Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Statistics & Data Analysis 47 (2004) 745–756

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Comparing the asymptotic power of exact tests in 2 × 2 tables

A. Martín Andrés[a,*], A. Silva Mato[b], J.M. Tapia García[a], M.J. Sánchez Quevedo[c]

[a] *Bioestadística, Facultad de Medicina, Universidad de Granada, Granada 18071, Spain*
[b] *Bioestadística, Facultad de Medicina, Universidad de Alcalá, Madrid, Spain*
[c] *Estadística, Universidad de Cádiz, Cádiz, Spain*

Received 19 August 2003; received in revised form 18 November 2003

Abstract

A 2 × 2 table may arise from three types of sampling, depending on the number of previously fixed marginals, and may yield three possible, differing, probabilistic models. From the unconditional point of view each model requires a specific solution but, within each model, the calculation time increases as the test procedure chosen is more powerful and, between the models, the calculation time decreases in the number of marginals fixed. Moreover, each model yields a test which is generally more powerful than the test of any other model with a larger number of marginals fixed. The condition under which a less powerful test, of the same or a different model, can substitute a more powerful test with a loss of power lower than 2% is determined. It is concluded that the Fisher exact test can be used as an approximation to Barnard's exact test for a table with 0 or 1 fixed marginals, when the sample size is $\geqslant 100$ or when the smaller sample size is $\geqslant 80$, respectively. Similarly, Barnard's exact test for a table with 1 fixed marginal can be used as an approximation of the same test for a table with 0 fixed marginals, when the sample size is $\geqslant 50$.

Keywords: Barnard's exact test; Conditional test; Fisher's exact test; Power; Unconditional test; 2 × 2 tables

1. Introduction

The test of independence in a 2 × 2 table is a classic problem in statistics and about which hundreds of articles have been published (Martín Andrés, 1997). The test is very

* Corresponding author.
E-mail address: amartina@ugr.es (A. Martín Andrés).

0167-9473/$ - see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/j.csda.2003.11.012

Table 1
Presentation of results in the form of a 2 × 2 table

	Characteristic A		
	YES	NO	Total
Characteristic B			
YES	x_1	y_1	n_1
NO	x_2	y_2	n_2
Total	a_1	a_2	n

common in applied statistics, especially in the field of clinical trials (the comparison of the proportions of successes for two treatments) and of epidemiology (the analysis of the influence of the presence of a risk factor in the presentation of an illness). Whatever the case, one should apply the most powerful test possible, especially in clinical trials. However, this is often difficult to achieve, due to the existence of conceptual and computational problems. For this reason we decided to analyze the problem in detail.

A 2×2 table is a presentation of results as seen in Table 1, but these may have been obtained under three types of sampling (among others): Samplings I, II or III according to whether the value of n, n_i, or n_i and a_i respectively, has been fixed. Each sampling produces a different possible model (Models I, II or III, respectively) based on the fact that the sole random variable(s) (r.v. in the following) in the problem are (x_1, y_1, x_2), (x_1, x_2) or (x_1), a multinomial, double-binomial or generalized hypergeometric r.v., respectively. Therefore, the probability of obtaining results like those of Table 1 (for each model) will be:

$$P(x_1, y_1, x_2) = C(n; x_1, y_1, x_2, y_2) \times p_{11}^{x_1}\ p_{12}^{y_1}\ p_{21}^{x_2}\ p_{22}^{y_2},$$

$$P(x_1, x_2) = C(n_1; x_1) \times C(n_2; x_2) \times p_1^{x_1}\ (1-p_1)^{y_1}\ p_2^{x_2}\ (1-p_2)^{y_2},$$

$$P(x_1) = C(n_1; x_1) \times C(n_2; x_2)\varphi^{x_1} \Bigg/ \sum_{h=r}^{s} C(n_1; h) \times C(n_2; a_1 - h) \times \varphi^h\ , \tag{1}$$

where $C(n; x_1, y_1, x_2, y_2) = n!/\{x_1!y_1!x_2!y_2!\}$, $C(a; b) = a!/\{b!(a-b)!\}$, $r = \max(0; a_1 - n_2)$, $s = \min\ (a_1; n_1)$, and p_{ij}, p_i and φ represent the parameters of each model.

If the aim of the experiment is to contrast hypothesis H: “the characteristics A and B are independent” against an alternative (one- or two-tailed) hypothesis K, then, under each model, H is $p_{ij} = p_i.p._j$, $p_1 = p_2$ and $\varphi = 1$, respectively. Therefore, if we agree that $p_1. = q$, $p._1 = p$ and $p_1 = p_2 = p$, then

$$P(x_1, y_1, x_2|\mathrm{H}) = C(n; x_1, y_1, x_2, y_2) \times p^{a_1}(1-p)^{a_2} q^{n_1}(1-q)^{n_2},$$

$$P(x_1, x_2|\mathrm{H}) = C(n_1; x_1) \times C(n_2; x_2) \times p^{a_1}(1-p)^{a_2},$$

$$P(x_1|\mathrm{H}) = C(n_1; x_1) \times C(n_2; x_2)/C(n; a_1). \tag{2}$$

In order to perform the test to an target error α, it is necessary to define a critical region (CR in the following). Let CR(α|I), CR(α|II) and CR(α|III) be the CR for each model. The real error of the test will then be

