ELSEVIER Computational Statistics & Data Analysis 40 (2002) 339–354

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Asymptotical tests in 2×2 comparative trials (unconditional approach)

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

[a] *Bioestadística, Facultad de Medicina, Universidad de Granada, 18071 Granada, Spain*
[b] *Estadística, Universidad de Cádiz, 11003 Cádiz, Spain*
[c] *Bioestadística, Facultad de Medicina, Universidad de Alcalá, 28871 Alcalá de Henares (Madrid), Spain*

Received 1 February 2001; received in revised form 1 November 2001; accepted 1 November 2001

Abstract

The unconditional Barnard's test for the comparison of two independent proportions is difficult to apply even with moderately large samples. The alternative is to use a χ^2 type, arc sine or mid-p asymptotic test. In the paper, the authors evaluate some 60 of these tests, some new and others that are already familiar. For the ordinary significances, the optimal tests are the arc sine methods (with the improvement proposed by Anscombe), the χ^2 ones given by Pearson (with a correction for continuity of 2 or of 1 depending on whether the sample sizes are equal or different) and the mid-p-value ones given by Fisher (using the criterion proposed by Armitage, when applied as a two-tailed test). For one-(two) tailed tests, the first method generally produces reliable results $E > 10.5$ ($E > 9$ and unbalanced samples), the second method does so for $E > 9$ ($E > 6$) and the third does so for all cases, although for $E \leqslant 6$ ($E \leqslant 10.5$) it usually gives too many conservative results. E refers to the minimum expected quantity. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Arc sine transformation; Barnard's test; Binomial proportions; Continuity correction; Fisher's exact test; Mid-p-value; 2×2 tables; Unconditional test; Validity conditions

1. Introduction

Let $x_i \sim B(n_i, p_i)$, $i = 1, 2$, be two random independent binomial variables. In practically all Experimental Sciences one quite frequently has to test H_0: $p_1 = p_2$ $(=p)$ against an alternative with one or two tails (the classic comparison of two proportions). In the Health Sciences, this is customarily referred to as a *Comparative* 2×2

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

PII: S0167-9473(02)00005-1

Trial. The way to solve the problem is twofold: by using a conditional test (Fisher's exact test (Fisher, 1935)) or by using an unconditional one (Barnard's test (Barnard's, 1947)). The former is based on the conditional random variable $(x_1|x_1+x_2=a_1)$ that follows the hypergeometric distribution under H_0; because of this, if $\mathrm{CR}(x_1|a_1)$ is a critical region formed by a group of values x_1 that fall between $r=\max(0;a_1-n_2)$ and $s=\min(a_1,n_1)$, then the error α of the conditional test will be $\alpha_{\mathrm{C}}=\Sigma_{\mathrm{CR}(x_1|a_1)}P(i)$, where $P(i)=C(n_1;i)C(n_2;a_1-i)/C(n;a_1)$. The unconditional test is based on the bidimensional random variable (x_1,x_2) following a double binomial; because of this, if $\mathrm{CR}(x_1,x_2)$ is a critical region formed by a group of pairs (x_1,x_2), where $0\leqslant x_i\leqslant n_i$, then the error $\alpha(p)$ of the unconditional test will be $\alpha(p)=\Sigma_{\mathrm{CR}(x_1,x_2)}P(x_1,x_2)$, where $P(x_1,x_2)=C(n_1,x_1)C(n_2,x_2)p^{a_1}(1-p)^{n_1+n_2-a_1}$, and its size is $\alpha_1=\mathrm{Max}_{0<p<1}\,\alpha(p)$. The reasons put forward to support one or other methodology can be found in Yates (1984) and Martín Andrés (1991). This article adopts the unconditional point of view, which is licit, since it produces a test which generally is more powerful than the conditional one (Haber, 1987).

The unconditional test has two drawbacks. On the one hand, there are a great number of versions (ways of constructing the critical region), but Martín Andrés et al. (1998) select the best of these: their programs are available at URL http://www.ugr.es/~bioest/Software.htm. Also, the versions usually take a long time to compute (Silva Mato and Martín Andrés, 1997)—basically due to the maximization which the determination of α_{I} implies—making the test impossible to apply even for moderate values of n_i. When it is not possible to apply the unconditional exact test, one has no choice but to use an asymptotic test.

