Computational Statistics & Data Analysis 19 (1995) 579–591
North-Holland

Is Fisher's exact test very conservative?

A. Martín Andrés
Bioestadística, Facultad de Medicina, Universidad de Granada, Granada, Spain

I. Herranz Tejedor
Bioestadística, Facultad de Medicina, Universidad Complutense, Madrid, Spain

Received April 1993
Revised September 1993

Abstract: There are various two-tailed test versions of Fisher's exact test for analyzing a 2×2 table. In this paper, the optimal version is selected on the basis of the concept of mean power (arranging in order from the smallest to the largest hypergeometrical probability, and in the case of a tie, arranging in order from the largest to the smallest value of the odds-ratio), and this selection is as valid when considering it as a conditional test as it is when considering it as an unconditional test. The comparison of the power of the version selected (with one and two tails), with that of the more common unconditional tests (Barnard, 1947, and McDonald et al., 1977), shows that the loss of power produced by using Fisher's test is very slight in the majority of situations, and this is acceptable in return for the greater ease of computation and a more generic validity (for all types of sample).

Keywords: Fisher's exact test; Power; 2×2 Tables; Unconditional tests.

1. Introduction

A 2×2 table is the distribution of sample results like those shown in Table 1, but which essentially may be produced by three types of sample (Barnard, 1947): no fixed marginal (multinomial distribution), one fixed marginal (two binomials distribution) or two fixed marginals (hypergeometric distribution). Whatever the case, by conditioning on the obtained marginals and in the verification of H_0

Correspondence to: A. Martín Andrés, Catedra de Bioestadística, Facultad de Medicina, 18071 Granada, Spain.

SSDI 0167-9473(94)00013-9

Table 1
Results presented in the form of a 2×2 table

	A	$\overline{A}$	Totals
B	x_1	y_1	n_1
$\overline{B}$	x_2	y_2	n_2
Totals	a_1	a_2	n

(independence, homogeneity of proportions or randomness, respectively), the probability of a table like that given is:

$$P(X_1 = x_1 \mid n_i, a_i, H_0) = P(x_1) = \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{n}{a_1}}, \tag{1}$$

where

$$r = \max(0; a_1 - n_2) \leqslant x_1 \leqslant \min(a_1; n_1) = s. \tag{2}$$

If conditioning is done only in the previously fixed marginals, the form of probability in the table varies according to the type of sampling; so, in the third case equation (1) remains the same, but in the second case it will be:

$$P(X_1 = x_1, X_2 = x_2 \mid n_i, H_0 \equiv p_1 = p_2 = p) = P(x_1, x_2)$$
$$= \binom{n_1}{x_1}\binom{n_2}{x_2} p^{a_1}(1-p)^{a_2}, \tag{3}$$

where $p_i (i = 1, 2)$ are the proportions of individuals which possess characteristic A in the populations from which the n_i size samples are drawn. Henceforth, $\hat{p}_i = x_i/n_i$ and $\hat{p} = a_1/n$, and this case, because of its special importance (and because it is the most studied), will serve as a model in the rest of the paper.

In order to test $H_0 \equiv p_1 = p_2$, statisticians are strongly divided between the proponents of the *conditional principle* (always to apply equation (1)) and those who prefer the *unconditional principle* (to apply the most suitable equation for each case). The former group (Yates, 1984; Barnard, 1989; Upton, 1992) holds the view that the only valid test is the well-known Fisher's exact test (1935). The second group (Barnard, 1947; McDonald et al., 1977; Liddell, 1978; Upton, 1982; Haber, 1987) feels that each model requires a different test (those of Barnard, 1947, in the first two models; that of Fisher in the third). The theoretical reasons for supporting one or the other methodology are many, and are not our objective in this paper (see Martin, 1991; Barnard, 1989 and Upton, 1992; especially the last two, in which the writers justify their change of opinion in favour of conditioning). On a practical level, those who defend the unconditional test have a strong argument when they point out that it has greater power than Fisher's exact test, which they accuse of being too conservative. However, this criticism is based on very limited tables and partially deficient procedures,

where the gain in power is not quantified and the size of the same compared to the much greater time of computation is not evaluated. The principal objective of this paper is to effect a wide-ranging study of all these aspects.

