The Annals of Statistics
1987, Vol. 15, No. 2, 670–693

THE ROBUSTNESS AND SENSITIVITY OF THE MIXED-DIRICHLET BAYESIAN TEST FOR "INDEPENDENCE" IN CONTINGENCY TABLES[1]

BY I. J. GOOD AND J. F. CROOK

Virginia Polytechnic Institute and State University and Winthrop College

A mixed-Dirichlet prior was previously used to model the hypotheses of "independence" and "dependence" in contingency tables, thus leading to a Bayesian test for independence. Each Dirichlet has a main hyperparameter κ and the mixing is attained by assuming a hyperprior for κ. This hyperparameter can be regarded as a flattening or shrinking constant. We here review the method, generalize it and check the robustness and sensitivity with respect to variations in the hyperpriors and in their hyperhyperparameters. The hyperpriors examined included generalized log-Students with various numbers of degrees of freedom ν. When ν is as large as 15 this hyperprior approximates a log-normal distribution and when $\nu = 1$ it is a log-Cauchy. Our experiments caused us to recommend the log-Cauchy hyperprior (or of course any distribution that closely approximates it). The user needs to judge values for the upper and lower quartiles, or any two quantiles, of κ, but we find that the outcome is robust with respect to fairly wide variations in these judgments.

1. The hierarchical Bayesian approach to testing "independence" in contingency tables. Consider a contingency table with r rows and s columns, with cell or category entries n_{ij}, $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, s$, row totals $n_{i.} = \sum_j n_{ij}$, column totals $n_{.j} = \sum_i n_{ij}$ and sample size $N = \sum_{ij} n_{ij}$. Let the corresponding unknown physical (or material) probabilities be denoted by p_{ij}, $p_{i.}$, $p_{.j}$ and, of course, 1.

Three familiar procedures for sampling a contingency table are as follows:

Procedure 1, *or multinomial sampling*, where the sampling is from the population as a whole;

Procedure 2, *or product-multinomial sampling*, *or stratified sampling*, where the row (or column) totals are fixed by the experimenter; and

Procedure 3, where both the row and column totals are fixed.

For some further general discussion of these procedures, with citations, see Good (1976, page 1161).

A hierarchical Bayesian approach for all three procedures, based on a mixed-Dirichlet prior, was used or discussed by Good (1965, 1976, 1980a, 1980b and

Received September 1985; revised October 1986.

[1]This work was supported by a grant from the National Institutes of Health.

AMS 1980 *subject classifications*. Primary 62H17; secondary 62H15, 62F15.

Key words and phrases. Bayes factors against independence, Bayesian robustness, Bayesians (averaging over), Dirichlet-multinomial distribution, flattening constants, hierarchical Bayes, hyperhyperparameters, log-Cauchy distribution, log-normal distribution, log-Student distribution, mixtures of conjugate priors, multinomial significance tests, multivariate Bayesian methods, shrinking constants.

1983) and Crook and Good (1980, 1982) for testing the null hypothesis H that $p_{ij} = p_{i.}p_{.j}$, that is, "independence" of row and column categorizations. [For unmixed Dirichlet priors see Gûnel and Dickey (1974) and references therein.] We here adopt a more general procedure and provide some new explanations.

Our main aim is to investigate "Bayesian robustness" of a test of the null hypothesis H when the sampling is by Procedure 1 or 2.

As suggested by a referee, we at once present the results of our calculations regarding 21 contingency tables. (See Tables 1–20.) These results will not be immediately intelligible even to those who know our earlier work because our present model is somewhat different from what we have used in the past. The full meanings of the results will emerge later. For the present, observe that the numbers in the body of each table are Bayes factors against H, for a variety of Bayesian models, and at first the reader should look at the average factor in the row labelled $\nu = 1$. [The Bayes factor against H provided by evidence E is defined as $O(H)/O(H|E)$, a ratio of *odds*, and is equal to $P(E|H')/P(E|H)$ where H' denotes the negation of H. We have previously written $\overline{H}$ for H'.] For example, in Table 1, this average factor is 64 and it can be compared with the P-value of 1/2150 obtained by Fisher's "exact test" for two-by-two contingency tables. It should be held in mind, for each of the 21 cases, that the Pearson chi-squared, χ^2, and also Fisher's exact test when applicable, are conditional on all the marginal totals $(n_{i.})$ and $(n_{.j})$ and so are most appropriate under sampling Procedure 3. For two-by-two tables, χ'^2 denotes χ^2 with Yates' continuity correction, and $P(\chi^2)$ and $P(\chi'^2)$ denote the right-hand P-values for χ^2 and χ'^2.

Each contingency table, except number 20, is denoted by a self-explanatory abbreviated notation; for example, [10, 3; 2, 15] denotes the criminal two-by-two table whose rows are [10, 3] and [2, 15].

When $P(\chi^2)$ is not close to 1, the Bayes factors (in the row labelled $\nu = 1$) are always smaller than $1/P(\chi^2)$, and often much smaller. This is a typical and important relationship or "discrepancy" between Bayes factors and P-values. [See, for example, Good and Crook (1974), formula (3.1), and Berger and Sellke (1986).] We were somewhat surprised by the Bayes factors for Table 20, the horse-kick data. We shall give reasons near the end of the paper why our assumptions are probably inapplicable in this case.

We now begin to list our assumptions, with special emphasis on where they differ from those in our earlier work, and we shall return to the discussion of the numerical results in Section 10.

ASSUMPTION 1 ("Ancillarity" of the row totals). The row totals alone (or the column totals alone) convey no evidence for or against H under Sampling Procedure 1.

This assumption goes without saying under Sampling Procedure 2 because in this case the row totals are fixed before the sample is taken. Another way of stating Assumption 1 is that the Bayes factor against H provided by (n_{ij}) must

be unaffected by knowing $(n_{i.})$ in advance and hence that the Bayes factor will be the same whether the sampling is by Procedure 1 or Procedure 2. Therefore, in the following theory, we shall think mainly in terms of Procedure 1 for the sake of definiteness and because it is less difficult to do so.

Since the Bayes factor against H provided by the row totals $(n_{i.})$ is equal to $P[(n_{i.})|H']/P[(n_{i.})|H]$, we have, under Sampling Procedure 1,

$$P\big[(n_{i.})|H'\big] = P\big[(n_{i.})|H\big]. \tag{1.1}$$

Under Sampling Procedure 2, this equation is a truism. Similarly (under Sampling Procedure 1), we have

$$P\big[(n_{.j})|H'\big] = P\big[(n_{.j})|H\big], \tag{1.1A}$$

but we do *not* assume that $P[(n_{i.}),(n_{.j})|H'] = P[(n_{i.}),(n_{.j})|H]$.

For testing the null hypothesis H by a "sharp" Bayesian method (that is, with sharp or precise priors) we need to assume some definite prior for the p_{ij} given the nonnull hypothesis H', and for the $p_{i.}$ and $p_{.j}$ given H.

ASSUMPTION 2 ((n_{ij}) as a multinomial). We regard H' as the hypothesis that the frequencies (n_{ij}) have a multinomial distribution with physical cell probabilities (p_{ij}), and where (p_{ij}) has a prior distribution. Of course, under H', we are not assuming that $p_{ij} = p_{i.}p_{.j}$.

Given H', the prior epistemic (= personal or logical) expectation of p_{ij} is called q_{ij}. (We shall usually write "subjective" although "epistemic" would be more precise.) Of course, given H', the prior subjective expectations of $p_{i.}$ and $p_{.j}$ are $q_{i.}$ and $q_{.j}$ where $q_{i.} = \Sigma_j q_{ij}$ and $q_{.j} = \Sigma_i q_{ij}$.

ASSUMPTION 3 (The marginal totals as multinomials). Given H, both $(n_{i.})$ and $(n_{.j})$ have multinomial distributions with, of course, physical cell probabilities $(p_{i.})$ and $(p_{.j})$, respectively, and where $(p_{i.})$ and $(p_{.j})$ have prior distributions. The corresponding property, given H', follows from Assumption 2.