$$\alpha^*(p,q) = \sum_{\mathrm{CR}(\alpha|\mathrm{I})} P(x_1,y_1,x_2|\mathrm{H}), \quad \alpha^*(p) = \sum_{\mathrm{CR}(\alpha|\mathrm{II})} P(x_1,x_2|\mathrm{H}),$$

$$\alpha^* = \sum_{\mathrm{CR}(\alpha|\mathrm{III})} P(x_1|\mathrm{H}), \tag{3}$$

where p and q are two nuisance parameters. One way of eliminating these is by maximization (Barnard, 1947), which produces the *unconditional test.* Then, the size of the test will be

$$\alpha_{\mathrm{I}}^* = \max_{p,q} \alpha^*(p,q), \quad \alpha_{\mathrm{II}}^* = \max_{p} \alpha^*(p), \quad \alpha_{\mathrm{III}}^* = \alpha^*, \tag{4}$$

where α_{I}^*, α_{II}^*, $\alpha_{\mathrm{III}}^* \leqslant \alpha$. (As can be seen, maximization is not necessary in Model III because there are no nuisance parameters, so size and real error coincide.) Another way of eliminating nuisance parameters is by conditioning in the really observed marginals (Fisher, 1935), which produces the conditional test. Here, the three models have the same solution (α_{III}^*): the well-known Fisher's exact test. This is not the place to argue the appropriateness of one solution or the other: the reader wishing to know more may refer to the discussion in Yates (1984) and the review by Martín Andrés (1991).

In order to form the CR it is necessary to define a criterion for ordering the points in the sample space. To this end, it is sufficient to supply a direction-sensitive statistic $T = T(x_1, y_1, x_2, y_2)$, so that the points in the sample space are introduced into the CR from the smallest to the largest value of T. If T_0 is the value of T in the last point introduced into the CR, then $\mathrm{CR}(\alpha|\mathrm{I}) = \{x_1, y_1, x_2 | T \leqslant T_0\}$ and α_{I}^* is the p-value of the table yielding the value T_0. It is similarly for other cases.

Generally speaking, unconditional tests are more powerful than conditional ones: generally α_{I}^*, $\alpha_{\mathrm{II}}^* \leqslant \alpha_{\mathrm{III}}^*$. Unfortunately, the difficulties of calculation increase proportionately as one goes down the models. Thus, today there is no problem in calculating α_{III}^* for any table (see the StatXact package for example), but α_{II}^* (α_{I}^*) can only be calculated in moderate (small) samples. What is more, for each of the models there are various possible T-statistics which give rise to tests that are either more or less powerful; unfortunately, the more powerful a test is, the greater the calculation difficulties it produces. (Although no UMP tests exist in these models, "optimal" tests do exist in the sense that they are generally more powerful.) The aim of this paper is to reply to the general question: when can a non-optimal test be used without obtaining a excessive loss of power? The question must be answered "within" each model (by evaluating the optimal T versions) and "between" the models (by evaluating a totally or partially conditional test against an unconditional one).

2. Long-term power

Let there be two test procedures: O (the optimal or generally more powerful) and A (the alternative or less powerful, but with fewer calculation difficulties). Let $\Theta(O)$ and $\Theta(A)$ be their powers for a given alternative. The loss of power for not using the optimal test will be $\Delta\Theta(O,A)=\Theta(O)-\Theta(A)$. When $\Delta\Theta(O,A)\geqslant 0$ is small ($\leqslant 2\%$, for example) it will be understood that A can substitute O. The aim then is to determine the circumstances where this can occur. Unfortunately the answer will be confusing unless a simpler definition of power than the traditional one is adopted. In effect, given an error α, the power Θ for each model and for a given alternative is

$$\Theta(p_{11},p_{12},p_{21})=\sum_{\mathrm{CR}(\alpha|\mathrm{I})}P(x_1,y_1,x_2),\qquad \Theta(p_1,p_2)=\sum_{\mathrm{CR}(\alpha|\mathrm{II})}P(x_1,x_2),$$

$$\Theta(\varphi)=\sum_{\mathrm{CR}(\alpha|\mathrm{III})}P(x_1) \tag{5}$$

with the probabilities $P(\cdot)$ given by expression (1). Because in Model II (for example), Θ depends on (p_1,p_2), then $\Delta\Theta$ will reach a different value for each value of (p_1,p_2), and this means the conclusions will be confused (because, as we have said, there are no UMP tests).

