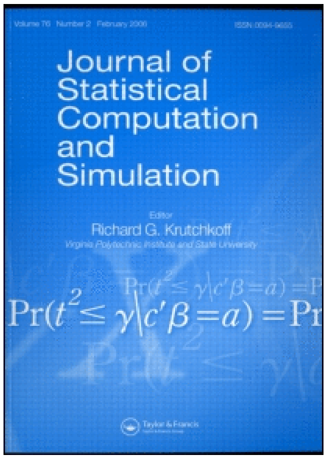
Volume 78 Number 2 February 2008
ISSN 0094-9655
Journal of Statistical Computation and Simulation
Editor
Richard G. Krutchkoff
Virginia Polytechnic Institute and State University
$\Pr(t^2 \leq \gamma \backslash c'\beta = a) = \Pr$
Taylor & Francis

Journal of Statistical Computation and Simulation

Exact tests for equality of two proportions: fisher v. bayes

Telba Z. Ironyt [a]; Carlos A. B. Pereirat [a]
[a] Universidade de Sao Paulo Institute de Matema'tica e Estatistica C. Postal, Sao Paulo, Brazil

Online Publication Date: 01 August 1986
To cite this Article: Ironyt, Telba Z. and Pereirat, Carlos A. B. (1986) 'Exact tests for equality of two proportions: fisher v. bayes', Journal of Statistical Computation and Simulation, 25:1, 93 - 114
To link to this article: DOI: 10.1080/00949658608810926
URL: http://dx.doi.org/10.1080/00949658608810926

J. Statist. Comput. Simul., 1986, Vol. 25, pp. 93–114
0094-9655/86/2502-0093 $18.50/0

Exact Tests for Equality of Two Proportions: Fisher v. Bayes

TELBA Z. IRONY†§ and CARLOS A. B. PEREIRA‡§
*Universidade de São Paulo, Instituto de Matemática e Estatística
C. Postal 20.570-01498, São Paulo, Brazil*

(*Received October 9, 1985*)

Yates (1984) using theoretical and philosophical arguments claims to have proved that the Fisher exact test for comparing the proportions of two binomial experiments is the best exact test. The present article uses objective and practical arguments to confront the Fisher exact test with a Bayes exact test. Using simulated samples we claim to have proved here the inferiority of the Fisher exact test in relation to a Bayes exact test. The comparison is based on the quality concept of Dawid (1982).

KEY WORDS: Fisher exact test, Bayes exact test, stated error probabilities, actual error frequencies (or probabilities), estimated error frequencies.

AMS Classification: 62F03 62A20 62A15.

1. INTRODUCTION

This article is devoted to a practical confront between a Bayes exact test (*BE* test) and the Fisher exact test (*FE* test) for comparing the proportions, p and q, of two binomial experiments, X and Y, with sample sizes m and n, respectively. The data set may be displayed in the following 2×2 contingency table.

†Itau Informática Ltd, São Paulo, Brazil.

‡CNPq Brasilia, Brazil.

§Present address: University of California, Operations Research Center, Berkeley, California 94720, U.S.A.

TABLE I
Sample displayed in a 2 × 2 table

Experiment	Success	Failure	Sample size
X	x	$m-x$	m
Y	y	$n-y$	n
Total	t	$M-t$	$M=n+m$

In the same manner, the parametric structure is displayed as:

TABLE II
Parametric structure

Experiment	Success	Failure	
X	p	$1-p$	$0<p<1$ and
Y	q	$1-q$	$0<q<1$

The problem is to test $H{:}p=q$, the null hypothesis, against $A{:}p\neq q$, the alternative hypothesis, when m and n are considered too small to permit the adoption of an asymptotic method.

The classical treatment for this problem is the Fisher exact test, that has received a great deal of attention through discussion papers about its validity. In a long article on the history and the analysis of these discussions, Yates (1984) shows that, under the classical point of view, there is no better test than the *FE* test. Hence, for one who rejects this test (Basu, 1979), the natural alternative is to look for a Bayesian solution. However, this is a difficult attitude to be taken by a classical statistician because the literature is flooded of radical papers rejecting one point of view in favor of the other. Hence, in principle, classical statisticians must reject Bayesian methods and Bayesians must reject classical ones. For a complete analysis of these two viewpoints see Kempthorne (1980a, 1980b) and Lindley (1982).

In spite of the final conclusion on the superiority of the *BE* test, discussions on philosophical aspects are not the purpose of the present note. The intention here is to look at the *FE* test and at the *BE* test as two simple decision rules and then, objectively, to verify which of them is the best. What is meant by a rule that is objectively good is explained next and follows the lines of Dawid (1982).

A decision rule to test H against A is a binary function on the sample space associated with a statement, (α, β), about the actual values, ALFA and BETA, of the probabilities of the two kinds of errors. A good test must satisfy the following conditions:

i) α=ALFA and β=BETA. That is, the test states correctly the error probabilities.

ii) For a fixed $k>0$, the value of ALFA+kBETA is minimized by the test. Note that k establishes an order of importance between α and β (for a technical discussion see De Groot, 1975).

Using a representative number of simulated samples, Section 3 shows that the *FE* test, unlike the *BE* test, fails to follow both (i) and (ii). Sections 4 and 5 analyse the reasons for this failure. The two tests are described in the sequel.

2. BAYES AND FISHER EXACT TESTS

The *BE* test and the *FE* test are defined in this section in order to simplify the understanding of the notation considered.