ASSUMPTION 4. Under H, where by definition $p_{ij} = p_{i.}p_{.j}$, the prior distributions of $(p_{i.})$ and $(p_{.j})$ are independent. Therefore,

$$E(p_{i.}p_{.j}|H) = E(p_{i.}|H)E(p_{.j}|H),$$

where the expectations are subjective, and also the *prior* distributions of $(n_{i.})$ and $(n_{.j})$ are independent given H. (We do *not* assume this independence given H'.)

From Assumptions 3 and 4, combined with (1.1) and (1.1A), it follows that the prior subjective expectations of $p_{i.}$ and $p_{.j}$, given H, being N^{-1} times the prior expectations of $(n_{i.})$ and $(n_{.j})$, must be the same as when H' is given, that is, they must be equal to $q_{i.}$ and $q_{.j}$. Therefore, given H, the prior (subjective)) expectation of p_{ij} which, given H, is the prior expectation of $p_{i.}p_{.j}$ (by the definition of H), must be $q_{i.}q_{.j}$.

ASSUMPTION 5 (Making H' "close" to H). The prior subjective expectation of p_{ij}, given H', is also equal to $q_{i.}q_{.j}$, that is, $q_{ij} = q_{i.}q_{.j}$.

In other words, we are assuming that the prior for (p_{ij}), given H', is in a sense as close as possible to that given H, subject to certain other natural assumptions that prevent the two priors from coinciding. This is a standard policy when nonnull hypotheses are specified in any statistical model although it is often done "unconsciously." In short, we choose our model so that it is difficult to get strong support for the null hypothesis.

Although we have just assumed that $q_{ij} = q_{i.}q_{.j}$ we shall usually write q_{ij} for the sake both of making the formulas a little shorter and for the sake of greater generality.

We think of $(q_{i.})$ and $(q_{.j})$ as the subjective expectations of $(p_{i.})$ and $(p_{.j})$ before the interior (n_{ij}) of the contingency table is known, but after the marginal totals $(n_{i.})$ and $(n_{.j})$ are known. This makes sense under Sampling Procedure 1.

ASSUMPTION 6 (Determination of $q_{i.}$ and $q_{.j}$). We determine $(q_{i.})$ and $(q_{.j})$ in the manner of Good (1967) (see Section 2 below). That is, we take, for example,

$$q_{i.} = \frac{n_{i.} + k_0}{N + rk_0}, \tag{1.2}$$

where k_0 is a multinomial "flattening constant" that depends on $(n_{i.})$ according to formula (24) of that paper, namely,

$$k_0 = \int_0^\infty \psi(k)F(k)\frac{k\,dk}{N + rk} \Big/ \int_0^\infty \psi(k)F(k)\frac{dk}{N + rk},$$

in which

$$F(k) = r^N\Gamma(rk)\{\Pi\Gamma(n_{i.} + k)\}/\{[\Gamma(k)]^r\Gamma(N + rk)\},$$

while ψ is a hyperprior, which we take as a log-Cauchy with lower and upper quartiles as $10/r$ and $50/r$. (A log-Cauchy distribution is determined by its quartiles: see Section 8.) There is one flattening constant k_0 for the row totals and another one for the column totals. These estimates of the q's are robust, especially when the frequencies $n_{i.}$ and $n_{.j}$ are not as small as 2.

Our fixing of $(q_{i.})$ and $(q_{.j})$ in this manner combines Bayesian and empirical Bayesian methods in that $(q_{i.})$, for example, is the posterior expectation of $(p_{i.})$, given $(n_{i.})$. We then use $(q_{i.})$ and $(q_{.j})$, in the empirical Bayes spirit, as hyperparameters in *priors* for $(p_{i.})$, $(p_{.j})$ and (p_{ij}). We could go fully Bayesian by ascribing hyperpriors to $(q_{i.})$ and $(q_{.j})$, but we think this would be too complicated and would have little effect on the results. In our earlier work we took $q_{i.} = 1/r$, $q_{.j} = 1/s$ and $q_{ij} = 1/(rs)$, but we now believe it is better to determine $(q_{i.})$, $(q_{.j})$ and (q_{ij}) in the manner just described. Note that we treat rows and columns on a par as is appropriate under Sampling Procedure 1.

ASSUMPTION 7 (No "structure"). Apart from the specification of the prior expectation of (p_{ij}), H' asserts no further structure related to the specific pattern of rows and columns such as, for square contingency tables, symmetry about the leading diagonal. In other words, we treat (n_{ij}), given H', as a multinomial sample having rs categories where each category has its own initial expectation and where the classification into rows and columns is otherwise irrelevant.

Assumption 7 affects, but does not determine, how we assign a prior distribution to (p_{ij}), given H'; so far we have specified only the prior *expectation* of (p_{ij}).

In any real-life situation we recommend that the statistician and his client should study their contingency table, or the scientific knowledge underlying it, to see if it seems to have any clear-cut pattern or patterns. Our advice is only a special case of the (somewhat controversial) need to examine almost any data in an exploratory manner to see if it has any interesting features that we had not taken into account in previous modelling.

In spite of these cautions we shall adopt our "no structure" assumption, given H', and even make the following slightly stronger

ASSUMPTION 7′ (Stronger "no structure"). Given the categorization into r cells corresponding to the sample $(n_{i.})$, then, for each integer $s > 1$, the sample $(n_{i.})$ could constitute the row totals of some r by s contingency table having physical probabilities (p_{ij}), for which, under the nonnull hypothesis, the "no-structure" assumption applies. We may call this our "stronger no-structure" assumption.

We need now to discuss multinomials before completing the specification of the Bayesian model.

2. The mixed-Dirichlet Bayesian model. We first discuss multinomials, partly because this discussion will lead to a constraint on the Bayesian model for contingency tables. The discussion also leads to a constraint on the model for multinomials themselves.

Let $(p_1, p_2, \ldots, p_t)$ be the t physical probabilities corresponding to t categories, where $p_1 + p_2 + \cdots + p_t = 1$. The subjective probability that the next item sampled will belong to the ith category is equal to the subjective expectation of p_i. Suppose that we have a multinomial sample $(m_1, m_2, \ldots, m_t)$, denoted by (m_i), where $\Sigma m_i = M$, the sample size. Suppose further that, whatever the sample (m_i), the subjective expectation of p_i, given the data, is $(m_i + k_i)/(M + \kappa)$ where $\kappa = \Sigma k_i$, so that the numbers κ_i can be regarded as flattening constants. As, for example, in Good (1965, pages 22–25) this assumption can be seen to be *equivalent* to assuming the prior Dirichlet density

$$D(t, \mathbf{p}, \mathbf{k}) = \Gamma(\kappa) \prod_{i=1}^{t} \left\{ \frac{p_i^{k_i - 1}}{\Gamma(k_i)} \right\}, \tag{2.1}$$

for the p_i's, where $\mathbf{p}$ means $(p_1, p_2, \ldots, p_t)$ and $\mathbf{k}$ means $(k_1, k_2, \ldots, k_t)$.

When there is no sample, that is, when $M = 0$, the initial subjective probability of the ith category is k_i/κ which we call π_i. Thus $D(t, \mathbf{p}, \mathbf{k}) = D(t, \mathbf{p}, \kappa\pi)$ where $\pi = (\pi_1, \pi_2, \ldots, \pi_t)$. [Cf. Good (1967), formula (25), where κ was called kt but the misprint of a plus sign for t was undetected.] We may picture the prior probabilities π_i as the lengths of consecutive segments of a unit interval. For the sake of consistency, that is, invariance under lumping of categories, the flattening constants must be "additive." For example, if the first two categories are combined, thus reducing the number of categories to $t - 1$, the corresponding flattening constant is $k_1 + k_2$, that is $(\pi_1 + \pi_2)\kappa$. This is an intuitive reason for the lumping property of the Dirichlet distribution: An analytic proof is given by Wilks (1962, page 179). Note that it would contradict previous usage to call κ a flattening constant, and we call it a shrinking constant or shrinker because it shrinks the observed proportions $\mathbf{m}/M$ towards π.

Zabell (1982) generalized the argument of W. E. Johnson (1932) to the case where the prior expectations of the p_i's are unequal, so in this case it is again natural to use a mixture of general Dirichlet distributions, as proposed for this case by Good (1967, page 409) without noticing that Johnson's argument could be generalized.

