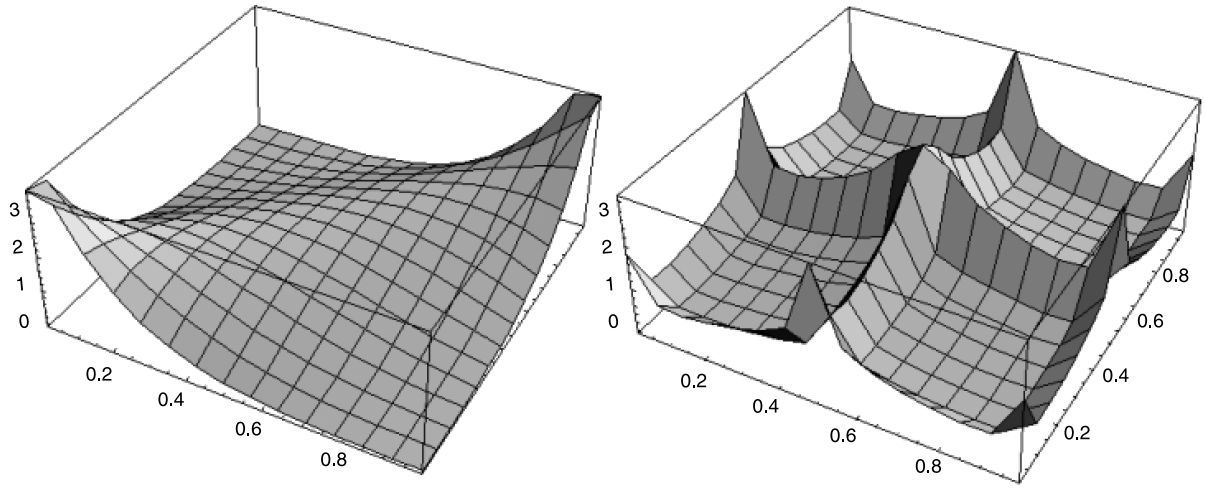

Figure 2. Intrinsic prior for (p_1, p_2) for $t_1 = 10$ and $t_2 = 10$ (left) and the log-Cauchy mixture of Dirichlets of Good and Crook (1987) (right). The intrinsic prior concentrates mass symmetrically around the line $p_1 = p_2$. The recommended prior of Good and Crook is also symmetric around the line $p_1 = p_2$ with a somewhat usual shape, moving mass to the boundaries of the parameter space.

For $t_1 = t_2 = 10$, Figure 2(left) displays the intrinsic prior. The probability mass is concentrated around the line $p_1 = p_2$, and the prior is symmetric around this line. The figure also shows the recommended prior of Good and Crook (1987), a log-Cauchy mixture of Dirichlets. Although this prior is also symmetric around the line $p_1 = p_2$, its shape is somewhat unusual. In contrast to the intrinsic prior, it does not concentrate its mass in a neighborhood of the line $p_1 = p_2$, but rather puts more mass on the boundaries.

The posterior probability of the null for the intrinsic priors $(\pi^U(p_0), \pi^I(p_1, p_2|t_1, t_2))$, conditional on the sample (y_1, y_2), is $P(M_0|y_1, y_2, t_1, t_2) = 1/(1 + B_{10}(y_1, y_2, t_1, t_2))$, where

$$B_{10}(y_1, y_2, t_1, t_2) = \left[\frac{n_1 + n_2 + 1}{(n_1 + t_1 + 1)(n_2 + t_2 + 1)}\right]\left[\frac{(t_1 + 1)(t_2 + 1)}{t_1 + t_2 + 1}\right] \times \binom{n_1 + n_2}{y_1 + y_2} \sum_{i=0}^{t_1} \sum_{j=0}^{t_2} \frac{\binom{t_1}{i}^2 \binom{t_2}{j}^2}{\binom{t_1+t_2}{i+j}\binom{n_1+t_1}{y_1+i}\binom{n_2+t_2}{y_2+j}}. \quad (6)$$

3.3 Consistency

When the sample information is weak, the posterior probability of the models involved varies as the intrinsic prior varies through the training sample size t. But as the sample information becomes stronger, as it does when the sample size n increases, we expect the posterior probability of the models to be more robust with respect to t. In particular, as the sample size n tends to infinity, the sampling distribution should overwhelm any prior information. Thus we should be able to prove consistency of the intrinsic Bayesian procedure for any finite training sample size t. Specifically, for any finite t, we want to ensure that $\lim_{n\to\infty} P(M_0|y, t) = 1$, when sampling from the null and $\lim_{n\to\infty} P(M_1|y, t) = 1$ when sampling from the alternative. We first consider the case of testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, where θ is a binomial success probability. We have the following theorem.

Theorem 1. For testing $M_0 : B(y|\theta_0)$ versus $M_1 : \{B(y|\theta), \pi^I(\theta|\theta_0, t)\}$, the intrinsic posterior probability is consistent for any finite training sample size t.

The proof of this theorem is given in Appendix, where we also extend the result to other cases considered in this article.

3.4 Computational Issues

Note that calculating the intrinsic priors necessitates summing over all tables with table total t. Although this sometimes can be done for the 2×2 case, the calculation quickly becomes impossible in the general case; for example, for the seventh table in Table 3 (the Mendel data) in Section 4.2, there are 162,750,684,200,297,895 tables with the same table total and 2,689,129,357,824 tables with the same row totals. Thus to calculate the intrinsic priors, we use a Monte Carlo sum, which we now explain.

Because the space of tables is so large, generating tables uniformly will not be efficient, as most of the posterior probability will be close to the observed table. Thus we use an importance sampling strategy, taking as a candidate distribution a multinomial with cell probabilities equal to the observed table. (In theory, the choice of candidate distribution has no bearing on the resulting calculation; however, choosing the candidate to have high probability near the observed table will help the Monte Carlo convergence.)