2.1 The *BE* test

The Bayesian procedure considered here is like the Jeffreys' test for sharp hypothesis (Jeffreys, 1961). Note that H defines the set $\Omega_0=\{(p,q);\ 0<p=q<1\}$, that is, a line in the parametric space $\Omega=\{(p,q);\ 0<p<1, 0<q<1\}$. Hence, Ω and Ω_0 have different dimensions.

Since the objective here is to confront the two tests, only uniform priors (on Ω_0 and on $\Omega_1=\Omega-\Omega_0$) are considered because no prior knowledge, as in the *FE* test, must be used. With these priors, the predictive probabilities for (x,t), the data, under H and A are, respectively,

$$f_H(x,t)=\frac{\binom{t}{x}\binom{M-t}{m-x}}{\binom{M}{m}}\frac{1}{M+1}$$

and

$$f_A(x,t)=\frac{1}{(m+1)(n+1)}. \tag{1}$$

In order to give the same importance for H and A, we consider the prior probabilities $\Pr(H)=\Pr(A)=\frac{1}{2}$. The Bayes factor in favor of H is then

$$b(x,t)=\frac{f_H}{f_A}=\frac{(m+1)(n+1)}{(M+1)}\frac{\binom{t}{x}\binom{M-t}{m-x}}{\binom{M}{m}}. \tag{2}$$

It is interesting to notice that $b(x,t)$ is the ratio of the likelihood averages on Ω_0 and on Ω_1. For a fixed positive constant k, the function test, B_k, of the BE test is

$$B_k(x,t)=0 \quad \text{if} \quad b(x,t)\geqq k$$

and

$$B_k(x,t)=1 \quad \text{if} \quad b(x,t)<k.$$

The main property of this test (De Groot, 1975) is to minimize $\alpha+k\beta$, where $\alpha=\sum f_H(x,t)$ and $\beta=1-\sum f_A(x,t)$; the sum is over the critical region $CB_k=\{(x,t);b(x,t)<k\}$. Since we consider no preference between the two hypotheses, to minimize $\alpha+\beta$ we must take $k=1$.

Finally, we notice that α and β defined here are the averages of the two kinds of errors taken over Ω_0 and Ω_1, respectively.

2.2 The *FE* test

The FE test considers $\theta=[p(1-q)/q(1-p)]$, the cross-product, as the parameter of interest. The marginals, m, n, t, and $M-t$, carry no information about θ in Fisher's opinion (Fisher, 1935). Hence, in the FE test these marginals are considered fixed before the data collection. Consequently, the likelihood functions used by the FE test are

$$L(\theta, t|x) = \frac{\binom{m}{x}\binom{n}{y}\theta^x}{\sum_i \binom{m}{i}\binom{n}{t-i}\theta^i}$$

where $\sum_i$ is the sum over all possible values that x can assume with the fixed marginals.

The hypotheses H and A are equivalent to $H':\theta=1$ and $A':\theta\neq 1$, respectively. Hence, the function $L(\theta, t|x)$ at $\theta=1$ is expressed as

$$L(1, t|x) = \frac{\binom{m}{x}\binom{n}{y}}{\binom{M}{t}} = \frac{\binom{t}{x}\binom{M-t}{m-x}}{\binom{M}{m}}. \tag{3}$$

In the set of all values that x can assume with the fixed marginals, let us consider the set C_x^t of all points d such that $L(1,t|x) \geqq L(1,t|d)$; that is $C_x^t = \{d; L(1,t|x) \geqq L(1,t|d)\}$. Representing the sum over C_x^t by $\sum_d$ and the fixed level of significance by α_0, the test function, F_{α_0}, of the FE test is

$$F_{\alpha_0}(x,t) = 0 \quad \text{if} \quad \sum_d L(1,t|d) \geqq \alpha_0$$

and (4)

$$F_{\alpha_0}(x,t) = 1 \quad \text{if} \quad \sum_d L(1,t|d) < \alpha_0.$$

Looking at expressions (2) and (3), we can write

$$b(x,t) = \frac{(m+1)(n+1)}{M+1} L(1,t|d)$$

that suggests a close relation between $B_k(x,t)$ and $F_{\alpha_0}(x,t)$, the two exact tests in confront. Note, however, that $L(1,t|x)$ is used differently by the two tests.

Finally, we emphasize that two decision rules have been constructed. Objectively, two different 'partitions of the sample space were defined. We claim that the comparison of these two partitions, described in the sequel, is objective without any appeal to philosophical and theoretical arguments.

3. *BE* TEST VERSUS *FE* TEST

3.1 Preliminaries

A test of H against A is an accept/reject rule associated with a statement about the values of the error probabilities. To represent the values stated by the BE test we write $\alpha(B_k)$ and $\beta(B_k)$ for the probabilities of first and second kind of errors, respectively. Analogously, for significance level equal α_0, $\alpha(F_{\alpha_0})=\alpha_0$ and $\beta(F_{\alpha_0})$ are the corresponding values stated by the FE test. These stated values, however, may not be equal to the actual values of the frequencies in which the two kinds of errors are commited by using these tests. Respectively to the first and second kind of errors, let us represent the true value of the error frequencies by

$$\text{ALFA}B_k \text{ and BETA}B_k \text{ for the } BE \text{ test}$$

and

$$\text{ALFA}F_{\alpha_0} \text{ and BETA}F_{\alpha_0} \text{ for the } FE \text{ test.}$$

The way of comparing the tests and how they were constructed is explained next.

With the order of importance between the errors fixed by the value of k, we state the criterion of choice. This value of k specifies the BE test function B_k and its error probabilities

$$\alpha_k=\alpha(B_k) \quad \text{and} \quad \beta(B_k).$$

With $\alpha_0=\alpha_k$, we define the test function F_{α_k}. After fixing the values of k and α_0, we name the tests as B_kE test and $F_{\alpha_0}E$ test or $F_{\alpha_k}E$ test if $\alpha_0=\alpha_k=\alpha(B_k)$. The following definition states our choice criterion.