We may write our mixed-Dirichlet prior in the form

$$\int_0^\infty \Gamma(\kappa) \prod_{i=1}^{t} \left\{ \frac{p_i^{\kappa\pi_i - 1}}{\Gamma(\kappa\pi_i)} \right\} h(t, \kappa)\, d\kappa = \int_0^\infty D(t, \mathbf{p}, \kappa\pi) h(t, \kappa)\, d\kappa, \tag{2.2}$$

where h can be regarded as a hyperprior. We shall now argue that h can be taken as a function of κ alone, that is, that it is mathematically independent of t. The argument will be more general and more rigorous than the one that was given by Good (1980a) and in more detail by Crook and Good (1980, page 1201).

Imagine that, for each i, the ith category is further categorized into s subcategories having prior probabilities π_{ij}, $j = 1, 2, \ldots, s$; $\sum_j \pi_{ij} = \pi_{i\cdot} = \pi_i$, where s is the same for each i. We are, so to speak, treating our one-dimensional sequence of probabilities π_i as the row totals of a "population contingency table" of s columns. We think of contingency tables as r by s, so we temporarily replace t by r in this argument.

The physical probabilities p_{ij} in the contingency table, given H', can also be thought of as those for a multinomial of rs categories in accordance with Assumption 2. They have the prior density

$$\int_0^\infty D\left[rs, (\kappa\pi_{ij}), (p_{ij})\right] h(rs, \kappa)\, d\kappa.$$

Therefore, by the lumping property of the Dirichlet distribution, the prior for the row totals $(p_{i\cdot})$ (the original multinomial) must be

$$\int_0^\infty D\left[r, (\kappa\pi_i), (p_{i\cdot})\right] h(rs, \kappa)\, d\kappa,$$

because $\pi_i = \sum_j \pi_{ij}$. But s could be any integer (in accordance with our stronger no-structure assumption) so $h(rs, \kappa)$ must be mathematically independent of s, for each r. Given two values of r, say r_1 and r_2, it follows, by taking s as r_2 and

r_1 in turn, that $h(r_1, \kappa) = h(r_1 r_2, \kappa)$ and $h(r_2, \kappa) = h(r_2 r_1, \kappa)$. Therefore, $h(r_1, \kappa) = h(r_2, \kappa)$; in other words, $h(r, \kappa)$ is mathematically independent of r as we claimed. Hence, we may now write $h(\kappa)$ instead of $h(r, \kappa)$. Of course, rows and columns can be interchanged in this discussion. The stronger "no-structure" assumption thus forces us to conclude that the same function h is to be used for the vectors $(p_{i\cdot})$ and $(p_{\cdot j})$, and, given H', for the "vector" (p_{ij}). This property was assumed by Good (1980a, page 31) and Crook and Good (1980, page 1201) but with less justification than we have now supplied.

To recapitulate, when we are considering a multinomial of t categories, for which no structure is assumed beyond the prior expectations π_i of the physical probabilities p_i, we take as our prior density for (p_i)

$$\int_0^\infty D(t, \mathbf{p}, \kappa\boldsymbol{\pi}) h(\kappa)\, d\kappa, \tag{2.3}$$

for some hyperprior density $h(\cdot)$. We shall use (2.3) with $t = r$, s and rs according to context. That is, the prior density of $(p_{i\cdot})$, given H, is expression (2.3) with $t = r$, $\mathbf{m} = (n_{i\cdot})$ and $\boldsymbol{\pi} = (q_{i\cdot})$, and similarly for $(p_{\cdot j})$; and these two priors are independent by Assumption 4. Again, the prior density of (p_{ij}), given H', is expression (2.3) with $t = rs$, $\mathbf{m} = (n_{ij})$ and $\boldsymbol{\pi} = (q_{ij})$. These two sentences, combined with Assumption 6, complete the description of the Bayesian model except that the choice of the hyperpriors will be discussed in Section 7.

3. The Dirichlet-multinomial distribution. If the parameters $(p_1, p_2, \ldots, p_t)$ in the multinomial probability $M!\Pi(p_i^{m_i}/m_i!)$ are assumed to have the Dirichlet prior $D(t, \mathbf{k}, \mathbf{p})$, then the "marginal" probability mass function of (m_i) is the compound Dirichlet-multinomial probability

$$\begin{aligned} &\frac{M!\Gamma(\kappa)}{\Gamma(M+\kappa)} \prod_{i=1}^{t} \frac{\Gamma(m_i + k_i)}{\Gamma(k_i) m_i!} \qquad (k_i > 0;\ i = 1, 2, \ldots, t) \\ &= \frac{M!\Gamma(\kappa)}{\Gamma(M+\kappa)} \prod_{i=1}^{t} \frac{\Gamma(m_i + \kappa\pi_i)}{\Gamma(\kappa\pi_i) m_i!}, \qquad (\kappa > 0). \end{aligned} \tag{3.1}$$

[See for example, Johnson and Kotz (1969), page 309, Good (1957), page 862, and Mosimann (1962).] We denote this expression by $DM[t, (m_i), (\kappa\pi_i)]$ or $DM(t, \mathbf{m}, \kappa\boldsymbol{\pi}]$. A four-level apparently non-Bayesian hierarchical model based on the Dirichlet-multinomial distribution was used recently by Goodhardt, Ehrenberg and Chatfield (1984). We take a Bayesian point of view and regard κ as a hyperparameter. It acts as a shrinker.

4. The doubly compounded distributions. Since we are assuming a hyperprior density h for κ, the frequencies (m_i) have a marginal distribution

$$\int_0^\infty DM(t, \mathbf{m}, \kappa\boldsymbol{\pi}) h(\kappa)\, d\kappa. \tag{4.1}$$

This can be regarded as a *doubly compounded* probability mass function which we call the h-Dirichlet-multinomial or h-$DM(t, \mathbf{m}, \boldsymbol{\pi})$. For example, if h is a

log-normal density, then (4.1) would define a log-normal-Dirichlet-multinomial distribution. (A doubly compounded distribution can be "de-Bayesianized" by saying it is capable of describing data, just as singly compounded distributions are sometimes so regarded.)

Just as (2.3) expresses the three relevant priors succinctly, (4.1) expresses the relevant marginal distributions of $(n_{i.})$ and $(n_{.j})$, given H, and of (n_{ij}), given H'. The distribution of (n_{ij}) conditional on $(n_{i.})$ and $(n_{.j})$, given H, is given by the Fisher–Yates formula quoted at (5.2) below.

5. The Bayes factor against "independence" under sampling procedures 1 and 2. Good (1976) gave formulas, based on symmetric Dirichlets, for the Bayes factors against H, assuming sampling Procedures 1 and 2, and called these factors F_1 and F_2. Crook and Good (1980), following a method in Good (1980a), modified the hyperprior to make $F_1 = F_2$, because the row totals alone were assumed to contain no evidence (or at least negligible evidence) for or against H. Let us verify that this condition is satisfied with our present model, as defined following formula (2.3). This verification is not strictly necessary but it might give the reader confidence that no mistakes have occurred.

By (2.3) the prior density of (p_{ij}), given H', is

$$\int D\left[rs, (p_{ij}), (\kappa q_{ij})\right] h(\kappa)\, d\kappa$$

and that of $(p_{i.})$ is, therefore,

$$\int D\left[r, (p_{i.}), (\kappa q_{i.})\right] h(\kappa)\, d\kappa,$$

by the lumping property of Dirichlet distributions. (There is an implicit reversal of the order of integrations in this argument.) But, again by (2.3), this is the same as the prior density of $(p_{i.})$, given H. Hence,

$$P\left[(n_{i.})|H\right] = P\left[(n_{i.})|H'\right],$$

so our Assumption 1 is verified. Thus the Bayes factor against H provided by (n_{ij}), given Sampling Procedure 1, is

$$\begin{aligned} F_1 &= P(E_1|H')/P(E_1|H) \quad \left(\text{where } E_1 \text{ denotes } (n_{ij})\right) \\ &= P(E_1 \,\&\, E_2|H')/P(E_1 \,\&\, E_2|H) \quad \left(\text{where } E_2 \text{ denotes } (n_{i.})\right) \\ &= \frac{P(E_2|H')P(E_1|E_2 \,\&\, H')}{P(E_2|H)P(E_1|E_2 \,\&\, H)} \\ &= P(E_1|E_2 \,\&\, H')/P(E_1|E_2 \,\&\, H) = F_2. \end{aligned}$$

Thus Assumption 1, which we have just verified, that the row totals alone give no evidence for or against H implies that $F_1 = F_2$ in the sense that F_1 and F_2 are the same function of (n_{ij}) and (q_{ij}), that is, of (n_{ij}), $(q_{i.})$ and $(q_{.j})$ because $q_{ij} = q_{i.}q_{.j}$.