To this end, in Section 2, we give different classic versions of one- and two-tailed tests. In Section 4, we select the most powerful version of the Fisher's exact test (something that has not been satisfactorily dealt with in the relevant literature). In Section 5 the powers of the optimal Fisher's test are compared to those of the optimal unconditional test (Barnard) and the most common unconditional test (McDonald et al.), all with one and two tails, and over a wide range of situations. All the selections and/or comparisons are perfomed on the basis of the "long-term power" and "mean power" described in Section 3.

2. Arrangement criteria and versions of tests

2.1. In Fisher's exact test

If Table 1 is the observed table, its *P*-value is:

$$P_F(x_1, x_2, y_1, y_2) = \sum_{T(X_1) \geq T(x_1)} P(X_1), \tag{4}$$

where $T(\cdot)$ is a given rule of arrangement which makes the points enter the critical region (CR) one by one (or more than one in the case of ties). When the test is one-tailed ($H_1 \equiv p_1 > p_2$ for example), there is only one possible arrangement (Davis, 1986): $T(x_1) = x_1$. When the test has two tails ($H_1 \equiv p_1 \neq p_2$), there are various possible arrangements (it is assumed that $\hat{p}_1 > \hat{p}_2$ and that $\hat{q}_1 = 1 - \hat{p}_i$); the most frequent are:

H $\equiv T_1(x_1) = -P(x_1) \equiv$ from smallest to largest probability.
D $\equiv T_2(x_1) = \hat{p}_1 - \hat{p}_2 \equiv$ from largest to smallest difference in proportions.
R $\equiv T_3(x_1) = \hat{p}_1 \div \hat{p}_2 \equiv$ from largest to smallest relative risk.
O $\equiv T_4(x_1) = \hat{p}_1\hat{q}_2 \div \hat{p}_2\hat{q}_1 \equiv$ from largest to smallest odds-ratio.
I $\equiv T_5(x_1) = -\sum_{i=x_1}^{s} P(i) \equiv$ from smallest to largest sum of the probabilities of the tail.

(The traditional criterion of arranging in order from the largest to the smallest value of the chi-squared statistic is equivalent to T_2.) Any of the 5 rules divides the sample space in 2 parts (which form the CR): from r to x_1' and from x_1 to s, where x_1' is such that $T_i(x_1') > T_i(x_1)$, but $T_i(x_1' + 1) < T_i(x_1)$. Occasionally, when $T_i(x_1) = T_i(x_1')$, there is a tie, for which reason, by complementing any of the above rules by one of the others, the tie between points x_1 and x_1' may be broken, and point x_1 would then be included, so coming nearer the target error α. As there are 5 rules (each complemented by the four others in the case of tie-breaks), this gives a total of 20 different arrangement criteria (in fact they are less, as a tie in the principal criterion may produce a tie in the secondary one).

In symbolic form, the criterion **H** (when it is the principal criterion) complemented by **D**, for a tie-break, gives us the criterion **HD**.

Whatever the $T(\cdot)$ arrangement rule is, a CR to target error α will be built by entering points in it – in order of T – until $P_F(x_1, x_2, y_1, y_2)$ is the nearest possible to the error α without exceeding it.

2.2. *In Barnard's test*

If Table 1 is the observed table, its P-value is:

$$P_I(x_1, x_2) = \max_{0<p<1} \sum_{T(X_1, X_2) \geqslant T(x_1, x_2)} P(X_1, X_2), \tag{5}$$

where $T(\cdot, \cdot)$ is a given arrangement rule which makes the points enter the CR one by one (or by more than one in the case of ties). Here, also, there may be various versions of criterion T. The most powerful is that of Barnard, but, because of the complexity of computing it, it is rarely used. Of the remaning versions, the most common and one of the most powerful (Martín and Silva, 1994) is that of McDonald et al. (1977), and we shall examine this (arranging in order from the smallest to the largest value of the P-value of Fisher's exact test with one tail). Here, the arrangement order will affect both the one-tailed and the two-tailed tests.

Whatever the $T(\cdot, \cdot)$ rule of arrangement is selected, a CR to target error α will by built by entering points in it, in the order of T, until $P_I(x_1; x_2)$ is the nearest possible to the error α without exceeding it.