Definition The B_kE test (the $F_{\alpha_0}E$ test) is better than the $F_{\alpha_0}E$ test (the B_kE test) if the following conditions are satisfied:

i) $$|\alpha(B_k)-\text{ALFA}\,B_k| \underset{(\geqq)}{\leqq} |\alpha(F_{\alpha_0})-\text{ALFA}\,F_{\alpha_0}|$$

and

ii) $$\text{ALFA}\,B_k+k\text{BETA}\,B_k \underset{(\geqq)}{\leqq} \text{ALFA}\,F_{\alpha_0}+k\text{BETA}\,F_{\alpha_0}.$$

In case of $\alpha_0=\alpha(B_k)$, F_{α_k} substitutes F_{α_0}.

Note that the distance between the true frequency and the stated probability of the second kind of error was not included in the definition. The reason for this is that the *FE* test does not specify $\beta(F_{\alpha_0})$. Condition (i) indicates which one of the two tests lies to a less extent. Condition (ii) indicates the test with less error frequency.

To proceed with our analysis, we consider the particular case where the two kinds of error have the same importance. That is, $k=1$ is going to be used in the sequence. Recall that in this case the two test functions are represented by B_1 and F_{α_1} and write the B_1E test and the $F_{\alpha_1}E$ test.

The test functions B_1 and F_{α_1} are characterized by the critical regions CB_1 and CF_{α_1} defined below. If (x, t) is the data observed let us consider

$$F(x|M,m,t)=\sum_{d=0}^{x} L(1,t|d) \quad\text{and}\quad \min[F(1,t|d)\colon 1-F(1,t|d)]=f(x,t).$$

The two critical regions are

$$CB_1=\{(x,t); b(x,t)<1\}$$

and

$$CF_{\alpha_1}=\left\{(x,t); f(x,t)<\frac{\alpha(B_1)}{2}\right\}$$

where

$$\alpha(B_1)=\sum f_H(x,t)=\sum\frac{\binom{t}{x}\binom{M-t}{m-x}}{\binom{M}{m}(M+1)},$$

the sum covering all points of CB_1. The value of $\beta(B_1)$ is evaluated as

$$\beta(B_1)=1-\frac{\text{number of elements of } CB_1}{(m+1)(n+1)}.$$

To illustrate the evaluation of the above entities, we consider the following example.

Example 1 Consider the case where $m=n=5$. The values taken by the two functions, B_1 and F_{α_1}, in each sample point are given by the following decision boards. Here,

$$\alpha_1=\alpha(B_1)=0.2251 \quad \text{and} \quad \beta(B_1)=1-\frac{20}{36}=\frac{4}{9}.$$

TABLE III

Values assumed by the test function $B_1(x,t)$

x $t-x$	0	1	2	3	4	5
0	0	0	1	1	1	1
1	0	0	0	1	1	1
2	1	0	0	0	1	1
3	1	1	0	0	0	1
4	1	1	1	0	0	0
5	1	1	1	1	0	0

TABLE IV

Values assumed by the test function $F_{\alpha_1}(x,t)$ where $\alpha_1=\alpha(B_1)=0.2251$

x $t-x$	0	1	2	3	4	5
0	0	0	0	1	1	1
1	0	0	0	0	1	1
2	0	0	0	0	0	1
3	1	0	0	0	0	0
4	1	1	0	0	0	0
5	1	1	1	0	0	0

Note that both tests state the value 0.2251 for the first-kind-of-error probability. However, the boards show the difference between B_1 and F_{α_1}. Hence, since $CF_{\alpha_1} \neq CB_1$ and $CF_{\alpha_1} \subset CB_1$, at least one of the tests is making a wrong statement about the first kind of error.

3.2 The simulation

In order to estimate the true values of the error frequencies, a large number of simulated samples have been considered.

Remember that if $p=q$, then one who rejects $H{:}p=q$ is committing the first kind of error. To estimate the values of ALFAB_1 and ALFAF_{α_1} we consider nine pairs of points (p,q) where $p=q$. These nine values of $p=q$ are $0.1, 0.2, \ldots, 0.9$. For each one of these parametric points, a thousand samples (like Table I) were simulated. For each one of these simulated samples we have calculated the values (0 or 1) of the test functions B_1 and F_{α_1}. For each one of the nine parametric points, (p,p), we estimate the error frequencies by

$$\frac{\text{number of samples at } (p,p) \text{ where } B_1(x,t)=1}{1{,}000} = \text{alfa}\, B_1|p$$

and

$$\frac{\text{number of samples at } (p,p) \text{ where } F_{\alpha_1}(x,t)=1}{1{,}000} = \text{alfa}\, F_{\alpha_1}|p.$$

Example 2 (continuation) Table V presents the values of alfa$B_1|p$ and alfa$F_{\alpha_1}|p$ in the case of $m=n=5$.

It is interesting to note from Table V that alfa$B_1|p >$ alfa$F_{\alpha_1}|p$ for all nine values of p.

Since we want to characterize the best test independently of (p,q), we estimate the true first-kind-of-error frequencies, ALFAB_1 and ALFAF_{α_1}, by

$$\text{alfa}\, B_1 = \frac{\text{total number of samples in which } B_1(x,t)=1}{9{,}000}$$

and

$$\text{alfa}\, F_{\alpha_1} = \frac{\text{total number of samples in which } F_{\alpha_1}(x,t)=1}{9{,}000},$$

the averages of alfa$B|p$ and alfa$F_{\alpha_1}|p$.