When an asymptotic test is used as an approximation to the conditional test, it is customary to apply the classic χ^2 test with the continuity correction (c.c. in the following) given by Yates (1934) with the precaution proposed by Mantel (1974) when the test is a two-tailed one. Other possible versions of c.c.'s were considered by Martín Andrés et al. (1992) and the validity conditions (v.c. from here on) of the optimal methods can be consulted in Martín Andrés and Herranz Tejedor (1997, 2000). However, where the asymptotic test is intended to be used as an approximation to the unconditional test, the problem has not been sufficiently studied. It is true that Martín Andrés and Silva Mato (1996) analyse 20 c.c.'s to the classic χ^2 test, χ^2 and obtain the optimal c.c.'s which are quoted below. However, the behaviour of other asymptotic methods proposed in the relevant literature remains to be evaluated. This is the case of the various methods mentioned by Martín et al. (1998), which belong to the non-classic χ^2, arc sine or mid-p types.

The use of the arc sine transformation in this context is fairly customary when the aim is to determine the sample size, but its performance as an approximation to the unconditional exact test has not yet been evaluated.

The use of Fisher's mid-p-value as an approximation to the Barnard test is quite surprising and should be justified. Given that the Fisher exact test is very conservative (compared to Barnard's test), Plackett, in his discussion of Yates (1984), proposed Fisher's mid-p-value as a means of reducing its conservatism. The idea was favourably received by Barnard (1989), Routledge (1992), Upton (1992) and Agresti (2001) because it was a way of terminating the conditional vs. unconditional argument

(Haber, 1992). Haber (1986) was the first to propose mid-p as an approximation to the unconditional test, one that was described by Hirji et al. (1991) as a quasi-exact test. Both the authors and Davis (1993) agree that mid-p is generally conservative, but quite less so than Fisher's exact test, and behaves in a very similar fashion to the χ^2 test without c.c. Note that although one needs to use a computer to apply the mid-p, actually obtaining it presents no problem (no matter what the value of n_i may be).

The aim of this article is to make a comparative evaluation of all the referred methods, with the intention of selecting the optimal and obtaining its v.c. This involves the evaluation of $\frac{55}{57}$ new asymptotic methods for tests of $\frac{\text{one}}{\text{two}}$ tails, plus the 4 optimal given by Martín and Silva. The mentioned comparison cannot be made through the power of different tests, because the real error of type I of each one of them is different. It is for this that the approach of Martin and Silva is adopted (see Section 3.1).

2. Asymptotic method to be studied

2.1. Type χ^2 methods

The most usual χ^2 statistic without c.c. ("uncorrected") is the χ^2_{U}, although frequently the version χ^2_{P} given by Pearson (1947) is used, where

$$\chi^2_{\mathrm{U}} = \frac{(x_1y_2 - x_2y_1)^2}{a_1a_2n_1n_2}\,n, \quad \chi^2_{\mathrm{P}} = \frac{(x_1y_2 - x_2y_1)^2}{a_1a_2n_1n_2}\,(n-1) \tag{1}$$

and with $y_i = n_i - x_i$, $a_1 = x_1 + x_2$, $a_2 = n - a_1 = y_1 + y_2$ and $n = n_1 + n_2$. It has been fairly well established (Pearson, 1947; Cox, 1970) that the asymptotic statistics based on discrete random variables require a c.c. consistent with adding to or subtracting from the experimental value, half the jump between that and the immediately following value (Kendall and Stuart, 1973). Conover (1974) observed that "the following value" is difficult to obtain in the present case, and therefore Haber (1982) proposed adding or subtracting "half the average jump" of the implied variable. This means that for any statistic of the shape $\chi^2 = N^2/D$ -where $N = x_1y_2 - x_2y_1$ and D refers to any function of the data—the "corrected" statistic will have a format which depends on what is considered to be the variable base of the problem. If the variable is χ^2 (Martín and Silva), χ (Haber) or N (Yates), then the corrected statistics will be $\chi^2 - c$, $(\chi - c)^2$ or $(|N| - c)^2/D$, which will be referred to as being type M, H or T. In these statistics, c is the c.c. and its value will be the average jump of χ^2, χ or N in the whole of the sample space. In the case of χ^2_{U} and χ^2_{P}, the value of c for the corrected statistics of type M or H was obtained by Haber (1982) and Martín and Silva (1996), thus yielding the following statistics UM, UH, PM and PH:

$$\chi^2_{\mathrm{UM}} = \chi^2_{\mathrm{U}} - \frac{n}{n_0}, \quad \chi^2_{\mathrm{UH}} = \left(\chi_{\mathrm{U}} - \frac{\sqrt{n}}{n_0}\right)^2,$$

$$\chi^2_{\mathrm{PM}} = \chi^2_{\mathrm{P}} - \frac{n-1}{n_0}, \quad \chi^2_{\mathrm{PH}} = \left(\chi_{\mathrm{P}} - \frac{\sqrt{n-1}}{n_0}\right)^2, \tag{2}$$

where $n_0 = (n_1+1)(n_2+1) - 1 - \text{hcf}(n_1, n_2)$ if $n_1 \neq n_2$, $n_0 = [(n+2)^2/8]^-$ if $n_1 = n_2$, hcf refers to the "highest common factor" and $[x]^-$ refers to the integer part of x. For the type T statistic, and because $|N|$ takes values between 0 and $n_1 n_2$, it is immediately clear (using the arguments of Martín and Silva) that the new corrected statistics will be the following UT and PT:

$$\chi^2_{\text{UT}} = \frac{(|x_1 y_2 - x_2 y_1| - \frac{n_1 n_2}{n_0})^2}{a_1 a_2 n_1 n_2} n, \quad \chi^2_{\text{PT}} = \frac{(|x_1 y_2 - x_2 y_1| - \frac{n_1 n_2}{n_0})^2}{a_1 a_2 n_1 n_2} (n-1). \tag{3}$$

When $n_1 \to \infty$, then $c \to 0$ in cases (2), but not in cases (3), which suggests these other new asymptotic c.c.'s (methods UTA and PTA):

$$\chi^2_{\text{UTA}} = \frac{(|x_1 y_2 - x_2 y_1| - c_0)^2}{a_1 a_2 n_1 n_2} n, \quad \chi^2_{\text{PTA}} = \frac{(|x_1 y_2 - x_2 y_1| - c_0)^2}{a_1 a_2 n_1 n_2} (n-1), \tag{4}$$

where $c_0 = 1$ if $n_1 \neq n_2$ and $c_0 = 2$ if $n_1 = n_2$. It is clear that $\chi^2_{\text{UTA}} < \chi^2_{\text{UT}}$ and $\chi^2_{\text{PTA}} < \chi^2_{\text{PT}}$, so that the first methods will give somewhat higher p-values than those of the second ones.

For the corrected type T statistics, Martín and Silva proposed that c should not be "the average jump", but rather "half the approximate jump". Among the various possible versions, those that behaved best were of types S and C which, based on $c = \min(n_1, n_2)/2$ and $c = \text{hcf}(n_1, n_2)/2$, respectively, were given by Schouten et al. (1980) and Cook (1981). For example, $\chi^2_{\text{US}} = \{|x_1 y_2 - x_2 y_1| - \min(n_1 n_2)/2\}^2 n / \{a_1 a_2 n_1 n_2\}$.

This same reasoning can be applied to any other type χ^2 statistic that is defined, and in this way each one of them yields 6 statistics with c.c. (those of types M, H, T, TA, S and C). As Martín et al. (1998) proposed 7 different χ^2 statistics to the χ^2_{U} and χ^2_{P}—expressions (5)–(7) and (10)–(13) in the article mentioned—this gives $7 \times 6 = 42$ different methods (generally with differing c values). To these must be added the methods UT, PT, UTA and PTA (which are new proposals) and the methods UH, UM, PH and PM (which were selected as optimal by Martín and Silva), giving a total of 50 methods to be analysed comparatively. For example, for the statistic χ^2_{D} of D'Agostino et al. (1988), the method DT would be based on $\chi^2_{\text{DT}} = \{|x_1 y_2 - x_2 y_1| - n_1 n_2 / n_0\}^2 (n-2) / [n\{n_2 x_1 y_1 + n_1 x_2 y_2\}]$, where $n_0 = (n_1+1)(n_2+1) - 3 - \text{hcf}(n_1, n_2)$ if $n_1 \neq n_2$ and $n_0 = 2[(n_1+1)(n_2+1)/4]^- - 2$ if $n_1 = n_2$. The remaining cases may be requested from the authors.