In order to avoid this, in what follows the definition of long-term power by Martín Andrés and Tapia Garcia (1999) and Martín Andrés and Silva Mato (1994) will be adopted for the case of Models I and II, respectively (for a more detailed justification of this concept, see the articles cited). The definition is based on the assumption that the unknown parameters (p_i or p_{ij}) are distributed in the long term as a uniform r.v. at [0,1]. In the case of a two-tailed test, this makes the definition of long-term power very "intuitive":"number of points in the CR"/ "number of points in the sample space (SS)". Written in symbols, if X is the test procedure that has been chosen, then

$$\Theta(X)=\text{Card CR}(X)/\text{Card SS}\quad\text{(two-tailed test)}, \tag{6}$$

where $\mathrm{CR}(X)$ refers to the CR which yields procedure X and Card SS $=(n+1)(n+2)(n+3)/6$ in Model I or Card SS $=(n_1+1)(n_2+1)$ in Model II. In the case of a one-tailed test for the alternative with positive association (K: odds-ratio >1), the expression is more complicated, so now

$$\Theta(X)=2\times\sum_{\mathrm{CR}(X)}P_{\bar{\mathrm{F}}}(x_1,y_1,x_2,y_2)/\text{Card SS}\quad\text{(one-tailed test)}, \tag{7}$$

where $P_{\bar{\mathrm{F}}}\,(x_1,y_1,x_2,y_2)=\sum_{h=r}^{x_1}C(n_1+1;h)\times C(n_2+1;a_1+1-h)/C(n+2;a_1+1)$ is the p-value of Fisher's exact test for the alternative with negative association (K: $p_1<p_2$) in the table of frequencies x_1, y_1+1, x_2+1, y_2. Finally, the increase in power obtained by using method O in place of method A will be

$$\Delta\Theta(O,A)=\{\text{Card CR}(O)-\text{Card CR}(A)\}/\text{Card SS}\quad\text{(two-tailed)},$$

$$\Delta\Theta(O,A) = 2 \times \left\{ \sum_{\mathrm{CR}(O)-\mathrm{CR}(A)} P_{\bar{\mathrm{F}}}(x_1, y_1, x_2, y_2) - \sum_{\mathrm{CR}(A)-\mathrm{CR}(O)} P_{\bar{\mathrm{F}}}(x_1, y_1, x_2, y_2) \right\} \Big/ \mathrm{Card\ SS\ (one\text{-}tailed)}\,. \tag{8}$$

For the following, the comparisons of power will always be effected for the classic error $\alpha = 5\%$. The results for $\alpha = 1\%$ and 10% may be requested from the authors.

3. Analysis within Model III

The most usual ordering statistic in the case of Model III is $T = P(x_1|\mathrm{H})$. This yields the generally most powerful two-tailed test, although there are other statistics which are almost equally as powerful (Martín Andrés and Herranz Tejedor, 1995); for tests with one tail, all the non-randomized definitions yield the same solution. Given that in the present model the intensity of the calculation is the same for any definition of T, for the following it will be understood that Model III is applied under the previous definition (which we shall call procedure P).

4. Analysis within Model II

Martín Andrés et al. (1998) showed that the T statistics which are generally more powerful in Model II are, in order from best to worst, those we shall call B, B' and FM. Order B refers to the original order of Barnard (1947): $B = \alpha^*_{\mathrm{II}}$, where α^*_{II} refers to the p-value of the table (the property of convexity which will be mentioned later means that the starting point must be $x_1 = 0$ and $x_2 = n_2$ in the one-tailed test for the negative association alternative). Order B' is an approximation to the order of Barnard because $B' = \alpha^*(\hat{p})$, with $\hat{p} = a_1/n$. Order $\mathrm{FM} = \alpha^*_{\mathrm{III}}$ (p-mid) refers to the mid-p-value of the Fisher exact test: the value α^* of expression (3) less half the probability of the last point introduced. Unfortunately, calculation times are very high for B, high for B' and moderate for FM. When can B' or FM be used instead of B?

It has been shown (Martín Andrés and Silva Mato, 1994) that $\Theta(X)$ varies strongly with the values of n_1, $K = n_2/n_1 \geqslant 1$, α and with the number of tails in the test. Because of this $\Delta\Theta(B,A)$ has been calculated—where $A = B'$ or $A = \mathrm{FM}$—for $n_1 = 10(10)80$, $K = 1(0.5)3$ and 4, $\alpha = 5\%$, and one- (two-) tailed test. Table 2 gives the value of $\Delta\Theta$ for the case of $\alpha = 5\%$. It can be seen that the methods B' and FM can be used always, because the loss of power is small ($\Delta\Theta \leqslant 2\%$, frequently quite smaller).

5. Analysis within Model I

Here something similar to Model II occurs: Martín Andrés and Tapia Garcia (1999) showed that the generally most powerful statistics T are, in order from best to worst,

