Assessing Robustness of Intrinsic Tests of Independence in Twoway Contingency Tables

George Casella
University of Florida

Elías Moreno
University of Granada, Spain

Abstract: A condition needed for testing nested hypotheses from a Bayesian viewpoint 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. Further, the degree of concentration of the priors is controlled by a discrete parameter m, the training sample size, which plays an important role in the resulting answer.

In this paper we study, for small or moderate sample sizes, robustness of the tests of independence in contingency tables with respect to intrinsic priors with different degrees of concentration around the null. We compare these tests with frequentist tests and the robust Bayes tests of Good and Crook. For large sample sizes robustness is achieved since the intrinsic Bayesian tests are consistent.

We also discuss conditioning issues and sampling schemes, and argue that conditioning should be on either one margin or the table total, but not on both margins.

Examples using real are simulated data are given.

Key words and phrases: Bayesian inference, Bayes factors, Monte Carlo integration, Monte Carlo methods, chi-squared tests.

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); there is also the textbook Agresti(1996).

Exact frequentist inference in contingency tables can be done by applying the same test statistic to all tables with the same marginals, and 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). The properties of these tests have been investigated by Mehta and co-authors (see, for example, Mehta *et al.* 2000). An interesting alternative approach, called a volume test, was proposed by Diaconis and Efron (1985), which also contains a discussion on the meaning of the chi squared test.

Early Bayesian analyses of contingency tables were done by Altham (1969, 1971), where a product of binomials and beta priors was considered, but the focus was not on testing but rather on estimation of the difference of binomial parameters. Leonard (1975) and Nazarret (1987) use normal priors on the log functions of the cell probabilities for tables with fixed margins and Kadane et al. (2002) study robustness of some unimodal classes of priors. Gunel and Dickey (1974), Good (1976), Good and Crook (1987), and Albert and Gupta (1982,1983) use Dirichlet priors for the vector of parameters $p = (p_1, ..., p_k)$ of multinomial distributions. The hyperparameters in the Dirichlet are either subjectively determined or integrated out with the help of a hyperprior distribution.

1.1 Priors for Testing Hypothesis

Some developments have tried to be robust. For instance, Good and Crook (1987) start with a Dirichlet prior for p with hyperparameters $\alpha_1 = \cdots = \alpha_k = \alpha$, and then assume that α follows a hyperprior ranging from the log Normal distribution to the log Cauchy distribution. Their conclusion was that the log Cauchy hyperprior gives results that are the most robust.

In general, many priors that might be appropriate for estimation purposes cannot be recommended as priors for constructing Bayesian tests. This is because the null hypothesis is not taken into account in the formulation of the prior. Without doing so, it is impossible to guarantee that the prior distribution will concentrate around the null hypothesis, a condition that is widely accepted (see, for example, Jeffreys 1961, Chapter 5, Berger 1994, Berger and Sellke 1987, Casella and Berger 1987, Morris 1987), and should be required of a prior for testing a hypothesis.

Gunel and Dickey (1974), in discussing Bayes factors for contingency tables, note the importance of the *Savage continuity condition*. They argue that the most realistic priors will be "continuous probability densities with high concentrations on small neighborhoods of η_H" (in their notation η_H is the parameter value in the null). This is exactly what the intrinsic prior does here. Starting from a default prior, which will not concentrate probability near H_0, but instead will spread it out in H_1 giving high probability to models far from H_0, the intrinsic prior construction creates a new prior that (i) concentrates probability near H_0 and (ii) does it in a way that maintains consistency of the tests. That is, as the sample size becomes infinite, the test will always make the correct decision.

It is important to realize that if a prior on 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. This is important because, if H_0 is reasonable, it is important to be able to distinguish H_0 from reasonable alternatives, which will be close. Putting high prior probability on extreme models, far from H_0, is wasteful. If such models are truly generating the data, 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.

Lastly, we note that when the row or column totals are fixed, Howard (1998) argues that the prior parameters should not be independent.

1.2 Intrinsic Priors

Intrinsic priors were introduced in hypothesis testing in order to convert improper priors into proper ones (Berger and Pericchi 1996, Moreno 1997, Moreno *et al.* 1998) but there is no inherent limitation in using them when the default prior is proper. For testing

$$H_0 : \{f_0(x|p_0), \pi_0(p_0)\} \text{ vs. } H_1 : \{f_1(x|p_1), \pi_1(p_1)\}, \tag{1}$$

where $f_0(x|p_0)$ is nested in $f_1(x|p_1)$, $\pi_0(p_0)$ and $\pi_1(p_1)$ are default estimation priors, the intrinsic prior for p_1 conditional on H_0 is given by $\pi^I(p_1|H_0) = \pi_1(p_1)E_{p_1}[m_0(x)/m_1(x)]$ where $m_i(x), i = 0, 1$ are the respective marginals and the expectation is taken with respect to $f_1(x|p_1)$. This calculation can be done

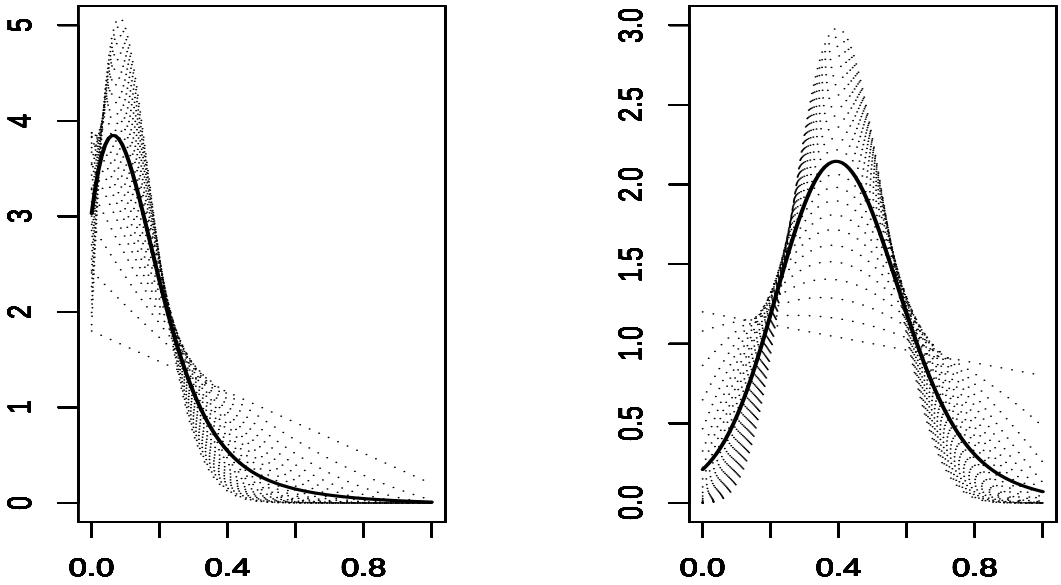

Figure 1: Intrinsic priors from the uniform prior for $p_0 = .1$ (left) and $p_0 = .4$ (right) for $m = 1, 2, \ldots, 25$. As m increases the prior concentrates more probability mass in the neighborhood of p_0. The solid curve is the average intrinsic prior.

whether or not any of the priors in (1) are proper or improper. When the priors are improper, it is typical to choose the sample size for x so that $m_i(x)$ is greater than zero and finite.

