International Journal of Statistics and Probability; Vol. 6, No. 2; March 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Tests of Independence for a 2×2 Contingency Table with Random Margins

Yuan Yu[1], Dhiman Bhadra[2], & Balgobin Nandram[1]

[1] Department of Mathematical Sciences, WPI, Worcester, MA 01609, USA

[2] Production and Quantitative Methods Area, Indian Institute of Management Ahmedabad, Gujarat 380015, India

Correspondence: Dhiman Bhadra, Production and Quantitative Methods Area, Indian Institute of Management Ahmedabad, Gujarat 380015, India. E-mail: dhiman@iima.ac.in

Received: December 26, 2016 Accepted: January 24, 2017 Online Published: February 21, 2017

doi:10.5539/ijsp.v6n2p106 URL: https://doi.org/10.5539/ijsp.v6n2p106

The research is financed by (Sponsoring information)

Abstract

Fisher's exact test is commonly used for testing the hypothesis of independence between the row and column variables in a $r \times c$ contingency table. It is a "small-sample" test since it is used when the sample size is not large enough for the Pearsonian chi-square test to be valid. Fisher's exact test conditions on both margins of a 2×2 table leading to a hypergeometric distribution of the cell counts under independence. Moreover, it is conservative in the sense that its actual significance level falls short of the nominal level. In this paper, we modify Fisher's exact test by lifting the restriction of fixed margins and allow the margins to be random. In doing so, we propose two new tests - a likelihood ratio test in a frequentist framework and a Bayes factor test in a Bayesian framework, both of which are based on a new multinomial distributional framework. We apply the three tests on data from the Worcester Heart Attack study and compare their power functions in assessing gender difference in the therapeutic management of patients with acute myocardial infarction (AMI).

Keywords: Bayes factor, Fisher's exact test, hypergeometric, likelihood ratio, power function

1. Introduction

Analysis of contingency tables is an important area in the statistical analysis of data. The most common form of a contingency table is probably the 2×2 one although more general $r \times c$ dimensional tables are often used to cross-classify two variables. Analysis of contingency tables usually revolves around testing for independence between the row and column variables. There are various procedures (or tests) of achieving the same but the one that is commonly used is the celebrated Pearsonian chi-square test formulated by the late Karl Pearson in 1900. This test compares the observed frequencies (of each cell of the table) to those expected under independence through a test statistic which can be shown to follow a chi-square distribution under the null hypothesis of independence. However, this approximation is only valid for large samples (expected frequencies in each cell at least 5 in one formulation). However, in many instances, where data is sparse, the above sample size restriction may not be satisfied and in that case the Pearsonian test statistic cannot be approximated by a chi-square distribution. In those cases, Fisher's exact test is generally used.

Fisher's exact test of independence conditions on both margins of a 2×2 table (which are fixed) leading to a hypergeometric distribution of the cell counts under the null hypothesis of independence. The "exactness" of this test is attributable to the fact that the probabilities of the hypergeometric distribution can be calculated exactly under the null hypothesis of independence. Thus the significance of deviation from null hypothesis can be calculated exactly rather than relying on an approximation that becomes more precise (or "exact") only in the limit as the sample size tends to infinity which is the case for the "large sample" Pearsonian test. In fact, for small, sparse and unbalanced data, the exact and asymptotic p-values can be quite different and may even lead to contradictory conclusions. In contrast, Fisher's test is "exact" as long as the experimental procedure keeps the row and column totals fixed and hence it is applicable regardless of the sample characteristics. Having said that, Fisher's exact test, although being a "small sample" test, is valid for all sample sizes.

Despite the fact that Fisher's exact test provides exact p-values, one feature that has always come under scrutiny is its "conservativeness". This is due to the discreteness of the hypergeometric distribution due to which its actual rejection rate (probability of type I error) is usually less than the nominal significance level. The Fisherian basis of this test, namely that the marginal totals are ancillary statistics and hence provide no information regarding the table 1 has been shown to be incorrect by Berkson (1978) who compared the exact test with the normal (Z) test at nominal significance levels $\alpha = 0.01, 0.05$ and showed that the effective level (i.e. probability of type I error) is closer to the nominal level (α) for the latter test. The normal test was also found to have considerably larger power than Fisher's test. In view of these findings, it was proposed that Fisher's exact test may not be used in preference to the normal test (Berkson, 1978).

In this paper, we have put forward two distinct tests of independence as alternatives to Fisher's exact test in the context of a 2×2 contingency table. These are a likelihood ratio test formulated in a frequentist framework and a test based on Bayes factor formulated in a Bayesian framework. In doing so, we have relaxed the fixed-margin assumption of the Fisher's test and have allowed the margins of the 2×2 table to be random. Instead of the hypergeometric distribution which results from the fixed-margin restriction, the proposed tests are based on a multinomial distribution which adds to their flexibility. We apply our proposed tests on data obtained from the Worcester Heart Attack study and compare their power functions (against the Fisher's test) at different significance levels. An extension of the tests to a more general $r \times c$ contingency table is also shown.

The rest of the paper is organized as follows. In Section 2, we elaborate on the characteristics of the Fisher's exact test, specifically as it relates to its conservativeness. In Section 3, we explain the proposed likelihood ratio and the test based on Bayes factor in detail including the multinomial distribution on which these tests are based. Section 4 provides a brief overview of the Worcester Heart Attack study and a comparative analysis of the performance of the above tests in the context of the study. We end with a discussion in Section 5.

2. Conservativeness of Fisher's Test

Tests of independence of the row and column variables in a 2×2 contingency table is synonymous to testing for equality of two independent binomial proportions. The three well known tests for this purpose are respectively (a) Pearson chi-squared test for large samples, (b) Chi-square test with Yates' continuity correction for intermediate sized samples (Yates, 1984) and (c) the Fisher's exact test for small samples. A considerable body of research have done a comparative analysis of these tests with respect to various characteristics and have concluded that Fisher's exact test and the Yates correction to Pearson's test are both extremely conservative (Berkson, 1978; Conover, 1974; Liddell, 1976; Grizzle, 1967; Kempthorne, 1979; Upton, 1982). An extensive study comparing the above two tests with 22 other alternative tests of independence corroborated the same (D'Agostino, 1988). It was also noted in the same study that the uncorrected chi-square test performs relatively well in comparison and that the scaled Pearson chi-square test $(N - 1)X^2/N$ (where N is the combined sample size and X^2 is the Pearson chi-square test statistic) can be used as a preferred alternative.

While the "exactness" of Fisher's test may seem to outweigh its shortcomings Upton (Upton, 1982) was of the opinion that its conservativeness and inferior power characteristic are of greater concern than its so-called "exactness". This was confirmed by D'Agostino *et al.* (D'Agostino, *et al.*, 1988) who further demonstrated that the uncorrected chi-squared test and the two-sample t test with pooled variance are robust since their actual levels of significance are, in most situations, close to or smaller than the nominal levels which is not the case for Fisher's test.

Little (Little, 1989) analysed two 2×2 tables of widely varying sizes, the smaller one being $T_1 = \{a, b, c, d\} = \{3, 0, 0, 3\}$ (provided by Barnard, 1945), and the larger being $T_2 = \{a, b, c, d\} = \{170, 2, 162, 9\}$. In both examples, the idea was to compare two treatments (each yielding a binomial sample) with respect to their success probabilities, say π_1 and π_2. Thus, the null hypothesis was that the treatments are equally effective ($H_0 : \pi_1 = \pi_2$) while the alternative can be either one ($H_a : \pi_1 > \pi_2$) or two ($H_a : \pi_1 \neq \pi_2$) sided. Table 1. depicts the above setup.

Table 1. General notation of a 2×2 table

	Outcome		
Treatment	Success	Failure	Total
1	a	b	n_1
2	c	d	n_2
Total	m_1	m_2	$n = n_1 + n_2$

On applying the three tests, namely Pearsonian (P), Yates correction (Y) and Fisher's (F) on each of the above tables, it was observed that the p-values of the last two tests are close and considerably larger than that for the Pearsonian test for both the set-ups. Table 2. has the details (p-values are in percentages).