Table 2
Model II vs. Models II and III

One tail							Two Tails						
n_1	$K = 1.0$	$K = 1.5$	$K = 2.0$	$K = 2.5$	$K = 3.0$	$K = 4.0$	n_1	$K = 1.0$	$K = 1.5$	$K = 2.0$	$K = 2.5$	$K = 3.0$	$K = 4.0$
10	45.76	49.54	52.33	53.39	54.68	56.31	10	38.02	43.18	45.89	48.25	49.27	51.00
	0	0	0.84	0	0.57	0		0	0	0	0	0	0
	1.57	0	0.81	−0.02	0.55	0.84		0	0	1.60	0.70	1.18	0.89
	7.94	9.91	7.50	7.42	6.78	6.40		9.92	4.54	4.33	5.59	5.28	3.55
20	60.32	63.2	65.34	66.29	67.23	68.19	20	53.97	58.06	60.16	61.44	62.45	63.49
	0	−0.29	0.23	0.36	1.19	0.79		0	0	0	0	0	−0.12
	0.85	0.87	0.22	−0.01	0.30	0.12		0	0.30	1.16	0.19	0.47	0.59
	6.54	5.62	5.13	4.85	4.50	3.95		5.44	2.45	2.78	3.36	2.81	2.11
30	67.05	69.78	71.37	72.26	72.85	73.91	30	61.60	65.08	66.84	68.00	68.70	69.74
	0	−0.13	0	0.81	0.74	0.61		0	0.28	0	0	0	0
	−0.01	0	0.30	0.16	0.20	0.16		1.04	0.56	0.53	0.94	0.36	0.43
	5.21	3.64	3.65	3.35	3.13	3.18		4.78	2.53	1.69	2.04	1.99	1.76
40	71.4	73.59	75.04	75.88	76.44	77.23	40	66.39	69.49	71.06	72.01	72.69	73.53
	0.22	−0.07	0.80	0.60	0.54	0.41		0	0.16	0.06	0	0	−0.03
	0.45	0.38	0.23	0.18	0.12	0.06		0.71	0.48	0.66	0.43	0.33	0.60
	4.00	2.77	2.71	2.69	2.36	2.38		3.33	1.68	1.62	1.64	1.29	1.33
50	74.28	76.34	77.58	78.39	78.85	79.57	50	69.9	72.5	73.97	74.82	75.42	76.19
	0	−0.05	0.63	0.57	0.37	0.24		0	0	0	−0.03	0	0
	0.37	0.24	0.11	0.33	0.42	0.08		0.54	0.21	0.51	0.37	0.23	0.43
	3.32	2.27	2.19	2.24	2.04	2.00		3.00	1.19	1.36	1.21	0.99	1.15
60	76.47	78.43	79.5	80.17	80.68	81.3	60	72.29	74.87	76.11	76.95	77.47	78.21
	0	−0.03	0.49	0.36	0.34	0.22		0	0.04	0	0	0	0
	0.31	0.21	0.26	0	0.31	0.20		0.21	0.22	0.35	0.26	0.39	0.31
	2.27	2.11	2.00	1.79	1.73	1.62		2.58	1.19	1.19	0.93	0.99	1
70	78.09	80.01	81.00	81.66	82.1	82.67	70	74.35	76.64	77.83	78.62	79.10	79.78
	0	0	0.46	0.40	0.32	0.18		0	0.03	0	0.02	−0.01	0
	0.41	0.17	0.13	0.02	0.22	0.09		0.52	0.19	0.34	0.26	0.33	0.31
	2.16	1.88	1.72	1.60	1.52	1.50		2.26	0.98	1.01	0.82	0.92	0.79
80	79.52	81.23	82.22	82.84	83.23	83.97	80	76.02	78.11	79.24	79.95	80.42	81.04
	0.05	0.52	0.43	0.34	0.27	0.18		0.06	0.02	−0.02	0.03	0	0
	0.17	0.05	0.12	0.29	0.08	0.08		0.18	0.22	0.35	0.23	0.29	0.23
	2.04	1.62	1.52	1.44	1.35	1.37		2.13	0.95	0.90	0.73	0.81	0.70

Long-term power (Θ) for method B in Model II (1st entry) and absolute increase of power of the same one ($\Delta\Theta$) regarding the methods B' (2nd entry) and FM (3rd entry) in Model II and regarding the method FC in Model III (4th entry) for $\alpha = 5\%$.

$B = \alpha_{\text{I}}^*, B' = \alpha_{\text{I}}^*(\hat{p}, \hat{q})$—where $\hat{p} = a_1/n$ and $\hat{q} = n_1/n$—, and $\text{FM} = \alpha_{\text{III}}^*$ (p-mid). However the calculation difficulties are now much more serious, and this means we can only study the cases where $n = 10(10)50$. Again B is much slower than B', which in turn is slower than FM. Table 3 shows the values of $\Delta\Theta(B, X)$ for $X = B'$ and $X = \text{FM}$. It can be seen that for $n = 50$, $\Delta\Theta(B, B')$ continues to be $> 2\%$ in the one-tailed test, and continues to grow in the two-tailed test, for which reason it is difficult to say when it will perform acceptably. On the other hand, and also for $n = 50$, the growth of $\Delta\Theta(B, \text{FM})$ has slowed and is very small for the one-tailed test, while in the two-tailed test it actually decreases, and in both cases $\Delta\Theta(B, \text{FM}) \ll 2\%$. Therefore one can tentatively conclude that in the one- (two-) tailed test method FM can always be used (if $n \geqslant 50$) with a loss of power lower than 2%.

Table 3
Model I vs. Models I, II and III

n	10	20	30	40	50
One tail	21.90	39.21	49.16	55.17	59.14
	−2.61	0.24	1.84	2.32	2.22
	−1.40	1.04	1.17	1.28	1.34
	3.27	2.12	2.75	2.22	1.79
	11.47	11.00	9.62	8.37	7.24
Two tails	16.80	34.90	43.73	49.59	53.87
	−2.08	1.36	1.83	1.52	2.47
	0.02	2.49	2.27	2.61	1.18
	2.11	2.26	2.20	1.83	1.47
	5.59	8.81	6.89	5.79	4.45

Long-term power (Θ) for method B in Model I (1st entry) and absolute increase of power of the same one ($\Delta\Theta$) regarding the methods B' (2nd entry) and FM (3rd entry) in Model I, regarding the method BC in Model II (4th entry) and regarding the method FC in Model III (5th entry) for $\alpha = 5\%$.