Given the *h-DM* distribution for the three relevant frequency counts, $(n_{i.})$, $(n_{.j})$ and (n_{ij}), we have, when sampling by Procedure 1,

$$P\left[(n_{ij})|H'\right] = h\text{-}DM\left[rs,(n_{ij}),(q_{ij})\right] = \int_0^\infty DM\left[rs,(n_{ij}),(\kappa q_{ij})\right]h(\kappa)\,d\kappa,$$

and

$$P\left[(n_{i.})|H'\right] = P\left[(n_{i.})|H\right]$$

(because the row totals are assumed to give no evidence for or against H)

$$= h\text{-}DM\left[r,(n_{i.}),(q_{i.})\right].$$

In the future we will use the less explicit symbol f (for "friendlier") in place of *h-DM*. We then have

$$P\left[(n_{ij})|(n_{i.}),H'\right] = \frac{f\left[rs,(n_{ij}),(q_{ij})\right]}{f\left[r,(n_{i.}),(q_{i.})\right]}. \tag{5.1}$$

This formula generalizes formula (4.7) of Good (1976). We shall sometimes take $q_{i.}$ and $q_{.j}$ for granted and omit them from some notations. We determine them as described in Section 1: see (1.2).

We must now write down the probabilities given H. We recall first the Fisher–Yates formula [or Fisher–Yates–Irwin–Mood formula: see Good (1984a) and (1984b) and Barnard (1984) for a historical discussion]

$$\text{F.–Y.} = P\left[(n_{ij})|(n_{i.}),(n_{.j}),H\right] = \frac{\Pi n_{i.}!\,\Pi n_{.j}!}{N!\,\Pi n_{ij}!}. \tag{5.2}$$

Now

$$\begin{aligned} P\left[(n_{ij})|(n_{i.}),H\right] &= P\left[(n_{ij}),(n_{.j})|(n_{i.}),H\right] \quad \text{(trivially)} \\ &= P\left[(n_{.j})|(n_{i.}),H\right]P\left[(n_{ij})|(n_{i.}),(n_{.j}),H\right] \\ &= P\left[(n_{.j})|H\right]\text{F.–Y.} \quad \text{(see Assumption 4)} \\ &= \text{F.–Y.}\,f\left[s,(n_{.j}),(q_{.j})\right]. \end{aligned} \tag{5.3}$$

On taking the ratio of (5.1) to (5.3) we have the Bayes factor against H,

$$F_1 = F_2 = \frac{f\left[rs,(n_{ij}),(q_{ij})\right]}{f\left[r,(n_{i.}),(q_{i.})\right]f\left[s,(n_{.j}),(q_{.j})\right]\text{F.–Y.}}, \tag{5.4}$$

which is symmetrical with respect to rows and columns, as it had to be because it is equal to F_1. For an algorithm for computing F_1, see Crook and Good (1985).

It is of interest to generalize the remark made by Good (1976, Section 9). In fact, F_1 can be written in the form

$$\frac{F\left[H_{00}\colon(n_{ij})\right]}{F\left[H_{0.}\colon(n_{i.})\right]F\left[H_{.0}\colon(n_{.j})\right]},$$

where H_{00} denotes the hypothesis that $p_{ij} = q_{i.}q_{.j}$, etc., and $F[H_{00}\colon (n_{ij})]$ denotes the Bayes factor against H_{00} provided by (n_{ij}), and so on. The three Bayes factors all refer to multinomial hypotheses.

In the particular case where $h(\kappa)$ is a Dirac function, that is, then κ is assumed to have a specific value, (5.4) reduces to

$$F_1(\kappa) = F_2(\kappa) = \frac{DM\left[rs, (n_{ij}), (\kappa q_{ij})\right]}{DM\left[r, (n_{i.}), (\kappa q_{i.})\right] DM\left[s, (n_{.j}), (\kappa q_{.j})\right]}\text{F.–Y.}$$

$$= \frac{\prod_{ij}\prod_{a=0}^{n_{ij}-1}(a + \kappa q_{ij})\prod_{a=0}^{N-1}(a + \kappa)}{\prod_{i}\prod_{a=0}^{n_{i\cdot}-1}(a + \kappa q_{i.})\prod_{j}\prod_{a=0}^{n_{\cdot j}-1}(a + \kappa q_{.j})}, \tag{5.5}$$

where empty products, if they occur, are interpreted as unity in accordance with the customary convention. We conjecture that $F_1(\kappa)$ is always unimodal (a property that is useful when we are interested in finding its maximum value).

It is possible to interpret formula (5.5) in a non-Bayesian manner, by regarding the compounded distributions as physical distributions. Then (5.5), for any fixed κ, can be interpreted as a simple likelihood ratio instead of as a Bayes factor.

6. What we vary in the robustness study. Our robustness study deals with the variability of F_1 or F_2 when the assumptions are varied in two ways:

(i) variations in the choice of the quartiles of κ, as in earlier work;
(ii) variations from one hyperprior h to another.

7. Hints on the choice of the hyperprior or mixing function h. Our numerical robustness or sensitivity study will involve a number of different hyperpriors h, and we must consider what kinds are reasonable. In previous work we assumed that h was a log-Cauchy density and questions of robustness involved only the variation of its two hyperhyperparameters or equivalently its two quartiles. For the sake of completeness, we shall redefine the log-Cauchy distribution later and mention some of its properties.

It needs to be emphasized that the hyperprior is apt to depend on the application and on the judgment of the statistician and of his client. Much of the subjectivistic aspect can be taken into account by making judgments of the upper and lower quartiles, q_U and q_L, of the hyperparameter or shrinker κ. (Any two quantiles would do and checks for consistency can be obtained by judging additional quantiles.) Tests for robustness can be performed both by varying q_U and q_L, and by varying the functional form of h. In the longer version of this paper, available from the authors, there are a number of suggestions concerning judgments about κ. We mention here just one idea which we express in terms of a multinomial population with unknown physical probabilities (p_i) whose prior subjective expectations are (π_i). Then it can be shown that an approximation to κ is

$$\sum \frac{p_i(1 - p_i)}{\pi_i} \Bigg/ \left(\sum \frac{p_i^2}{\pi_i} - 1 \right), \tag{7.1}$$

so that judgments about the quartiles of κ could depend on those judged for this expression.

We next turn our attention to the functional form for h. Observe first that h must be "proper at infinity" otherwise we would find that $F_1 = F_2 = 1$, that is, the evidence for or against H would be obliterated [cf. Good (1965, pages 38–39)]. To see this, note that when $\kappa \to \infty$, DM tends to the multinomial form $M!\Pi(\pi_i^{m_i}/m_i!)$ (as one would expect) so that, from (5.4), we find that $F_1 = F_2 = 1$ as we just said. On the other hand, we want to be permissive towards large values of κ because they correspond to believing that our initial estimates of the $p_{i\cdot}$, $p_{\cdot j}$ and p_{ij} are fairly accurate, and we do not want to eliminate this possibility by fiat. Hence a hyperprior $h(\kappa)$ that is, for example, exponentially small for large κ would in our opinion be unreasonable. A numerical example given later, where h is taken as a Weibull distribution, will presumably convince the reader of this point.

To discuss the form of $h(\kappa)$ for $\kappa \ll 1$ note that

$$(7.2) \qquad DM(t, \mathbf{m}, \kappa\boldsymbol{\pi})/DM(t, \mathbf{m}, \boldsymbol{\pi}) \sim M\kappa^{t'-1}\Pi_i'\left\{\frac{\Gamma(m_i)\Gamma(1+\pi_i)}{\Gamma(m_i+\pi_i)}\right\},$$

where $\kappa \to 0$, where i runs through the values of i for which $m_{i\cdot} > 0$, and where t' is the number of such i's. If $t' \geq 2$, which is true in all cases of interest, the contribution to the integral in (4.1), from $\kappa \ll 1$, will be negligible if $\kappa^{1+\varepsilon}h(\kappa) \to 0$ when $\kappa \to 0$ for some $\varepsilon > 0$. This condition is satisfied for the forms of h that we shall entertain: We allow $h(\kappa)$ to tend to infinity at the origin, but not as fast as $\kappa^{-1-\varepsilon}$.