For example, as detailed in the Supplementary Materials, the intrinsic Bayes factor for an $a \times b$ table is given by

$$B_{10}(\mathbf{y}, t) = \frac{\Gamma(t + ab)}{\Gamma(t + n + ab)}\left[\frac{\Gamma(n + a)\Gamma(n + b)}{\Gamma(t + a)\Gamma(t + b)}\right] \times \sum_{\mathbf{x}:\sum x_{ij}=t} \binom{t}{\mathbf{x}} \frac{(\prod r_i(\mathbf{x})!)(\prod c_j(\mathbf{x})!)}{(\prod r_i(\mathbf{y})!)(\prod c_j(\mathbf{y})!)} \times \frac{\prod(x_{ij} + y_{ij})!}{\prod x_{ij}!}. \quad (7)$$

Table 2. p-values and posterior probabilities for selected tables from Efron (1996)

Table	Data	p-value	Uniform	Intrinsic $t=1$	Intrinsic $t=n$
34	{20, 0; 18, 5}	0.051	0.215	0.215	0.215
1	{8, 7; 2, 11}	0.054	0.170	0.170	0.253
18	{30, 1; 23, 4}	0.173	0.551	0.551	0.406
38	{43, 4; 14, 5}	0.106	0.395	0.395	0.340
16	{7, 4; 4, 6}	0.395	0.451	0.451	0.497

NOTE: The tables are ordered by p-values, which are calculated using Fisher's exact test. Posterior probabilities are also shown for the uniform prior and both ends of the intrinsic range, $t=1$ and $t=n$. Note that the value for $t=1$ is identical to that of the uniform prior (which corresponds to $t=0$).

For observed data $\mathbf{y}=\{y_{ij}\}$ with $\sum_{ij} y_{ij}=n$, we take a candidate distribution

$$\begin{aligned}\mathbf{x}=(x_{ij}) &\sim \text{Multinomial}(n,\hat{\theta}_{11},\ldots,\hat{\theta}_{ab}),\\ \hat{\theta}_{ij}&=\frac{y_{ij}+1}{n+ab}, \qquad i=1,\ldots,a, j=1,\ldots,b,\end{aligned} \tag{8}$$

where the cell probabilities are slightly modified to avoid zero entries. We then generate $\mathbf{x}_k, k=1,\ldots,M$, and use the Monte Carlo average,

$$B_{10}(y,t)=\frac{(t+ab-1)!}{(t+n+ab-1)!}\frac{1}{M}\sum_{k=1}^{M}\frac{\binom{t}{\mathbf{x}_k}(\prod r_i(\mathbf{x}_k)!)(\prod c_j(\mathbf{x}_k)!)}{(\prod r_i(\mathbf{y})!)(\prod c_j(\mathbf{y})!)} \times\frac{\prod(x_{k_{ij}}+y_{ij})!}{\prod x_{k_{ij}}!}\frac{1}{\binom{t}{\mathbf{x}_k}\prod_{ij}\hat{\theta}_{ij}^{x_{k_{ij}}}},$$

to calculate the Bayes factor (7). Calculation of the Monte Carlo sum is typically fast, and 30,000 random vectors are sufficient for most tables.

4. EXAMPLES AND EVALUATIONS

In this section we evaluate the performance of the intrinsic posterior probabilities with a simulation study and a number of examples. We pay particular attention to the range of posterior answers to check when robustness is present.

Recall from Section 3 that we derived the intrinsic posterior probability of the null under two different sampling models, P_1 and P_2. Operationally, we found that for the most part, the posterior probabilities tend to be similar under these two models. In what follows, we compute all of the posterior probabilities under the sampling model P_1, assuming only that the table total is fixed. In the absence of firm information to the contrary, this model seems to be the most likely sampling model under which contingency table data are collected.

The training sample size t has a natural range from 1 to n, because taking t larger than n concentrates more mass near the null. Moreover, as $t\to\infty$, the posterior probability of H_0 goes to 1. Thus the behavior of the posterior probability for the range of t from 1 to n is of interest; if this probability remains flat, then we interpret this as evidence of robustness.

4.1 2 × 2 Tables

Efron (1996) analyzed data from a multicenter trial to see whether a new surgical method for ulcers was superior to an older method (see also Casella 2001). For each of 39 hospitals, Efron provided a 2 × 2 table, along with the successes and failures for each of the hospitals. (The notation $\{a,b;c,d\}$ denotes a two-way table with first row $\{a,b\}$ and second row $\{c,d\}$, with the rows corresponding to the treatments. Thus in the table {8, 7; 2, 11}, one treatment had a success rate of 8/15 and the other treatment had a success rate of 2/13.)

Inspection of these tables reveals much variability in both the number of patients and the success rates of the table. The first two tables in Table 2, 34 and 1, suggest an association, a conclusion strongly supported by the intrinsic prior analysis. Throughout the entire range of t, the posterior probability of H_0 remains <0.5.

The next table, 18, suggests moderate deviation from the null, and the range of intrinsic posterior probabilities crosses 0.5, indicating nonrobustness of the inference. That is, the data are not conclusive in either direction, and a firm conclusion cannot be drawn here. (Recall that we interpret p-values and posterior probabilities on different scales. Typically, posterior probabilities of $H_0<0.5$ are considered evidence against H_0, while p-values of ≤ 0.05 are considered evidence against H_0.) Note that both the uniform posterior probability and the intrinsic with $t=1$ accept the null hypothesis. This illustrates a property of priors, such as the uniform, that put a lot of mass at the extremes of the parameter space. We have observed that such priors tend to be biased toward H_0, but documenting this bias is difficult.

The final two tables, 38 and 16, also represent robust cases. The intrinsic posterior probabilities are quite flat and never cross 0.5. Table 38 presents stronger evidence against the null, while table 16 presents stable but weak evidence against the null.