6. Model II vs. Model III

It is usually said that the Fisher exact test (Model III), from the unconditional point of view, is a valid but conservative test. From the perspective of Model II, this was justified by Pearson (1947), Martín Andrés (1991) and Silva Mato and Martín Andrés (1995), but here the problem will again be analyzed, with some modifications.

The SS of Model II is formed by all the tables (x_1, x_2) where $0 \leqslant x_i \leqslant n_i$. For each value $0 \leqslant a_1 \leqslant n$, the Fisher exact test (to target error α) yields a $\mathrm{CR}(\alpha|\mathrm{II}) = \mathrm{CR}(a_1)$ with a real error given by expression (4) and which, because it varies with a_1, we shall call $\alpha^*_{\mathrm{III}}(a_1) \leqslant \alpha$. Let $\mathrm{CR(FC)} = \bigcup_{a_1} \mathrm{CR}(a_1)$ and $\alpha^*_{\mathrm{III}} = \max_{a_1} \alpha^*_{\mathrm{III}}(a_1) \leqslant \alpha$. The aim is to demonstrate that CR(FC) is a valid CR for Model II. The proof is based on the fact that $P(x_1, x_2|\mathrm{H}) = P(a_1|\mathrm{H}) \times P(x_1|\mathrm{H})$, where $P(a_1|\mathrm{H}) = C(n; a_1) \times p^{a_1}(1-p)^{a_2}$. The real error $\alpha^*(p)$ for the previous CR(FC) will be, using expression (3):

$$\alpha^*(p) = \sum_{a_1 \in \mathrm{CR(FC)}} P(a_1|\mathrm{H}) \sum_{x_1 \in \mathrm{CR}(a_1)} P(x_1|\mathrm{H}) = \sum_{a_1 \in \mathrm{CR(FC)}} P(a_1|\mathrm{H}) \alpha^*_{\mathrm{III}}(a_1)$$

$$\leqslant \alpha^*_{\mathrm{III}} \sum_{a_1 \in \mathrm{CR(FC)}} P(a_1|\mathrm{H}) \leqslant \alpha^*_{\mathrm{III}} \leqslant \alpha \quad (\forall p) \Rightarrow \alpha^*_{\mathrm{II}} \leqslant \alpha^*_{\mathrm{III}} \leqslant \alpha, \tag{9}$$

so implying that CR(FC) is a valid CR for Model II (because $\alpha^*_{\mathrm{II}} \leqslant \alpha$). On the other hand, let a_1^* be the value of a_1 in which $\alpha^*_{\mathrm{III}}(a_1^*) = \alpha^*_{\mathrm{III}}$, and let x_1^* be the last point introduced into $\mathrm{CR}(a_1^*)$ (it is also the last point introduced into CR(FC)). The p-value for the said point is α^*_{III} (α^*_{II}) according to Model III (II), and $\alpha^*_{\mathrm{II}} \leqslant \alpha^*_{\mathrm{III}}$ according to expression (9). Hence the Fisher exact test is conservative. But the result needs to be modified:

(1) The way to build the previous CR(FC) gives rise to an ordering method which we shall call FC. A classic order in Model II (Boschloo, 1970) is given by the

statistic $F=\alpha^*_{\mathrm{III}}$ (order from the smallest to the largest p-value of the Fisher exact test). As CR(FC) is licit, then CR(F)⊂CR(FC)—where CR(F) is the CR which produces the order F—and therefore F will produce a uniformly more powerful test than the one produced by the order FC. Consequently, it can be said that the unconditional test F for Model II is UMP with regard to the one-tailed Fisher exact test (excluding the two-tailed test for the reason given below).

(2) However, the above result is not correct for any other order in Model II (not even for B, which is generally the most powerful), because the order which induces B (for example) is not the one which induces FC, and so CR(FC) and CR(B) are not in an inclusion relation. Hence, it can only be stated that B is generally more powerful than FC (because B is generally more powerful than F and the latter is UMP with regard to FC).

(3) Moreover, the conclusion obtained in (1) is only valid for the one-tailed test. The reason is that for a test to be valid in Model II, it must be inferentially logical. Barnard (1947) indicated that the test should respect the properties of:

$$\text{Symmetry: if } (x_1,x_2)\in \mathrm{CR} \Rightarrow (y_1,y_2)\in \mathrm{CR} \text{ (for } K: p_1 \neq p_2\text{)};$$

$$\text{Convexity: if } (x_1,x_2)\in \mathrm{CR} \Rightarrow (x_1,x_2+1),(x_1-1,x_2)\in \mathrm{CR} \quad \text{(for } K\text{: } p_1 < p_2\text{)}.$$