TABLE V

Estimates of the true first-kind-of-error frequencies in some parametric points for $m=n=5$

p	alfa$B_1\|p$	alfa$F_{\alpha_1}\|p$
0.1	0.092	0.009
0.2	0.211	0.046
0.3	0.305	0.082
0.4	0.340	0.108
0.5	0.354	0.108
0.6	0.366	0.119
0.7	0.310	0.071
0.8	0.216	0.050
0.9	0.124	0.013

Example 3 (continuation) For the case of $m=n=5$ we obtain

$$\alpha_0=\alpha(B_1)=0.2251,\ \text{alfa}B_1=0.2576,\ \text{and alfa}F_{\alpha_1}=0.0673.$$

Although alfa$F_{\alpha_1}<$alfaB_1, we have $|\alpha(B_1)-\text{alfa}B_1|=0.0325<|\alpha_0-\text{alfa}F_{\alpha_0}|=0.1578$, which means that the B_1E test lies less than the $F_{\alpha_1}E$ test. This conclusion uses the fact that alfaB_1 and alfaF_{α_1} are good estimates of ALFAB_1 and ALFAF_{α_1}.

Recall now that if $p\neq q$, then one who does not reject $H{:}p=q$ is committing the second kind of error.

To estimate the values of BETAB_1 and BETAF_{α_1} we consider 36 pairs of points (p,q) where $p<q$; p taking the values on $\{0.1,0.2,\ldots,0.8\}$ and q taking the values on $\{0.2,0.3,\ldots,0.9\}$. (The symmetry of the problem allowed us to consider no point (p,q) where $p>q$.) To each of these pairs a thousand samples were simulated. To each one of these simulated samples we calculated the values (0 or 1) of the test functions B_1 and F_{α_1}. For each of the 36 points, (p,q), we estimate the error frequency by

$$\frac{\text{number of samples at } (p,q) \text{ where } B_1(x,t)=0}{1{,}000}=\text{beta}B_1|(p,q)$$

and

$$\frac{\text{number of samples at } (p,q) \text{ where } F_{\alpha_1}(x,t)=0}{1{,}000}=\text{beta}F_{\alpha_1}|(p,q).$$

Example 4 (continuation) Table VI presents the values of $\text{beta}\,B_1(p,q)$ and $\text{beta}\,F_{\alpha_1}|(p,q)$ in the case of $m=n=5$.

Table VI indicates that $\text{beta}\,B_1|(p,q)<\text{beta}\,F_1|(p,q)$, for all (p,q), in the contrary direction of the alfa values of Table V.

TABLE VI

Estimates of the true second-kind-of-error frequencies in some parametric points for $m=n=5$

p	q	$\text{beta}\,B_1$	$\text{beta}\,F_{\alpha_1}$	p	q	$\text{beta}\,B_1$	$\text{beta}\,F_{\alpha_1}$
0.1	0.2	0.792	0.962	0.3	0.7	0.339	0.646
0.1	0.3	0.655	0.882	0.3	0.8	0.218	0.476
0.1	0.4	0.491	0.796	0.3	0.9	0.109	0.323
0.1	0.5	0.351	0.634	0.4	0.5	0.608	0.865
0.1	0.6	0.221	0.491	0.4	0.6	0.568	0.803
0.1	0.7	0.105	0.328	0.4	0.7	0.481	0.742
0.1	0.8	0.062	0.117	0.4	0.8	0.325	0.620
0.1	0.9	0.016	0.070	0.4	0.9	0.208	0.481
0.2	0.3	0.717	0.921	0.5	0.6	0.618	0.875
0.2	0.4	0.603	0.856	0.5	0.7	0.554	0.828
0.2	0.5	0.481	0.755	0.5	0.8	0.466	0.745
0.2	0.6	0.340	0.616	0.5	0.9	0.333	0.628
0.2	0.7	0.234	0.478	0.6	0.7	0.664	0.874
0.2	0.8	0.115	0.322	0.6	0.8	0.591	0.853
0.2	0.9	0.057	0.171	0.6	0.9	0.485	0.799
0.3	0.4	0.655	0.900	0.7	0.8	0.715	0.928
0.3	0.5	0.564	0.822	0.7	0.9	0.640	0.879
0.3	0.6	0.463	0.724	0.8	0.9	0.777	0.963

Analogous to the estimation of ALFA, we estimate the true second-kind-of-error frequencies by

$$\text{beta}\,B_1=\frac{\text{Total number of samples in which } B_1(x,t)=0}{36{,}000}$$

and

$$\text{beta}\,F_{\alpha_1}=\frac{\text{Total number of samples in which } F_{\alpha_1}(x,t)=0}{36{,}000},$$

the averages of $\text{beta}\,B_1|(p,q)$ and $\text{beta}\,F_{\alpha_1}|(p,q)$.

Example 5 (continuation) For the case of $m=n=5$, we obtain

$$\beta(B_1)=1-\frac{20}{36}=\frac{4}{9}=0.4444, \text{ beta}B_1=0.4339, \text{ and beta}F_{\alpha_1}=0.6731.$$

On the other hand we notice that $\alpha(B_1)+\beta(B_1)=0.6696$ and that alfaB_1 + beta$B_1=0.6915$. Since alfaF_{α_1} + beta$F_{\alpha_1}=0.7404$, we estimate that, for $m=n=5$,

$$\text{ALFA}B_1+\text{BETA}B_1<\text{ALFA}F_{\alpha_1}+\text{BETA}F_{\alpha_1},$$

and, from the previous example, that

$$|\alpha(B_1)-\text{ALFA}B_1|<|\alpha(F\alpha_1)-\text{ALFA}F_{\alpha_1}|.$$

Hence, using Definition 1 we conclude the superiority of the B_1E test for $m=n=5$. Figures 1 and 2 illustrate this fact by presenting the power functions of the two tests. Note that the power function of the B_1E test takes higher values, for all (p, q), than that of the $F_{\alpha_1}E$ test.