3. Criteria for comparing different versions of tests

3.1. *Power*

A common criterion in statistics for comparing tests is that of their power, although with discrete variables (as in this case) the difficulty arises that the target error α is almost never reached (and so the sizes of the tests are different). This difficulty may be avoided substantially if the study is performed with a wide range of values for α (as will be seen below). Moreover, as the CRs yielded by each version of the test are not contained one within the other (Upton, 1982), there is no version which is uniformly better than another, and so the powers will have to be compared point by point (in pairs of values of p_1 and p_2). Thus, given the target error α, the $CR(\alpha)$ which it produces, and the values p_1 and p_2, the power for Fisher's exact test is:

$$\theta(p_1, p_2, \alpha) = \frac{\sum_{CR(\alpha)} \binom{n_1}{x_1}\binom{n_2}{x_2} e^{\lambda x_1}}{\sum_{i=r}^{s} \binom{n_1}{i}\binom{n_2}{a_1 - i} e^{\lambda i}} \tag{6}$$

(where $\lambda = \ln\{p_1 q_2 \div p_2 q_1\}$ and $q_i = 1 - p_i$) while that for the unconditional test is:

$$\theta(p_1, p_2, \alpha) = \sum_{CR(\alpha)} \binom{n_1}{x_1}\binom{n_2}{x_2} p_1^{x_1} q_1^{y_1} p_2^{x_2} q_2^{y_2}. \tag{7}$$

Martín and Silva (1994) point out the problem that when one compares the powers of the two versions of the test (say A and B), one will be larger than the other in some values of (p_1, p_2) but not in others. With the aim of globalizing the results, Haber (1987) compares

$$\min_{p_1} \theta(p_1, p_2, \alpha \mid | p_1 - p_2 | = \Delta)$$

in A and B for various values of Δ, while Eberhardt and Fligner (1977) compare the area of the parametric space (p_1, p_2) in which $\theta_A > \theta_B$, with the area where the opposite takes place $(\theta_A < \theta_B)$. Martín and Silva (1994) argue that such comparisons are deficient because everything depends on the relative abundance of each pair (p_1, p_2), that is, on the "a priori" distribution assigned to it. Thus, Eberhardt and Fligner's criterion implies presuming that p_1 and p_2 are independent uniform random variables and that each point (p_1, p_2) of the parametric space is assigned a complementary weight of 1 or 0 depending on whether θ_A is larger or smaller (respectively) than θ_B in it. It seems more reasonable to assign to each point the weight which it has: $\theta(p_1, p_2, \alpha)$.

All these criteria are conventional, as are those that follow, but the ones defended in this paper seem to us to be more acceptable for the reasons we have alleged.

3.2. Long term power

The authors mentioned above indicate that, assuming that p_i follows a uniform distribution during that length of the experimenter's life (with independent p_1 and p_2), the long-term power $\theta(\alpha)$ is given by the double integration of equations (6) and (7), which in each case yields:

$$\theta(\alpha) = \sum_{CR(\alpha)} \binom{n_1}{x_1}\binom{n_2}{x_2} \int_0^1 \int_0^1 \frac{dp_1\, dp_2}{\sum_{i=r}^{s} \binom{n_1}{i}\binom{n_2}{a_1 - i}\left(\frac{q_2}{p_2}\right)^{i-x_1}\left(\frac{p_1}{q_1}\right)^{i-x_1}} \tag{8}$$

(to be determined by numerical integration), and:

$$\theta(\alpha) = \frac{\text{no. of points in the } CR(\alpha)}{\text{no. of points in the sample space}}, \tag{9}$$

both of which are for two-tailed tests. For one-tailed tests, the only equation which we shall specify is that parallel to (9), and so for $H_1 \equiv p_1 > p_2$, we shall have:

$$\theta(\alpha) = \frac{2}{(n_1 + 1)(n_2 + 1)} \sum_{CR(\alpha)} P_F(x_1, x_2 + 1, y_1 + 1, y_2), \tag{10}$$

where $P_F(\cdot)$ is given by (4) for the alternative $H_1 \equiv p_1 < p_2$; see Martín and Silva (1994).