As a one-tailed test, order FC respects convexity (Hajek and Havranek, 1978), but not as a two-tailed test (there are counter examples to that effect), and so the two-tailed Fisher exact test is not only generally conservative, but is also incoherent (from the perspective of Model II). Precisely for this reason, the condition of convexity is previously imposed on order F (as a two-tailed test). However order F as a two-tailed test (and thus, the method FC it induces) respects the new (and desirable) property of:

$$\begin{aligned}\text{Diagonality: } & \text{if } (x_1|a_1,n_1)\in \mathrm{CR} \Rightarrow (x_1-1|a_1,n_1)\in \mathrm{CR} \text{ (for } \hat{p}_1 < \hat{p}_2\text{)},\\ & \text{if } (x_1|a_1,n_1)\in \mathrm{CR} \Rightarrow (x_1+1|a_1,n_1)\in \mathrm{CR} \text{ (for } \hat{p}_1 > \hat{p}_2\text{)},\end{aligned}$$

which is a consequence of the fact that order P for Model III does respect the property of convexity. This indicates that the CR which they produce does not present gaps (even when it is not convex).

In order to evaluate how much power is lost by using test FC instead of test B, one proceeds as in Section 4. Table 2 shows the results. It can be seen that the increase in power decreases when n_1 increases, and that the case of $K=n_2/n_1=1$ behaves worse than the others (especially in the two-tailed test). If a loss of power of approximately 2% or less is considered irrelevant, then method FC can substitute B where $n_1 \geqslant 80$ ($n_1 \geqslant 70$ if $K>1$) in the one-tailed test, or where $n_1 \geqslant 80$ ($n_1 \geqslant 40$ if $K>1$) in the two-tailed test.

7. Model I vs. Model III

For each value $0 \leqslant n_1 \leqslant n$, in the previous section the Fisher exact test produced a CR(FC) and an error $\alpha^*_{\rm III} \leqslant \alpha$. Given that they both vary with n_1, let us note them here as CR(n_1) and $\alpha^*_{\rm III}\ (n_1) \leqslant \alpha$. For the present Model I, the new CR will be $\mathrm{CR(FC)} = \bigcup_{n_1} \mathrm{CR}(n_1)$ with error $\alpha^*_{\rm III} = \max_{n_1} \alpha^*_{\rm III}(n_1)$. Pearson (1947) indicated that $P(x_1, y_1, x_2|\mathrm{H}) = P(n_1|\mathrm{H}) \times P(x_1, x_2|\mathrm{H})$, where $P(n_1|\mathrm{H}) = C(n; n_1) \times q^{n_1}(1-q)^{n_2}$, so that the error $\alpha^*(p,q)$ for the new CR(FC), using expression (3), will be

$$\alpha^*(p,q) = \sum_{n_1 \in \mathrm{CR(FC)}} P(n_1|\mathrm{H}) \sum_{(x_1,x_2) \in \mathrm{CR}(n_1)} P(x_1,x_2|\mathrm{H}) = \sum_{n_1 \in \mathrm{CR(FC)}} P(n_1|\mathrm{H}) \alpha^*(p|n_1)$$

$$\leqslant \sum_{n_1 \in \mathrm{CR(FC)}} P(n_1|\mathrm{H}) \alpha^*_{\rm III}(n_1) \leqslant \alpha^*_{\rm III} \leqslant \alpha\ (\forall p) \Rightarrow \alpha^*_{\rm I} \leqslant \alpha^*_{\rm III} \leqslant \alpha, \qquad (10)$$

where $\alpha^*(p|n_1)$ refers to expression (9). Therefore, as in the previous section, the CR(FC) is a licit CR (because $\alpha^*_{\rm I} \leqslant \alpha$) and the Fischer exact test is conservative (because $\alpha^*_{\rm I} \leqslant \alpha^*_{\rm III}$). The rest of the observations for Section 6 can also be applied here. In fact, the properties to be verified by the CR of Model I are more numerous (Martín Andrés and Tapia Garcia, 1998).

Table 3 also yields the values of $\Delta\Theta(B, \mathrm{FC})$, which are obtained similarly to those of Section 5. It can be seen that, even for $n = 50$, the absolute increase in power ($\Delta\Theta$) is of the order of 7.2% (one tail) and 4.5% (two tails), while the relative increase—$\Delta\Theta/\Theta(B)$—is 12.2% and 8.3%, respectively, quantities that are too high to be assumed. Let us accept that the loss $\Delta\Theta = 2\%$ is reasonable. As $\Delta\Theta$ decreases in $n = 50$, the extrapolation of the results (which is the only thing feasible with our present calculation capacity) allows one to deduce that for $n = 100$ (one-tailed test) or $n = 70$ (two-tailed test) the Fisher exact test is acceptable (although as a two-tailed test it performs incoherently by not verifying the property of convexity).

8. Model I vs. Model II

It has already been said that the exact test for Model I is much more complex to calculate than the exact test for Model II. It is therefore advisable to study under what conditions the latter can substitute the former (without excessive loss of power). To this end, let us look at test B for Model II (which is generally the more powerful).