Table 2. P-values (in %) corresponding to one and two sided alternatives for Pearson, Yates and Fisher's tests

	T_1		T_2	
Test	One-sided	Two-sided	One-sided	Two-sided
Pearson (P)	0.7	1.4	1.6	3.2
Yates (Y)	5.1	10.2	3.2	6.4
Fisher (F)	5.0	10.0	3.0	3.5

It is clear that for one-sided tests, p-values of Y and F are quite close. Moreover, the p-values for Y and F are considerably larger than that of P, even for the larger sized table T_2. Overall, the results illustrate that even in quite respectable-sized samples, choice of the three methods matter. Although conventional wisdom favors the use of Y in moderate samples and F in small samples, this position has been challenged . Building on earlier work (Berkson, 1978; Grizzle, 1967; Upton, 1982) it has been claimed that Y and F are much too conservative in small samples and hence the usual pooled t-test for independent normal samples is preferable. The t-test is essentially a studentized version of P, and hence will perform like P unless the combined sample size is less than, say, 30.

Table 3. depicts the empirical rejection rates for Pearsonian and Fisher's tests corresponding to the one-sided alternative ($H_a : \pi_1 > \pi_2$) at the nominal 5% level with null values of $\pi = .2$, .5, and .8, and samples sizes ranging from 5 to 200 in each group.

Table 3. Rejection probabilities under H_0 (in %) of Pearson and Fisher's tests at $\alpha = 0.05$ for a one-sided alternative

	$\pi = 0.5$		$\pi = 0.2, 0.8$	
$n_1 = n_2$	Pearson	Fisher	Pearson	Fisher
5	5.5	1.1	2.2	.2
10	5.8	2.1	4.6	1.5
20		2.1	5.1	2.3
50	4.5	2.9	5.0	2.9
100	5.2	3.8	4.9	3.5
200	4.9	4.0	-	-
∞	5	5	5	5

Based on the results above, it is clear that Fisher's test is conservative since its empirical rejection rates are much smaller than the nominal level (5%). Chi-square test with Yates correction (not shown in the table) is similarly conservative. On the other hand, the empirical rejection rate of the Pearsonian test is close to the nominal level, even for small samples (except for $n_1 = n_2 = 5$). Previous research has shown that rejection rates for the studentized version of P are close to the nominal levels (D'Agostino, 1988). Overall, the above findings are impressive evidences against the use of Fisher's exact test or the chi-square test with Yates' correction. There are two reasons why Fisher's test is so conservative. Firstly, the test statistic has a discrete null distribution, yielding discontinuities at the rejection rates. Secondly, the test conditions on both fixed margins. While the former reasoning applies to any discrete statistic, the fixed-margin restriction compounds the conservativeness issue (Mehta *et al.*, 2003). It is to be noted, however, that the computations in Table 3. are based on conditioning on one margin as dictated by the sampling design.

3. Proposed Methodology

In this section we will lay out two tests allowing the margins of a contingency table to be random, rather than fixed as is the case for Fisher's exact test. In doing so, we will compare these tests against Fisher's in terms of power function and conservativeness. However, we will first develop a novel framework based on the multinomial distribution, on which these tests will be based.

3.1 Multinomial Distribution

We start with a motivating example taken from the Worcester Heart Attack study which deals with analyzing gender differences in receiving Lidocaine therapy in Acute Myocardial Infarction (AMI) patients older than 75 years who have a history of hypertension and stroke. There were 81 patients in total (30 males and 51 females). Moreover, the design does not accommodate fixed margins. Table 4. has the details.

Table 4. Lidocaine therapy data

	Therapy		
Gender	Receive	Not receive	Total
Males	8	22	30
Females	16	35	51
Total	24	57	81

Clearly, the hypothesis of independence (between gender and receipt of lidocaine therapy) is synonymous to that of equality of the success probabilities of two binomial distributions corresponding to males and females respectively (receiving therapy being success and vice versa) (Hashemi *et al.*, 1997; Nandram *et al.*, 2005).

Let I and J be random variables representing the gender of a patient (1 for male and 0 for female) and whether a patient has received the therapy (1 received and 0 did not receive) respectively. Let $P(I = 1) = p$ and $P(J = 1) = q$. Clearly, the duplet (I, J) can determine all the possible combinations of gender and therapy. Now assuming X and Y to be the number of patients who are males and the number of patients receiving Lidocaine therapy respectively, we have

$$X = \sum_{k=1}^{N} I_k,\ Y = \sum_{k=1}^{N} J_k,\ Z = \sum_{k=1}^{N} I_k J_k,$$

where N is the total number of patients. Clearly, X and Y follow binomial distributions with success probabilities p and q respectively. Also, defining π_{ij} as $P(I = i, J = j); (i, j = 0, 1)$, we have the following table,

		J		
		1	0	
I	1	π_{11}	π_{10}	p
	0	π_{01}	π_{00}	$1-p$
		q	$1-q$	1

where $p = \pi_{11} + \pi_{10}$ and $q = \pi_{11} + \pi_{01}$. For the counts, we have the following 2×2 contingency table

	Receive	Not receive	
Males	Z	X-Z	X
Females	Y-Z	N-X-Y+Z	N-X
	Y	N-Y	N

with probability mass function as follows

$$P(z, x-z, y-z, N-x-y+z) = \frac{n!}{z!(x-z)!(y-z)!(n-x-y+z)!} \cdot {\pi_{11}}^{z} {\pi_{10}}^{x-z} {\pi_{01}}^{y-z} {\pi_{00}}^{n-x-y+z}$$

where $0 \leq x \leq n$, $0 \leq y \leq n$, $\max(0, x+y-n) \leq z \leq \min(x, y)$. Let γ denote the correlation between I and J. It is straightforward to show that

$$\gamma = \frac{\pi_{11} - pq}{\sqrt{p(1-p)q(1-q)}}. \tag{1}$$

The following simultaneous equations represent the relationships between $\pi_{ij}(i, j = 0, 1)$ and (p, q) defined in the table above

$$\begin{aligned} \pi_{11} + \pi_{10} &= p, \\ \pi_{11} + \pi_{01} &= q, \\ \pi_{11} + \pi_{10} + \pi_{01} + \pi_{00} &= 1. \end{aligned} \tag{2}$$

Using (2) and (3), we have the following expressions

$$\begin{aligned}
\pi_{11} &= pq + \gamma\sqrt{p(1-p)q(1-q)},\\
\pi_{10} &= p(1-q) - \gamma\sqrt{p(1-p)q(1-q)},\\
\pi_{01} &= q(1-p) - \gamma\sqrt{p(1-p)q(1-q)},\\
\pi_{00} &= (1-p)(1-q) + \gamma\sqrt{p(1-p)q(1-q)}. \qquad (3)
\end{aligned}$$

Furthermore, the marginal distribution of X can be shown to be

$$\begin{aligned}
P(X=x) &= \sum_{z=0}^{x}\sum_{y=z}^{n-x+z} P(X,Y,Z)\\
&= \sum_{z=0}^{x}\sum_{y=z}^{n-x+z} \frac{n!{\pi_{11}}^{z}(p-\pi_{11})^{x-z}(q-\pi_{11})^{y-z}(1-p-q+\pi_{11})^{n-x-y+z}}{z!(x-z)!(y-z)!(n-x-y+z)!}\\
&= \sum_{z=0}^{x}\frac{n!{\pi_{11}}^{z}(p-\pi_{11})^{x-z}}{z!(x-z)!}\sum_{y=z}^{n-x+z}\frac{(p-\pi_{11})^{x-z}(q-\pi_{11})^{y-z}(1-p-q+\pi_{11})^{n-x-y+z}}{(y-z)!(n-x-y+z)!}\\
&= \frac{n!p^{x}(1-p)^{n-x}}{x!(n-x)!}\sum_{z=0}^{x}\frac{x!(\frac{\pi_{11}}{p})^{z}(1-\frac{\pi_{11}}{p})^{x-z}}{z!(x-z)!}\sum_{y-z=0}^{n-x}\frac{(n-x)!(\frac{q-\pi_{11}}{1-p})^{y-z}(1-\frac{q-\pi_{11}}{1-p})^{n-x-y+z}}{(y-z)!(n-x-y+z)!}\\
&= \binom{n}{x}p^{x}(1-p)^{n-x}. \qquad (4)
\end{aligned}$$

So that $X \sim \text{Binomial}(n,p)$. Similarly, the marginal distribution of Y can be shown to be Binomial(n,q). It follows immediately that $E(X)=np$, $E(Y)=nq$, $V(X)=np(1-p)$ and $V(Y)=nq(1-q)$.