In our view, examining the range of the probabilities corresponding to the intrinsic priors is more informative than just using the uniform prior. The variability of the posterior probability as a function of t provides much information about the robustness of our conclusion.

4.2 The Tables of Good and Crook (1987)

Good and Crook (1987) analyzed 21 contingency tables, many drawn from the literature and some artificial. We reanalyzed these tables to demonstrate the performance of our procedure and also to contrast it with frequentist p-values and the robust procedure of Good and Crook. Table 3 summarizes the results for the 21 tables, showing the exact p-values (for 2 × 2 tables, the p-values are calculated using Fisher's exact test. For larger tables, the "exact" calculation generates a large sample (we used 100,000) from all tables with the same margins to use

Table 3. The 21 tables of Good and Crook (1987)

		p-value				Post. prob. of H_0^*		
							Intrinsic	
Table	Data	Exact (U)	Exact (W)	Volume (U)	Volume (W)	Good/ Crook	$t=1$	$t=n$
1	{10, 3; 2, 15}	0.001	0.000	0.088	0.000	0.015	0.003	0.016
2	{29, 33; 131, 78}	0.028	0.001	0.450	0.027	0.217	0.312	0.257
3	{200, 8; 182, 20}	0.018	0.000	0.488	0.020	0.156	0.361	0.179
4	{105, 5; 88, 11}	0.116	0.012	0.578	0.097	0.294	0.609	0.371
5	{409, 3; 174, 8}	0.005	0.000	0.071	0.005	0.083	0.284	0.078
6	{225, 53, 206; 3, 1, 12}	0.000	0.000	0.000	0.000	0.125	0.933	0.241
7	{38, 60, 28; 65, 138, 68; 35, 67, 30}	0.764	0.452	0.998	0.765	0.123	0.997	0.823
8	{61, 12, 60; 17, 6, 1; 39, 22, 7}	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	{17, 4, 8; 5, 12, 0; 10, 3, 13}	0.000	0.000	0.087	0.000	0.000	0.000	0.000
10	{58, 52, 1; 26, 58, 3; 8, 12, 9}	0.000	0.000	0.000	0.000	0.303	0.000	0.000
11	{2, 2, 2; 2, 2, 2; 2, 2, 2}	1.00	0.846	1.00	0.995	0.450	0.648	0.701
12	{6, 6, 6; 6, 6, 6; 6, 6, 6}	1.00	0.977	1.00	1.00	0.327	0.891	0.839
13	{1, 2, 3; 1, 2, 3; 1, 2, 3}	1.00	0.817	1.00	0.995	0.500	0.696	0.740
14	{1, 5, 20; 1, 5, 20; 1, 5, 20}	1.00	0.948	1.00	1.00	0.294	0.988	0.855
15	{5, 0, 0; 5, 0, 0; 5, 0, 0}	1.00	1.00	1.00	0.985	0.520	0.964	0.872
16	{6, 0, 0; 0, 6, 0; 0, 0, 6}	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	{5, 1, 0; 4, 0, 2; 2, 4, 0}	0.033	0.001	0.301	0.026	0.200	0.080	0.121
18	{68, 119, 26, 7; 20, 84, 17, 94; 15, 54, 14, 10; 5, 29, 14, 16}	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18*A*	{4, 1, 1, 0; 2, 3, 0, 3; 1, 2, 2, 0; 0, 0, 0, 1}	0.113	0.009	0.603	0.123	0.277	0.101	0.064
19	Income and no. of children	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20	Horsekick data	0.279	0.000	1.00	0.279	0.062	0.021	0.999

NOTE: The p-values were computed using the exact formula (1), adapted to unweighted (U) or weighted (W) p-values or weighted or unweighted volume tests. All simulations used 100,000 iterations. The Good–Crook probabilities are from Good and Crook (1987). The uniform prior corresponds to the intrinsic prior with $t = 0$. In all calculations, this posterior probability was identical to that of the intrinsic prior with $t = 1$, so the column of uniform posterior probabilities is not shown.

as a reference distribution. This is easily done with the `R` function `chisq.test`), the posterior probabilities from Good and Crook, and the resulting ranges of the posterior probabilities of the null for intrinsic priors when the concentration parameter t varies from $t = 1$ to $t = n$.

We include the p-values from the frequentist tests described in Section 1.1, along with a modification of the volume test, The weights in the volume test tend to correct the excessive conservativeness of the unweighted test.

4.2.1 Comparison With Frequentist p-Values. Of the four p-values computed, the volume-unweighted is the more conservative test, accepting the null in all tables except 6, 8, 10, 16, 18, and 19. Both the exact-unweighted and the volume-weighted provide essentially the same p-value, and they reject the null in the preceding six tables plus in the additional six tables 1, 2, 3, 5, 9, and 17. The exact-weighted is the less-conservative test, which rejects the null in the preceding 12 tables plus in tables 4, 18*A*, and 20.

In all cases where the p-values of the exact-unweighted and volume-weighted were significant, the intrinsic posterior probabilities were <0.5, except in tables 6 (Fienberg data), 18*A*, and 20. In tables 6 and 20, robustness of the posterior answers was not present, and for the sparse table 18*A*, the disagreement was not only between the p-values and intrinsic posterior probabilities, but also among the p-values. This suggests that the answer of the exact-unweighted and volume-weighted tests may be very close to that of the objective Bayesian intrinsic prior.

But this is not the case, and in fact the p-value can be a poor tool as a measure of the uncertainty of an hypothesis. For instance, tables 11 and 12 have the same structure; all cells contain the same counts, clearly favoring the independence between rows and columns, and the p-values of exact-unweighted and volume-weighted are equal to 1, even when the sample sizes are different, and our uncertainty on the null hypotheses should be larger in table 11 than in table 12. Nevertheless, the intrinsic posterior probabilities of the null in table 11 are in the interval $(0.648, 0.701)$, and those in table 12 are in the interval $(0.839, 0.891)$. This behavior seems more reasonable than that of the p-values.