2.2. Methods based on the arc sine transformation

Martín et al. (1998) refer to the two traditional statistics A1 and A2

$$\chi^2_{\text{A1}} = \frac{(\sin^{-1}\sqrt{\hat{p}_1} - \sin^{-1}\sqrt{\hat{p}_2})^2 4 n_1 n_2}{n},$$

$$\chi^2_{\text{A2}} = \frac{(2n_1+1)(2n_2+1)(\sin^{-1}\sqrt{\hat{p}'_1} - \sin^{-1}\sqrt{\hat{p}'_2})^2}{n+1}, \tag{5}$$

where $\hat{p}_i = x_i/n_i$ and $\hat{p}'_i = (x_i+3/8)/(n_i+3/4)$. It is only meaningful to perform the types H and M c.c.'s on these. For example, it can be shown that $\chi^2_{\mathrm{A1H}} = 4n_1n_2\{|\sin^{-1}\hat{p}_1^{0.5} - \sin^{-1}\hat{p}_2^{0.5}| - \pi/2n_0\}^2/n$, where $n_0 = (n_1+1)(n_2+1) - 1 - \mathrm{hcf}(n_1, n_2)$ if $n_1 \neq n_2$ and $n_0 = 2[(n_1+1)(n_2+1)/4]^-$ if $n_1 = n_2$. In exceptional circumstances, and given that it has never been evaluated, a classic type Y c.c. (Yates' classic correction), which in reality is a conditional c.c., can be performed. For example, if $\hat{p}_1 > \hat{p}_2$, then this will give $\chi^2_{\mathrm{A1Y}} = 4n_1n_2\{\sin^{-1}(\hat{p}_1 - 1/2n_1)^{0.5} - \sin^{-1}(\hat{p}_2 + 1/2n_2)^{0.5}\}^2/n$. The remaining expressions may be obtained on request from the authors. This produces 8 new methods to be analysed: the methods A1, A1H, A1M, A1Y and their homonyms based on A2. In this article, we consider A1 and A2 (which have no c.c.) because they have not been previously evaluated in the relevant literature.

2.3. *Methods based on Fisher's mid-p-value*

If we reorder the samples so that $\hat{p}_1 > \hat{p}_2$ (for which, if necessary, it is sufficient to permute the values of x_i and y_i), then $x_1 > E_{11} = a_1n_1/n$ and Fisher's mid-p-value for the observed value x_1 and for the alternative H_1: $p_1 > p_2$ produces the method FM

$$\mathrm{FM}(x_1) = \sum_{i=x_1+1}^{s} P(i) + \frac{1}{2}P(x_1), \tag{6}$$

where $P(i)$ and s are as indicated in the introduction. For a two-tailed test there are more possibilities. Hirji et al. (1991) mention the following two options:

$$\begin{aligned}\mathrm{FMH}(x_1) &= \sum_{i=r}^{x'_1} P(i) + \sum_{i=x_1+1}^{s} P(i) + \begin{cases} P(x_1)/2 & \text{if } P(x_1) > P(x'_1), \\ 0 & \text{if } P(x_1) = P(x'_1), \end{cases} \\ \mathrm{FMA}(x_1) &= 2\mathrm{FM}(x_1),\end{aligned} \tag{7}$$

where $x'_1 < E_{11}$ so that $P(x'_1) \leqslant P(x_1)$ and $P(x'_1+1) > P(x_1)$. The FMH method is based on the direct application of the mid-p concept; the FMA is based on Armitage's criterion that the p-value of a two-tailed test is double that of a test with one tail. A third option is to use the criterion proposed by Mantel (1974), who defined it with reference to Yates' χ^2 test used as an approximation to the conditional test, that is, the p-value of two tails is the sum of the p-values of one tail for the original data (x_1) and those of the other tail (x'_1). This produces the method FMM defined by

$$\begin{aligned}\mathrm{FMM}(x_1) &= \mathrm{FM}(x_1) + 1 - \mathrm{FM}(x'_1 - 1) \\ &= \sum_{i=r}^{x'_1-1} P(i) + \sum_{i=x_1+1}^{s} P(i) + \frac{1}{2}\{P(x'_1) + P(x_1)\}.\end{aligned} \tag{8}$$

As one can see, in all the cases the mid-p constitutes a sort of c.c. to Fisher's statistic. All of this implies that a further $\frac{1}{3}$ method should be analysed for $\frac{\text{one}}{\text{two}}$-tailed tests.