It is important to note that here we are using a *theoretical* x in that no actual data are used in the construction of the intrinsic prior. In our calculations, x will be distributed according to either $f_0(x|p_0)$ or $f_1(x|p_1)$, with sample size m. Furthermore, to avoid ambiguities in notation, for the data we will use sample size n and observations y, and for the theoretical training sample we will have sample size m and variables x.

As an example, consider the simple case of sampling from a binomial distribution $B(y|n, p)$ with n known. A default prior for estimation of the parameter p is usually chosen from one of the following distributions: the uniform (Bayes 1783, Laplace 1812), the Jeffreys's prior (Jeffreys 1961, Bernardo 1979), Zellner's prior (Zellner 1977), or that of Novik and Hall (Novik and Hall 1965). The first two are the most popular and they are proper. The third is proper and the fourth is improper. Any of these distributions can be used as reasonable default priors

for estimating p in the absence of subjective prior information (see, for example, Berger 1985, page 89). However, these priors are not appropriate as they do not concentrate mass around a null hypothesis, so they are not suitable for testing hypotheses.

For example, for the testing problem $H_0 : p = p_0$ *versus* $H_1 : p \neq p_0$, where p is the probability of success in a Bernoulli trial, 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* m, is

$$\begin{aligned} \pi^I(p|p_0, m) &= E^{M_1} \frac{B(x|m, p_0)}{\int_0^1 B(x|m, p)dp} \\ &= \frac{1}{m+1} \sum_{i=0}^{m} Be(p|i+1, m-i+1) Be(p_0|i+1, m-i+1) \end{aligned}$$

where the expectation is taken with respect to the binomial $B(x|m, p)$, and $Be(p|a, b)$ represents the beta distribution for p with parameters a and b. Figure 1 shows intrinsic priors for two values of p_0 and $m = 1, 2, \ldots 25$. The prior is always unimodal, and for $m = 1$ it is a linear function of p, but as m increases it concentrates more probability mass in the neighborhood of p_0. When we start with the Jeffreys prior and $m \geq 2$, the resulting intrinsic priors are very close to those obtained from the uniform. Thus, in this simple case, the intrinsic prior has provided a default prior that is centered on the null hypothesis. We will see that this is the case in more complex examples.

We note that the training sample size m plays an important role, in that it controls the concentration of the prior for the alternative parameter values around H_0. For small m, the intrinsic prior will remain close to the default prior. For large m, the intrinsic prior is nicely centered on H_0, and concentrates higher mass close to H_0. The point is to know how sensitive is the resulting Bayesian test to variation of the concentration parameter m, a point we will discuss later. Finally, we note that the construction of the prior is *fully automatic*.

The approach taken in this paper falls under what has come to be known as "objective" Bayes analysis, which typically results in the use of priors that may depend on the model structure. Objective Bayes analyses have become very popular (see Berger 2000, Clyde and George 2000, Kim and Sun 2000, Wasserman 2000, Berger *et al* 2001, Sweeting 2001, Berger and Pericchi 2006, Girón

et al. 2006), but are still not universally accepted. In the problem that we are considering, the use of the intrinsic prior within the objective Bayes framework produces an answer that is free from shortcomings of some previous approaches.

1.3 Summary

The remainder of the paper is organized as follows. In Section 2 we discuss the question of the appropriate model for sampling and inference, and conclude that the most appropriate model is the one with no restrictions on the marginal totals. 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 contained in Appendix A.) Consistency is looked at in Section 3.3, with technical details relegated to Appendix B. In Section 3.4 we also discuss calculation of the intrinsic priors, which requires summing over all possible tables with table total m, or row totals m_i. Section 4 evaluates the performance of intrinsic posterior probabilities with a number of examples, both real and artificial. Section 5 contains a concluding discussion.

2 Sampling Models and Inference

An $a \times b$ two-way contingency table is a vector of observations $\mathbf{y} = (y_{11}, \ldots, y_{1b}, y_{21}, \ldots, y_{2b}, y_{a1}, \ldots, y_{ab})$ of length ab satisfying $\sum y_{ij} = n$. We denote the row totals and columns totals by $r_i(\mathbf{y})$ and $c_j(\mathbf{y})$, respectively. The standard frequentist analysis of the table is a test of independence, with the classical test being Pearson's chi squared test. The justification of the test is asymptotic, and is based on the assumption of having a table with fixed margins, with a multinomial distribution conditional on the margins.

With small samples or sparse tables, Pearson's chi squared test is unreliable. Alternatives are based on considering sets of tables with the same margins as the observed table, and creating a p-value by counting how many of these alternate tables have test statistics more extreme that the observed. These so-called exact tests, which can be seen as generalizations of Fisher's exact test for a 2×2 table, have many variations (see Agresti 1992 for a review). The statistical package *StatXact* calculates many exact tests.

What concerns us about this approach is the restriction to the set of tables with the same margins as the observed table. As we note in Section 2.1, and as is well-known, there is virtually no realistic sampling scheme that will result in a table with fixed marginal totals (other than the Lady Tasting Tea). The most reasonable sampling scheme, which would only condition on the table total, is rarely used. The reasons for this are mostly technical, not statistical. This is a problem with a long history which we do not repeat here; the review by Agresti (1992) discusses the issues involved.

2.1 Sampling Models

Data like those in Table 1 lead us to question the validity of the usual technique of conditioning on both margins. To obtain data such as these, there are three possible sampling models that could have been used:

1. Continue sampling until a fixed total number, n, of patients is reached, allocating patients at random to the treatments. This is multinomial sampling.

2. Fix the number of patients to be allocated to each of the two treatments, and continue sampling until these numbers are reached, allocating patients at random to the treatments. This is binomial sampling.

3. Fix the number of successes for each of the treatments, and continue sampling until these numbers are reached, allocating patients at random to the treatments. This is negative binomial sampling.

Note that none of these sampling models would yield as data this contingency table with *both margins fixed*.

Although we do not know the sampling rule used to obtain this sample, it is fairly certain that patients were *not* sampled until there were 36 cases in which the cancer was controlled and 5 in which it was not. It is possible that it was decided to allocate 23 patients to surgery and 18 to radiation therapy, but what is most likely is the experiment was run with some kind of random allocation, and was stopped, for some reason, when there were 41 patients. Thus, if we were to consider a repetition of this experiment, it is reasonable to condition on the table total 41, or on the margin $(23, 18)$.

Table 1: Results of a study comparing radiation therapy with surgery in treating cancer of the larynx (see Agresti 1996, page 50)

	Cancer Controlled	Cancer not Controlled	
Surgery	21	2	23
Radiation Therapy	15	3	18
	36	5	41

Good and Crook (1987) talk about 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,

and they note that P_3 is not a very common sampling model, with P_1 and P_2 being more useful. They derive Bayesian tests under P_1 and P_2, using a prior based on a mixture of Dirichlet distributions. They illustrate the performance of their method on a number of example tables, both real and artificial. In general, their answers are reasonable, indicating that calibration of the set of all tables is possible. However, there are some disturbing anomalies. 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 report an average Bayes factor of 2.1, which would lead to a posterior probability of the null hypothesis of $1/(1 + 2.1) = .327$. (Our intrinsic procedure yields posterior probabilities of the null that varies between 0.839 and 0.891 as the concentration parameter m varies).