Furthermore, it is well known that the meaning of the p-values is affected by the dimension of the parameter space of the models. To correct this, some dimension corrections have been introduced for testing hypotheses in linear models (e.g., Mallows C_p, adjusted R^2, Akaike information criterion). (For discussion on this topic, see Sellke, Bayarri, and Berger 2001; Girón et al. 2006; Moreno and Girón 2008.) Therefore, the general rule of rejecting the null when the p-value is <0.05 does not take into account that the p-values must be calibrated before such a rule can be set. For instance, the p-values of the exact-unweighted and volume-weighted tests for tables 4 and 18*A* are approximately 0.11, and the dimensions of the tables are

2×2 and 4×4, respectively. The question is whether a p-value of 0.11 has the same meaning for both tables. We see that the intrinsic analysis does not provide the same answer for both tables; for table 4, the answer is nonrobust, but for table 18*A*, the rejection of the null is robust.

4.2.2 Comparison With the Good–Crook Procedure. The Good–Crook robust procedure rejects the null in 20 out of 21 cases, while the intrinsic prior rejects the null in 12 of 21 cases. The Good–Crook procedure supports only the null in table 15, for which the intrinsic posterior probabilities are in the interval $(0.872, 0.964)$, giving a robust acceptance of the null. Moreover, all four p-values are almost 1.

Of particular interest are the artificial tables 11–15, where the results of the Good–Crook robust analysis are surprising and not in agreement with our results. In each table the rows are exactly the same, and thus all provide evidence for the null, which is stronger with increasing sample size, and the ranges of the intrinsic posterior probabilities suggest robustness in accepting the null for all of them. The Good–Crook robust analysis rejects the null for tables 11, 12, and 14; provides no conclusion for table 13; and weakly accepts H_0 in table 15. Good and Crook explained that "*the data suggest that the two-way characterization is irrelevant; all* 9! *permutations of the interior of the table are the same*." We do not fully understand this reasoning.

4.2.3 Assessing Robustness. To illustrate the effect of varying the parameter t, we look at Figure 3, which examines the behavior of four interesting tables from Table 3.

The four tables of Figure 3 illustrate the entire range of possibilities. Some are robust either for or against H_0, whereas some are nonrobust, with conclusions that are dependent on the tails of the prior. In such cases, (e.g., Fienberg or horsekick), it is important to reassess the prior, because clearly the data alone cannot yield a conclusive decision.

An interesting case is provided by the Mendel data (table 7 in Table 3). The intrinsic Bayesian tests are robust, the range of intrinsic posterior probabilities is quite small and accept the null,

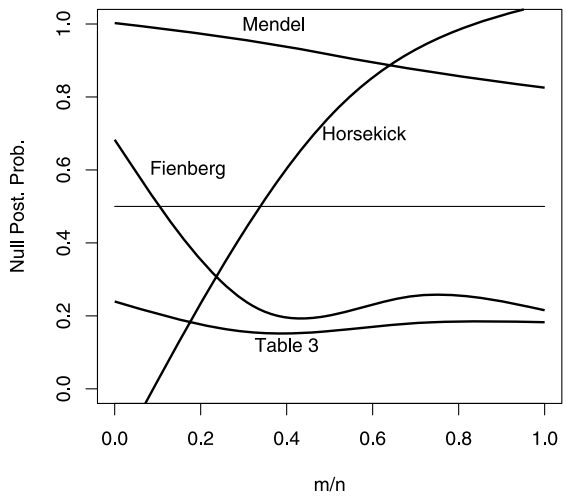

Figure 3. Ranges of posterior probabilities of four tables from Table 3. The Mendel data are table 7 (robust, evidence in favor of null), the Fienberg data are table 6 (nonrobust, evidence against the null for moderate t), the horsekick data are table 20 (nonrobust, evidence in favor of the null for large t), and data with no other name are table 3 (robust, evidence against the null).

and the p-values strongly support the null hypothesis. The historical consensus supports the null hypothesis (ignoring the debate about "cooked" data). But the Good–Crook robust analysis strongly rejects the null hypothesis, in opposition to what is commonly concluded about this data set. Good and Crook defended these conclusions, citing problems with computation and "flatness" of the margins.

As mentioned earlier, another instance in which the intrinsic posterior analysis leads to different conclusions is in table 6 (Bishop, Fienberg, and Holland 1978, p. 387), where the counts are very unbalanced. Here both the p-value and the Good–Crook robust analysis reject the null hypothesis, while the intrinsic posterior analysis accepts the null when t is small ($1 \leq t \leq 0.1n$) and rejects the null when t is large ($0.1n \leq t \leq n$). Bishop, Fienberg, and Holland (1978) presented three analyses of this table, all of which suggested some deviation from the null.

In contrast, the ranges of the intrinsic posterior probabilities of the null show robustness for most of the tables. Exceptions are some small unbalanced tables (4 and 6) and large-dimensional tables with small sample sizes (e.g., the horsekick table), where robustness is not present. For these situations, the message is that the data themselves are not conclusive, and thus we either need to add subjective information on the concentration parameter t or to collect more data.

5. DISCUSSION

The analysis of contingency tables is somewhat unique because of the discrepancy between the sampling model and the commonly used model for analysis. Specifically, calculating a test statistic conditional on both margins being fixed is the most common analysis, but the corresponding sampling model is almost impossible to realize. There have been many arguments both for and against the practice of conditioning on both margins, and we are not, in any way, joining that discussion. However, we note that from the frequentist standpoint, one reason for conditioning on both margins of a table is to obtain a reasonable reference set of tables for comparison with the observed table. Specifically, not only can the number of unconditional tables be prohibitively large (as can the number of conditional tables), but also the unconditional set can contain tables that are so extreme as to be impossible to ever observe. In our approach, this problem is handled by the fact that the intrinsic priors give little weight to such tables.