If it were not improper, the Jeffreys–Haldane density $1/\kappa$ would be a candidate for being h but we must achieve propriety, and a way to do so, while approximating proportionality to $1/\kappa$ almost as much as possible, is to use distributions asymptotically proportional to $1/[\kappa(\log\kappa)^2]$ for large κ. The log-Cauchy distribution, used in earlier work, has this property. (The tendency to infinity when $\kappa \to 0$ is not really desirable but, for the reason just mentioned, this end of the distribution is unimportant, so we have preferred the elegance of a simple functional form for h having convenient properties.) We shall also consider generalized log-Student distributions. They raise $(\log\kappa)^{-1}$ to higher powers (than does the log-Cauchy) for large κ, and we believe this is a disadvantage. We have also considered the log-normal distribution (which is the limiting form of the generalized log-Students). We next list these distributions explicitly.

8. The hyperpriors tried. Each of our hyperpriors will have two hyper-hyperparameters which can be determined if two quantiles are assumed for the hyperprior.

If the logarithm of a random variable κ has a Cauchy distribution, we say that the random variable has a log-Cauchy distribution. We slightly generalize this distribution by giving it an additional parameter, but we continue to refer to

the distribution as log-Cauchy as in Good (1969, page 45). The density is

$$\frac{\lambda}{\kappa\pi\left\{\lambda^2 + \left[\log(\kappa/\mu_1)\right]^2\right\}}, \qquad \lambda > 0,\ \mu_1 > 0. \tag{8.1}$$

This distribution has median μ_1, and upper and lower quartiles

$$q_U = \mu_1 e^{\lambda}, \qquad q_L = \mu_1 e^{-\lambda}, \tag{8.2}$$

so that

$$\mu_1 = (q_L q_U)^{1/2}, \qquad \lambda = \tfrac{1}{2}\log(q_U/q_L). \tag{8.3}$$

Let $q(\alpha)$ denote the 100αth percentile. For example, $q(0) = 0$, $q(\frac{1}{4}) = q_L$, $q(\frac{1}{2}) = \mu_1$, $q(\frac{3}{4}) = q_U$ and $q(1) = \infty$. (8.2) and (8.3) can be generalized corresponding to any two quantiles $q(\alpha)$ and $q(\beta)$.

The log-Cauchy density is strictly decreasing if $\lambda \geq 1$, otherwise it has a local minimum at $\mu_1\exp(-1 - \sqrt{1 - \lambda^2})$ and a local maximum at

$$\mu_1\exp(-1 + \sqrt{1 - \lambda^2}).$$

Some other properties of the log-Cauchy distribution are mentioned by Good (1969, pages 46: a minus sign is omitted in the formula for the first percentile).

We call the density function

$$\frac{\lambda^{\nu}}{\kappa\nu^{1/2}B(\frac{1}{2}, \frac{1}{2}\nu)}\left\{\lambda^2 + \nu^{-1}\left[\log\frac{\kappa}{\mu_1}\right]^2\right\}^{-(\nu+1)/2}, \qquad \nu > 0,\ \lambda > 0,\ \mu_1 > 0, \tag{8.4}$$

the generalized log-Student density. When $\nu = 1$ it reduces to the log-Cauchy density. The median of (8.4) is the geometric mean of its upper and lower quartiles, a property shared by any distribution whose random variable has a logarithm with a density symmetric about its median. When $\nu \to \infty$, the generalized log-Student distribution tends to the log-normal for $\kappa > 0$. In our numerical work we considered the values $\nu = 1$, 2 and 15, and regarded the case $\nu = 15$ as close enough to being log-normal. In other words, $\nu = 15$ is virtually the same as $\nu = \infty$. The density function of a log-normal variable goes through the origin although each log-Student density has the y axis as an asymptote. The apparent paradox is resolved by a glance at Figure 1.

Finally, we considered the Weibull distribution as our hyperprior although, as stated in Section 7, it cannot be expected to lead to reasonable results when the null hypothesis is (approximately) true. The Weibull density is of the form

$$abx^{a-1}e^{-bx^a}. \tag{8.5}$$

If two quantiles $q(\alpha)$ and $q(\beta)$ are judged, corresponding to the 100α and 100β percentiles, then one can compute a and b from the equations

$$a = \log\left[\frac{\log(1-\alpha)}{\log(1-\beta)}\right] + \log\left[\frac{q(\alpha)}{q(\beta)}\right] \tag{8.6}$$

and

$$b = [q(\beta)]^{\alpha'}/[q(\alpha)]^{\beta'}, \tag{8.7}$$

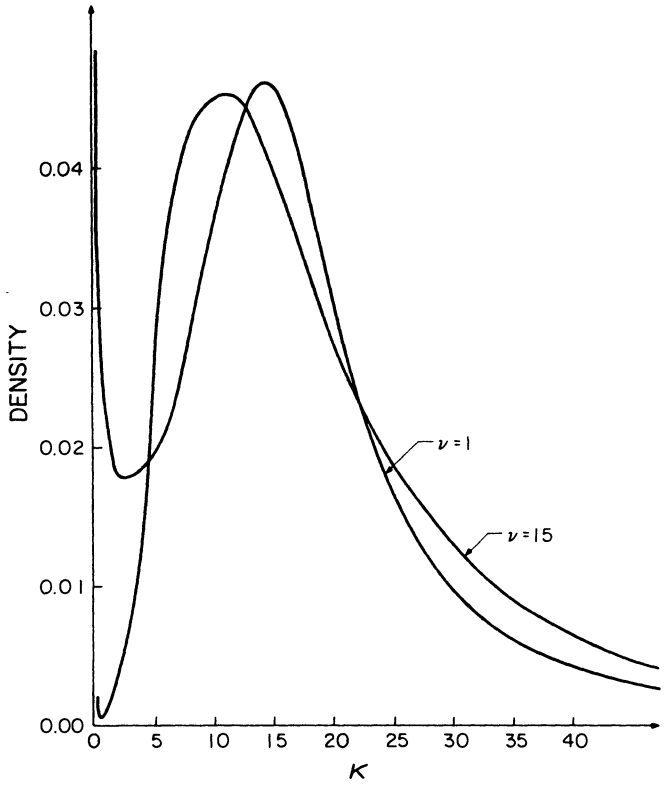

FIG. 1. *The generalized log-Student densities with* $\nu = 1$ *and* $\nu = 15$, *and* $q_L = 10$, $q_U = 25$. *The y axis is an asymptote for all (finite) values of* ν. *The graph is strictly decreasing if* $4\nu\lambda^2 \geq (\nu + 1)^2$. *For* $\nu = 1$ *(the log-Cauchy), the condition is* $\lambda \geq 1$ *while if* $\lambda < 1$ *the local maxima and minima are at* $\mu_1 \max[-1 \pm (1 - \lambda^2)^{1/2}]$ *where* μ_1 *and* λ *are given by* (8.3).

where

$$\alpha' = \log\log[1/(1 - \alpha)], \qquad \beta' = \log\log[1/(1 - \beta)].$$

9. Known marginal probabilities. In census work, one often has large multinomial samples from which $(p_{i\cdot})$ and $(p_{\cdot j})$ can be estimated with considerable accuracy, and can be regarded as known, but a much smaller sample for the corresponding contingency table. See, for example, Deming and Stephan (1940), Bishop, Fienberg and Holland (1975, page 84ff.) and Pelz (1977, pages 46–54). Most of the statistical literature dealing with such situations is concerned with the estimation of the probabilities p_{ij} but we here discuss how it affects the Bayes factor against H. We assume Sampling Procedure 1 because we already know that $F_2 = F_1$.