2.2 Sufficiency and Ancillarity

The procedure P_3 should be discounted as a practical sampling procedure, leaving us with P_1 and P_2. These latter procedures 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. There are $ab - 1$ free parameters, the cell probabilities. If sampling procedure

P_2 is used, where we fix the a row totals, then the distribution of the frequencies is that of a independent multinomials, each with $b-1$ free parameters and 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 we look at here is whether is it desirable to have such an equality.

Good and Crook (1987) state the following assumption

> Assumption 1: ("Ancillarity of the row totals"). The row totals alone (or the column totals alone) convey no evidence for or against H_0 under P_1.

They then argue that this assumption should be reflected in the chosen statistic, and they choose their prior to *force* this to be the case, although it can result in somewhat unreasonable results, as discussed in Sections 2.1 and 4.2. However, we think that P_1 and P_2 are distinct procedures with distinct structures, and consequently should have distinct statistics.

To defend our position, we look at the 2×2 case (although we could argue in the general case) and consider the joint density of x_{11}, r, and c, the $(1,1)$ observation, the total of the first row, and the total of the first column, respectively. Using p to denote the parameter, a direct factorization yields

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

Assumption 1 above requires that $f(r|p) \propto f(r|p_0)$, where p_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, $r \sim$ binomial$(n, \theta_{11} + \theta_{12})$, 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 do this simple case

to better understand the workings of the priors and the resulting probabilities; the general case is treated in Appendix A.

3.1 Margins Unrestricted

We start with a 2×2 contingency table with n individuals classified into four cells each having 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 $\theta_{ij}=p_iq_j$, 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

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

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 $M(\mathbf{y}|n,\theta_{ij})$. A default prior for the parameters (p_1,q_1) is $\pi(p_1,q_1)$=Uniform$(p_1|0,1)\times$ Uniform$(q_1|0,1)$.

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

$$M_0:\{f_0(\mathbf{x}|n,p_1,q_1),\pi(p_1,q_1)\} \text{ and } M_1:\{M(\mathbf{x}|n,\theta_{ij}),\mathcal{D}_3(\theta_{ij}|1,1,1,1)\}. \quad (2)$$

Notice 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 will then substitute the intrinsic prior $\pi^I(\theta_{ij}|m)$ for $\mathcal{D}_3(\theta_{ij}|1,1,1)$ in (2), where m is the training sample size.

It is straightforward to see that based on a training sample size m, the intrinsic prior for θ_{ij} is

$$\pi^I(\theta_{ij}|m) = \frac{(m+3)!}{[(m+1)!]^2}\sum_{\mathbf{x}:\sum_{ij}x_{ij}=m}\binom{m}{\mathbf{x}} \quad (3)$$

$$\times \quad \left(\prod_{i=1}^{2} r_i(x)!\right)\left(\prod_{j=1}^{2} c_j(x)!\right)\left(\prod_{i,j}\frac{\theta_{ij}^{x_{ij}}}{x_{ij}!}\right). \tag{4}$$

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, respectively. For a data set $y = y_{ij}$, the Bayes factor $B_{10}(y)$, for (3) versus a uniform prior for p_1 and q_1, is equal to $m_1^I(y|n)/m_0(y|n)$, where

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

and

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

If a priori we assume that $P(M_0) = P(M_1) = 1/2$, then for any training sample size m the posterior probability of the null is given by

$$P(M_0|y,m) = \frac{1}{1+B_{10}(y)}. \tag{5}$$

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

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

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

m_1 and m_2, the intrinsic prior is a convex combination of the product of beta distributions, that is

$$\begin{aligned}\pi^I(p_1,p_2|m_1,m_2) &= \sum_{i=0}^{m_1}\sum_{j=0}^{m_2}\binom{m_1}{i}\binom{m_2}{j}\frac{\Gamma(i+j+1)\Gamma(m_1+m_2-i-j+1)}{\Gamma(m_1+m_2+2)} \\ &\times \; Be(p_1|i+1,m_1-i+1)Be(p_2|j+1,m_2-j+1).\end{aligned}$$

We see that, under the intrinsic prior, the parameters p_1 and p_2 are not *a priori* independent.

For $m_1 = m_2 = 10$ Figure 2, left figure, displays the intrinsic prior. Note that the probability mass is concentrated around the line $p_1 = p_2$, and the prior is symmetric around this line. For contrast, we also show 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 ($\pi^U(p_0)$, $\pi^I(p_1,p_2|m_1,m_2)$), conditional on the sample (y_1,y_2), is given by

$$P(M_0|y_1,y_2,m_1,m_2) = \frac{1}{1+B_{10}(y_1,y_2)},$$

where

$$\begin{aligned}B_{10}(y_1,y_2) &= \left[\frac{n_1+n_2+1}{(n_1+m_1+1)(n_2+m_2+1)}\right]\left[\frac{(m_1+1)(m_2+1)}{m_1+m_2+1}\right] \\ &\times\binom{n_1+n_2}{y_1+y_2}\sum_{i=0}^{m_1}\sum_{j=0}^{m_2}\frac{\binom{m_1}{i}^2\binom{m_2}{j}^2}{\binom{m_1+m_2}{i+j}\binom{n_1+m_1}{y_1+i}\binom{n_2+m_2}{y_2+j}}. \qquad (6)\end{aligned}$$

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 m. However, 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. In particular, as the sample size n tends to infinity, the sampling distribution should overwhelm any prior information. Thus, we should be able to

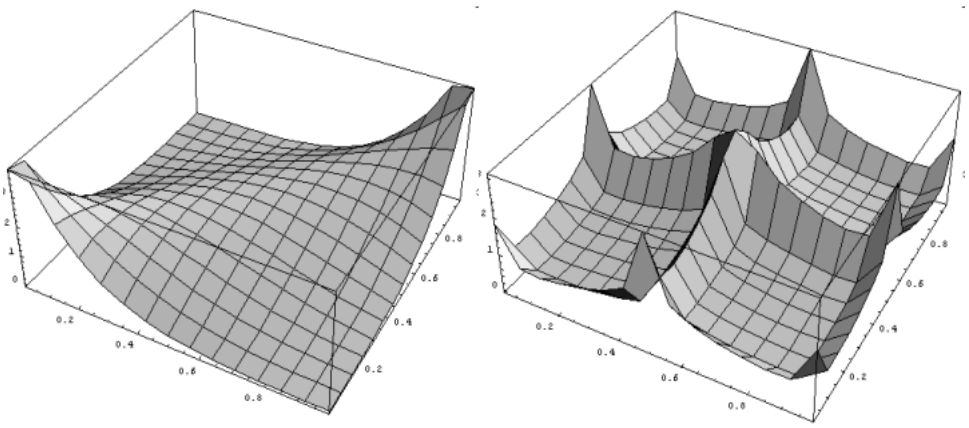

Figure 2: Intrinsic prior for (p_1, p_2) for $m_1 = 10$ and $m_2 = 10$ (left) and the log-Cauchy mixture of Dirichlets of Good and Crook (1987) (right panel)

prove consistency of the intrinsic Bayesian procedure for any finite training sample size m. Specifically, for any finite m, we want to insure that when sampling from the null

$$\lim_{n\to\infty} P(M_0|y, m) = 1,$$

and when sampling from the alternative

$$\lim_{n\to\infty} P(M_1|y, m) = 1.$$

Indeed, 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, discussed in Section 1.2. We can establish the following theorem:

Theorem 1 *For testing*

$$M_0 : B(y|\theta_0) \quad \textit{vs.} \quad M_1 : \{B(y|\theta), \pi^I(\theta|\theta_0, m)\},$$

the intrinsic posterior probability is consistent, for any finite training sample size m.