For each value $0 \leqslant n_1 \leqslant n$, the test B for Model II (to the target error α) gives a $\mathrm{CR}(\alpha|\mathrm{II}) = \mathrm{CR}(n_1)$ with a real error given by expression (4) and which, because it varies with n_1, we shall call $\alpha^*_{\rm II}\ (n_1) \leqslant \alpha$. Let $\mathrm{CR(BC)} = \bigcup_{n_1} \mathrm{CR}(n_1)$ and $\alpha^*_{\rm II} = \max_{n_1} \alpha^*_{\rm II}\ (n_1) \leqslant \alpha$. The aim is to demonstrate that CR(BC) is a valid CR for Model I. From the previous section, $P(x_1, y_1, x_2|\mathrm{H}) = P(n_1|\mathrm{H}) \times P(x_1, x_2|\mathrm{H})$, and so the real error $\alpha^*(p,q)$ for CR(BC) will be, using expression (3):

$$\alpha^*(p,q) = \sum_{n_1 \in \mathrm{CR(BC)}} P(n_1|\mathrm{H}) \sum_{(x_1,x_2)\in \mathrm{CR}(n_1)} P(x_1,x_2|\mathrm{H}) \leqslant \sum_{n_1 \in \mathrm{CR(BC)}} P(n_1|\mathrm{H})\alpha^*_{\mathrm{II}}(n_1)$$

$$\leqslant \alpha^*_{\mathrm{II}} \sum_{n_1 \in \mathrm{CR(BC)}} P(n_1|\mathrm{H}) \leqslant \alpha^*_{\mathrm{II}} \Rightarrow \alpha^*_{\mathrm{I}} \leqslant \alpha^*_{\mathrm{II}} \leqslant \alpha, \tag{11}$$

which implies that CR(BC) is a valid CR for Model I (because $\alpha^*_{\mathrm{I}} \leqslant \alpha$) and that the Barnard test for model II is conservative from the perspective of Model I (because $\alpha^*_{\mathrm{I}} \leqslant \alpha^*_{\mathrm{II}}$). For this reason, Barnard (1947) proposed using $\alpha^*_{\mathrm{II}}\ (n_1)$ as a handy solution to Model I, adding that not much is lost by doing so. Boschloo (1970) was of the same opinion, but modified this by saying that one could condition in n_1 (as before) or in a_1, so that there are really two possible sizes $\alpha^*_{\mathrm{II}}(n_1)$ and $\alpha^*_{\mathrm{II}}(a_1)$; hence he proposed that the value to be used should be

$$\alpha^*_{\mathrm{II}}\ (a_1,n_1) = \min\{\alpha^*_{\mathrm{II}}(a_1); \alpha^*_{\mathrm{II}}(n_1)\}. \tag{12}$$

In reality Boschloo proposed expression (12) for the case where order F is used, and Martín and Silva (1995) proposed this for the order B.

The proof for expression (11) is correct, but conclusion (12) is not necessarily so. To see this, let us look at a table with a value n_1. In it, expression (12) will have yielded a CR formed by the points in CR(n_1)—because, in these points $\alpha^*_{\mathrm{II}}(n_1) \leqslant \alpha$ and thus $\alpha^*_{\mathrm{II}}(a_1,n_1) \leqslant \alpha$—with a few extras: those where $\alpha^*_{\mathrm{II}}(n_1) > \alpha$ but $\alpha^*_{\mathrm{II}}(a_1) \leqslant \alpha$ or the set CR($+|n_1$). By moving this to the boundaries of expression (11):

$$\alpha^*(p,q) \leqslant \alpha^*_{\mathrm{II}} + \sum_{n_1} P(n_1|\mathrm{H}) \sum_{(x_1,x_2)\in \mathrm{CR}(+|n_1)} P(x_1,x_2|\mathrm{H}) \tag{13}$$

and so there is no guarantee that $\alpha^*(p,q) \leqslant \alpha$ and the test in expression (12) can be liberal.

The result for expression (11) is thus valid if one conditions in the rows or in the columns, that is, if one always uses $\alpha^*_{\mathrm{II}}(n_1)$ or if one always uses $\alpha^*_{\mathrm{II}}(a_1)$. However, in Model I there is no pre-arrangement as to what characteristic is situated in the rows, and so the researcher may choose $\alpha^*_{\mathrm{II}}(n_1)$ or $\alpha^*_{\mathrm{II}}(a_1)$ as he/she prefers. The only way to avoid this is to propose the solution:

$$\alpha^*_{\mathrm{II}}(a_1,n_1) = \max\{\alpha^*_{\mathrm{II}}(a_1);\ \alpha^*_{\mathrm{II}}(n_1)\}, \tag{14}$$

which guarantees that the CR obtained consists of a few points less than CR(n_1) and so expression (11) will be valid (although a more conservative test is obtained). We shall call the resulting method BC. Martín Andrés and Tapia Garcia (1998) indicated that every order defined in Model I should verify the properties of equivalence, convexity and symmetry (the last in the case of a two-tailed test). It is easy to see that order BC respects the first and last properties, but it has no need to respect the second, and so the outcome is the same with BC as with FC: it yields an incoherent test. However, order BC (again like FC) does respect the property of diagonality described in Section 6 (Sánchez Quevedo, 2002).

Table 3 also contains the values of $\Delta\Theta(\mathrm{B},\mathrm{BC})$ under the present Model I and with the criteria of Section 5. From this it can be deduced that the BC method can substitute B, with a loss of power lower than 2%, when $n \geqslant 50$ ($n \geqslant 40$) in the one- (two-) tailed test.