Since we are now regarding $p_{i\cdot}$ and $p_{\cdot j}$ as known (for all i and j) we have

$$P[(n_{ij})|H] = \frac{N!}{\Pi n_{ij}!}\Pi p_{ij}^{n_{ij}} = \frac{N!}{\Pi n_{ij}!}\Pi p_{i\cdot}^{n_{i\cdot}} \cdot \Pi p_{\cdot j}^{n_{\cdot j}}. \tag{9.1}$$

Moreover, $q_{i\cdot}$ and $q_{\cdot j}$ are now equal to $p_{i\cdot}$ and $p_{\cdot j}$ so that

$$P\left[(n_{ij})|H'\right] = f\left[(n_{ij}),(p_{i\cdot}p_{\cdot j})\right], \tag{9.2}$$

and F_1 is the ratio of the two probabilities (9.2) and (9.1). For Sampling Procedure 2 we have to divide the probabilities in both (9.1) and (9.2) by the same expression namely $N!\Pi(p_{i\cdot}^{n_{i\cdot}}/n_{i\cdot}!)$. This verifies that $F_2 = F_1$ and we have

$$F_1 = F_2 = \frac{f\left[(n_{ij}),(p_{i\cdot}p_{\cdot j})\right]\Pi n_{ij}!}{N!\Pi p_{i\cdot}^{n_{i\cdot}}\Pi p_{\cdot j}^{n_{\cdot j}}}. \tag{9.3}$$

If only $(p_{i\cdot})$ is known, then

$$P\left[(n_{ij})|H\right] = f\left[(n_{\cdot j}),(q_{\cdot j})\right]N!\Pi(p_{i\cdot}^{n_{i\cdot}}/n_{i\cdot}!)\text{F.–Y.}, \tag{9.4}$$

while

$$P\left[(n_{ij})|H'\right] = f\left[(n_{ij}),(p_{i\cdot}q_{\cdot j})\right] \tag{9.5}$$

and F_1 and F_2 are equal to the ratio of the probabilities given by (9.5) and (9.4).

10. Numerical results and recommendations. We now return to the numerical results in Tables 1 to 20.

In each example we have eye-balled the contingency table to see if it has any special structure, and have found none.

Apart from "a priori" arguments for choosing one hyperprior rather than another (or a prior in other circumstances), it is desirable to look at numerical results because some implications might seem more reasonable than others. It is not necessary that the examples should be real ones. When they are not real, the method is an application of what physicists would call Gedanken experiments, that is, of the device of imaginary results. They have latterly been given a more preposterous name. It is also interesting, as in most of the following examples, to apply the methods to real data.

We present the values of F_1 $(= F_2)$ for 21 contingency tables of which 7 are artificial. The meanings of $P(\chi^2)$, $P(\chi'^2)$ and P("exact") were mentioned in Section 1 where it was also mentioned that these are most appropriate under Sampling Procedure 3. The rows labelled 1, 2 and 15 correspond to the generalized log-Student hyperprior for κ with $\nu = 1$, 2 and 15. When $\nu = 15$ the hyperprior is a very close approximation to a log-normal distribution and we could have written ∞ in place of 15. The first nine columns of values of F_1 are headed by the corresponding values of q_L and q_U, the judged lower and upper quartiles of κ. We have selected the nine pairs (q_L, q_U) to give adequate variety. The tenth column gives the average F_1 for these nine cases, or, as one might say, the average over nine Bayesians. This tenth column gives a rough idea of what F_1 might be if we had gone up another level in the hierarchy, and put a bivariate hyperhyperprior on (q_L, q_U). The three values of these averages, corresponding to $\nu = 1$, 2 and 15 (or ∞), agree with one another within a factor of 2 except for cases where $P(\chi^2) = 1$ or $P(\chi^2) < 10^{-4}$, or for the 14×20 horse-kick data (where the ratio is about $3\frac{1}{2}$ corresponding to $\nu = 1$ and $\nu = 15$). Thus it seems

TABLE 1
Criminality of monozygotic twins of criminals (two-by-two)

ν \ (q_L, q_U)	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	140	116	65	77	55	31	47	31	17	64	8
2	172	136	66	81	58	32	49	32	18	72	10
15	201	159	67	86	60	33	51	34	19	79	11

[Fisher (1938), page 99, who cites Lange (1931). See also Good (1978).] [10,3; 2,15]. $\chi^2 = 13.0$, $\chi'^2 = 10.5$; $1/P(\chi^2) = 3270$, $1/P(\chi'^2) = 1640$, $1/P(\text{"exact"}) = 1/2150$.

TABLE 2
Parental decision-making and political affiliation (two-by-two)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	4.9	4.7	4.0	3.2	3.3	3.2	3.0	3.1	2.8	3.6	2
2	3.8	4.5	4.4	3.1	3.5	3.4	3.1	3.3	2.9	3.6	1.5
15	1.6	3.1	4.9	2.7	3.5	3.6	2.9	3.5	3.0	3.2	3

[Bishop, Fienberg and Holland (1975) (= BFH), page 380.] [29,33; 131,78]. $\chi^2 = 5.0$, $\chi'^2 = 4.4$, $1/P(\chi^2) = 39$, $1/P(\chi'^2) = 27$.

TABLE 3
Inoculations during an epidemic (two-by-two)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	8.0	7.2	5.6	4.9	5.0	4.5	4.6	4.7	4.0	5.4	2
2	6.8	7.7	6.7	5.0	5.6	5.1	4.8	5.2	4.3	5.7	2
15	2.2	5.1	8.2	4.2	5.8	5.8	4.5	5.7	4.5	5.1	4

[Greenwood and Yule (1915).] Tables 3, 4 and 5 refer to various locations. [200,8; 182,20]. $\chi^2 = 5.9$, $\chi'^2 = 5.0$; $1/P(\chi^2) = 66$, $1/P(\chi'^2) = 39$.

TABLE 4
See the caption of Table 3 (two-by-two)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	3.6	3.4	2.8	2.0	2.1	2.1	1.8	1.9	1.9	2.4	2
2	2.3	2.8	2.9	1.7	2.0	2.1	1.6	1.9	1.9	2.1	2
15	0.7	1.6	2.9	1.4	1.9	2.2	1.4	1.9	1.9	1.8	4

[105, 5; 88, 11]. $\chi^2 = 3.2$, $\chi'^2 = 2.3$; $1/P(\chi^2) = 13$, $1/P(\chi'^2) = 7.8$.

TABLE 5
See the caption of Table 3 (two-by-two)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	16	14	10	10	10	9	9	9	8	11	2
2	17	18	14	11	12	11	10	11	9	13	2
15	4	12	19	9	13	13	9	13	10	11	5

[409, 3; 174, 8]. $\chi^2 = 9.3$, $\chi'^2 = 7.4$; $1/P(\chi^2) = 450$, $1/P(\chi'^2) = 160$; $1/P(\text{"exact"}) = 217$.

TABLE 6
Political preferences (two-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	15	11	6	6	5	4	5	4	4	7	4
2	19	14	8	6	6	5	4	5	4	8	5
15	1	5	9	2	5	6	2	5	4	4	9

(BFH, page 387.) [225, 53, 206; 3, 1, 12]. $\chi^2 = 6.7$, $1/P(\chi^2) = 28$.

TABLE 7
Mendel's data on garden peas (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	18	12	5.8	7.7	5.5	3.4	5.2	3.6	2.3	7.1	8
2	28	16	5.9	5.8	4.5	3.0	2.9	2.7	2.1	7.9	13
15	0.08	1.0	3.5	0.1	0.7	1.9	0.06	0.7	1.6	1.1	58

(BFH, page 328.) [38, 60, 28; 65, 138, 68; 35, 67, 30]. $\chi^2 = 1.9$, $1/P(\chi^2) = 1.4$.

TABLE 8
Distribution of votes (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	2(6)	1(6)	0.8(6)	2(6)	2(6)	0.8(6)	3(6)	2(6)	0.4(6)	1.6(6)	7.5
2	8(6)	5(6)	1(6)	5(6)	3(6)	0.9(6)	5(6)	2(6)	0.4(6)	3.4(6)	20
15	14(6)	19(6)	3(6)	11(6)	6(6)	1(6)	8(6)	3(6)	0.4(6)	7.3(6)	48

(BFH, page 99.) [61, 12, 60; 17, 6, 1; 39, 22, 7]. $\chi^2 = 42$, $1/P(\chi^2) = 50{,}000{,}000$. 3(6), for example, means 3×10^6.