3. Selection of the optimal method

3.1. Previous selection

The aim of an asymptotic method is to give a p-value (P_{A}) which is approximately equal to that of the exact method (P_{E}). The maximum difference $P_{\mathrm{A}} - P_{\mathrm{E}}$ that is admissible is a matter of opinion. For the conditional test, Martín et al. (1992) welcomed the criterion proposed by Cochran (1954) requiring that $|P_{\mathrm{A}} - P_{\mathrm{E}}| \leqslant \delta P_{\mathrm{E}}$, where δ is of the order of 20% or 50% for values of P_{E} of 5% or 1%, respectively. When Martín and Silva (1996) tried to repeat the criterion for the unconditional test, they encountered a double problem. On the one hand, the values of P_{A} and P_{E} differ much more here; on the other, computer capacity prevents the computations going further than $n = 50$. For this reason, the evaluation of the asymptotic methods was carried out using wider values of δ than the previous ones, and showed that the optimal methods (of those evaluated) were UM, UH, PM and PH already mentioned.

As shown above, $\frac{55}{57}$ new asymptotic methods for $\frac{\text{one}}{\text{two}}$-tailed tests ($\frac{46}{46}$ of type χ^2, $\frac{8}{8}$ of type arc sine and $\frac{1}{3}$ of type mid-p) were proposed. In order to concentrate on those that perform best, a preliminary critical study was carried out based on the criteria proposed by Martín and Silva (1996). The complete results may be obtained from the authors on request. The study shows that the best performing methods are UT, PT, UTA, PTA, A2, FM, FMH, FMA and FMM. The rest are either obviously inferior or are similar to one of those selected (but more difficult to compute the evaluation). Therefore, these 9 methods are those which must be evaluated, together with the 4 optimal given by Martín and Silva. Method A1 will also be considered because it is quite well-known and has never been studied in this context. This makes a total of $\frac{11}{13}$ $\frac{\text{one}}{\text{two}}$-tailed methods to be studied.

3.2. Evaluation of the selected asymptotic methods

To evaluate and compare the $\frac{11}{13}$ proposed asymptotic methods (UM, UH, UT, UTA, PM, PH, PT, PTA, A1, A2, FM, FMH, FMA and FMT), we adopt similar (but improved) criteria to those used by Martín and Silva (1996). According to these authors, the behaviour of an asymptotic test depends fundamentally on the real p-value (P_{E}), on the unbalance of samples ($K = n_2/n_1 \geqslant 1$) and the minimum expected quantity $E = \min(a_i) \times \min(n_i)/n$. Hence, in the following, all the samples (x_1, x_2, n_1, n_2) with $n = 20(1)100$ that satisfy each of the following combinations of P_E, K and E, will be generated:

P_{E}: 0.001-0.01; 0.01-0.1 (2 groups),
K : 1; 1(.25)2(1)3; > 3 (7 groups),
E : $\leqslant 1.5$; 1.5(1)4.5(1.5)12; > 12 (10 groups),

which gives $2 \times 7 \times 10 = 140$ groups of $(P_{\mathrm{E}}; K; E)$ values for the one-tail tests (and similarly for two-tail tests). A given group has N samples $(x_1; x_2; n_1; n_2)$. For example, the first value of $N = 82$ in Table 1was obtained as follows: (1) As $K = 1$, all the samples with $n = 20(2)100$ are considered, and this yields 40 sample spaces where $n_1 = n_2 = n/2$.