3.1.1 Maximum Likelihood Estimation

In order to formulate the likelihood ratio test, we need to obtain the maximum likelihood estimates of p, q and γ. Towards that end, let us first consider the following 2×2 table

	Received	Not Received	
Male	n_{11}	n_{10}	$n_{1.}$
Female	n_{01}	n_{00}	$n_{0.}$
	$n_{.1}$	$n_{.0}$	n

where $n_{1.} = n_{11} + n_{10}$ and $n_{.1} = n_{11} + n_{01}$. In order to proceed with a statistical optimization, we can set the initial value of the parameters as

$$\hat{p} = \frac{n_{11}+n_{10}}{n} \quad \hat{q} = \frac{n_{11}+n_{01}}{n} \quad \hat{\gamma} = \frac{n_{11}/n - \hat{p}\hat{q}}{\sqrt{\hat{p}(1-\hat{p})\hat{q}(1-\hat{q})}}. \qquad (5)$$

From (4), we have

$$\begin{aligned}
pq + \gamma\sqrt{p(1-p)q(1-q)} &\geq 0,\\
p(1-q) - \gamma\sqrt{p(1-p)q(1-q)} &\geq 0,\\
q(1-p) - \gamma\sqrt{p(1-p)q(1-q)} &\geq 0,\\
(1-p)(1-q) + \gamma\sqrt{p(1-p)q(1-q)} &\geq 0. \qquad (6)
\end{aligned}$$

So, given the value of any two parameters, we can get an interval for the third parameter. We can then calculate the respective estimates from these intervals. Accordingly, the intervals of p corresponding to the range of γ are

$$\begin{aligned}
\text{if } \gamma &> 0, & \frac{\gamma^2 q}{1-q+\gamma^2 q} \leq p \leq \frac{q}{\gamma^2(1-q)+q}\\
\text{if } \gamma &\leq 0, & \frac{\gamma^2(1-q)}{q+\gamma^2(1-q)} \leq p \leq \frac{1-q}{\gamma^2 q+1-q}. \qquad (7)
\end{aligned}$$

Similarly, the intervals for q can be shown to be

$$\begin{aligned}\text{if } \gamma &> 0, \qquad \frac{\gamma^2 p}{1-p+\gamma^2 p} \leq q \leq \frac{p}{\gamma^2(1-p)+p} \\ \text{if } \gamma &\leq 0, \qquad \frac{\gamma^2(1-p)}{p+\gamma^2(1-p)} \leq q \leq \frac{1-p}{\gamma^2 p+1-p}.\end{aligned} \tag{8}$$

Finally, the interval for γ will be

$$\max\left\{\frac{-pq}{\sqrt{p(1-p)q(1-q)}}, \frac{-(1-p)(1-q)}{\sqrt{p(1-p)q(1-q)}}\right\} \leq \gamma \leq \min\left\{\frac{p(1-q)}{\sqrt{p(1-p)q(1-q)}}, \frac{q(1-p)}{\sqrt{p(1-p)q(1-q)}}\right\},$$

which can be simplified to

$$\max\left\{-\sqrt{\frac{pq}{(1-p)(1-q)}}, -\sqrt{\frac{(1-p)(1-q)}{pq}}\right\} \leq \gamma \leq \min\left\{\sqrt{\frac{p(1-q)}{(1-p)q}}, \sqrt{\frac{q(1-p)}{(1-q)p}}\right\}.$$

It is clear from the above simplification that the two terms in the left and right limits of (10) have a reciprocal relationship (i.e., the maximum is greater than -1 and minimum is less than 1). So, on choosing the initial values of p, q and γ, we can get a new set of parameter values by successive random selection from within their ranges (i.e., stochastic optimization). For example, we generate 1100 sets of parameter values and plug each set into the likelihood function $f(p, q, \gamma)$ given by

$$\begin{aligned}f(p,q,\gamma) = C \cdot \left(pq + \gamma\sqrt{p(1-p)q(1-q)}\right)^{n_{11}} \left(q - pq - \gamma\sqrt{p(1-p)q(1-q)}\right)^{n_{10}} \cdot \\ \left(p - pq - \gamma\sqrt{p(1-p)q(1-q)}\right)^{n_{01}} \left(1 - p - q + pq + \gamma\sqrt{p(1-p)q(1-q)}\right)^{n_{00}}.\end{aligned} \tag{9}$$

where $C = \frac{n!}{n_{11}!n_{10}!n_{01}!n_{00}!}$ is the coefficient of the mass function of p, q.

With the above restrictions, we can apply a grid method to perform a statistical optimization of the parameters in the likelihood function above by drawing a large set of values of the triplet (p, q, γ). (In our application, we have drawn 1100 values with a burn-in of 100 values). Considering the fact that these samples may not be stable, we drop the first set of values and then plug in the remaining sets to the likelihood function to obtain the values for which the function is maximized. For our example, we generated 1100 values and dropped the first 100 to obtain the likelihood maximizer.

3.2 Likelihood Ratio Test

The first test that we develop as an alternative to the Fisher's test (in terms of power and conservative properties) is the likelihood ratio test. Below, we first go over the power and p-value calculations for this test followed by that of Fisher's under the multinomial distributional framework described above.

Let us consider the following hypotheses,

$$H_0 : \gamma = 0 \qquad \text{versus} \qquad H_a : \gamma > 0.$$

For testing the above hypotheses, we can define the likelihood ratio test statistic as

$$T = \frac{L(x, y, z|H_0)}{L(x, y, z|\Omega)} = \frac{L(x, y, z|\hat{p}_0, \hat{q}_0, \gamma = 0)}{L(x, y, z|\hat{p}, \hat{q}, \hat{\gamma})},$$

where $L(x, y, z|H_0)$ and $L(x, y, z|\Omega)$ represent the likelihood function under the null hypothesis and the whole parameter space respectively. Here $(\hat{p}_0, \hat{q}_0)$ and $(\hat{p}, \hat{q}, \hat{\gamma})$ are the maximum likelihood estimates of the parameters under the null and alternative hypothesis respectively. The maximum likelihood estimates are calculated as mentioned in Section 3.1 (i.e., stochastic optimization). For calculating the p-value of T, we proceed as follows:

1. Under the null hypothesis, draw 1000 samples of the quadruplet $(n_{11}, n_{10}, n_{01}, n_{00})$ from a multinomial distribution with parameter $\hat{p} = \dfrac{n_{11} + n_{10}}{n}, \hat{q} = \dfrac{n_{11} + n_{01}}{n}, \gamma = 0$.
2. For the h^{th} sample drawn above i.e $(n_{11}^{(h)}, n_{10}^{(h)}, n_{01}^{(h)}, n_{00}^{(h)})$, calculate the parameter values under H_0 i.e $(p_0^{(h)}, q_0^{(h)}, \gamma^{(h)} = 0)$ and H_a i.e $(p^{(h)}, q^{(h)}, \gamma^{(h)})$.