TABLE 9
Supervisors' ratings of student teachers (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	1500	1200	600	1500	1000	400	1300	600	200	900	7.5
2	3500	2400	800	2000	1100	400	1500	700	200	1400	18
15	5900	5100	1100	2700	1400	400	1800	700	200	2100	30

(BFH, page 402.) [17, 4, 8; 5, 12, 0; 10, 3, 13]. $\chi^2 = 28$, $1/P(\chi^2) = 70{,}000$.

TABLE 10
Lambs born in consecutive years (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	3(4)	3(4)	1(4)	3(4)	2(4)	1(4)	4(4)	3(4)	1(4)	2.3(4)	4
2	12(4)	8(4)	3(4)	7(4)	5(4)	2(4)	6(4)	4(4)	1(4)	5.3(4)	12
15	12(4)	22(4)	5(4)	11(4)	9(4)	2(4)	10(4)	5(4)	1(4)	8.6(4)	22

(BFH, page 288.) [58, 52, 1; 26, 58, 3; 8, 12, 9]. $\chi^2 = 50$, $1/P(\chi^2) = 2{,}000{,}000{,}000$.

TABLE 11
Artificial (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	2.5	2.0	1.6	0.7	0.9	1.0	0.6	0.8	1.0	1.2	4
2	1.4	1.4	1.4	0.5	0.7	1.0	0.5	0.7	1.0	1.0	3
15	0.2	0.5	1.2	0.3	0.6	0.9	0.4	0.6	1.0	0.6	4

[2, 2, 2; 2, 2, 2; 2, 2, 2]. $\chi^2 = 0$, $1/P(\chi^2) = 1$.

TABLE 12
Artificial (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	5.5	4.0	2.4	1.4	1.3	1.3	0.8	1.0	1.1	2.1	7
2	3.6	2.9	2.0	0.8	1.0	1.2	0.5	0.8	1.1	1.5	7
15	0.1	0.5	1.5	0.2	0.5	1.0	0.2	0.5	1.0	0.6	15

[6, 6, 6; 6, 6, 6; 6, 6, 6]. $\chi^2 = 0$, $1/P(\chi^2) = 1$.

TABLE 13
Artificial (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	1.8	1.6	1.3	0.6	0.7	0.9	0.5	0.7	0.9	1	4
2	0.9	1.0	1.2	0.4	0.6	0.9	0.4	0.6	0.9	0.8	3
15	0.1	0.4	1.0	0.3	0.5	0.8	0.4	0.6	0.9	0.6	10

$[1,2,3;\ 1,2,3;\ 1,2,3]$. $\chi^2 = 0$, $1/P(\chi^2) = 1$.

TABLE 14
Artificial (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	6.7	4.7	2.6	1.7	1.5	1.4	0.9	1.0	1.1	2.4	7
2	4.8	3.6	2.2	0.9	1.1	1.2	0.5	0.8	1.1	1.8	10
15	0.1	0.5	1.5	0.2	0.5	1.0	0.2	0.5	1.0	0.6	15

$[1,5,20;\ 1,5,20;\ 1,5,20]$. $\chi^2 = 0$, $1/P(\chi^2) = 1$.

TABLE 15
Artificial (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	0.88	0.88	0.88	0.94	0.93	0.92	0.96	0.95	0.94	0.92	1
2	0.91	0.90	0.89	0.95	0.94	0.93	0.97	0.95	0.94	0.93	1
15	0.94	0.91	0.89	0.96	0.95	0.93	0.97	0.96	0.95	0.94	1

$[5,0,0;\ 5,0,0;\ 5,0,0]$. $\chi^2 = 0$, $1/P(\chi^2) = 1$.

TABLE 16
Artificial (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	4(6)	2(6)	32(4)	14(4)	10(4)	5(4)	4(4)	3(4)	2(4)	7(5)	200
2	6(6)	2(6)	36(4)	17(4)	11(4)	6(4)	4(4)	3(4)	3(4)	10(5)	200
15	10(6)	3(6)	41(4)	18(4)	13(4)	8(4)	2(4)	3(4)	3(4)	15(5)	500

[6, 0, 0; 0, 6, 0; 0, 0, 6]. $\chi^2 = 36$, $1/P(\chi^2) = 3{,}000{,}000$.

TABLE 17
Artificial (three-by-three)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	7	6	4	4	4	3	3	3	2	4	3.5
2	8	7	4	4	4	3	3	3	2	4	4
15	10	8	4	5	4	3	4	3	2	5	5

[5, 1, 0; 4, 0, 2; 2, 4, 0]. $\chi^2 = 10.5$, $1/P(\chi^2) = 30$.

TABLE 18
Eye color versus hair color (four-by-four)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	0.9(22)	0.7(22)	0.4(22)	1.5(22)	1.1(22)	0.5(22)	2.7(22)	1.4(22)	0.4(22)	1(22)	7
2	12(22)	5(22)	1.1(22)	12(22)	5(22)	1.0(22)	18(22)	4(22)	0.5(22)	7(22)	24
15	690(22)	415(22)	7(22)	198(22)	55(22)	2.4(22)	137(22)	19(22)	0.6(22)	169(22)	1150

[Snee (1974) and Diaconis and Efron (1985).] [68, 119, 26, 7; 20, 84, 17, 94; 15, 54, 14, 10; 5, 29, 14, 16]. $\chi^2 = 138$, $1/P(\chi^2) = 4 \times 10^{24}$.

TABLE 18A
Subsample of eye color versus hair color (four-by-four)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	5.3	4.0	2.7	2.0	2.0	1.9	1.8	1.8	1.6	2.6	3
2	5.0	4.1	2.8	2.0	2.0	1.9	1.8	1.7	1.6	2.5	3
15	2.9	3.8	2.9	2.0	2.1	1.9	1.8	1.7	1.6	2.3	2

[4, 1, 1, 0; 2, 3, 0, 3; 1, 2, 2, 0; 0, 0, 0, 1]. $\chi^2 = 13.7$, $1/P(\chi^2) = 7.4$.

TABLE 19
Income and number of children (five-by-four)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	6(106)	4(106)	2(106)	7(106)	5(106)	3(106)	10(106)	6(106)	3(106)	5(106)	5
2	14(107)	6(107)	2(107)	16(107)	7(107)	2(107)	26(107)	8(107)	2(107)	9(107)	7
15	3(113)	7(112)	9(109)	3(112)	9(111)	5(109)	8(111)	4(111)	7(108)	5(112)	40,000

[Cramér (1946), page 444, and Diaconis and Efron (1985).] [2161, 3577, 2184, 1636; 2755, 5081, 2222, 1052; 936, 1753, 640, 306; 225, 419, 96, 38; 39, 98, 31, 14]. $\chi^2 = 568.6$, $1/P(\chi^2) = 1.5(136)$.

TABLE 20
Bortkiewicz's horse-kick data (14-*by*-20)

ν	(1,3)	(1,5)	(1,20)	(5,15)	(5,25)	(5,75)	(10,25)	(10,50)	(10,200)	Aver.	Max / Min
1	30	20	9	21	13	7	21	11	5	15	6
2	120	52	13	56	24	8	54	16	5	39	24
15	30400	556	18	209	31	7	59	10	4	3500	7600
Weib.	9(−46)	4(−16)	0.05	6(−18)	1(−6)	0.1	9(−17)	1(−4)	0.6	0.08	7(44)

[BFH, page 326, Winsor (1947) and Bortkiewicz (1878).] $\chi^2 = 240$, $1/P(\chi^2) = 1.6$.

fair to say that these average values of F_1 have Bayesian robustness except in extreme circumstances.