The theorem is proved in Appendix B. There we also extend the result to other cases considered in this paper.

3.4 Computational Issues

Note that calculating the intrinsic priors in Sections 3 and Appendix A necessitates summing over all tables with table total m. Although this can sometimes 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) there are $162,750,684,200,297,895$ tables with the same table total, and $2,689,129,357,824$ with the same row totals. Thus, to calculate the intrinsic priors we use a Monte Carlo sum.

Since 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, to calculate (11) for an $a \times b$ table, the Bayes factor for observed data $\mathbf{y} = \{y_{ij}\}$ with $\sum_{ij} y_{ij} = n$, we take a candidate distribution

$$\mathbf{x} = (x_{ij}) \quad \sim \quad \text{Multinomial } (n, \hat{\theta}_{11}, \ldots \hat{\theta}_{ab}), \tag{7}$$

$$\hat{\theta}_{ij} \quad = \quad \frac{y_{ij}+1}{n+ab}, \quad i = 1, \ldots a, \quad j = 1, \ldots, b \tag{8}$$

where the cell probabilities are slightly modified to avoid zero entries. We then generate $\mathbf{x}_k, \quad k = 1, \ldots M$ and calculate the Bayes factor (11) as

$$\begin{aligned} B_{10}(y) &= \frac{(m+ab-1)!}{(m+n+ab-1)!} \\ &\times \frac{1}{M} \sum_{k=1}^{M} \frac{\binom{m}{\mathbf{x}_k} \left(\prod r_i(\mathbf{x}_k)!\right) \left(\prod c_j(\mathbf{x}_k)!\right)}{\left(\prod r_i(\mathbf{y})!\right) \left(\prod c_j(\mathbf{y})!\right)} \frac{\prod (x_{k_{ij}} + y_{ij})!}{\prod x_{k_{ij}}!} \frac{1}{\binom{m}{\mathbf{x}_k} \prod_{ij} \hat{\theta}_{ij}^{x_{k_{ij}}}}. \end{aligned}$$

Calculation of the Monte Carlo sum is typically fast, and between 5000 and 30,000 random vectors have been sufficient for most tables. Even tables with

huge marginal totals, such as the income data in Appendix C, can be easily accommodated.

To calculate the intrinsic priors for the case of one margin fixed, as in (6), the random variables are generated in each row according to (7).

4 Examples and Evaluations

In this section we evaluate the performance of the intrinsic posterior probabilities both 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 2 that we have derived the posterior probability of the intrinsic procedure under two different sampling models, P_1 and P_2. Operationally, we have found that, for the most part, the posterior probabilities under these two models tend to be similar. In the following we have computed all of the posterior probabilities under sampling model P_1, in which we only assume 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 m has a natural range from 1 to n, as taking m larger than n will result in concentrating more mass close to the null. Moreover, as $m \to \infty$, the posterior probability of H_0 goes to 1. So what is of interest is the behavior of the posterior probability for the range of m from 1 to n, and if this probability remains flat we interpret this as evidence of robustness.

4.1 2×2 Tables

Efron (1985) analyzed data from a multicenter trial to see if a new surgical method for ulcers was superior to an older method (see also Casella 2001). In each of 39 hospitals a 2×2 table was reported, with the successes and failures for each of the hospitals.

Inspection of the tables[1] shows a good deal of variability, both in the number of patients and the success rates of the table. The first two tables in Table 2 (34

[1]The notation $\{a, b; c, d\}$ denotes a 2×2 table with first row $\{a, b\}$ and second row $\{c, d\}$, where here the rows correspond to the treatments. Thus in the table $\{8, 7; 2, 11\}$ one treatment had success rate 8/15 and the other had success rate 2/13.

and 1) suggest that there is association, a conclusion that is strongly supported by the intrinsic prior analysis. We see that throughout the entire range of m, the posterior probability of H_0 remains below .5.

Table 2: P-values and Posterior Probabilities for Selected Tables from Efron(1985) - Tables ordered by p-values, which are calculated using Fisher's exact test. The intrinsic posterior probability is calculated for both ends of the range; $m = 1$ and $m = n$. Note that the value for $m = 1$ is identical to that of the uniform prior (which corresponds to $m = 0$.

Table	Data	p-value	Uniform	Intrinsic $m = 1$	Intrinsic $m = n$
34	$\{20, 0; 18, 5\}$	.051	.215	.215	.215
1	$\{8, 7; 2, 11\}$	.054	.170	.170	.253
18	$\{30, 1; 23, 4\}$	.173	.551	.551	.406
38	$\{43, 4; 14, 5\}$	.106	.395	.395	.340
16	$\{7, 4; 4, 6\}$	.395	.451	.451	.497

The next table (18) suggests moderate deviation from the null, and we see that the range of intrinsic posterior probabilities crosses .5, indicating nonrobustness of the inference. That is, the data are not conclusive in either direction, and a firm conclusion cannot be drawn here.[2] Note that both the uniform posterior probability and the intrinsic with $m = 1$ 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 documentation of this bias is difficult.

The final two tables, (38 and 16) both represent robust cases. The intrinsic posterior probabilities are quite flat throughout the range, and never cross .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 m gives us a lot of information

[2]Recall that we interpret p-values and posterior probabilities on different scales. Typically, posterior probabilities of H_0 less than .5 are considered evidence against H_0, while p-values in the rage of .05 or less are considered evidence against H_0.

about the robustness of our conclusion.

4.2 The Tables of Good and Crook (1987)

Good and Crook (1987) analyze 21 contingency tables, many drawn from the literature and some that are artificial. We reanalyze those tables to both show how our procedure performs, and to contrast it with the robust procedure of Good and Crook. Table 3 in Appendix C summarizes the results of the 21 tables, showing the exact p-values[3], 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 m varies from $m = 1$ to $m = n$.

It is interesting to note that of the twenty-one tables, the Good-Crook robust posterior probability rejects the null hypothesis in 20 out of 21 cases, only supporting the null in Table 15, for which the intrinsic posterior answers are robust, they are in the interval $(.872, .964)$, and the p-value is 1.

On the other hand, 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), or large dimensional tables with small sample sizes, as the horsekick table, in which case robustness is not present. For these situations we either need to add subjective information on the concentration parameter m or to collect more data; the message is that the data themselves are not conclusive.

To illustrate the effect of varying the parameter m, we look at Figure 3, which examines the behavior of four interesting tables from Appendix C

In the four tables of Figure 3 we see the entire range of possibilities. Some tables are robust either for or against H_0, while some a nonrobust, having conclusions that are dependent on the tails of the prior. In such cases, (for example, Fienberg or Horsekick) it is important to reassess the prior, for it is clear that the data alone cannot yield a conclusive decision.