3. Plug in each of the above drawn sample parameter estimates, as well as the original table values into the likelihood ratio test statistic and count the number of samples for which the test statistic value is smaller than that corresponding to the original table values. Thus, the p-value will be

$$P(T < T_{obs}|H_0) = \#(T < T_{obs})/1000$$

where T_{obs} is the test statistic value obtained by plugging in the original table values.

3.2.1 Power Calculation

One of the most critical element to look at when comparing two or more tests is the power function. Towards that end, we will calculate the power functions of the proposed likelihood ratio test (LRT) and Fisher's exact test under various significance levels, specifically $\alpha = 0.01, 0.05$. In order to evaluate the power of the LRT, we need to calculate the probability of rejection of the null hypotheses under the alternative parameter space (i.e $\gamma > 0$) but conditional on the restriction imposed by p and q in (10). In doing so, we divide the range into thin slices of width .01, the i^{th} slice being $\gamma_i, (i = 1, ..., s)$ where s is the number of slices and $\gamma_1 = 0$. For each γ_i, we evaluate the power in the same way as we calculated the p-value in Section 3.2. However, instead of drawing samples from the null hypothesis, we draw a fixed number (say 1000) samples corresponding to each $\gamma_i > 0$ and then calculate the LR statistics for each of the drawn samples. Thus, the p-value can be expressed as

$$p = \#(T_i < T_\alpha)/1000,$$

where T_α is the α^{th} quantile of the 1000 likelihood ratio test statistic values obtained from the samples generated above.

For calculating the power function of Fisher's exact test, we follow the same procedure outlined above (i.e., we compute the power at each γ_i). For that purpose, we need to use the non-central hypergeometric distribution. Let us consider the following table

	Receive	Not receive	
Males	x	$m_1 - x$	m_1
Females	$n - x$	$m_2 - n + x$	m_2
	n	$N - n$	N

where $N = m_1 + m_2$. Thus, the probability mass function of x will be

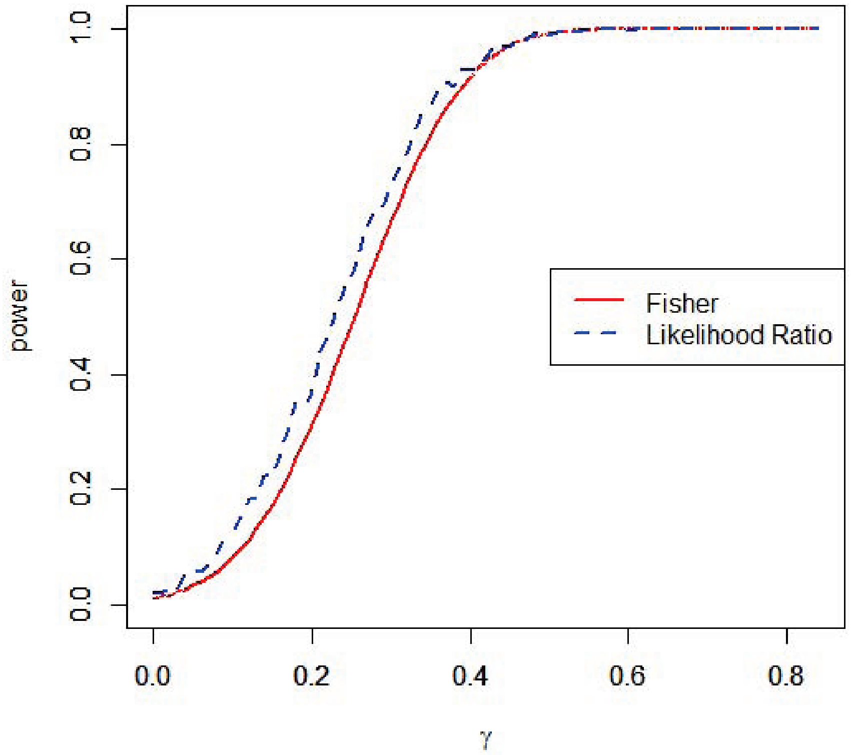
$$f(x|n, m_1, N, w) = \frac{\binom{m_1}{x}\binom{m_2}{n-x} w^x}{\sum_{y=x_{min}}^{x_{max}} \binom{m_1}{y}\binom{m_2}{n-y} w^y}, \tag{10}$$

where $x \in [x_{min}, x_{max}]$ represent the number of the male patients who received the therapy, and $x_{min} = \max(0, n - m_2), x_{max} = \min(n, m_1)$ while w is the odds ratio of receiving therapy for males versus females and is given by

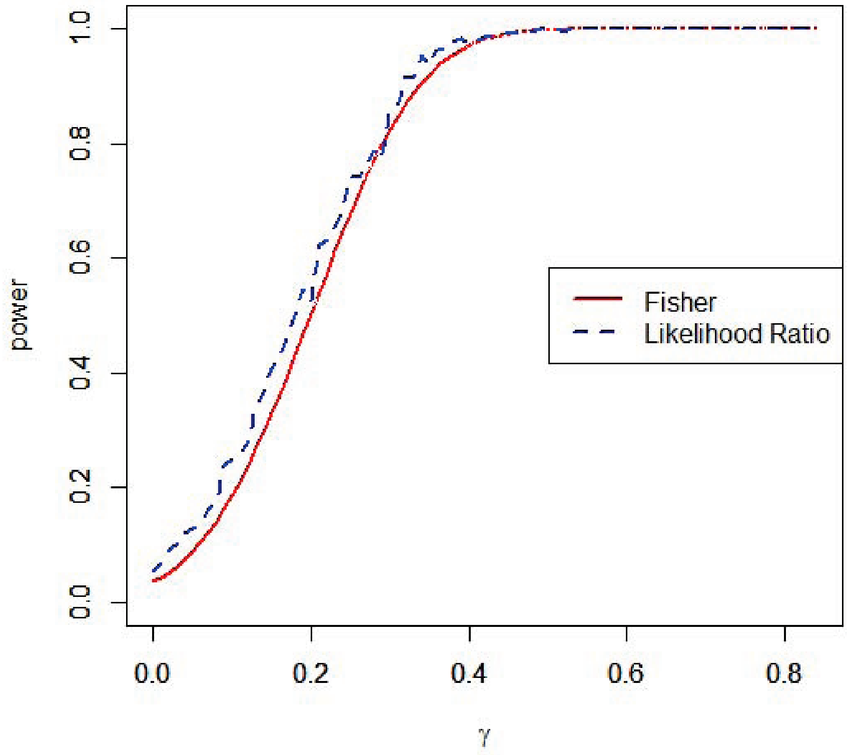
$$w = \frac{\pi_{11}}{\pi_{11} + \pi_{10}} \frac{\pi_{01} + \pi_{00}}{\pi_{01}} = \frac{pq + \gamma\sqrt{p(1-p)q(1-q)}}{q(1-p) - \gamma\sqrt{p(1-p)q(1-q)}}.\frac{1-p}{p}.$$

The above odds ratio can be expressed as a function of γ since $p = m_1/N$ and $q = n/N$, each of which can easily be obtained from the table above. Thus, we can map a specific γ to a specific non-central hypergeometric distribution via the odds ratio w. To obtain the power, it is imperative to find c_α, the upper $(1 - \alpha)\%$ critical value of X. However, this is not possible due to the discrete nature of the distribution of X as we cannot strictly satisfy the equality $Pr(x \geq c_\alpha|H_0) = \alpha$. Hence we choose c_α such that $c_\alpha = \{x : Pr(x \geq c_\alpha|H_0) - \alpha$ is minimum$\}$. For this value of c_α, the power function is given by

$$\begin{aligned} \text{power}(\gamma) &= P(x \geq c_\alpha|H_a) = Pr(x \geq c_\alpha|\gamma \neq 0) \\ &= \sum_{x=c_\alpha}^{n} f(x|n, m_1, N, w(\gamma)). \end{aligned} \tag{11}$$



(a) $\alpha = .01$



(b) $\alpha = .05$

Figure 1. Power functions of likelihood ratio and Fisher's exact tests at $\alpha = .01$ and $\alpha = .05$

Using the data in Table 5 we have calculated the power functions for all age groups. Figures 1(a) and 1(b) depict the power functions of the proposed LRT and Fisher's exact test at $\alpha = 0.01$ and $\alpha = 0.05$ respectively for patients older than 75 years in the Lidocaine therapy group. It is clear that, in both cases the power curve of the likelihood ratio test is consistently above that of the Fisher's exact test. Thus we are less likely to make a type II error with the LRT than that with Fisher's exact test. Additionally, the fact that at $\gamma = 0$, the power function degenerates to the test size validates our methodology of computing the power functions.