The last (eleventh) column gives the ratio of the largest to the smallest value of F_1 corresponding to each value of ν. This is a measure of the Bayesian sensitivity or robustness for that value of ν. Where the ratio is large there is low robustness. The greater sensitivity for $\nu = 15$ (or $\nu = \infty$) (see especially Tables 7 and 20) is because the log-normal hyperprior has an upper tail that is too thin and is therefore too committal (too "informative"). The upper tail of the log-Cauchy is about as thick as any tail can be for a proper distribution. For this hyperprior, the ratios of the largest to smallest values of F_1, when the quartiles are varied widely (as shown), are, for the 21 tables, nearly always smaller than for $\nu = 2$ or $\nu = 15$, so $\nu = 1$ gives results that are the most robust. We therefore recommend the use of $\nu = 1$, the log-Cauchy hyperprior. Furthermore, if the statistician and his client have difficulty in choosing q_L and q_U, we suggest that they should report the values of F_1 for about nine pairs (q_L, q_U), and the average, as we have done, though not necessarily for the pairs we have chosen.

In Table 20, the Weibull hyperprior gave such absurd results that we decided, for the other tables, not to report the results of this assumption although it *did* give reasonable results for most of the tables. The Weibull hyperprior has a very thin right tail, far thinner than any sensible person could judge; but, when the maximum of $F_1(\kappa)$ does not occur for a large value of κ, the Weibull's tail does not wag the dog.

For the artificial contingency Table 12, where $n_{ij} = 6$ for all i and j, so that $\chi^2 = 0$, most of the Bayes factors favor the nonnull hypothesis. The intuitive explanation is that the data suggest that the two-way categorization is irrelevant: all 9! permutations of the interior of the table are the same.

For Mendel's data on garden peas (Table 7), and for Bortkiewicz's data on horse-kicks (Table 20), χ^2 is again less than the number of degrees of freedom, whereas our model supports H'. Both cases can be better understood when we recall that χ^2 depends on Sampling Procedure 3. Unfortunately, the sample sizes of both tables preclude known methods for the accurate computation of the Bayes factor F_3. Our guess is that they would both "subceed" 1. We have the following further thoughts about Table 20.

The unbiased estimate of the "repeat rate" $\Sigma p_{i\cdot}^2$ is $\Sigma n_i(n_i - 1)N^{-1}(N - 1)^{-1}$, which is equal to $1.0733/14$ and that for $\Sigma p_{\cdot j}^2$ is $1.0947/20$. Hence $(n_{i\cdot})$ and $(n_{\cdot j})$ are both rather "flat" (although both differ significantly from complete flatness, having chi-squared values of 27.3 and 37.5 with 13 and 19 degrees of freedom, respectively). In accordance with Crook and Good (1980, page 1214), in these circumstances the row and column totals jointly "somewhat undermine" H. Although that conclusion was based on small values of r and s, it helps to explain why the Bayes factors against H are appreciable in Table 20. But one could have guessed *in advance*, in this example, that $(n_{i\cdot})$ and $(n_{\cdot j})$ would both be fairly flat so the marginal totals cannot contain much evidence for or against H; in other words F_3 would be more relevant than F_1. The ability to guess, from exogenous knowledge, that the margins would be flat was not built in to our model.

Leaving aside Mendel and Bortkiewicz, we conclude by emphasizing what we said earlier; that if F_1 is large we can reject H without necessarily accepting the particular form of H' assumed in the model. If H can be clearly rejected as compared with a noncomplicated and reasonable hypothesis H' that is "close" to it, then surely H can be rejected, "period," even if H' also comes under suspicion. Indeed, if there is a discrepancy (in ratio) between F_1 and $1/P(\chi^2)$ of as much as 1000, as in Table 10 where the ratio is 10,000, or in Table 19 which is a highly extreme case, then perhaps H' (as formulated with its "priors") is ruled out in addition to H. This comment assumes that the chi-squared approximation to $P(\chi^2)$ is adequate for such large values of χ^2. The Bayes factors do not depend on asymptotic theory.

Acknowledgments. We are grateful to the referees and an Associate Editor for their conscientious suggestions, especially regarding the removal of diversions. A longer version of the paper is available from the authors.

REFERENCES

BARNARD, G. A. (1984). The early history of the Fisher–Yates–Irwin formula. *J. Statist. Comput. Simulation* **20** 153–155.

BERGER, J. and SELLKE, T. (1986). Testing a point null hypothesis: the irreconcilability of *P*-values and evidence. *J. Amer. Statist. Assoc.* To appear.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis.* M.I.T. Press, Cambridge, Mass.

BORTKIEWICZ, L. VON (1878). *Das Gesetz der kleinen Zahlen.* Teubner, Leipzig.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

CROOK, J. F. and GOOD, I. J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. II. *Ann. Statist.* **8** 1198–1218.

CROOK, J. F. and GOOD, I. J. (1982). The powers and "strengths" of tests for multinomials and contingency tables. *J. Amer. Statist. Assoc.* **77** 793–802.

CROOK, J. F. and GOOD, I. J. (1985). The computation of a Bayes factor against independence in contingency tables. Unpublished.

DEMING, W. E. and STEPHAN, F. W. (1940). On a least squares adjustment of a sampled frequency table. *Ann. Math. Statist.* **11** 427–444.

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

FISHER, R. A. (1938). *Statistical Methods for Research Workers*, 7th ed. Oliver and Boyd, Edinburgh.

GOOD, I. J. (1957). Saddle-point methods for the multinomial distribution. *Ann. Math. Statist.* **28** 861–881.

GOOD, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press, Cambridge, Mass.

GOOD, I. J. (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. Ser. B* **29** 399–431.

GOOD, I. J. (1969). A subjective evaluation of Bode's law and an "objective" test for approximate numerical rationality (with discussion). *J. Amer. Statist. Assoc.* **64** 23–66.

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. (1978). Monozygotic criminals. *J. Statist. Comput. Simulation* **8** 161–162.

GOOD, I. J. (1980a). The contributions of Jeffreys to Bayesian statistics. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.) 21–34. North-Holland, Amsterdam.

GOOD, I. J. (1980b). Some history of the hierarchical Bayesian methodology (with discussion). In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 489–519. Univ. Press, Valencia. Also in *Trabajos Estadist. Investigacion Oper.* **31** 489–504 (1980).

GOOD, I. J. (1983). The robustness of a hierarchical model for multinomials and contingency tables. In *Scientific Inference, Data Analysis and Robustness* (G. E. P. Box, T. Leonard and C.-F. Wu, eds.) 191–211. Academic, New York.

GOOD, I. J. (1984a). The early history of the Fisher–Yates–Irwin formula and Fisher's "exact test". *J. Statist. Comput. Simulation* **19** 315–319.

GOOD, I. J. (1984b). A further note on the history of the Fisher–Yates–Irwin formula. *J. Statist. Comput. Simulation* **20** 155–159.

GOOD, I. J. and CROOK, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69** 711–720.

GOODHARDT, G. J., EHRENBERG, A. S. C. and CHATFIELD, C. (1984). The Dirichlet: a comprehensive model of buying behaviour (with discussion). *J. Roy. Statist. Soc. Ser. A* **147** 621–655.

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

GÛNEL, E. and DICKEY, J. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61** 545–557.

JOHNSON, N. J. and KOTZ, S. (1969). *Distributions in Statistics: Discrete Distributions*. Wiley, New York.

JOHNSON, W. E. (1932). Appendix to Probability: deductive and inductive problems (R. B. Braithwaite, ed.). *Mind* **41** 421–423.

LANGE, J. (1931). *Crime and Destiny*. Allen and Unwin, London. (Translated by C. Haldane.)

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate beta distribution and correlations among proportions. *Biometrika* **49** 65–82.

PELZ, W. (1977). Topics on the estimation of small probabilities. Ph.D. thesis, Dept. of Statistics, Virginia Polytechnic Institute and State Univ., Blacksburg, Va.

SNEE, R. (1974). Graphical display of two-way contingency tables. *Amer. Statist.* **28** 9–12.

WILKS, S. S. (1962). *Mathematical Statistics*. Wiley, New York.

WINSOR, C. P. (1947). Quotations: Das Gesetz der kleinen Zahlen. *Human Biology* **19** 154–161.

ZABELL, S. (1982). W. E. Johnson's "sufficientness postulate." *Ann. Statist.* **10** 1091–1099.

DEPARTMENT OF STATISTICS
VIRGINIA POLYTECHNIC INSTITUTE
BLACKSBURG, VIRGINIA 24061

DEPARTMENT OF STATISTICS
WINTHROP COLLEGE
ROCK HILL, SOUTH CAROLINA 29730