3.3. Mean power

For comparing two tests, what has been written above still has the disadvantage that the conclusions attained depend on the error α used. On the other hand, what the researcher wants is the optimal test for the data obtained, and the first significance in it is reached in an error α (its P-value) that is not one of the traditional ones. All this means, as authors remind us repeatedly, that one should establish the conclusions on the basis of the mean power attained in reasonable intervals of α. Thus, assuming that any value of Type I error between α and α' is equally important, the authors themselves define the concept of mean power for the test for the interval (α, α'):

$$\bar{\theta}(\alpha, \alpha') = \frac{\alpha' A(\alpha') - \alpha A(\alpha)}{\alpha' - \alpha}, \tag{11}$$

where:

$$A(\alpha) = \frac{1}{\alpha}\int_0^{\alpha} \theta(t)\, dt = \bar{\theta}(0, \alpha). \tag{12}$$

The value for $A(\alpha)$ varies according to whether $\theta(\alpha)$ is given by (8), (9) or (10). To obtain it, let $CR_0 = \emptyset$, CR_1, $CR_2, \ldots$, $CR_t = CR(\alpha)$ the successive CRs, which consist of N_i points, obtained on incorporating one by one (sometimes more, in the case of ties) the points of the sample space under the T-arrangement criterion chosen. Each of them is differentiated from the previous one in the set of points D_i $(CR_i - CR_{i-1} = D_i)$, whose number is Δ_i $(N_i - N_{i-1} = \Delta_i)$, and has been obtained from a working error of $\alpha_0 = 0$, $\alpha_1 < \alpha_2 < \cdots < \alpha_t \leqslant \alpha$, yielding a power of $\theta(\alpha_0) = 0$, $\theta(\alpha_1) < \theta(\alpha_2) < \cdots < \theta(\alpha_t) = \theta(\alpha)$. Thus, $\theta(\cdot)$ is a function in steps with jumps in the values of α_i, the area below it is $\alpha A(\alpha)$ and finally,

$$A(\alpha) = \theta(\alpha) - \frac{1}{\alpha}\sum_{i=1}^{t} \{\theta(\alpha_i) - \theta(\alpha_{i-1})\}\alpha_i. \tag{13}$$

In the case of (8), the conditional method with two-tailed test, one only has to substitute in (13) the values obtained. In the case of (9), unconditional method with two-tailed test, the same authors obtain an explicit equation:

$$A(\alpha) = \frac{N\alpha - \sum_{i=1}^{t} \Delta_i \alpha_i}{\alpha(n_1 + 1)(n_2 + 1)}, \tag{14}$$

where $N = N_t$ is the total number of points in $CR(\alpha)$. In the case of (10), unconditional method with one-tailed test, they obtain:

$$A(\alpha) = \frac{2}{\alpha(n_1+1)(n_2+1)} \sum_{i=1}^{t} (\alpha - \alpha_i) \times \sum_{(x_1, x_2) \in D_i} P_F(x_1, x_2+1, y_1+1, y_2), \tag{15}$$

where $P_F(\cdot)$ is as in (10).

Although the symbols in all the expressions remain the same, one must not forget that $\theta(\alpha)$ refers to $\theta(\alpha \mid a_i, n_i)$ in equation (8), and to $\theta(\alpha \mid n_i)$ in (9) and (10).

4. Choosing the most powerful Fisher's test

4.1. General remarks

As already pointed out above, Fisher's exact test only admits one version as a one-tailed test, and so, in this case, no choice is required. When the test has two tails, the most frequent possible versions are the twenty cited above, and a choice must be made between them. There is abundant literature on partial selections (see Martín y Luna, 1987), but there has never been such a wide selection as the present one. In all the cases, the choice will be made on the basis of the mean power in the intervals of α from 0% to 1%, from 1% to 5% and from 5% to 10%, the first for those using Bonferroni's method, the second for the usual significances, the third for the indications of significance. In all the cases too, the choice will be effected in two stages: (i) Selecting the optimal tie-break rule for each arrangement criterion; (ii) Selecting the optimal criterion between the 5 of the previous stage. It is a good idea to choose the optimal from among the simple criteria (**H**, **D**, **R**, **O** and **I**) so us to evaluate the gain produced when the tie-break rule is introduced.