Table VII presents the results obtained for different values of $m=n$. For all sample sizes included in the table, the conclusion favors the B_1E test in prejudice of the $F_{\alpha_1}E$ test. The values of $\alpha(B_1)=\alpha(F_{\alpha_1})$ are closer to the corresponding values of alfaB_1 than to the values of alfaF_{α_1} and alfaB_1 + betaB_1 < alfaF_{α_1} + betaF_{α_1} indicating that the B_1E test commits less error than the $F_{\alpha_1}E$ test.

The results presented in Table VII suggest that the $F_{\alpha_1}E$ test must be substituted by the B_1E test. This conclusion is in the direction of Definition 1. Example 7 presents the same results in two cases where $m\neq n$.

Example 6 Table VIII presents the results for $(m,n)=(5,3)$ and $(m,n)=(8,14)$.

Again, the superiority of the B_1E test over the $F_{\alpha_1}E$ test is suggested. The decision boards are presented next in order to give the opportunity to the reader to analyse the conservativeness of the F_{α_1} function.

We claim that the results presented in this section permit us to conclude that, in the sense of Definition 1, the BE test is better than the FE test. In the next section, we analyse the construction of the FE test.

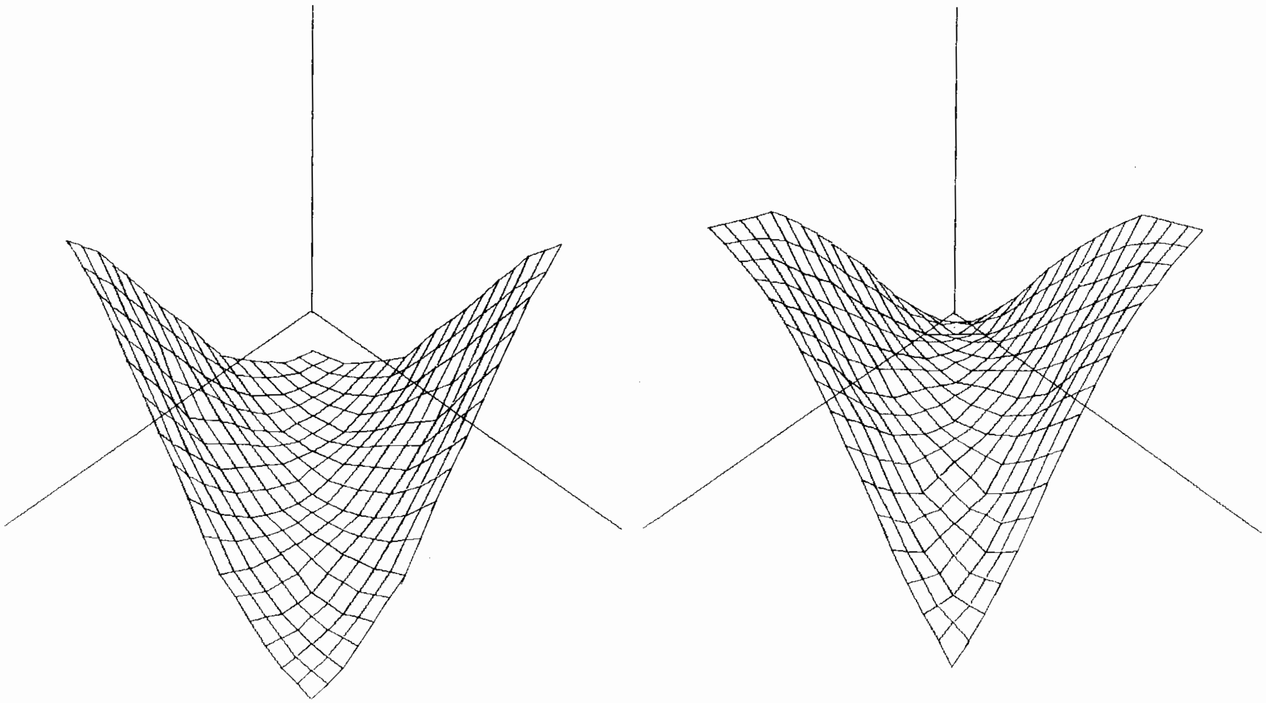

FIGURE 1 Frontal view of the power functions of the $F_{\alpha_1}E$ and B_1E tests for $m = n = 5$.