An interesting case is provided by the Mendel data (Table 7, Appendix C). The intrinsic Bayesian tests are robust, the range of intrinsic posterior probabilities are quite small, and the p-value strongly support the null hypothesis.

[3]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 $10,000$) from all tables with the same margins to use as a reference distribution. This is easily done with the `R` function `chisq.test`.

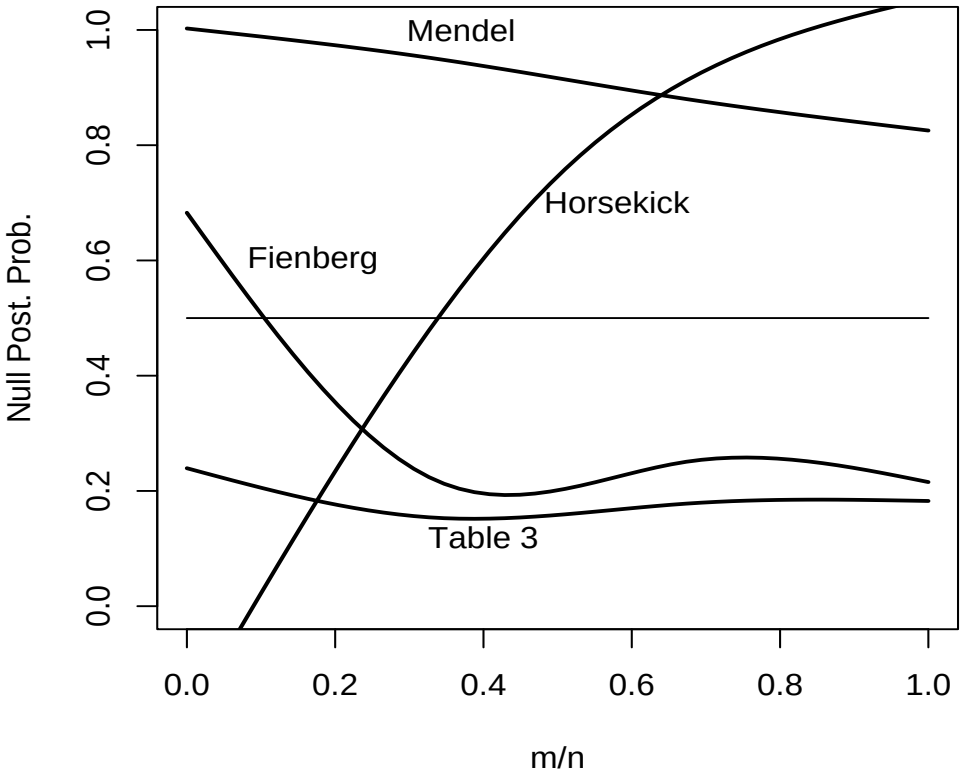

Figure 3: Ranges of posterior probabilities of four tables from Appendix C. The Mendel data is table 7 (robust, evidence in favor of null), the Fienberg data is table 6 (nonrobust, evidence against the null for moderate m), the horsekick data is table 20 (nonrobust, evidence in favor of the null for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

The historical consensus supports the null hypothesis (ignoring the debate about "cooked" data). However, the GC robust analysis strongly rejects the null hypothesis, which is in opposition to what is commonly concluded about these data sets. GC defend these conclusions, citing problems with computation, and "flatness" of the margins.

In the artificial Tables 11-15 (Appendix C), the results of the GC robust analysis are also not in agreement with our results. These tables all provide strong evidence for the null (in each table the rows are exactly the same), and the ranges of the intrinsic posterior probabilities suggest robustness in accepting the null for all of these tables. The GC 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. GC explain: "*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, but note that the ranges of intrinsic posterior probabilities are calculated conditional only on the table total. We must conclude that the price GC pays to achieve robustness is an increased tendency to reject the null hypothesis.

As we mentioned above, another instance where the intrinsic posterior analysis leads to different conclusions is in Table 6 (Appendix C, political preferences, Bishop, Fienberg and Holland 1978, page 387), where the counts are very unbalanced. Here both the p-value and the GC robust analysis rejects the null hypothesis, while the intrinsic posterior analysis will accept the null when m is small ($1 \leq m \leq 0.1\ n$), and will reject when m is large ($.1\ n \leq m \leq n$). Bishop, Fienberg and Holland (1978) give three analyses of this table, and all suggest that there is some deviation from the null.

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.

We observe that, from the frequentist point of view, 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 the unconditional set can also 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, and the Monte Carlo calculations are tailored to ignore these tables of low probability.

A Bayesian analysis of a contingency table starts with a likelihood and a prior, where the likelihood reflects the sampling model actually used, the prior typically reflects crude prior beliefs, and evaluates the performance of the result-

ing procedure. It appears to be widely accepted that prior beliefs should be such that the parameters should be *a priori* dependent, as is emphasized strongly by Howard (1998). This property is typically not satisfied by the usual default prior for estimation (for example, 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 is also accomplished by the intrinsic priors and the Good/Crook priors, as can be seen in Figure 2, although the Good/Crook priors also put high mass to extreme tables.

The priors of Good and Crook (1987) are mixtures of Dirichlets. 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 achieve robustness. The intrinsic priors are also mixtures, see (9), and the degree of concentration around the null is accommodated by a discrete parameter m. To complete this analogy we could also mix the parameter m with respect to a hyperprior with a heavy tail. However, the price GC has paid for the robustification of the procedure is to have a procedure which can result in unreasonable conclusions; see Tables 11,12,14,and 15 in Appendix C.

We have learned an important lesson from this work, that in many cases our conclusion is more sensitive to the prior than we suspect, even when we consider our prior to be "noninformative" or "robust". The range of answers that we have seen from the intrinsic analysis is startling - for a large number of tables the conclusion can turn from "accept" to "reject". Unfortunately, we are not able to classify the types of tables that lead to this nonrobust behavior, although we suspect that imbalance and sparseness in the cell sizes will contribute to the sensitivity. However, we do have a diagnostic that can alert us to situations when consideration of the prior information results in an important factor in the inference.

We also note that, by construction, the range of the intrinsic priors, from $m = 1$ to $m = n$ is a very reasonable range. In terms of the tails of the prior, we range from extremely flat tails ($m = 1$) to tails that are equal to those of the data ($m = n$). This is a natural bound, as the situation where one would give more weight to the prior than the data is an extremely rare situation. Thus, we

have a natural range 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 informative enough and, when they are weak, which is often reflected in imbalance or sparseness of the table, we obtain the warning that the resulting tests are not robust, thus requiring more prior information or more data.

A General $a \times b$ Tables

In this appendix we generalize the calculations of Section 3 to the case of an $a \times b$ contingency table. The calculations are similar to the previous ones, and hence are only summarized.

A.1 Margins Unrestricted

We suppose that n individuals have been classified in one of the cells with unknown probability $\theta = \{\theta_{ij}\}$, $i = 1, ..., a, j = 1, ..., b$, and $\sum_{ij} \theta_{ij} = 1$. Under this sampling model the distribution of the table $\mathbf{y} = \{y_{ij}\}$ is the multinomial distribution $M(\mathbf{y}|n, \theta)$. A default prior for $\theta = \theta_{ij}$ can be either an $(a \times b - 1)-$dimensional Dirichlet with all parameters equal to $1/2$ (the Jeffreys prior) or a Dirichlet with all parameters equal to 1.