A Bayesian analysis of independence in a contingency table starts with a likelihood and a prior for the unrestricted table, where the likelihood reflects the sampling model. The prior typically reflects crude prior beliefs and evaluates the performance of the resulting procedure. It appears to be widely accepted that prior beliefs allow the parameters to be a priori dependent, as emphasized by Howard (1998). This property typically is not satisfied by the usual default prior for estimation (e.g., uniform), but is enjoyed by the Good and Crook (1987) priors and the intrinsic priors. Another important property, noted by Gunel and Dickey (1974), is that a prior should give mass to alternatives that are close to the null. This also can be accomplished with the intrinsic priors.

The Good–Crook mixtures of Dirichlet priors give high mass to extreme tables, as shown in Figure 2. To achieve symmetry, they take the mixing parameter to be α, the common exponent in the Dirichlet prior, and mix the parameter of the range $(0, \infty)$ using a heavy-tailed density to robustify the mixture. The intrinsic priors are mixtures as well [see (14) in Supplementary Materials], and the degree of concentration around the null is accommodated by a discrete parameter, t. To complete this analogy, we also could mix the parameter t with respect to a hyperprior with a heavy tail; however, the price that Good–Crook pays for the robustification of the procedure is to have a procedure that can result in unreasonable conclusions; this is apparent in tables 11, 12, 14, and 15.

We have learned the important lesson that our conclusions regarding independence between rows and columns are more sensitive to the prior than we suspected. Given the ranges of intrinsic posterior probabilities, it follows that for some tables, the conclusion can turn from "accept" to "reject" depending on the degree of concentration of the prior around the null. Unfortunately, we cannot classify the types of tables leading to this nonrobust behavior, although we suspect that imbalance and sparseness in cell sizes contribute to the sensitivity. Nonetheless, we do have a diagnostic that alerts us to situations where consideration of the prior information is an important factor in the inference.

We also note that, by construction, the range of the concentration parameters t of the intrinsic prior class, from $t = 1$ to $t = n$, is very reasonable. In terms of the tails of the prior, we range from extremely flat tails ($t = 1$) to tails equal to those of the likelihood of the data ($t = n$). This is a natural bound, because situations in which more weight would be given to the prior than to the data are very rare. Thus we have a natural class of priors for assessing robustness.

The performance of the intrinsic posterior probabilities, when starting with the unconditional likelihood, is extremely attractive. It seems to be robust when the data are sufficiently informative, and when the information that they provide is weak (often reflected in imbalance or sparseness of the table), we receive a warning that the resulting tests are not robust, and that more prior information or more data are needed.

APPENDIX: CONSISTENCY

Here we give a detailed proof of Theorem 1, consistency of the intrinsic posterior for the binomial and multinomial cases, and indicate how the proof extends to more general cases.

A.1 Proof of Theorem 1

For the case of Theorem 1,

$$H_0 : f(y|\theta_0) \quad \text{vs.} \quad H_1 : \{f(y|\theta), \pi^I(\theta|t)\},$$

the default marginal distributions are

$$\begin{aligned} m_0(y) &= \binom{n}{y}\theta_0^y(1-\theta_0)^{n-y} \qquad \text{and} \\ m_1(y) &= \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\,d\theta = \frac{1}{n+1}, \end{aligned} \tag{8}$$

leading to the intrinsic prior

$$\pi^I(\theta|t) = (t+1)\sum_{x=0}^{t}\binom{t}{x}\theta_0^x(1-\theta_0)^{t-x}\binom{t}{x}\theta^x(1-\theta)^{t-x}$$

and intrinsic marginal

$$m^I(y) = \int \binom{n}{y}\theta^y(1-\theta)^{n-y}\pi^I(\theta)\,d\theta. \tag{9}$$

We want to show that the Bayes factor $B_{10} = m^I(y)/m_0(y)$ goes to 0 under H_0 and ∞ under H_1. To show that B_{10} goes to ∞ under H_1, first note that

$$\begin{aligned} \pi^I(\theta|t) &\geq (t+1)\sum_{x=0}^{t}\binom{t}{x}\theta_0^x(1-\theta_0)^{t-x}\theta^x(1-\theta)^{t-x} \\ &= (t+1)[\theta\theta_0 + (1-\theta)(1-\theta_0)]^t \\ &\geq (t+1)\min(\theta_0, 1-\theta_0)^t = K, \end{aligned}$$

where we have used the fact that $\binom{t}{x} \geq 1$. Thus

$$\begin{aligned} B_{10} &\geq K\frac{\int \binom{n}{y}\theta^y(1-\theta)^{n-y}\,d\theta}{\binom{n}{y}\theta_0^y(1-\theta_0)^{n-y}} \\ &= \frac{K}{n+1}\frac{\binom{n}{y}^{-1}}{\theta_0^y(1-\theta_0)^{n-y}}. \end{aligned} \tag{10}$$

Stirling's approximation yields

$$\begin{aligned} \binom{n}{y}^{-1} &\approx n^{1/2}\left(\frac{y}{n}\right)^{y+1/2}\left(\frac{n-y}{n}\right)^{n-y+1/2} \\ &\approx n^{1/2}\theta^{n\theta+1/2}(1-\theta)^{n(1-\theta)+1/2}, \end{aligned}$$

because $y \approx n\theta$ as $n \to \infty$. Substituting into (10) and rearranging terms yields

$$B_{10} \geq \frac{Kn^{1/2}}{n+1}\left[\frac{\theta^\theta(1-\theta)^{1-\theta}}{\theta_0^\theta(1-\theta_0)^{1-\theta}}\right]^n. \tag{11}$$

Finally, note that $a(\theta) = \frac{\theta^\theta(1-\theta)^{1-\theta}}{\theta_0^\theta(1-\theta_0)^{1-\theta}}$ is minimized (and equal to 1) at $\theta = \theta_0$. Thus, for any fixed θ in H_1, we have that $a(\theta) = 1 + \varepsilon$ for some $\varepsilon > 0$, and thus

$$B_{10} \geq \frac{Kn^{1/2}}{n+1}(1+\varepsilon)^n \to \infty,$$

as $n \to \infty$. Thus, for any θ in H_1, the Bayes factor goes to infinity, and the posterior probability of H_0 goes to 0.