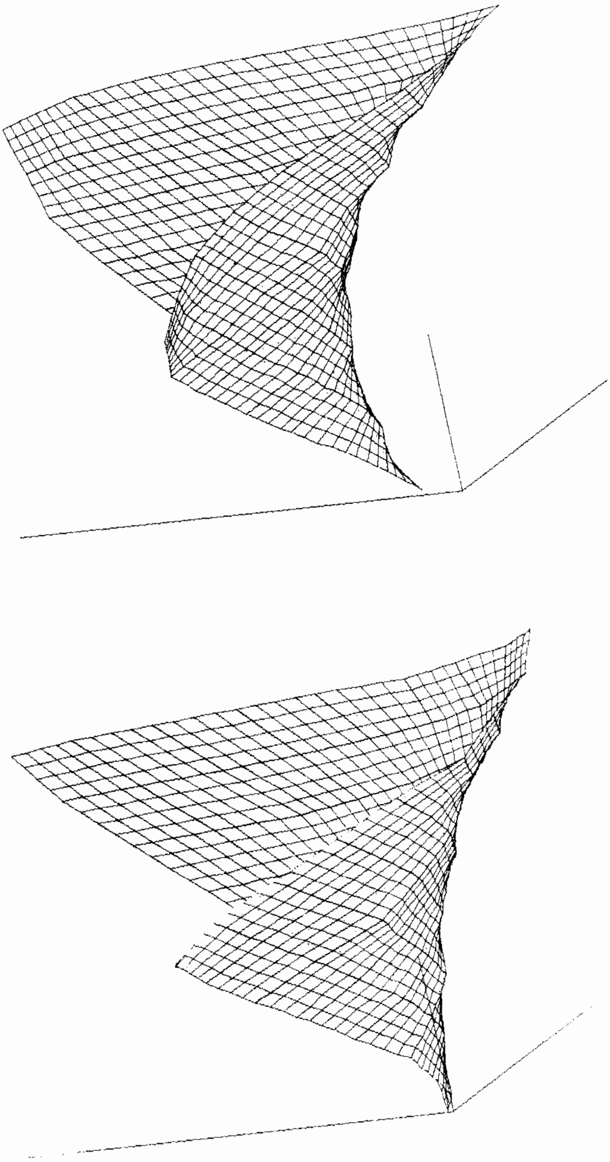

FIGURE 2 Lateral view of the power functions of the $F_{\alpha_1}E$ and BE tests for $m=n=5$.

TABLE VII

Stated and actual error frequencies in the case of $m=n$.

			$\alpha(B_1)+$			Alfa B_1 +			alfa F_{α_1} +
$m=n$	$\alpha(B_1)$	$\beta(B_1)$	$\beta(B_1)$	alfa B_1	beta B_1	beta B_1	alfa F_{α_1}	beta F_{α_1}	beta F_{α_1}
1	0.3333	0.5000	0.83333	0.3622	0.4836	0.84578	0.0000	1.0000	1.00000
2	0.4667	0.3333	0.80000	0.5203	0.3151	0.83539	0.0750	0.7842	0.85917
3	0.1286	0.6250	0.75357	0.1478	0.6222	0.76997	0.0153	0.8870	0.90233
4	0.1810	0.5200	0.70095	0.1999	0.5136	0.71344	0.0369	0.7737	0.81058
5	0.2251	0.4444	0.66955	0.2474	0.4361	0.68358	0.0677	0.6719	0.73956
6	0.2627	0.3878	0.65049	0.2896	0.3742	0.66378	0.0856	0.5946	0.68011
7	0.1647	0.4688	0.63344	0.1839	0.4766	0.66053	0.0337	0.6970	0.73069
8	0.1245	0.4815	0.60601	0.1387	0.4722	0.61083	0.0346	0.6691	0.70361
9	0.1448	0.4400	0.58477	0.1578	0.4242	0.58194	0.0370	0.6326	0.66964
10	0.1639	0.4050	0.56881	0.1837	0.3868	0.57050	0.0432	0.5969	0.64011
11	0.1818	0.3750	0.55681	0.1992	0.3535	0.55269	0.0498	0.5633	0.61306
12	0.1245	0.4201	0.54466	0.1342	0.4164	0.55064	0.0411	0.5545	0.59558
13	0.1125	0.4184	0.53087	0.1266	0.4114	0.53792	0.0442	0.5376	0.58178
14	0.1227	0.3956	0.51827	0.1373	0.3832	0.52056	0.0486	0.5176	0.56617
15	0.1328	0.3750	0.50777	0.1440	0.3613	0.50531	0.0503	0.4958	0.54614
16	0.1426	0.3564	0.49904	0.1614	0.3393	0.50075	0.0528	0.4729	0.52569
17	0.1145	0.3765	0.49102	0.1279	0.3552	0.48306	0.0456	0.4867	0.53225
18	0.1079	0.3740	0.48185	0.1204	0.3527	0.47314	0.0476	0.4619	0.50950
19	0.1028	0.3700	0.47284	0.1218	0.3511	0.47286	0.0456	0.4584	0.50394
20	0.0995	0.3651	0.46461	0.1141	0.3455	0.45964	0.0434	0.4552	0.49867

TABLE VIII

Stated and actual error frequencies for $(m, n) = (5, 3)$ and $(m, n) = (8, 14)$

m	n	$\alpha(B_1)$	$\beta(B_1)$	$\alpha(B_1)+$ $\beta(B_1)$	alfaB_1	betaB_1	alfaB_1+ betaB_1	alfaF_{α_1}	betaF_{α_1}	alfa$F_{\alpha_1}+$ betaF_{α_1}
5	3	0.222	0.500	0.722	0.247	0.485	0.732	0.095	0.679	0.774
8	14	0.144	0.422	0.566	0.163	0.414	0.577	0.035	0.612	0.647

TABLE IX

Decision board of the B_1E test for $m=5$ and $n=3$

$t-x$ \ x	0	1	2	3	4	5
0	0	0	1	1	1	1
1	0	0	0	0	1	1
2	1	1	0	0	0	0
3	1	1	1	1	0	0

TABLE X

Decision board of the $F_{\alpha_1}E$ test for $m=5$ and $n=3$

$t-x$ \ x	0	1	2	3	4	5
0	0	0	0	1	1	1
1	0	0	0	0	0	1
2	1	0	0	0	0	0
3	1	1	1	0	0	0

TABLE XI

Decision board of the B_1E test for $m=8$ and $n=14$

x \ $t-x$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
2	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1
3	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1
4	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1
5	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1
6	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1
7	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
8	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0

TABLE XII

Decision board of the $F_{\alpha_1}E$ test for $m=8$ and $n=14$

x \ $t-x$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
2	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
3	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1
4	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1
5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
6	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0

4. THE *BE* WITH *k* SATISFYING $\alpha(B_k)=\text{alfa}\,F_{\alpha_1}$

Although our discussion now restricts itself to the particular case of $m=n=5$, the conclusions are general. The intention here is to make the analysis as understandable as we can.