4.2. As a conditional test

The mean power, given by (8), (11) and (13), apart from depending on the interval of values of α chosen, now also depends on the values of a_i and n_i. So as not to repeat the tables, let us agree that $a_1 = \min(a_i, n_i)$ and $n_1 = \min(n_1, n_2)$; with this, the mean power will depend on n, a_1 and n_1, and these parameters will have to be given a wide range of values so as to obtain the optimal method in a wide range of situations. Equation (8) has been obtained by numerical intergation (by Simpson's method in MapleV software).

The results of all the selection may be obtained from the authors. The results of the last selection, for the interval 1% to 5%, are presented in Table 2, in which only the methods without tie-break rules are referred to, because the

Table 2
Mean powers for the two-tailed Fisher's exact test (as a conditional test) for the arrangement criteria indicated (first row) in the tables shown (first three columns) and for the interval of values of α from 1% to 5%

n	a_1	n_1	**I**	**H**	**D**	**R**	**O**
10	3	3	18.2	18.2	18.2	0.0	0.0
10	5	5	35.3	35.3	35.3	35.3	35.3
30	3	3	20.5	20.5	20.5	0.0	0.0
30	3	7	19.1	19.1	19.1	0.0	0.0
30	3	11	24.1	24.1	24.1	0.0	0.0
30	3	14	5.8	5.8	5.8	0.0	0.0
30	8	8	37.8	37.8	37.8	42.7	42.7
30	8	12	46.7	46.7	46.7	44.4	47.5
30	8	15	50.2	50.2	50.2	50.2	50.2
30	15	15	44.7	44.7	44.7	44.7	44.7
50	4	6	24.4	24.4	24.4	0.0	0.0
50	4	12	25.7	25.7	25.7	0.0	0.0
50	4	18	21.2	21.2	21.2	0.0	0.0
50	4	24	72.2	72.2	72.2	0.0	0.0
50	10	10	22.5	23.0	22.5	16.0	16.0
50	10	20	53.7	53.7	51.0	53.7	53.7
50	15	15	49.1	49.1	49.8	46.7	47.7
50	15	22	55.4	55.4	55.4	55.4	55.5
50	25	25	47.7	47.7	47.7	47.7	47.7

inclusion of a tie-break rule has been found to affect the power to an insignificant extent (see the following subsection). The conclusions (for the whole body of the data) are that the best method is **H**, but that **I** and **D** (in that order) are practically equivalent to it, while the worst methods are **O** and **R** (in that order), which differ slighty between themselves.

4.3. As an unconditional test

The proponents of the unconditional method have the difficulty of the large amount of time needed to carry out calculations for this method, which makes it well-nigh impracticable for ordinary use. Fisher's method, which Pearson (1947) shows is also valid as an unconditional method, is known to have less power, but being relatively easy to compute, it is advisable to give the optimal version of the same, as an unconditional test, and this implies that the power will be given by (11) and (14).

As Fisher's exact test acts on the diagonals of the sample space (x_1, x_2), that is, on constant values of $a_1 = x_1 + x_2$, and the CRs of the previous section were built for fixed values of a_1, the present CRs will now be the union of all the $CR(a_1)$ built to an equal error α. The choice of the optimal criterion will be made by a similar procedure to that described above, but now the parameters to be fixed are only n and n_1. The values chosen for n are those of the intervals

6–14; 16–24; 27–33; 37–43; 48–52,

Table 3
Mean powers of Fisher's exact test (as an unconditional test) for the arrangement criteria indicated (first row), in the values for n shown (first column) and for each interval of values of α

n\Methods	H	HO	DH	RH	OH	IH
α: 0%–1%						
6–14	2.3	2.3	2.3	2.1	2.3	2.3
16–24	10.8	10.8	10.8	7.0	9.9	10.8
27–33	18.5	18.6	18.5	10.5	15.6	18.6
37–43	24.7	24.7	24.6	13.8	20.0	24.7
48–52	29.5	29.5	29.5	16.9	23.8	29.6
α: 1%–5%						
6–14	9.3	9.3	9.3	7.5	9.2	9.3
16–24	22.1	22.1	22.1	14.3	19.9	22.1
27–33	30.8	30.8	30.7	19.3	26.4	30.8
37–43	36.9	36.9	36.9	23.9	31.4	36.9
48–52	41.6	41.6	41.6	27.9	35.4	41.6
α: 5%–10%						
6–14	17.3	17.3	17.3	13.3	16.9	17.3
16–24	31.2	31.2	31.2	21.1	28.1	31.2
27–33	39.9	39.9	39.9	27.4	35.0	39.9
37–43	45.7	45.7	45.7	32.5	39.9	45.7
48–52	50.1	50.1	50.1	36.7	43.7	50.0