To establish consistency if θ_0 is the true parameter, we can bound $\pi^I(\theta|t)$ from above, and arrive at (10) as an upper bound (with a different value of K that will depend on t and θ_0 but not n or y). Under H_0, $y \approx n\theta_0$, so we obtain the right side of (11) as an upper bound, but with the expression in square brackets equal to 1, showing that $B_{10} \to 0$ as $n \to \infty$. Thus if the parameter value is in H_0, then the Bayes factor goes to 0 and the posterior probability of H_0 goes to 1, and the consistency is established.

A.2 Consistency for the Multinomial Case

For the multinomial case, the arguments are similar to those in Section A.1. The models are

$$\begin{aligned} M_0 &: M(\mathbf{x}|t, p_iq_j), \qquad p_iq_j \text{ fixed}, \\ M_1 &: M(\mathbf{x}|t, \theta_{ij}), \qquad \pi(\boldsymbol{\theta}) = \Gamma(ab), \\ t_1(\mathbf{x}) &= \frac{\Gamma(ab)}{\Gamma(t+ab)}\binom{t}{\mathbf{x}}\prod_{ij}x_{ij}!, \end{aligned}$$

where θ is the vector of θ_{ij}. The intrinsic prior is

$$\pi^I(\boldsymbol{\theta}) = \Gamma(t+ab)\sum_{x_{ij}}\frac{\prod_{ij}(p_iq_j)^{x_{ij}}}{\prod_{ij}x_{ij}!}M(\mathbf{x}|t, \theta_{ij}).$$

Using similar arguments to those in Section A.1, we can bound π^I either above or below (depending on what is needed) with a bound independent of θ and n. Denoting this bound by K, the Bayes factor is thus

$$B_{10} \approx K \frac{\int M(\mathbf{y}|n, \theta_{ij})\, d\boldsymbol{\theta}}{M(\mathbf{y}|n, p_i q_j)} = \frac{K}{\Gamma(n+ab)} \frac{\prod_{ij} y_{ij}!}{\prod_{ij} (p_i q_j)^{y_{ij}}}.$$

Using Stirling's approximation and replacing y_{ij} with $n\theta_{ij}$ yields

$$\frac{\prod_{ij} y_{ij}!}{\Gamma(n+ab)} \approx \frac{1}{n^{ab-1}} \prod_{ij} \theta_{ij}^{n\theta_{ij}},$$

giving the Bayes factor

$$B_{10} \approx \frac{K}{n^{ab-1}} \left[\prod_{ij} \left(\frac{\theta_{ij}}{p_i q_j} \right)^{\theta_{ij}} \right]^n.$$

Under H_0, the expression in square brackets is equal to 1, and $B_{10} \to 0$ as $n \to \infty$. So if H_0 is true, then the posterior probability of H_0 goes to 1. If H_1 is true, then the expression in square brackets is equal to $1+\varepsilon$, for some $\varepsilon > 0$, and $B_{10} \to \infty$ as $n \to \infty$. This follows from the fact that for positive constants $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ satisfying $\sum_i a_i = \sum_i b_i = 1$, $\prod_i (\frac{a_i}{b_i})^{a_i} \geq 1$, with strict inequality unless $a_i = b_i$ for all i. (This is readily verified by taking logs and using Jensen's inequality.) So if H_1 is true, then the posterior probability of H_0 goes to 0.

A.3 Extensions

So far, we have proved the consistency of the Bayes factor for testing sharp null hypothesis for models such as

$$M_0 : f(\mathbf{y}|\theta_0) \quad \text{vs.} \quad M_1 : \{f(\mathbf{y}|\theta), \pi^I(\theta|\theta_0, t)\}, \tag{12}$$

where $\pi^I(\theta|\theta_0, t)$ denotes the intrinsic prior for θ conditional on the null θ_0, on the training sample of size t, and $f(y|\theta)$ a binomial or multinomial sampling model. Here we extend consistency to the case where the null is not a point, but rather a subspace $H_0 : \theta_0 \in \Theta_0 \subset \Theta$.

The nested Bayesian models are now

$$M_0 : \{f(\mathbf{y}|\theta_0), \pi_0(\theta_0)\} \quad \text{vs.} \quad M_1 : \{f(\mathbf{y}|\theta), \pi^I(\theta|t)\}, \tag{13}$$

where $\pi_0(\theta_0)$ is a probability density, and the intrinsic prior for θ is given by

$$\pi^I(\theta|t) = \int_{\Theta_0} \pi^I(\theta|\theta_0, t) \pi_0(\theta_0)\, d\theta_0 = \pi_1(\theta) E_{\mathbf{x}|\theta} \frac{t_0(\mathbf{x})}{t_1(\mathbf{x})},$$

with $m_0(\mathbf{x}) = \int f(\mathbf{x}|\theta_0)\pi_0(\theta_0)\, d\theta_0$, $m_1(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)\, d\theta$, and $\pi(\theta)$ the default prior for $f(y|\theta)$.