3.3 Bayes Factor

A Bayesian alternative to the likelihood ratio test for testing the hypothesis of independence is the Bayes factor which is the ratio of the marginal posterior densities under the null ($H_0 : \gamma = 0$) and alternative ($H_a : \gamma \neq 0$) hypotheses, denoted by M_0 and M_1 respectively. Thus the Bayes factor can be viewed as a discriminator between the null and alternative models, M_0 and M_1. Since the essence of our problem is that of testing two hypotheses, Bayes factor is a natural choice and can be calculated through a sampling route. The Bayes factor has been used in various contexts of our own work (e.g., Nandram, Bhatta & Bhadra 2013, Nandram & Katzoff 2012, Nandram & Choi 2007, Nandram, Cox & Choi 2005, Nandram & Choi 2006, Nandram, Liu, Choi & Cox 2005, Hashemi, Nandram & Goldberg 1997) for various situations of Bayesian testing for two-way contingency tables.

Under the null hypothesis, the probability density function (pdf) of (X, Y, Z) is given by

$$p(x, y, z|p, q, M_0) = \frac{n!}{z!(x-z)!(y-z)!(n-x-y+z)!} p^x(1-p)^{n-x} q^y (1-q)^{n-y}$$

while, under the alternative, the pdf will be

$$p(x, y, z|p, q, \gamma, M_1) = \frac{n!}{z!(x-z)!(y-z)!(n-x-y+z)!} {\pi_{11}}^{z} {\pi_{10}}^{x-z} {\pi_{01}}^{y-z} {\pi_{00}}^{n-x-y+z},$$

where $0 \leq x \leq n$, $0 \leq y \leq n$ and $\max(0, x + y - n) \leq z \leq \min(x, y)$. As shown in Section 3.1, the following identities still hold

$$\begin{aligned}
\pi_{11} &= pq + \gamma \sqrt{p(1-p)q(1-q)}, \\
\pi_{10} &= p(1-q) - \gamma \sqrt{p(1-p)q(1-q)}, \\
\pi_{01} &= q(1-p) - \gamma \sqrt{p(1-p)q(1-q)}, \\
\pi_{00} &= (1-p)(1-q) + \gamma \sqrt{p(1-p)q(1-q)}.
\end{aligned}$$

Since p, q and γ range from 0 to 1, a reasonable assumption for the prior distributions of the parameters are $\pi(p) = 1, \pi(q) = 1$ and $\pi(\gamma) = 1$. Thus the marginal posterior under the null hypothesis is

$$\begin{aligned}
f(x, y, z|M_0) &= \int_0^1 \int_0^1 p(x, y, z|p, q)\pi(p)\pi(q)\, dp\, dq \\
&= \int_0^1 \int_0^1 C \cdot p^x(1-p)^{n-x} q^y(1-q)^{n-y} \cdot 1 \cdot 1\, dp\, dq \\
&= C \cdot B(x+1, n-x+1)B(y+1, n-y+1), \qquad (12)
\end{aligned}$$

where $B()$ represents the beta function, and $C = \frac{n!}{z!(x-z)!(y-z)!(n-x-y+z)!}$ is the normalization constant. Similarly, the marginal posterior under the alternative hypothesis is

$$f(x, y, z|M_1) = \int_{a_0(p,q)}^{a_1(p,q)} \int_0^1 \int_0^1 p(x, y, z|p, q, \gamma)\pi(p)\pi(q)\pi(\gamma)\, dp\, dq\, d\gamma, \qquad (13)$$

where $a_0(p, q) = 0$ and $a_1(p, q) = \min\left(\sqrt{\frac{(p(1-q)}{q(1-p)}}, 1/\sqrt{\frac{(p(1-q)}{q(1-p)}}\right)$ is a function of (p, q) indicates the upper bound of γ given p and q. So the Bayes factor is the ratio of (14) and (15). That is,

$$T_{BF} = \frac{f(x, y, z|M_1)}{f(x, y, z|M_0)}.$$

The Bayes factor can be calculated using Monte Carlo integration of (15) (since (14) is in closed form). The computation is explained in detail in the Appendix. Following are some thresholds for Bayes factors provided by Kass & Raftery, 1995:

$2\ln(B_{10})$	(B_{10})	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

where B_{10} is the Bayes factor of M_1 with respect to M_0.

We can also obtain the power function of Bayes factor test as was done for the likelihood ratio test statistics in Section 3.2. To obtain the critical value of the Bayes test, we drew 1000 samples under M_0, calculated the Bayes factor and obtained the $100(1-\alpha)$ quantile (e.g., $\alpha = .01, .025, .05$) for each of these 1000 samples. Then, to obtain the power function for each γ we drew 1000 samples under M_1 and calculated the proportion of Bayes factors larger than the critical value.

4. Data Description and Results

We have done a comparative analysis of the proposed likelihood ratio and Bayes factor tests with Fisher's test (with respect to their size and power characteristics) in the context of the Worcester Heart Attack study. In this section, we first provide a brief description of the heart attack data followed by the main results of our analysis.

4.1 Worcester Heart Attack Study

The Worcester Heart Attack Study is a population-based study examining the trends in incidence rates, death rates (hospital and post discharge) and occurrence of major clinical complications of patients hospitalized with acute myocardial infarction (AMI) at all metropolitan Worcester (Massachusetts) hospitals during the period 1975-2003. This study also examined the time trajectory of duration of pre-hospital delay following the onset of severe coronary symptoms and mortality related to coronary heart disease in the greater Worcester population. This study was funded by the National Heart, Lung and Blood Institute since the mid-1980's. A total of 12,760 patients (all residents of greater Worcester Metropolitan area) were incorporated in this study. Patients discharged from hospitals were also followed up through a variety of sources to ascertain their long-term survival status and possible changes in post discharge survival over time.

With regard to the principal study findings, an initial increase, followed by decline, and then relative stabilization in the incidence rates of initial AMI was observed. Furthermore, patients hospitalized with AMI during more recent study years were found to be increasingly older with a greater prevalence of comorbidities. Finally, a marked increase in the use of various medical treatment approaches and coronary reperfusion strategies was observed over time.

We have applied the three tests - Fisher's exact test, Likelihood ratio test and Bayes factor test to the Worcester Heart Attack data described above with a view to compare their relative conservativeness (as measured by the respective p-values) and power characteristics. In doing so, we consider data corresponding to gender difference in the receipt of Lidocaine, Anti-platelet and Beta-blocker therapies for AMI patients with a history of hypertension and stroke. In each case, the frequencies are classified according to age (in years) of the patients involved. The data is summarized in Table 5. ("M" denote Male, "F" denote Female, "R" denote received therapy while "NR" denote not received therapy).

Table 5. Gender difference in the receipt of Lidocaine, Anti-platelet and Beta-blocker therapies for AMI patients with hypertension & stroke (*M: male; F: female; R: receive; NR: not receive*)

	Lidocaine				Anti-platelet				Beta-blocker			
	M		F		M		F		M		F	
Age(years)	R	NR	R	NR	R	NR	R	NR	R	NR	R	NR
< 55	4	4	0	5	3	5	2	3	0	1	3	2
55-59	2	3	1	2	3	2	0	3	0	6	2	1
60-64	6	5	2	4	5	6	1	5	2	4	1	5
65-69	4	3	6	6	3	4	8	4	4	3	2	7
70-74	8	11	3	12	14	5	5	10	7	12	6	6
≥ 75	8	22	16	35	16	14	22	29	7	21	16	66

4.2 Data Analysis

Corresponding to each of the therapies, the p-values for Fisher's exact test and likelihood ratio test as also the Bayes factors are shown in Table 6.