Under the null hypothesis $H_0 : \theta_{ij} = p_i q_j$, the density of the table is

$$f_0(\mathbf{y}|n, \mathbf{p}, \mathbf{q}) = \binom{n}{\mathbf{y}} \prod_{i,j} (p_i q_j)^{y_{ij}},$$

and the intrinsic prior is given by

$$\begin{aligned} \pi^I(\theta|m) &= \frac{\Gamma(m+ab)\Gamma(a)\Gamma(b)}{\Gamma(m+a)\Gamma(m+b)} && (9) \\ &\times \sum_{\mathbf{x}:\sum x_{ij}=m} \binom{m}{\mathbf{x}} \frac{(\prod r_i(\mathbf{x})!)\,(\prod c_j(\mathbf{x})!)}{\prod x_{ij}!} \prod \theta_{ij}^{x_{ij}}, && (10) \end{aligned}$$

where we recall that we denote the row totals and columns totals by $r_i(\cdot)$ and $c_j(\cdot)$, respectively. For a data set $\mathbf{y} = (y_{ij})$ the Bayes factor for the above intrinsic prior is

$$\begin{aligned} B_{10}(\mathbf{y}) &= \frac{\Gamma(m+ab)}{\Gamma(m+n+ab)} \left[\frac{\Gamma(n+a)\Gamma(n+b)}{\Gamma(m+a)\Gamma(m+b)}\right] && (11) \\ &\times \sum_{\mathbf{x}:\sum x_{ij}=m} \binom{m}{\mathbf{x}} \frac{(\prod r_i(\mathbf{x})!)\,(\prod c_j(\mathbf{x})!)}{(\prod r_i(\mathbf{y})!)\,(\prod c_j(\mathbf{y})!)} \frac{\prod (x_{ij}+y_{ij})!}{\prod x_{ij}!}. && (12) \end{aligned}$$

A.2 One Margin Fixed

In the case of one marginal fixed, say the row totals are fixed, the sampling distribution is that of a independent multinomials. Define

$$\mathbf{y}_i = (y_{i1}, ..., y_{ib}), \quad \mathbf{p}_i = (p_{1j}, ..., p_{ib}), \quad \sum_{j=1}^{b} y_{ij} = n_i.$$

Then the variables $\mathbf{y}_i, i = 1, \ldots, a$ are independent with multinomial distributions $M(\mathbf{y}_i|n_i, \mathbf{p}_i)$. To test

$$H_0 : \mathbf{p}_1 = ... = \mathbf{p}_a, \quad vs. \quad H_1 : not\ H_0,$$

or to compare the models,

$$M_0 : \left\{ \prod_{i=1}^{a} M(\mathbf{y}_i|n_i, \mathbf{p}_0),\ \pi^{\mathcal{D}}(\mathbf{p}_0) = \mathcal{D}_{b-1}(\mathbf{p}_0|1, ...1) \right\},$$

and

$$M_1 : \left\{ \prod_{i=1}^{a} M(\mathbf{y}_i|n_i, \mathbf{p}_i),\ \pi^{\mathcal{D}}(\mathbf{p}) = \prod_{j=1}^{b} \mathcal{D}_{b-1}(\mathbf{p}_j|1, ...1) \right\}.$$

The default marginal distributions under these models are

$$m_0(\mathbf{y}) = \frac{\Gamma(b)}{\Gamma(n+b)} \prod_{i=1}^{a} \binom{n_i}{\mathbf{y}_i} \prod_{j=1}^{b} c_j(\mathbf{y})! \text{ and } m_1(\mathbf{y}) = \Gamma(b)^a \prod_{i=1}^{a} \binom{n_i}{\mathbf{y}_i} \prod_{j=1}^{b} \frac{y_{ij}!}{\Gamma(n_i+b)},$$

where $n = \sum_i n_i$. The intrinsic prior for $\mathbf{p} = (\mathbf{p}_1, ..., \mathbf{p}_a)$ is

$$\pi^I(\mathbf{p}) = \Gamma(b) \sum_{\substack{(\mathbf{x}_1,...,\mathbf{x}_a): \\ \sum_j x_{ij} = m_i}} \frac{\prod_{j=1}^{b} c_j(\mathbf{x})!}{\Gamma(m+b)} \frac{\prod_{i=1}^{a} \Gamma(m_i+b)}{\prod_{ij} x_{ij}!} \prod_{i=1}^{a} \binom{m_i}{\mathbf{x}_i} \prod_{j=1}^{b} p_{ij}^{x_{ij}},$$

leading to the intrinsic marginal

$$\begin{aligned} m^I(\mathbf{y}) &= \Gamma(b) \prod_{i=1}^{a} \binom{n_i}{\mathbf{y}_i} \frac{\prod_{i=1}^{a} \Gamma(m_i+b)}{\Gamma(m+b)} \\ &\quad \times \sum_{\substack{(\mathbf{x}_1, \ldots, \mathbf{x}_a) \\ \sum_j x_{ij} = m_i}} \frac{\prod_{j=1}^{b} c_j(\mathbf{x})!}{\prod_{ij} x_{ij}!} \prod_{i=1}^{a} \binom{m_i}{\mathbf{x}_i} \frac{\prod_{j=1}^{b} (x_{ij}+y_{ij})!}{\Gamma(m_i+n_i+b)}. \end{aligned}$$

For a sample $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_b)$ the Bayes factor of model M_1 against M_0 for the intrinsic priors $(\pi^{\mathcal{D}}(p_0), \pi^I(p))$ is given by $B_{10}(\mathbf{y}) = m^I(\mathbf{y})/m_0(\mathbf{y})$ with posterior probability $1/(1 + B_{10}(\mathbf{y}))$. The Bayes factor changes very little if the Jeffreys prior, a Dirichlet distribution with parameter $(1/2, ..., 1/2)$, is used instead of the Dirichlet prior with parameters $(1, 1, ..., 1)$.

B Consistency

Here we give a detailed proof of Theorem 1, consistency of the intrinsic posterior for the binomial and multinomial cases, and we indicate how the proof extends to more general cases. We start with a lemma that will be useful in establishing the results.

Lemma 1 *Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be positive constants satisfying $\sum_i a_i = \sum_i b_i = 1$. Then*

$$\prod_i \left(\frac{a_i}{b_i}\right)^{a_i} \geq 1, \tag{13}$$

with strict inequality unless $a_i = b_i$ for all i.

Proof: The log of (13) is $\sum_i a_i \log(a_i/b_i)$. Using the facts that $\sum_i a_i = 1$ and $-\log$ is convex, Jensen's inequality yields

$$\sum_i a_i \log(a_i/b_i) = -\sum_i a_i \log(b_i/a_i) \geq -\log\left(\sum_i a_i \frac{b_i}{a_i}\right) = -\log\left(\sum_i b_i\right) = 0,$$

establishing the inequality. The strictness also follows from Jensen's inequality.