Theorem 2. Assume the following for any $\theta_0 \in \Theta_0$:

(i) the Bayes factor

$$B_{10}(\mathbf{y}; \theta_0, t) = \frac{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|\theta_0, t)\, d\theta}{f(\mathbf{y}|\theta_0)}$$

is consistent for testing the sharp null hypothesis (12),

(ii) the function $f(y|\theta_0)$ is a continuous function of θ_0,

(iii) the set Θ_0 is a compact set,

(iv)

$$k_t' = \inf_{\mathbf{x}} \frac{t_0(\mathbf{x})}{f(\mathbf{x}|\theta_0)} > 0, \qquad k_t = \sup_{\mathbf{x}} \frac{t_0(\mathbf{x})}{f(\mathbf{x}|\theta_0)} < \infty.$$

Then the Bayes factor for testing (13)

$$B_{10}(\mathbf{y}, t) = \frac{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|t)\, d\theta}{\int_{\Theta_0} f(\mathbf{y}|\theta_0)\pi(\theta_0)\, d\theta_0},$$

is consistent.

Proof. Suppose first that we are sampling from a distribution $f(y|\theta_0^*)$, where θ_0^* is an arbitrary but fixed null point. For sufficiently large n, we have

$$B_{10}(\mathbf{y}, t) = \frac{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|t)\, d\theta}{\int_{\Theta_0} f(\mathbf{y}|\theta_0)\pi_0(\theta_0)\, d\theta_0} \approx \frac{1}{k} \frac{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|t)\, d\theta}{f(\mathbf{y}|\hat{\theta}_0)\pi_0(\hat{\theta}_0)},$$

where $k = \int_{\Theta_0} d\theta_0$, and $\hat{\theta}_0$ is the maximum likelihood estimator (MLE) of θ_0. Then the intrinsic prior can be bounded as

$$\pi^I(\theta|t) = \pi_1(\theta) E_{\mathbf{x}|\theta} \frac{f(\mathbf{x}|\hat{\theta}_0)}{m_1(\mathbf{x})} \frac{m_0(\mathbf{x})}{f(\mathbf{x}|\hat{\theta}_0)} < k_m \pi_1(\theta) E_{\mathbf{x}|\theta} \frac{f(\mathbf{x}|\hat{\theta}_0)}{m_1(\mathbf{x})} = k_m \pi^I(\theta|\hat{\theta}_0, t).$$

Substituting in B_{10}, we have

$$B_{10}(\mathbf{y}, t) < \frac{k_m}{k\pi_0(\hat{\theta}_0)} \frac{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|\hat{\theta}_0, t)\, d\theta}{f(\mathbf{y}|\hat{\theta}_0)} \approx \frac{k_m}{k\pi_0(\theta_0^*)} \frac{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|\theta_0^*, t)\, d\theta}{f(\mathbf{y}|\theta_0^*)} \to 0,$$

where the last expression tends to 0 because $B_{10}(\mathbf{y}; \boldsymbol{\theta}_0^*, t)$ is consistent. This proves consistency under the null.

Suppose that we are sampling from a distribution $f(y|\theta^*)$, where θ^* is an arbitrary but fixed alternative point. We have

$$B_{01}(\mathbf{y}, t) = \frac{\int_{\Theta_0} f(\mathbf{y}|\theta_0)\pi_0(\theta_0)\, d\theta_0}{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|t)\, d\theta} < \frac{f(\mathbf{y}|\hat{\theta}_0)}{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|t)\, d\theta}.$$

Let $\tilde{\theta}_0$ denote the limit of the MLE $\hat{\theta}_0$ when sampling from θ^*. Then the intrinsic prior can be written as

$$\pi^I(\theta|t) = \pi_1(\theta) E_{\mathbf{x}|\theta} \frac{f(\mathbf{x}|\hat{\theta}_0)}{m_1(\mathbf{x})} \frac{m_0(\mathbf{x})}{f(\mathbf{x}|\hat{\theta}_0)} > k_m' \pi_1(\theta) E_{\mathbf{x}|\theta} \frac{f(\mathbf{x}|\hat{\theta}_0)}{m_1(\mathbf{x})} = k_m' \pi^I(\theta|\hat{\theta}_0, t).$$

Substituting in B_{01}, we have, for large n,

$$B_{01}(\mathbf{y}, t) < \frac{1}{k_m'} \frac{f(\mathbf{y}|\hat{\theta}_0)}{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|\hat{\theta}_0, t)\, d\theta} \approx \frac{1}{k_m'} \frac{f(\mathbf{y}|\tilde{\theta}_0)}{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|\tilde{\theta}_0, t)\, d\theta} \to 0,$$

where the last expression tends to 0 because $B_{01}(\mathbf{y}; \tilde{\theta}_0, t)$ is consistent. This completes the proof of Theorem 2.

Both the binomial and multinomial distributions satisfy the conditions in Theorem 2.

SUPPLEMENTAL MATERIALS

Appendixes B and C: Details of the calculations of the Bayes factors for general $a \times b$ tables (B), and the data for tables 19 and 20 (C) (.pdf file).

[Received February 2008. Revised March 2009.]

REFERENCES

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables" (with discussion), *Statistical Science*, 7, 131–177.

——— (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley.

Agresti, A., and Hitchcock, D. B. (2005), "Bayesian Inference for Categorical Data Analysis," *Statistical Methods and Applications: Journal of the Italian Statistical Society*, 14, 297–330.

Albert, J. H. (1987), "Empirical Bayes Estimation in Contingency Tables," *Communications in Statistics—Theory and Methods*, 16, 2459–2485.

——— (1990), "A Bayesian Test for a Two-Way Contingency Table Using Independence Priors," *Canadian Journal of Statistics*, 18, 347–363.

——— (2004), "Bayesian Methods for Contingency Tables," in *Encyclopedia of Biostatistics*, New York: Wiley.

Albert, J. H., and Gupta, A. K. (1982), "Mixtures of Dirichlet Distributions and Estimation in Contingency Tables," *The Annals of Statistics*, 10, 1261–1268.