One may criticize the B_1E test, which minimizes ALFA+BETA, saying that it gives the same importance for both kinds of error. On the other hand, in the $F_{\alpha_1}E$ test the ALFA value is lower than in the B_1E test. This may lead one to the following wrong conclusion: "a researcher who considers that the first kind of error causes more damage than the second kind of error must prefer the $F_{\alpha_1}E$ test". That this is a wrong statement is suggested in the discussion below.

Recall that the choice of k in the construction of the *BE* test defines the degree of importance of one kind of error in relation to the other. The B_kE test is the test which minimizes ALFA+kBETA. In that way, when we take $k<1$ we are supposing that the first kind of error causes more damage than the second. Hence, the minimization power affects ALFA more intensively than it affects BETA. Contrarily, if $k>1$ ALFA and BETA change places. This means that, by choosing appropriately the value of k, the B_kE test will produce the desired value of ALFA.

In the case of $m=n=5$, we noticed that alfa$B_1=0.2576>0.0673=$ alfaF_{α_1} although the stated value $\alpha(F_{\alpha_1})$ is not close to alfaF_{α_1}. In order to have a *BE* test stating the first kind of error as alfa$F_{\alpha_1}=$

0.0673, let us find a value of k, if it exists, such that $\alpha(B_k)=0.0673$ and $\beta(B_k)=\text{beta}F_{\alpha_1}=0.6731$. That is, this value of k, k', has to satisfy the two following equations:

$$\alpha(B_{k'})=\sum_{CB_{k'}}\frac{\binom{t}{x}\binom{10-t}{5-x}}{\binom{10}{5}}\frac{1}{11}=0.0673 \qquad (5)$$

and

$$\beta(B_{k'})=1-\frac{\text{number of elements of } CB_{k'}}{6\times 6}=0.6731 \qquad (6)$$

From Eq. (6) we conclude that $CB_{k'}$ must have twelve sample points since the number of elements of a set must be integer. To determine these twelve sample points we settle the condition of the Bayes test that $\alpha(B_{k'})$ must take the lowest possible value. An extensive analysis of the values taken by f_H in the sample sapace permit us to conclude that the critical regions of the $F_{\alpha_1}E$ test and of the $B_{k'}E$ test are the same (see Table IV). That is, $CF_{\alpha_1}=CB_{k'}$ or the two test functions F_{α_1} and $B_{k'}$ are equal. Clearly this does not mean that the $F_{\alpha_1}E$ test and the $B_{k'}E$ test are the same. Although $F_{\alpha_1}=B_{k'}$ the stated values $\alpha(F_{\alpha_1})=0.2251$ and $\alpha(B_{k'})$ may be very different.

To evaluate the constant k' we recall that

$$B_{k'}(x,t)=1 \quad \text{if and only if} \quad \frac{\binom{t}{x}\binom{10-t}{5-x}}{\binom{10}{5}11}<k'.$$

The boundary points of $CB_{k'}$ are the points $(x,t)=(4,5)$ and $(x,t)=(1,5)$ which produce the following inequality

$$\frac{\binom{5}{4}\binom{5}{1}}{\binom{10}{5}11}<k'\rightarrow k'>0.3247.$$

On the other hand, the boundary points of the acceptation region are $(x,t)=(5,8)$ and $(x,t)=(0,2)$ which produce the following inequality

$$\frac{\binom{8}{5}\binom{2}{0}}{\binom{10}{5}11} > k' \Rightarrow k' < 0.7272.$$

The conclusion is that for any value of k' in the interval (0.3247, 0.7272), the test function $B_{k'}$ is equal to the test function F_{α_1}.

We notice now that $\alpha(B_{k'})=0.0577$ and $\beta(B_{k'})=0.6667$ meaning that neither Eq. (5) nor Eq. (6) are exactly satisfied. The reason for this is that the values of $\alpha(B_{k'})$ and $\beta(B_{k'})$ are obtained from the use of f_H and f_A, which are discrete probability functions, and alfa$B_k=$ 0.0673 and beta $B_{k'}=0.6731$ were obtained by simulation.

Comparing, in terms of Definition 1, the $F_{\alpha_1}E$ test and the $B_{k'}E$ test, for any $k' \in (0.3247, 0.7272)$, we obtain

i) $\quad 0.0096=|\alpha(B_{k'})-\text{alfa}\,B_{k'}|<|\alpha(F_{\alpha_1})-\text{alfa}\,F_{\alpha_1}|=0.1578$

and

ii) $\quad \text{ALFA}\,B_{k'}+k'\text{BETA}\,B_{k'}=\text{ALFA}\,F_{\alpha_1}+k'\text{BETA}\,F_{\alpha_1}.$

Again, this indicates that the BE test is better than the FE test.

It is interesting to emphasize that $0.6694=\alpha(B_1)+\beta(B_1)<\alpha(B_{k'})+\beta(B_{k'})=0.7244$ and that, for any $k' \in (0.3247, 0.7272)$, $\alpha(B_1)+k'\beta(B_1)>\alpha(B_{k'})+k'\beta(B_{k'})$.