and for each of them all the possible pairs of values of (n_1, n_2), where $n_1 \leqslant n_2$, have been contemplated. This gives values for n of approximately 10, 20, 30, 40 and 50, and for each of them the average of the mean powers is taken, because it has been found that the variability of the mean powers within each range of values of n is not great. In total 459 tables (n_1, n_2) and 117, 934 points (x_1, x_2) have been studied.

All the selection results may be obtained from the authors. The results of the optimal selections are given in Table 3, where method **H** (without a tie-break) is included as a basis for comparison. The conclusions for all the data are:

(1) If we do not take into account the tie-break rule (following what is usual in the other published papers) the worst criteria are **R** and **O** (in that order), while criteria **I** and **H** are equal and **D** is almost imperceptively worse than them.
(2) In all the partial selections of the optimal, criterion **H** (the most commonly found in publications) is always present. This is also true of criterion **D**, although, in the case of equivalence, in Table 3 we have chosen to load the choice in favour of **H**.
(3) The best criterion is **HO**, which is closely equivalent to **H** (which is equivalent to **I**) and **DH** (in that order).
(4) In the methods which work well (**I**, **H** and **D**) the tie-break rule barely affects power.

5. Fisher's exact test versus the optimal unconditional test

5.1. General remarks

Fisher's exact test has been repeatedly accused of being conservative from the point of view of unconditional tests, but that accusation (generally obtained from small values of n and not very high values of $K = n_1 \div n_2 \geqslant 1$) has not been sufficiently detailed nor evaluated with sufficient accuracy in terms of power. Schouten et al. (1980) and Martín and Silva (1994) point out the strong dependence on the unconditional tests of the factor K (which measures the imbalance in values of n_i), though it is advisable to introduce it into planned comparisons for power. In the following, the mean power $\bar{\theta}(\alpha, \alpha' \mid n, K)$ will be determined for each of the values of n indicated above, and for each of the following intervals of K: $K = 1.00$; $1.00 < K \leqslant 1.25$; $1.25 < K \leqslant 1.50$; $1.50 < K \leqslant 1.75$; $1.75 < K \leqslant 2.25$; $2.25 < K \leqslant 3.00$; $3.00 < K \leqslant 4.25$; $4.25 < K \leqslant 6.00$.

The mean power $\bar{\theta}_M$ for the test of McDonald et al. (1977) (which is the most common unconditional test) was computed by Martín and Silva (1994). The mean power $\bar{\theta}_F$ for Fisher's test was calculated by the authors. The relevant point of the comparison between them is the power gain, both absolute $(\bar{\theta}_M - \bar{\theta}_F)$ and relative $(\{\bar{\theta}_M - \bar{\theta}_F\} \div \bar{\theta}_F)$, which are of interest since each gives complementary information. The results for values of α between 1% and 5% (the most common) are given in Table 4. The rest of the results, including those which have Barnard's test for reference (which is the most powerful unconditional one), may be requested from the authors. The calculations for $\bar{\theta}_F$ are made taking as a base equations (11) and (14) or (15), depending on whether the test has one or two tails, that is, taking unconditional formulae as a basis. The reason for this is that in order to compare Fisher's test with that of McDonald et al., the former must submit to the same conditions as the latter: the unconditional principle; otherwise the results obtained cannot be compared.

5.2. In one-tailed tests

For one-tailed test we have already said that there is only one version of Fisher's exact test, and the results in the first part of Table 4 refer to it. The conclusions are:

(i) The absolute increases in power go from 3% to 6%, from 4% to 9% or from 4% to 14% for each of the three intervals of α except if $K \leqslant 1.25$ in which case the gains are quite considerable.
(ii) The relative increases in power are lower than 10% when n is large ($n \geqslant 50$) and K is moderate (from 1.5 to 3).
(iii) Both increments tend toward zero as n increases, except when $K \leqslant 1.25$, in which case this tendency will not make itself felt until n is much larger than the values dealt with in this study.