Table 1
Values of H^0 (% of failures) for the selected asymptotic methods ($N = n^0$ of tables for calculating the H^0 values)

0.01–0.10	One-tail								Methods			
$K = 1$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FM
$E \leqslant 1.5$	82	12	10	7	2	7	2	7	2	100	20	80
$1.5 < E \leqslant 2.5$	152	0	0	0	0	0	0	0	0	100	0	16
$2.5 < E \leqslant 3.5$	166	40	36	40	6	40	17	40	14	96	41	2
$3.5 < E \leqslant 4.5$	230	3	3	3	3	3	3	3	3	70	4	2
$4.5 < E \leqslant 6.0$	350	6	2	3	5	6	2	6	3	9	7	2
$6.0 < E \leqslant 7.5$	374	3	2	1	1	3	1	3	1	16	3	1
$7.5 < E \leqslant 9.0$	363	2	1	2	1	2	2	2	1	14	2	1
$9.0 < E \leqslant 10.5$	352	4	2	1	1	4	1	4	1	9	5	1
$10.5 < E \leqslant 12$	336	4	1	1	3	3	1	3	1	7	4	2
$E > 12$	1553	2	1	0	2	2	0	2	1	8	2	1
$1 < K \leqslant 1.25$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FM
$E \leqslant 1.5$	623	28	25	22	21	24	21	24	21	100	26	63
$1.5 < E \leqslant 2.5$	993	0	0	0	0	0	0	0	0	100	0	21
$2.5 < E \leqslant 3.5$	1104	40	25	23	18	29	20	29	20	87	28	2
$3.5 < E \leqslant 4.5$	1508	5	5	5	4	5	5	5	5	62	6	2
$4.5 < E \leqslant 6.0$	2399	7	6	5	4	7	4	7	4	17	5	2
$6.0 < E \leqslant 7.5$	2648	6	4	3	2	5	3	5	3	11	3	1
$7.5 < E \leqslant 9.0$	2782	5	3	3	1	5	3	5	3	9	3	2
$9.0 < E \leqslant 10.5$	2755	3	2	2	1	3	2	3	2	7	2	2
$10.5 < E \leqslant 12$	2753	3	2	1	1	1	2	3	1	7	2	2
$E > 12$	12 703	1	1	1	0	1	1	1	1	4	1	1
$1.25 < K < 1.5$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FM
$E \leqslant 1.5$	657	45	35	37	34	37	34	37	34	100	33	9
$1.5 < E \leqslant 2.5$	1006	4	4	4	3	4	4	4	4	82	11	17
$2.5 < E \leqslant 3.5$	1134	31	22	19	16	26	18	25	17	76	21	7
$3.5 < E \leqslant 4.5$	1576	11	10	9	9	11	9	11	9	51	9	6
$4.5 < E \leqslant 6.0$	2358	11	10	9	7	11	9	11	9	31	7	5
$6.0 < E \leqslant 7.5$	2588	6	5	4	3	6	4	5	4	15	3	2
$7.5 < E \leqslant 9.0$	2650	4	3	3	2	4	2	4	2	8	3	2
$9.0 < E \leqslant 10.5$	2572	2	2	1	1	2	1	2	1	4	1	1
$10.5 < E \leqslant 12$	2529	1	1	1	1	1	1	1	1	2	1	1
$E > 12$	8561	1	1	1	0	1	1	1	1	2	1	0
$1.5 < K \leqslant 1.75$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FM
$E \leqslant 1.5$	494	51	49	50	47	50	47	50	47	100	27	47
$1.5 < E \leqslant 2.5$	944	20	19	18	17	19	17	19	17	68	12	16
$2.5 < E \leqslant 3.5$	1097	30	26	24	24	29	24	29	24	65	23	9
$3.5 < E \leqslant 4.5$	1360	14	12	11	10	13	11	13	11	51	6	6
$4.5 < E \leqslant 6.0$	2138	6	4	4	3	6	4	6	4	22	3	0
$6.0 < E \leqslant 7.5$	2202	2	1	1	1	2	1	2	1	12	1	0
$7.5 < E \leqslant 9.0$	2222	1	1	1	1	1	1	1	1	8	1	0
$9.0 < E \leqslant 10.5$	2163	1	1	1	1	1	1	1	1	5	1	0
$10.5 < E \leqslant 12$	2046	1	1	1	1	1	1	1	1	2	1	0
$E > 12$	5145	1	1	1	0	1	1	1	1	2	1	0

Table 1 (*Continued*)