5. WHY IS THE *FE* TEST SO CONSERVATIVE?

Fisher (1939) has asserted that the marginals $(m,n,t,M-t)$ carries no information about the veracity of one of the alternative hypotheses, H or A. This reasoning leads us to consider the Hypergeometric distribution as the basic model, leaving out the probability distribution of the marginal t. From the definition of the test function F_α, given by expression (4) in Section 2, we realize that it strongly depends on the value of t through the function $p(x,t)=\sum_d L(1,t|d)$, where the sum is over the set $C_x^t=\{d; L(1,t|d)\leqq L(1,t|x)\}$. The value of $p(\cdot,\cdot)$ at the observed sample (x,t) is known as the p-value of (x,t). For each fixed value of t, we can compute the conditional probability of rejecting H given this value of t (Berkson, 1978). That is,

$$\alpha(F_\alpha|t)=\sum_x L(1,t|x)$$

where the sum is over the set $C^t=\{x; F_\alpha(x,t)=1\}$. The values of $\alpha(F_\alpha|t)$ for $\alpha=\alpha_1=0.2251$ and $m=n=5$ are presented in Table XIII.

TABLE XIII

Values of $\alpha(F_\alpha|t)$ in the case of $\alpha=0.2251$ and $m=n=5$

t	$\alpha(F_{\alpha_1}|t)$
0	0
1	0
2	0
3	0.1667
4	0.0476
5	0.2063
6	0.0476
7	0.1667
8	0
9	0
10	0
SUM	0.6349

None of the values of $\alpha(F_{\alpha_1}|t)$ presented in Table XIII are equal to $\alpha(B_{k'})=0.0577$, the value of ALFA stated by the $B_{k'}E$ test. However, the average of the values of Table XIII is

$$\bar{\alpha}(F_{\alpha_1})=\frac{1}{11}\sum_{t=0}^{10}\alpha(F_{\alpha_1}|t)=0.0577.$$

That is, $\bar{\alpha}(F_{0.2251})=\alpha(B_{k'})=0.0577$. Hence, if in the $F_{\alpha_1}E$ test we substitute the stated value of ALFA, α_1, by $\bar{\alpha}(F_{\alpha_1})$ we obtain the $B_{k'}E$

test. Since $\bar{\alpha} < \alpha_1$, we conclude that the *FE* test is conservative because it states a value for the first-kind-of-error probability higher than it really must be. To consider $\bar{\alpha}$ as the first-kind-of-error probability is equivalent to consider a discrete uniform distribution for t (Krewski, Brennan and Bickis, 1984) and state the first-kind-of-error probability as the mean of $\alpha(F_\alpha|t)$. To compute the second-kind-of-error probability Krewski, Brennan and Bickis (1984) considered a uniform prior in the unit square as a mixture distribution of (p, q) for

$$\binom{m}{n} p^x(1-p)^{m-x} \binom{n}{y} q^y(1-q)^{m-y}.$$

6. FINAL OBSERVATIONS

In spite of our analysis being restricted to particular sample sizes, the generalization of the conclusions is natural due to the general construction of the two tests. In fact, by adjusting the statement about the first kind of error in the *FE* test, there is no difference between the *FE* and the *BE* tests.

The difference between $\alpha(B_k)$ and alfaB_k observed is due to the decision of simulating samples for few parametric points. This difference will decrease if in place of $\{0.1, 0.2, \ldots, 0.9\}$ we consider $\{0.01, 0.02, \ldots, 0.99\}$ to choose the values of p and q. However, the time consuming in the computer would increase too much. By its own construction, the values stated by the *BE* test are in fact the actual value of the frequencies of the first and second kind of errors. This is the real advantage of using the *BE* test (equivalently, the *FE* test adjusted by $\bar{\alpha}$).

Acknowledgement

The authors are thankful to Professor Clovis A. Peres for his helpful suggestions, mainly in Sections 4 and 5.

References

Basu, D. (1979). Discussion of Joseph Berkson's paper "In dispraise of the exact test". *Journal of Statistical Planning and Inference* 3, 189–192.

Berkson, J. (1978). Do the marginal of the 2×2 table contain relevant information respecting the table proportions? *Journal of Planning and Inference* **3**, 193–197.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association* **77**(379), 605–613.

De Groot, M. H. (1975). *Probability and Statistics.* Addison-Wesley, London.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, A* **98**(1), 39–54.

Irony, T. Z. (1984). Testes exatos para tabelas 2×2: Bayes v. Fisher. São Paulo. 133p. Dissertação (Mestrado)-IME. Universidade de São Paulo.

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

Kempthorne, O. (1980a). Foundations of statistical thinking and reasoning: Part I. *DMS Newsletter* **68**(5) (published by CSIRO Division of Math. and Statist., Australia).

Kempthorne, O. (1980b). Foundations of statistical thinking and reasoning: Part II. *DMS Newsletter* **69**, 3–7.

Krewski, D., Brennan, J. and Bickis, M. (1984). The power of the Fisher permutation test in $2 \times k$ tables. *Communications in Statistics: Simulation and Computation* **13**(4), 433–448.

Lindley, D. V. (1982). The Bayesian approach to statistics. In: T. Oliveira and B. Epstein (eds.), *Some Recent Advances in Statistics.* vol. 2, pp. 65–87. Academic Press, New York.

Pereira, C. A. de B. (1984). Teste de hipóteses definidas em espaços de diferentes dimensões: Visão Bayesiana e interpretação clássica. São Paulo. 107p. Tese (Livredocência)-IME-Universidade de São Paulo.

Yates, F. (1984). Tests of significance for 2×2 contingency tables. *Journal of the Royal Statistical Society, A* **147**(3), 426–463 (with discussions).