Table 6. P-values for Likelihood ratio and Fisher's tests and Bayes factors (BF) corresponding to Lidocaine, Anti-platelet and Beta-blocker therapies

	Lidocaine			Anti-platelet			Beta-blocker		
Age	LRT	Fisher	BF	LRT	Fisher	BF	LRT	Fisher	BF
<55	0.01	0.098	2.65	0.803	0.751	2.666	0.910	1	0.202
55-59	0.456	0.71	0.376	0.036	0.179	1.89	0.798	1	0.102
60-64	0.214	0.373	0.548	0.12	0.261	0.903	0.283	0.5	0.474
65-69	0.355	0.57	0.315	0.979	0.933	0.127	0.094	0.182	1.197
70-74	0.104	0.159	0.796	0.012	0.022	5.554	0.902	0.863	0.132
≥ 75	0.54	0.756	0.092	0.178	0.255	0.304	0.288	0.356	0.226

Overall, we observe that the p-value of the Fisher's test tend to be higher than that of the likelihood ratio test although with some exceptions (e.g., patients younger than 55 years for Lidocaine and Anti-platelet therapy and patients with age between 60-64 years for Beta-blocker therapy etc). This corroborates the fact that Fisher's test is relatively more conservative than the likelihood ratio test. For 15 of the 18 (age) groups considered across the three types of therapies (last five for Lidocaine, 1st, 3rd, 4th, 6th for Anti-platelet and all six for Beta-blocker), we obtain a consistent conclusion of failure to reject the hypotheses of significant gender difference in receiving therapy. That is, no evidence that males are more likely to receive the therapy than females since the p-values (for both the likelihood ratio and Fisher's test) are greater than the nominal level of 0.05 while the Bayes factor is less than 1. On the other hand, for the first group in Lidocaine and the second group in Anti-platelet, both likelihood ratio test and Bayes factor reject the hypotheses of independence while Fisher's exact test fails to do so. For the 5th group in Beta-blocker, all three tests tend to reject while for the 4th group in Beta-blocker, only Bayes factor rejects.

Table 7. Power values of the three tests for varying γ at $\alpha = .01$ and 0.05

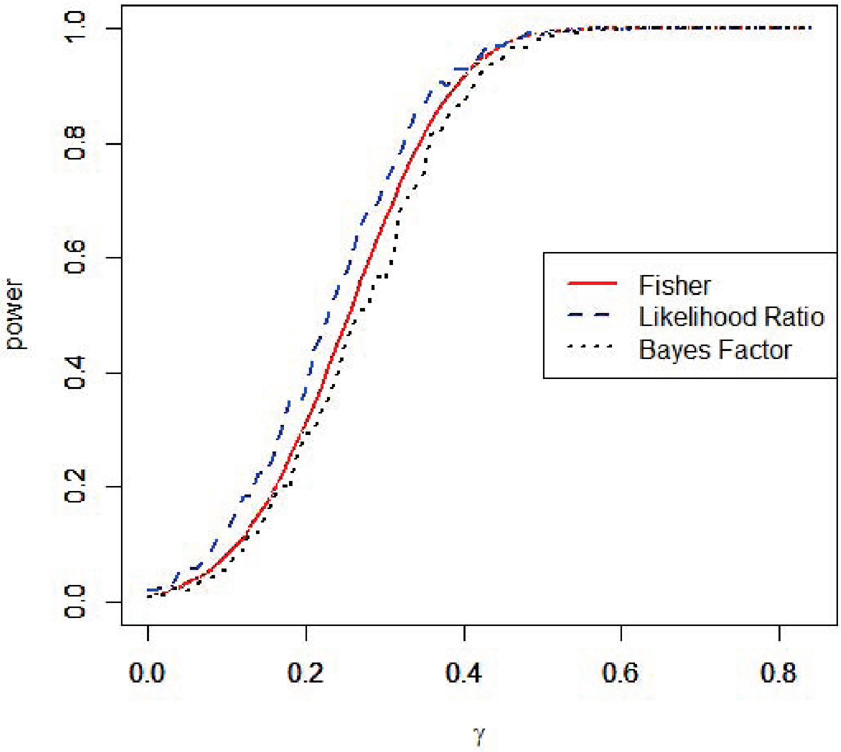
γ	LRT	Fisher's test	BF
0.00	0.022	0.011	0.008
0.05	0.060	0.032	0.018
0.10	0.128	0.082	0.057
0.15	0.224	0.173	0.152
0.20	0.368	0.312	0.292
0.25	0.573	0.486	0.453
0.30	0.731	0.631	0.561
0.35	0.868	0.816	0.749
0.40	0.927	0.917	0.869
0.45	0.973	0.971	0.951
0.50	0.990	0.992	0.986
0.55	0.997	0.9979	0.997
0.60	0.998	0.9998	0.996
0.65	1.000	0.999989	1.000
0.70	1.000	0.999999	1.000
0.75	1.000	1.000	1.000
0.80	1.000	1.000	1.000

(a) $\alpha = 0.01$

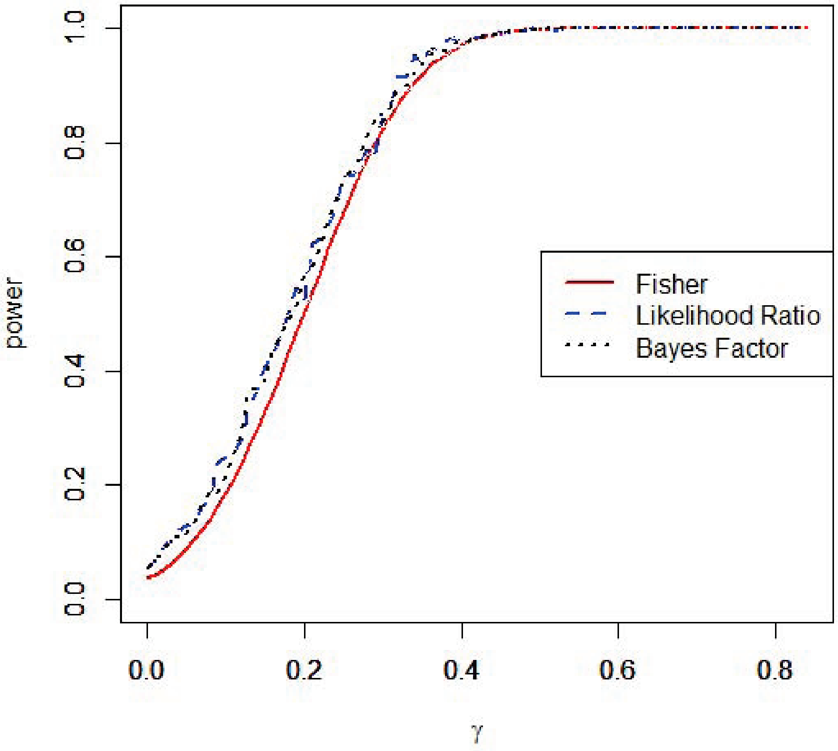
γ	LRT	Fisher's test	BF
0.00	0.054	0.0353	0.053
0.05	0.128	0.0887	0.113
0.10	0.249	0.1853	0.217
0.15	0.409	0.3287	0.384
0.20	0.526	0.5037	0.565
0.25	0.740	0.6795	0.736
0.30	0.860	0.8239	0.837
0.35	0.938	0.9199	0.940
0.40	0.974	0.9709	0.977
0.45	0.992	0.9918	0.994
0.50	1.000	0.9983	0.999
0.55	0.997	0.9997	1.000
0.60	0.999	0.99998	1.000
0.65	1.000	0.999999	1.000
0.70	1.000	0.9999999	1.000
0.75	1.000	1.000	1.000
0.80	1.000	1.000	1.000

(b) $\alpha = .05$

To understand this apparent contradiction, we construct the power functions of these tests. A higher power plot indicates a relatively smaller chance of failing to reject the null hypothesis when the alternative hypothesis is true. Figure 2 (patients



(c) α = .01



(d) α = .05

Figure 2. Power functions of the three tests at α = .01 and .05

older than 75 years using Lidocaine) shows that the power function of the likelihood ratio test generally goes above that of the Fisher's exact test at $\alpha = 0.01$ and 0.05. Table 7 depicts the power values for the three tests at $\alpha = 0.01$ and 0.05 for different values of the correlation parameter γ which corroborate the same. This superiority of the proposed likelihood ratio (in terms of power) may be due to the more flexible modeling framework (with respect to the Fisher's test) on which it is based. The power values of the test based on Bayes factor is generally between that of the likelihood ratio and Fisher's test.