0.01–0.10	One-tail								Methods			
$K = 1$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FM
$1.75 < K \leqslant 2$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FM
$E \leqslant 1.5$	608	50	50	50	50	50	50	50	50	99	25	31
$1.5 < E \leqslant 2.5$	915	27	29	31	32	28	32	29	32	59	5	6
$2.5 < E \leqslant 3.5$	1035	22	20	18	18	21	19	20	18	51	11	4
$3.5 < E \leqslant 4.5$	1379	6	5	5	5	5	5	5	5	50	3	1
$4.5 < E \leqslant 6.0$	2095	1	1	1	1	2	1	1	1	24	1	1
$6.0 < E \leqslant 7.5$	2137	1	1	1	0	1	1	1	1	11	1	0
$7.5 < E \leqslant 9.0$	2154	1	0	0	0	1	0	1	0	7	1	0
$9.0 < E \leqslant 10.5$	1961	1	1	1	0	1	1	1	1	5	1	0
$10.5 < E \leqslant 12$	1799	1	0	0	0	1	0	1	0	3	0	0
$E > 12$	3207	0	0	0	0	0	0	0	0	1	0	0
$2 < K \leqslant 3$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FM
$E \leqslant 1.5$	1672	72	71	71	70	72	71	71	70	57	19	18
$1.5 < E \leqslant 2.5$	3111	32	31	31	32	31	31	31	31	45	4	5
$2.5 < E \leqslant 3.5$	3354	28	26	25	21	28	24	27	24	44	4	0
$3.5 < E \leqslant 4.5$	4188	8	5	6	3	8	5	8	5	45	1	0
$4.5 < E \leqslant 6.0$	6341	4	2	2	1	4	2	4	2	26	1	0
$6.0 < E \leqslant 7.5$	6181	2	1	1	0	2	1	2	1	10	1	0
$7.5 < E \leqslant 9.0$	5792	1	1	1	0	1	1	1	1	5	0	0
$9.0 < E \leqslant 10.5$	4884	1	0	0	0	1	0	1	0	3	0	0
$10.5 < E \leqslant 12$	3685	0	0	0	0	0	0	0	0	2	0	0
$E > 12$	3019	0	0	0	0	0	0	0	0	2	0	0
$K > 3$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FM
$E \leqslant 1.5$	5354	91	89	90	87	91	90	91	90	23	9	1
$1.5 < E \leqslant 2.5$	7518	40	36	36	35	40	36	40	36	42	20	0
$2.5 < E \leqslant 3.5$	8656	23	21	21	19	23	21	23	21	43	5	0
$3.5 < E \leqslant 4.5$	9999	2	1	2	1	3	2	2	2	46	0	0
$4.5 < E \leqslant 6.0$	11 338	1	0	0	0	1	1	1	0	22	0	0
$6.0 < E \leqslant 7.5$	7144	0	0	0	0	0	0	0	0	6	0	0
$7.5 < E \leqslant 9.0$	3977	0	0	0	0	0	0	0	0	2	0	0
$9.0 < E \leqslant 10.5$	1663	0	0	0	0	0	0	0	0	2	0	0
$10.5 < E \leqslant 12$	371	0	0	0	0	0	0	0	0	1	0	0

0.01–0.10	Two-tails									Methods				
$K = 1$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$E \leqslant 1.5$	76	11	3	0	0	0	0	0	0	100	13	87	87	87
$1.5 < E \leqslant 2.5$	162	0	0	0	0	0	1	0	2	100	4	94	94	94
$2.5 < E \leqslant 3.5$	98	4	2	2	2	2	2	2	2	100	78	4	4	4
$3.5 < E \leqslant 4.5$	220	2	2	1	2	2	1	2	2	98	3	2	2	2
$4.5 < E \leqslant 6.0$	300	3	2	2	1	3	1	3	1	50	5	1	1	1
$6.0 < E \leqslant 7.5$	306	5	2	2	4	5	2	5	2	29	7	3	3	3
$7.5 < E \leqslant 9.0$	312	3	3	3	3	3	3	3	2	15	3	5	5	5
$9.0 < E \leqslant 10.5$	329	5	4	2	2	5	2	5	2	13	6	1	1	1
$10.5 < E \leqslant 12$	289	3	2	1	1	3	1	2	1	14	3	2	2	2
$E > 12$	1376	4	2	2	4	4	2	4	2	11	5	4	4	4

Table 1 (*Continued*)