9. Conclusions

A 2×2 table may have been obtained under three different sampling schemes (only the total size is previously fixed; the totals for the rows are previously fixed; the totals for rows and for columns are previously fixed) resulting in three possible different models (Models I, II and III, respectively). From the unconditional perspective, each model requires a different analysis: the maximization tests for Models I and II; the Fischer exact test for Model III. Unfortunately, each successive model presents greater calculation difficulties; in fact the generally most powerful test for Model I is impracticable today for $n > 50$. Fortunately, each model can be solved conservatively using a higher model (by conditioning in the first), although this is at the expense of accepting some inferential incoherence in the two-tailed tests: these do not verify the property of convexity (although they do verify the property of diagonality, which guarantees that the critical regions show no gaps).

On the other hand, within each model there are several possible procedures for testing. Unfortunately, the more powerful the procedure, the more difficult the calculation becomes.

Whatever the case, in this paper we analyze when a conservative method can be used without losing too much power. Table 4 gives a summary of the conclusions arrived at. It can be seen that, very generally, the Fischer exact test can be used as an approximation to Barnard's exact test for Model I (Model II) when $n \geqslant 100$ ($\min n_i \geqslant 80$). Similarly, Barnard's exact test for Model II can be used as an approximation of the same test for Model I when $n \geqslant 50$.

Table 4
Alternatives model and method to the optimal for use in a 2×2 table and conditions for this

	Alternative model		
	I	II	III
Real model			
I	FM always FM if $n \geqslant 50$	B by (14) if $n \geqslant 50$ B by (14) if $n \geqslant 40$	P if $n \geqslant 100^*$ P if $n \geqslant 70^*$
II	Model I is more complex than Model II	B' and FM always	P if $n_1 \geqslant 80$ and $K = 1$ or if $n_1 \geqslant 70$ and $K > 1$ P if $n_1 \geqslant 80$ and $K = 1$ or if $n_1 \geqslant 40$ and $K > 1$

Note: (1) In the upper (lower) part of each cell are given the results for the one- (two-) tailed test. (2) The notation for the method (inside each cell) is that used in the text in the corresponding section. (3) When the real Model is III, there are no calculation difficulties (use method P). (4) $K = n_2/n_1 \geqslant 1$.

*Result obtained by extrapolation.

The programs for the exact unconditional tests may be copied from http://www.ugr.es/∼bioest/Software.htm (programs TMP.EXE for Model I and SMP.EXE for Model II).

Acknowledgements

This research has been funded by the Dirección General de Investigación (Spain), Grant BFM2000-1472.

References

Barnard, G.A., 1947. Significance tests for 2 × 2 tables. Biometrika 34, 123–138.

Boschloo, R.D., 1970. Raised conditional level of significance for the 2 × 2 Table when testing the equality of two probabilities. Statist. Neerlandica 24 (1), 1–35.

Fisher, R.A., 1935. The logic of inductive inference. J. Roy. Statist. Soc. A 98, 39–54.

Hajek, P., Havranek, T., 1978. Mechanizing Hypothesis Formation. Springer, Berlin, Heidelberg, p. 163.

Martín Andrés, A., 1991. A review of classic non-asymptotic methods for comparing two proportions by means of independent samples. Comm. Statist. Simulation Comput. 20 (2,3), 551–583.

Martín Andrés, A., 1997. Entry Fisher's exact and Barnard's tests. In: Kotz, Johnson, Read (Eds.), Encyclopedia of Statistical Sciences, updated Vol. 2, Wiley-Interscience, New York, pp. 250–258.

Martín Andrés, A., Herranz Tejedor, I., 1995. Is Fisher's exact test very conservative? Comput. Statist. Data Anal. 19, 579–591.

Martín Andrés, A., Silva Mato, A., 1994. Choosing the optimal unconditioned test for comparing two independent proportions. Comput. Statist. Data Anal. 17, 555–574.

Martín Andrés, A., Tapia Garcia, J.M., 1998. On determining the *P*-value in 2 × 2 multinomial trials. J. Statist. Plann. Inference 69 (1), 33–49 (erratum in 79 (1999) 365).

Martín Andrés, A., Tapia Garcia, J.M., 1999. Optimal unconditional test in 2 × 2 multinomial trials. Comput. Statist. Data Anal. 31 (3), 311–321.

Martín Andrés, A., Sánchez Quevedo, M.J., Silva Mato, A., 1998. Fisher's mid-*p*-value arrangement in 2 × 2 comparative trials. Comput. Statist. Data Anal. 29 (1), 107–115.

Pearson, E.S., 1947. The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table. Biometrika 34, 139–167.

Sánchez Quevedo, M.J., 2002. Algunas cuestiones acerca de los ensayos comparativos 2 × 2. Ph.D. Thesis. Departamento de Estadística e I.O. Universidad de Cádiz, Spain.

Silva Mato, A., Martín Andrés, A., 1995. Optimal unconditional tables for comparing two independent proportions. Biometrical J. 37 (7), 821–836.

Yates, F., 1984. Test of significance for 2 × 2 contingency tables. J. Roy. Statist. Soc. A 147 (3), 426–463.