Table 4
Mean powers of Fisher's exact test (first entry) and absolute increases (second entry) and relative increases (third entry) respecting it of the unconditional test of McDonald et al., for errors α between 1% and 5%, various values of n (first column) and various values of $K = n_1/n_2 \geqslant 1$ (first row) in tests of one tail (first table) or two tails (second table, where the test used is the **HO** version of Fisher's test)

$n \setminus K$	$=1.00$	$\leqslant 1.25$	$\leqslant 1.50$	$\leqslant 1.75$	$\leqslant 2.25$	$\leqslant 3.00$	$\leqslant 4.25$	$\leqslant 6.00$
One tail								
6–14	11.5	13.7	13.2	12.5	11.1	8.0	8.4	6.3
	8.9	9.6	8.2	8.0	8.8	9.2	8.1	7.2
	77.5	70.4	62.2	64.3	79.7	114.5	96.4	114.6
16–24	16.9	20.7	30.0	29.5	27.5	25.4	21.6	16.4
	21.0	18.1	7.8	7.6	7.9	8.6	8.7	8.8
	124.1	87.2	26.0	25.9	28.8	33.9	40.3	53.4
27–33	10.4	22.8	41.1	40.3	38.7	36.0	32.2	26.8
	37.3	24.7	6.1	6.4	6.5	7.1	7.8	8.3
	358.5	108.3	14.8	15.9	16.8	19.8	24.3	30.9
37–43	10.5	27.3	48.3	47.5	46.2	43.6	39.8	34.4
	43.3	26.4	5.1	5.3	5.4	6.0	6.6	7.4
	412.7	96.6	10.6	11.3	11.7	13.8	16.7	21.5
48–52	8.5	24.9	53.6	52.7	51.5	49.0	45.4	40.1
	49.7	33.2	4.3	4.6	4.7	5.1	5.7	6.6
	584.2	133.4	8.0	8.7	9.1	10.5	12.6	16.4
Two tails								
6–14	7.4	12.8	11.9	12.0	10.5	7.8	8.4	6.3
	6.5	5.2	4.2	3.8	5.2	6.9	7.0	7.1
	87.8	40.6	35.3	31.7	49.5	88.5	83.3	112.7
16–24	24.6	29.0	27.8	27.5	25.8	24.0	21.0	16.4
	6.7	3.8	4.0	4.1	4.2	4.6	5.2	6.4
	27.2	13.1	14.4	14.9	16.3	19.2	24.8	39.0
27–33	35.8	39.0	38.5	37.9	36.5	34.2	30.9	25.9
	5.8	3.0	3.2	3.1	3.4	3.3	3.7	4.4
	16.2	7.7	8.3	8.2	9.3	9.6	12.0	17.0
37–43	43.3	46.0	45.7	45.0	43.7	41.5	37.9	33.3
	5.0	2.5	2.4	2.5	2.6	2.7	3.1	3.3
	11.5	5.4	5.3	5.6	5.9	6.5	8.2	9.9
48–52	48.9	51.2	50.9	50.1	49.0	46.6	43.3	38.4
	4.0	2.1	2.1	2.1	2.2	2.4	2.5	3.0
	8.2	4.1	4.1	4.2	4.5	5.2	5.8	7.8

(iv) The above conclusions are valid for the case of the test of McDonald et al.; in the case of Barnard's test the same conclusions hold, but the differences with Fisher's exact test are logically somewhat more marked.

5.3. In two-tailed tests

For two-tailed tests, the most powerful version of Fisher's test (version **HO**) has been chosen, although the conclusions do not change if any of the other

traditional versions is chosen. The second part of Table 4 refers to them. The conclusions are:

(a) The absolute increases of power always go from 2% to 5%, 7% or 10% for each of the three intervals of α.
(b) The relative increases of power are always lower than 10% when n is large ($n \geqslant 40$ or 50), and are quite frequently lower than 5%.
(c) Both increments tend toward zero as n increases.
(d) The increases are even less when K takes moderate ($1 < K \leqslant 3$) values.
(e) The above conclusions, valid for the test of McDonald et al., hold for Barnard's test, but logically with somewhat more marked differences.