——— (1983), "Bayesian Estimation Methods for 2×2 Contingency Tables Using Mixtures of Dirichlet Distributions," *Journal of the American Statistical Association*, 78, 708–717.

Altham, P. M. E. (1969), "Exact Bayesian Analysis of a 2×2 Contingency Table, and Fisher's Exact Significance Test," *Journal of the Royal Statistical Society, Ser. B*, 31, 261–269.

Bayes, T. (1783), "An Essay Towards Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society*, 53, 370–418.

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

——— (1994), "An Overview of Robust Bayesian Analysis" (with discussion), *Test*, 3, 5–124.

Berger, J. O., and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of p-Values and Evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112–122.

Bernardo, J. M. (1979), "Reference Posterior Distributions for Bayesian Inference" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 41, 113–147.

Bishop, Y. M. N., Fienberg, S. E., and Holland, P. W. (1978), *Discrete Multivariate Analysis*, Cambridge: MIT Press.

Casella, G. (2001), "Empirical Bayes Gibbs Sampling," *Biostatistics*, 2, 485–500.

Casella, G., and Berger, R. L. (1987), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem" (with discussion), *Journal of the American Statistical Association*, 82, 106–111.

Consonni, G., and La Rocca, L. (2008), "Tests Based on Intrinsic Priors for the Equality of Two Correlated Proportions," *Journal of the American Statistical Association*, 103, 1260–1269.

Crook, J. F., and Good, I. J. (1980), "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," *The Annals of Statistics*, 8, 1198–1218.

Diaconis, P., and Efron, B. (1985), "Testing for Independence in a Two-Way Table: New Interpretations of the Chi-Squared Statistic" (with discussion), *The Annals of Statistics*, 13, 845–913.

Efron, B. (1996), "Empirical Bayes Methods for Combining Likelihood" (with discussion), *Journal of the American Statistical Association*, 91, 538–565.

Epstein, L. D., and Fienberg, S. E. (1992), "Bayesian Estimation in Multidimensional Contingency Tables," in *Bayesian Analysis in Statistics and Econometrics*, New York: Springer, pp. 27–41.

Girón, J., Martinez, L., Moreno, E., and Torres, F. (2006), "Objective Testing Procedures in Linear Models: Calibration of the p-Values," *Scandinavian Journal of Statistics*, 33, 765–787.

Good, I. J. (1967), "A Bayesian Significance Test for Multinomial Distributions," *Journal of the Royal Statistical Society, Ser. B*, 29, 399–431.

——— (1976), "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," *The Annals of Statistics*, 4, 1159–1189.

Good, I. J., and Crook, J. F. (1987), "The Robustness and Sensitivity of the Mixed Dirichlet Bayesian Test for 'Independence' in Contingency Tables," *The Annals of Statistics*, 15, 670–693.

Gûnel, E., and Dickey, J. (1974), "Bayes Factors for Independence in Contingency Tables," *Biometrika*, 61, 545–557.

Howard, J. V. (1998), "The 2×2 Table: A Discussion From a Bayesian Viewpoint," *Statistical Science*, 13, 351–367.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), London: Oxford University Press.

Kadane, J. B., Moreno, E., Perez, M. E., and Pericchi, L. R. (2002), "Applying Nonparametric Robust Bayesian Analysis to Non-Opinionated Judicial Neutrality," *Journal of Statistical Planning and Inference*, 102, 425–439.

Kass, R., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–793.

Laird, N. M. (1978), "Empirical Bayes Methods for Two-Way Contingency Tables," *Biometrika*, 65, 581–590.

Laplace, P. S. (1812), *Theorie Analytique des Probabilities*, Paris: Courcier.

Leonard, T. (1975), "Bayesian Estimation Methods for Two-Way Contingency Tables," *Journal of the Royal Statistical Society, Ser. B*, 37, 23–37.

Mehta, C. R., Patel, N. R., and Senchaudhuri, P. (2000), "Efficient Monte Carlo Methods for Conditional Logistic Regression," *Journal of the American Statistical Association*, 95, 99–108.

Moreno, E. (1997), "Bayes Factor for Intrinsic and Fractional Priors in Nested Models. Bayesian Robustness," in *L_1-Statistical Procedures and Related Topics. Lecture Notes—Monograph Series*, Vol. 31, ed. D. Yadolah, Hayward, CA: Institute of Mathematical Statistics, pp. 257–270.

Moreno, E., and Girón, F. J. (2008), "Comparison of Bayesian Objective Procedures for Variable Selection in Linear Regression," *Test*, 3, 472–490.

Moreno, E., Bertolino, F., and Racugno, W. (1998), "An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing," *Journal of the American Statistical Association*, 93, 1451–1460.

Morris, C. N. (1987), Discussion of "Testing a Point Null Hypothesis: The Irreconcilability of p-Values and Evidence," by J. O. Berger and T. Sellke, *Journal of the American Statistical Association*, 82, 131–133.

Nazarret, W. (1987), "Bayesian Log-Linear Estimates for Three-Way Contingency Tables," *Biometrika*, 74, 401–410.

Novick, M. R., and Grizzle, J. E. (1965), "A Bayesian Approach to the Analysis of Data From Clinical Trials," *Journal of the American Statistical Association*, 60, 81–96.

Novick, M. R., and Hall, J. (1965), "A Bayesian Inference Procedure," *Journal of the American Statistical Association*, 60, 1104–1117.

Robert, C. P. (2001), *The Bayesian Choice* (2nd ed.), New York: Springer-Verlag.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of the p-Values for Testing Precise Null Hypotheses," *American Statistician*, 55, 62–71.

Zellner, A. (1977), "Maximal Data Information Prior Distributions," in *New Methods in the Application of Bayesian Methods*, eds. A. Aykac and C. Binmat, Amsterdam: North-Holland.