5. Discussion

Fisher's exact test is a well known small sample test used for testing the hypothesis of independence between the row and column variables in a 2 × 2 contingency table. It is generally used when the sample sizes are not large enough for the "large-sample" Pearsonian tests to be valid. However, Fisher's test conditions on both the margins of a contingency table resulting in a hypergeometric distribution of the cell counts. Moreover, it is conservative in the sense that its actual rejection rate (probability of type I error) falls short of the nominal significance level. In this paper, we have corrected the Fisher's exact test by lifting the restriction of fixed margins. That is, we have allowed the margins to be random and in doing so, have proposed two tests - the first based on likelihood ratio and the other based on the Bayes factor. We have performed some data analysis and have compared the conservativeness (probability of type I error) and power functions of these tests against that of the Fisher's exact test.

We have used data from the Worcester Heart Attack study. For the purpose of our study we considered data corresponding to gender difference in the receipt of Lidocaine, Anti-platelet and Beta-blocker therapies for patients suffering from acute myocardial infarction (AMI) who have a history of hypertension and stroke.

On applying the three tests on the above data, it was observed that the p-value of the likelihood ratio test tends to be lower than that of Fisher's test although with some exceptions. This corroborates the conservativeness property of the latter test. The test based on Bayes factor also gave similar conclusions as the likelihood ratio one. It was also observed that the power of the likelihood ratio test tends to be higher than that of the Fisher's test at different significance levels.

In this paper, we have restricted our discussion to 2 × 2 tables although it would be worthwhile to extend our proposed methodology to higher dimensional tables, specifically to $r \times c$ tables with random margins. (Fisher's exact test is known for $r \times c$ tables with fixed margins). One issue with higher dimensional tables is that we can no longer use the correlation coefficient (γ) between the rows and columns as a test parameter. This is one of our future work.

References

Berkson, J. (1978). Do the marginal totals of the 2 × 2 table contain relevant information respecting the table proportions? *Journal of Statistical Planning and Inference*, 43-44. https://doi.org/10.1016/0378-3758(78)90019-8

Barnard, G. A. (1945). A new test for 2 × 2 tables. *Nature, 156, 177* & *783*, 86-105. https://doi.org/10.1038/156177a0

Conover, W. J. (1974). Some reasons for not using the Yates continuity correction on 22 contingency tables. *Journal of the American Statistical Association, 69*, 374-376. https://doi.org/10.2307/2285661

D'Agostino, R. B., Chase, W., & Belanger, A. (1988) The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistician, 42*, 198-202. https://doi.org/10.2307/685002

Grizzle, J. E. (1967). Continuity correction in the chi-square test for 2 × 2 tables. *The American Statistician*, 28-32. https://doi.org/10.1080/00031305.1967.10479835

Hashemi, L., Nandram, B., & Goldberg, R. (1997). Bayesian analysis for a single 2 × 2 table. *Statistics in Medicine, 16*, 1311-1328. https://doi.org/10.1002/(SICI)1097-0258(19970630)16:12¡1311::AID-SIM568¿3.0.CO;2-3

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*, 773-795. https://doi.org/10.1080/01621459.1995.10476572

Kempthorne, O. (1979). Sampling inference, experimental inference and observation inference. *Sankhya, Series B, 40*, 115-145.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician, 43*, 283-288. https://doi.org/10.1080/00031305.1989.10475676

Liddell, D. (1976). Practical tests of 2 × 2 contingency tables. *Journal of the Royal Statistical Society. Series D (The Statistician), 25*, 295-304. https://doi.org/10.2307/2988087

Mehta, C. R., & Senchaudhuri, P. (2003). Conditional versus unconditional exact tests for comparing two binomials.

Cytel Software Corporation

Nandram, B., & Kim, H. (2002). Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical Computation and Simulation, 72*, 319-340. https://doi.org/10.1080/00949650212842

Nandram, B., Cox, L. & Choi, J. W. (2005). Bayesian Analysis of Nonignorable Missing Categorical Data: An Application to Bone Mineral Density and Family Income. *Survey Methodology, 31*, 213-225.

Nandram, B., Liu, N., Choi, J. W. & Cox, L. H. (2005). Bayesian Nonresponse Models for Categorical Data from Small Areas: An Application to BMD and Age. *Statistics in Medicine, 24*, 1047-1074. https://doi.org/0.1002/sim.1985

Nandram, B. & Choi, J. W. (2006). A Bayesian analysis of a two-way categorical table incorporating intra-class correlation. *Journal of Statistical Computation and Simulation, 76*, 233-249. https://doi.org/10.1080/10629360500108962

Nandram, B. & Choi, J. W. (2007). Alternative Tests of Independence in Two-Way Categorical Tables. *Journal of Data Science, 5*, 217-237.

Nandram, B. & Katzoff, M. (2012). A Hierarchical Bayesian Nonresponse Model for Two-Way Categorical Data from Small Areas with Uncertainty about Ignorability. *Survey Methodology, 38*, 81-93.

Nandram, B., Bhatta, D. R., Sedransk, J. & Bhadra, D. (2013). A Bayesian Test of Independence in a Two-Way Contingency Table Using Surrogate Sampling. *Journal of Statistical Planning and Inference, 143*, 1392-1408. https://doi.org/10.1016/j.jspi.2013.03.011

Nandram, B., Bhatta, D., & Bhadra, D. (2015). Likelihood ratio test of quasi-independence for sparse two-way contingency tables. *Journal of Statistical Computation and Simulation, 85*, 284-304. https://doi.org/10.1080/00949655.2013.815190

Upton, G. J. G. (1982). A comparison of alternative tests for the 2 × 2 comparative trial. *Journal of the Royal Statistical Society, Series A (General), 145*, 86-105. https://doi.org/10.2307/2981423

Yates, F. (1984). Tests of significance for 2 × 2 contingency tables. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 147*, 426-463. https://doi.org/10.2307/2981577

Appendix

Calculation of Bayes Factor

As shown in Section 3.3, the Bayes factor for our problem is given by

$$T_{BF} = \frac{f(x,y,z|M_1)}{f(x,y,z|M_0)}$$

where $f(x,y,z|M_0)$ and $f(x,y,z|M_1)$ are respectively given in (14) and (15). Thus, to calculate the Bayes factor, we need to evaluate the above expressions. The former is relatively straightforward since it is a product of two Beta functions. However, for approximating (15), we use Monte Carlo integration technique. Nandram and Kim (2002) used this procedure for studying the marginal likelihood for a class of Bayesian generalized models. In doing so, we rewrite (15) as

$$f(x,y,z|M_1) = \int_{a_0(p,q)}^{a_1(p,q)} \int_0^1 \int_0^1 \left\{ \frac{p(x,y,z|p,q,\gamma)\pi(p)\pi(q)\pi(\gamma)}{\pi_a(p,q,\gamma|x,y,z)} \right\} \pi_a(p,q,\gamma|x,y,z)\; dp\; dq\; d\gamma. \tag{14}$$