6. Conclusions

Fisher's exact test is the most widely known and accepted method for analysing a 2×2 table, but as a two-tailed test it has various versions. In this paper it has been shown, from both the conditional and unconditional points of view, that the traditional criteria of an ordered arrangement from the smallest to the largest probability of the hypergeometric (H), from the smallest to the largest probability of tail (I), and from the largest to the smallest difference of proportions (D), are practically equivalent (with a very slight advantage for the first two), and that the introduction of a subsidiary tie-breaking criterion does not improve the test to any practical purpose.

Moreover, Fisher's exact test has the advantage of being valid in the three classic cases of sampling in which a 2×2 table may arise, it is relatively simple to compute (it can often be done on a pocket calculator), it is easy to explain to students, it comes in almost all the statistical packages, and finally, it is valid from the unconditioning point of view. Its unconditional competitors (the tests of Barnard and McDonald et al., for example), on the contrary, require a specific version for each type of sampling, are difficult to compute and do not appear in any of the common statistics packages, although it is alleged that they are much more powerful than Fisher's exact test. In this paper, the authors have shown that even though Fisher's exact test is always somewhat less powerful than the unconditional tests mentioned, the difference is not great in the following cases:

(A) For one-tailed tests, if $n \geqslant 50$ and $1.5 \leqslant K \leqslant 3$,
(B) For two-tailed tests, if $n \geqslant 30$ and $1 < K \leqslant 3$, or if $n \geqslant 50$ and K is any other value,

(although for very low values of α the demands are somewhat greater), so that in these cases, and for routine experiments, the use of Fisher's exact test is justifiable (especially in the most common situation of two tails), as the saving in calculation time compensates for the slight loss of power it produces. The result is especially pleasing if we remember that the great majority of experiments verify condition (B), and that it is precisely when $n \geqslant 30$ or $n \geqslant 50$ when there are such great difficulties in computing in the usual unconditional tests. These

results are in accord with those obtained by Mehta and Hilton (1993) for wider tables than the 2×2 ones.

Acknowledgements

The authors wish to thank Professor Hardeo Sahai, University of Puerto Rico, for several editorial comments and suggestions which have greatly improved the clarity and exposition of the paper.

References

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

Barnard, G.A., On alleged gains in power from lower P-values, *Stat. in Medicine*, **8**(12) (1989) 1469–1477.

Davis, L.J., Exact tests for 2×2 contingency tables, *Amer. Stat.*, **40**(2) (1986) 139–141.

Eberhardt, R.A. and M.A. Fligner, A comparison of two tests for equality of two proportions, *Amer. Stat.*, **31** (1977) 151–155.

Fisher, R.A., The logic of inductive inference, *J. Royal Statist. Soc. A*, **98** (1935) 39–54.

Haber, M., A comparison of some conditional and unconditional exact tests for 2×2 contingency tables, *Commun. Statist. Simul.* **16**(4) (1987) 999–1013.

Luna del Castillo, J.D. and A. Martín Andrés, Tablas 2×2 y test exacto de Fisher, *Trabajos de Estadística* **2**(1) (1987) 15–43.

Liddell, D., Practical test of 2×2 tables, *Statist.* **27**(4) (1978) 295–304.

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

Martín Andrés, A. and A. Silva Mato, Choosing the optimal unconditioned test for comparing two independent proportions, to appear in: *Comp. Statist. and Data Anal.* (1994).

Mehta, C.R. and J.F. Hilton, Exact power of conditional and unconditional tests: Going beyond the 2×2 contingency table, *Amer. Statist.*, **47**(2) (1993) 91–98.

McDonald, L.L., B.M. Davis and G.A. Milliken, A non-randomized unconditional test for comparing two proportions in a 2×2 contingency table, *Technometrics* **19** (1977) 145–150.

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

Schouten, H.J.A., I.W. Molenaar, R. Van Strik and A. Boomsa, Comparing two independent binomial proportions by a modified chi-squared test, *Biometrical J.*, **22**(3) (1980) 241–248.

Upton, G.J.G., A comparison of alternative tests for the 2×2 comparative trial, *J. Roy. Statist. Soc. A*, **145**(1) (1982) 86–105.

Upton, G.J.G., Fisher's exact test, *J. Roy. Statist. Soc. A*, **155**(3) (1992) 395–402.

Yates, F., Test of significance for 2×2 contingency tables, *J. Roy. Statist. Soc. A*, **1473**) (1984) 426–463.