B.1 Proof of Theorem 1

For the case of Theorem 1,

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

the default marginal distributions are

$$m_0(y) = \binom{n}{y}\theta_0^y(1-\theta_0)^{n-y} \text{ and } m_1(y) = \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta = \frac{1}{n+1}, \tag{14}$$

leading to the intrinsic prior

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

and intrinsic marginal

$$m^I(y) = \int \binom{n}{y}\theta^y(1-\theta)^{n-y}\pi^I(\theta)\,d\theta. \tag{15}$$

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

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

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

$$\begin{aligned}B_{10} &\geq K\frac{\int \binom{n}{y}\theta^y(1-\theta)^{n-y}\,d\theta}{\binom{n}{y}\theta_0^y(1-\theta_0)^{n-y}}\\ &= \frac{K}{n+1}\frac{\binom{n}{y}^{-1}}{\theta_0^y(1-\theta_0)^{n-y}}.\end{aligned} \tag{16}$$

Stirling's approximation yields

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

since $y \approx n\theta$ as $n \to \infty$. Substituting into (16) and rearranging terms yields

$$B_{10} \geq \frac{Kn^{1/2}}{n+1}\left[\frac{\theta^\theta(1-\theta)^{1-\theta}}{\theta_0^\theta(1-\theta_0)^{1-\theta}}\right]^n \tag{17}$$

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 thus the posterior probability of H_0 goes to 0.

To establish consistency if θ_0 is the true parameter, we can bound $\pi^I(\theta|m)$ from above, and arrive at (16) as an upper bound (with a different value of K that will depend on m and θ_0 but not n or y). Under H_0, $y \approx n\theta_0$, so we obtain the right side of (17) 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, the Bayes factor goes to 0 and the posterior probability of H_0 goes to 1, and the consistency is established.

B.2 Consistency for the Multinomial Case

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

$$\begin{aligned} M_0 &: M(\mathbf{x}|m, p_i q_j), \quad p_i q_j \text{ fixed} \\ M_1 &: M(\mathbf{x}|m, \theta ij), \quad \pi(\theta) = \Gamma(ab), \quad m_1(\mathbf{x}) = \frac{\Gamma(ab)}{\Gamma(m+ab)} \binom{m}{\mathbf{x}} \prod_{ij} x_{ij}!, \end{aligned}$$

where θ is the vector of θij. The intrinsic prior is

$$\pi^I(\theta) = \Gamma(m+ab) \sum_{x_{ij}} \frac{\prod_{ij} (p_i q_j)^{x_{ij}}}{\prod_{ij} x_{ij}!} M(\mathbf{x}|m, \theta ij).$$

Using similar arguments to those in Section B.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

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

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, the posterior probability of H_0 goes to 1. If H_1 is true, Lemma 1 shows that the expression in square brackets is equal to $1 + \varepsilon$, for some $\varepsilon > 0$, and $B_{10} \to \infty$ as $n \to \infty$. So if H_1 is true, the posterior probability of H_0 goes to 0.

B.3 Extensions

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

$$M_0 : f(\mathbf{y}|\theta_0) \text{ vs. } M_1 : \{f(\ \mathbf{y}|\theta), \pi^I(\theta|\theta_0, m)\}, \tag{18}$$

where $\pi^I(\theta|\theta_0, m)$ denotes the intrinsic prior for θ conditional on the null θ_0, on the training sample of size m, 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 a subspace $H_0 : \theta_0 \in \Theta_0 \subset \Theta$.

The nested Bayesian models are now

$$M_0 : \{f(\mathbf{y}|\theta_0), \pi_0(\theta_0)\} \text{ vs. } M_1 : \{f(\mathbf{y}|\theta), \pi^I(\theta|m)\}, \tag{19}$$

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

$$\pi^I(\theta|m) = \int_{\Theta_0} \pi^I(\theta|\theta_0, m)\pi_0(\theta_0)d\theta_0 = \pi_1(\theta)E_{\mathbf{x}|\theta}\frac{m_0(\mathbf{x})}{m_1(\mathbf{x})},$$

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 that for any $\theta_0 \in \Theta_0$,*

(*i*) *the Bayes factor*

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

is consistent for testing the sharp null hypothesis (18),

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

(*iii*) *the set Θ_0 is a compact set, and*

(iv)

$$k'_m = \inf_{\mathbf{x}} \frac{m_0(\mathbf{x})}{f(\mathbf{x}|\theta_0)} > 0,\ k_m = \sup_{\mathbf{x}} \frac{m_0(\mathbf{x})}{f(\mathbf{x}|\theta_0)} < \infty.$$

Then, the Bayes factor for testing (19)

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

is consistent.

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

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

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

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

Substituting in B_{10} we have

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

where the last expression tends to zero because $B_{10}(\mathbf{y};\theta_0^*)$ is consistent. This proves consistency under the null.

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

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

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

$$\pi^I(\theta|m) = \pi_1(\theta)E_{\mathbf{x}|\theta}\frac{f(\mathbf{x}|\hat{\theta}_0)}{m_1(\mathbf{x})}\frac{m_0(\mathbf{x})}{f(\mathbf{x}|\hat{\theta}_0)} > k'_m\pi_1(\theta)E_{\mathbf{x}|\theta}\frac{f(\mathbf{x}|\hat{\theta}_0)}{m_1(\mathbf{x})} = k'_m\pi^I(\theta|\hat{\theta}_0, m).$$

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

$$B_{01}(\mathbf{y}) < \frac{1}{k'_m}\frac{f(\mathbf{y}|\hat{\theta}_0)}{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|\hat{\theta}_0, m)d\theta} \approx \frac{1}{k'_m}\frac{f(\mathbf{y}|\tilde{\theta}_0)}{\int_\Theta f(\mathbf{y}|\theta)\pi^I(\theta|\tilde{\theta}_0, m)d\theta} \to 0,$$

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

We note that both binomial and multinomial distributions satisfy the conditions in Theorem 2.

C The Good/Crook Data and Statistics

Table 3: The 21 Tables of Good and Crook (1987)