Thus, instead of integrating over its prior density, we do the integration over an approximation to the posterior density $\pi_a(p,q,\gamma|x,y,z)$. We can then implement the Monte Carlo method by drawing samples from this density. Towards this end, we need to obtain an approximation to the posterior density function $\pi_a(p,q,\gamma|x,y,z)$. Due to the conjugacy of the Beta-Binomial distribution, we approximate the posterior of p and q by the beta distribution. Thus, given the data (x,y,z), we can draw p from $\pi_a(p|x,y,z) = \text{beta}(x+1, n-x+1)$, and q from $\pi_a(q|x,y,z) = \text{beta}(y+1, n-y+1)$. Hence the approximate posterior density of (p,q,γ) can be expressed as :

$$\pi_a(p,q,\gamma|x,y,z) = \pi_a(p|x,y,z)\; \pi_a(q|x,y,z)\; \pi_a(\gamma|p,q,x,y,z) \tag{15}$$

We can draw p and q from the beta distributions and can certainly expect the posterior density of γ to be close to beta as well. In fact, by doing some minor adjustments, we can transform the range of γ to (0, 1) which is the admissible range for the beta distribution.

The detailed step-by-step procedure for obtaining the posterior density $\pi_a(\gamma|p,q,x,y,z)$ of γ is given below:

a) Given the count data (x, y, z), we draw $M = 1000$ sets of p and q from Beta$(x + 1, n - x + 1)$ and Beta$(y + 1, n - y + 1)$ respectively, the k^{th} sample being denoted as $p^{(k)}$ and $q^{(k)}$ respectively.
b) Generate $M = 1000$ sets of $(p^{(h)}, q^{(h)}, \gamma^{(h)})$ through the procedure (MLE) detailed in Section 3.1.1 and calculate $a_0(p^{(h)}, q^{(h)})$ and $a_1(p^{(h)}, q^{(h)})$ from the generated values, $(h = 1, 2, ..., 1000)$.
c) Fit the $\gamma^{(h)}$ values to the following equation:

$$\gamma_1^{(h)} = \frac{\gamma_{(h)} - a_0(p^{(h)}, q^{(h)})}{a_1(p^{(h)}, q^{(h)}) - a_0(p^{(h)}, q^{(h)})} \sim \text{Beta}(\mu\tau, (1-\mu)\tau), 0 < \mu < 1, \tau > 0. \tag{16}$$

Thus, we can estimate the beta parameters by equating the sample mean to the theoretical mean (μ) and sample variance to the theoretical variance $(\mu(1 - \mu)/(\tau + 1))$ which yields

$$\begin{aligned} E(\gamma_1^{(h)}) &= \frac{\sum_{h=1}^{M} \gamma_1^{(h)}}{M} = \mu \\ \text{Var}(\gamma_1^{(h)}) &= \frac{\sum_{h=1}^{M} (\gamma_1^{(h)} - E(\gamma_1^{(h)}))^2}{M-1} = \frac{\mu(1-\mu)}{\tau+1}. \end{aligned} \tag{17}$$

Solving the above equations we obtain the following estimates of μ and τ,

$$\begin{aligned} \hat{\mu} &= E(\gamma_1^h) \\ \hat{\tau} &= \frac{\mu(1-\mu)(M-1)}{\sum_{h=1}^{M} \left(\gamma_1^{(h)} - E(\gamma_1^{(h)})\right)^2} - 1 \end{aligned} \tag{18}$$

d) Draw $\gamma_1^{(k)}$ from Beta $(\hat{\mu}\hat{\tau}, (1 - \hat{\mu})\hat{\tau})$ and obtain $\gamma = h(\gamma_1^k) = [a_1(p^{(k)}, q^{(k)}) - a_0(p^{(k)}, q^{(k)})]/[\gamma_1 + a_0(p^{(k)}, q^{(k)})]$, where $p^{(k)}$ and $q^{(k)}$ are drawn from beta distribution in step (a)
e) The probability density function of γ_1 is given by

$$f(\gamma_1) = \frac{\gamma_1^{\hat{\mu}\hat{\tau}}(1-\gamma_1)^{(1-\hat{\mu})\hat{\tau}}}{B(\hat{\mu}\hat{\tau}, (1-\hat{\mu})\hat{\tau})}. \tag{19}$$

On transforming γ_1 to γ as $\gamma = h(\gamma_1)$, the probability density function of γ will be

$$\begin{aligned} \pi_a(\gamma|p, q, x, y, z) &= f(h^{-1}(\gamma)) \\ &= \frac{[h^{-1}(\gamma)]^{\hat{\mu}\hat{\tau}}[1 - h^{-1}(\gamma)]^{(1-\hat{\mu})\hat{\tau}}}{B(\hat{\mu}\hat{\tau}, (1-\hat{\mu})\hat{\tau})} \cdot \frac{d(h^{-1}(\gamma))}{d\gamma} \\ &= \frac{[\frac{\gamma - a_0(p,q)}{a_1(p,q) - a_0(p,q)}]^{\hat{\mu}\hat{\tau}}\{1 - [\frac{\gamma - a_0(p,q)}{a_1(p,q) - a_0(p,q)}]\}^{(1-\hat{\mu})\hat{\tau}}}{B(\hat{\mu}\hat{\tau}, (1-\hat{\mu})\hat{\tau})} \cdot \frac{1}{a_1(p, q) - a_0(p, q)} \end{aligned} \tag{20}$$

f) Together with

$$\begin{aligned} \pi_a(p|x, y, z) &= \frac{p^x(1-p)^{n-x}}{B(x+1, n-x+1)} \\ \pi_a(q|x, y, z) &= \frac{q^y(1-q)^{n-y}}{B(y+1, n-y+1)} \end{aligned}$$

we obtain

$$\begin{aligned} \pi_a(p, q, \gamma|x, y, z) &= \pi_a(p|x, y, z)\pi_a(q|x, y, z)\pi_a(\gamma|p, q, x, y, z) \\ &= \frac{p^x(1-p)^{n-x}}{B(x+1, n-x+1)} \frac{q^y(1-q)^{n-y}}{B(y+1, n-y+1)} \frac{[h^{-1}(\gamma)]^{\hat{\mu}\hat{\tau}}[1 - h^{-1}(\gamma)]^{(1-\hat{\mu})\hat{\tau}}}{B(\hat{\mu}\hat{\tau}, (1-\hat{\mu})\hat{\tau})} \cdot \frac{d(h^{-1}(\gamma))}{d\gamma} \end{aligned} \tag{21}$$

Based on the above derivations, our Bayes factor can now be expressed as

$$\begin{aligned} T_{BF} &= \frac{f(x, y, z|M_1)}{f(x, y, z|M_0)} \\ &= \frac{\int_{a_0(p,q)}^{a_1(p,q)} \int_0^1 \int_0^1 \left\{\frac{p(x, y, z|p, q, \gamma)\pi(p)\pi(q)\pi(\gamma)}{\pi_a(p, q, \gamma|x, y, z)}\right\} \pi_a(p, q, \gamma|x, y, z) \, dp \, dq \, d\gamma}{C \cdot B(x+1, n-x+1)B(y+1, n-y+1)} \\ &= \frac{\frac{1}{M} \sum_{k=1}^{M} \frac{C\pi_{11}^{z}\pi_{10}^{x-z}\pi_{01}^{y-z}\pi_{00}^{n-x-y+z}\pi(p)\pi(q)\pi(\gamma)}{\pi_a(p, q, \gamma|x, y, z)}}{C \cdot B(x+1, n-x+1)B(y+1, n-y+1)} \end{aligned} \tag{22}$$

g) We now plug in 1000 sets of $(p^{(k)}, q^{(k)}, \gamma^{(k)})$ in (24) to obtain samples from the Bayes factor. As we know, a Bayes factor value greater than 1 will indicate a rejection of null hypothesis and vice versa.

Copyrights