	Table	p-value	Posterior Probability of H_0*		
			(Good/Crook)	(Intrinsic) $m = 1$	(Intrinsic) $m = n$
1.	$\{10, 3; 2, 15\}$	.001	.015	.003	.016
2.	$\{29, 33; 131, 78\}$	.028	.217	.312	.257
3.	$\{200, 8; 182, 20\}$	.018	.156	.361	.179
4.	$\{105, 5; 88, 11\}$	.116	.294	.609	.371
5.	$\{409, 3; 174, 8\}$	.005	.083	.284	.078
6.	$\{225, 53, 206; 3, 1, 12\}$	.041	.125	.933	.241
7.	$\{38, 60, 28; 65, 138, 68;$ $35, 67, 30\}$	.763	.123	.997	.823
8.	$\{61, 12, 60; 17, 6, 1; 39, 22, 7\}$	.000	.000	.000	.000
9.	$\{17, 4, 8; 5, 12, 0; 10, 3, 13\}$	.000	.000	.000	.000
10.	$\{58, 52, 1; 26, 58, 3; 8, 12, 9\}$	.000	.303	.000	.000
11.	$\{2, 2, 2; 2, 2, 2; 2, 2, 2\}$	1.00	.450	.648	.701
12.	$\{6, 6, 6; 6, 6, 6; 6, 6, 6\}$	1.00	.327	.891	.839
13.	$\{1, 2, 3; 1, 2, 3; 1, 2, 3\}$	1.00	.500	.696	.740
14.	$\{1, 5, 20; 1, 5, 20; 1, 5, 20\}$	1.00	.294	.988	.855
15.	$\{5, 0, 0; 5, 0, 0; 5, 0, 0\}$	1.00	.520	.964	.872
16.	$\{6, 0, 0; 0, 6, 0; 0, 0, 6\}$	.000	.000	.000	.000
17.	$\{5, 1, 0; 4, 0, 2; 2, 4, 0\}$	.033	.200	.080	.121
18.	$\{68, 119, 26, 7; 20, 84, 17, 94;$ $15, 54, 14, 10; 5, 29, 14, 16\}$	.000	.000	.000	.000
18*A*.	$\{4, 1, 1, 0; 2, 3, 0, 3;$ $1, 2, 2, 0; 0, 0, 0, 1\}$	.113	.277	.101	.064
19.	Income and No.of Children	.000	.000	.000	.000
20.	Horsekick Data	.647	.062	.021	.999

* The uniform prior corresponds to the intrinsic prior with $m = 0$. In all calculations this posterior probability was identical to that of the intrinsic prior with $m = 1$, so the column of uniform posterior probabilities is not shown.

Table 19: Income and Number of Children, Diaconis and Efron (1985)

2161	3577	2184	1636
2755	5081	2222	1052
936	1753	640	306
225	419	96	38
39	98	31	14

Table 20: Horse Kick Data

1875	0	0	0	0	0	0	0	1	1	0	0	0	1	0	3
1876	2	0	0	0	1	0	0	0	0	0	0	0	1	1	5
1877	2	0	0	0	0	0	1	1	0	0	1	0	2	0	7
1878	1	2	2	1	1	0	0	0	0	0	1	0	1	0	9
1879	0	0	0	1	1	2	2	0	1	0	0	2	1	0	10
1880	0	3	2	1	1	1	0	0	0	2	1	4	3	0	18
1881	1	0	0	2	1	0	0	1	0	1	0	0	0	0	6
1882	1	2	0	0	0	0	1	0	1	1	2	1	4	1	14
1883	0	0	1	2	0	1	2	1	0	1	0	3	0	0	11
1884	3	0	1	0	0	0	0	1	0	0	2	0	1	1	9
1885	0	0	0	0	0	0	1	0	0	2	0	1	0	1	5
1886	2	1	0	0	1	1	1	0	0	1	0	1	3	0	11
1887	1	1	2	1	0	0	3	2	1	1	0	1	2	0	15
1888	0	1	1	0	0	1	1	0	0	0	0	1	1	0	6
1889	0	0	1	1	0	1	1	0	0	1	2	2	0	2	11
1890	1	2	0	2	0	1	1	2	0	2	1	1	2	2	17
1891	0	0	0	1	1	1	0	1	1	0	3	3	1	0	12
1892	1	3	2	0	1	1	3	0	1	1	0	1	1	0	15
1893	0	1	0	0	0	1	0	2	0	0	1	3	0	0	8
1894	1	0	0	0	0	0	0	0	1	0	1	1	0	0	4
	16	16	12	12	8	11	17	12	7	13	15	25	24	8	196

Data available at http://www.galton.uib.no/FordKurs/Datasets.html

Acknowledgment

Casella was supported by National Science Foundation Grants DMS-04-05543, DMS-0631632 and SES-0631588. Email: casella@stat.ufl.edu. Moreno's work was supported by the by Ministerio de Ciencia y Tecnología, Grant SEJ-02447, SEJ-65200 and Junta de Andaluciá Grant SEJ-02814. Email: emoreno@ugr.es. This work was started while Casella was on sabbatical at the University of Granada.

References

Agresti, A. (1992) A survey of exact inference for contingency tables (with discussion). *Statistical Science* **7** 131-177.

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

Albert, J. H. and Gupta, A. K. (1982). Mixtures of Dirichlet distributions and estimation in contingency tables. *Ann. Statist* **10** 1261-1268.

Albert, J. H. and Gupta, A. K. (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. *J. Roy. Statist. Soc. Ser. B* **31** 261-269.

Altham, P. M. E. (1971). The analysis of matched proportions. *Biometrika* **58** 561-576.

Bayes, T. (1783). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.* **53** 370-418.

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

Berger, J. O. (1994). An overview of robust Bayesian analysis (with discussion). *Test* **3** 5-124.

Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association* **95** 1269-1276.

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 Pericchi, L.R. (2006). Training samples in objective Bayesian model selection. *Ann. Statist.* **32**, 841-869.

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.

Berger, J. O. , De Oliveira, V. and Sanso, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association* **96** 1361-1374.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. 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.

Clyde, M. (2001). Discussion of Chipman, George and McCulloch. In *Model Selection*, (Ed. P. Lahiri), Hayward, CA: Institute of Mathematical Statistics, **38**, 117-124.

Clyde, M. and George, E. I. (2000). Flexible Empirical Bayes Estimation for Wavelets. *Journal of the Royal Statistical Society, Series B* **62** 681-698.

Diaconis, P. and Efron, B. (1985). Testing for independence in a two-way table: New interpretations of the chi-squared statistic (with discussion). *Ann. Statist* **13** 845-913.

Efron, B. (1996). Empirical Bayes methods for combining likelihood (with discussion). *Journal of the American Statistical Association* **91** 538-565.

Girón, J., Martinez, L., Moreno, E. and Torres, F. (2006). Objective testing procedures in linear models: Calibration of the p-values. *Scand. J. Statist.* **33** 765-787.

Good, I. J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist* **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. *Ann. Statist* **15** 670-693.

Greenwood, M. and Yule, G. U. (1915). The statistics of anti-typhoid and anti-cholera inoculations, and the interpretations of such statistics in general. *Proc. Roy. Soc. Medicine (Epidemiology)* **8** 113-190. (Reprinted in *Statistical Papers of George Udny Yule* (1971) Griffin, London.

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, Third Edition.* 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. *J. Statist. Plann. Inf.* **102** 425-439.

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

Kim, S. and Sun, D. (2000). Intrinsic priors for Model Selection Using an Encompassing Model. *Life Time Data Analysis* **6** 251-269.

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

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc. 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*, (Ed. D. Yadolah), Hayward, CA: Institute of Mathematical Statistics, vol. **29**, 257-270.

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 Berger/Sellke and Casella/Berger. *Journal of the American Statistical Association*, **82**, 106-111.

Nazarret, W. (1987). Bayesian log-linear estimates for three-way contingency tables. *Biometrika* **74** 401-410.

Novik and Hall (1965). A Bayesian Inference Procedure. *Journal of the American Statistical Association* **60** 1104-1117.

Sweeting, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* **88** 657-675.

Wasserman, L (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society, Series B* **62** 159-180.

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

George Casella, Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611.

E-mail: (casella@stat.ufl.edu)

Elías Moreno, Professor, Department of Statistics, University of Granada, 18071, Granada, Spain.

E-mail: (emoreno@ugr.es)