0.01–0.10	Two-tails									Methods				
$K = 1$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$1 < K \leqslant 1.25$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$E \leqslant 1.5$	324	29	35	40	44	36	45	37	48	100	39	19	100	15
$1.5 < E \leqslant 2.5$	845	17	19	22	26	20	27	20	28	99	27	30	76	15
$2.5 < E \leqslant 3.5$	916	9	7	7	5	8	6	8	6	91	17	28	18	44
$3.5 < E \leqslant 4.5$	1337	1	1	1	1	1	1	1	1	62	7	22	14	19
$4.5 < E \leqslant 6.0$	2186	2	2	1	1	2	1	2	1	35	3	21	9	23
$6.0 < E \leqslant 7.5$	2238	2	1	1	1	2	1	2	1	18	2	14	6	21
$7.5 < E \leqslant 9.0$	2453	2	1	1	1	1	1	1	1	10	2	11	4	22
$9.0 < E \leqslant 10.5$	2445	2	1	1	1	2	1	2	1	6	2	11	5	19
$10.5 < E \leqslant 12$	2424	1	1	1	1	1	1	1	1	5	2	9	4	21
$E > 12$	11 273	2	2	1	1	2	1	2	1	5	2	8	3	18
$1.25 < K \leqslant 1.5$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$E \leqslant 1.5$	388	58	61	68	72	63	75	64	77	100	95	19	100	19
$1.5 < E \leqslant 2.5$	766	14	14	17	17	15	18	15	18	73	45	43	60	30
$2.5 < E \leqslant 3.5$	1077	7	6	7	7	7	6	7	7	56	27	45	30	21
$3.5 < E \leqslant 4.5$	1252	4	4	4	4	4	4	4	4	48	13	23	17	22
$4.5 < E \leqslant 6.0$	2084	3	3	3	3	3	3	3	3	41	7	18	13	23
$6.0 < E \leqslant 7.5$	2308	2	2	2	2	2	2	2	2	25	4	16	10	22
$7.5 < E \leqslant 9.0$	2316	2	1	1	1	2	1	1	1	14	3	10	6	20
$9.0 < E \leqslant 10.5$	2263	1	1	1	1	1	1	1	1	7	2	10	4	20
$10.5 < E \leqslant 12$	2242	1	1	1	1	1	1	1	1	5	1	9	2	20
$E > 12$	7592	1	1	1	1	1	1	1	1	3	1	8	2	17
$1.5 < K \leqslant 1.75$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$E \leqslant 1.5$	358	9	21	30	40	28	46	31	48	95	93	37	100	37
$1.5 < E \leqslant 2.5$	675	6	5	5	8	5	7	5	7	54	85	40	54	8
$2.5 < E \leqslant 3.5$	981	12	12	14	13	13	14	13	14	54	43	29	28	21
$3.5 < E \leqslant 4.5$	1100	6	5	5	6	5	5	6	5	47	22	23	19	27
$4.5 < E \leqslant 6.0$	1900	3	3	2	2	3	2	3	2	42	9	18	13	22
$6.0 < E \leqslant 7.5$	1954	1	1	1	1	1	1	1	1	30	3	15	6	22
$7.5 < E \leqslant 9.0$	1954	1	1	0	1	1	0	1	0	16	3	14	4	18
$9.0 < E \leqslant 10.5$	1910	1	0	0	0	1	0	1	0	10	1	11	1	19
$10.5 < E \leqslant 12$	1793	0	0	0	1	0	0	0	0	5	1	9	2	18
$E > 12$	4576	0	0	0	0	0	0	0	0	4	0	8	2	17
$1.75 < K \leqslant 2$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$E \leqslant 1.5$	446	2	5	7	11	4	14	5	15	90	95	43	100	43
$1.5 < E \leqslant 2.5$	688	10	8	9	9	10	9	10	9	59	70	38	59	21
$2.5 < E \leqslant 3.5$	963	13	13	15	16	13	15	13	16	46	43	33	34	24
$3.5 < E \leqslant 4.5$	1106	2	1	1	2	2	1	2	1	44	21	25	21	24
$4.5 < E \leqslant 6.0$	1852	1	1	1	1	1	1	1	1	44	8	19	11	21
$6.0 < E \leqslant 7.5$	1912	1	1	0	0	1	0	1	0	35	3	16	3	20
$7.5 < E \leqslant 9.0$	1895	1	0	0	1	1	0	1	0	18	1	13	2	18
$9.0 < E \leqslant 10.5$	1733	1	1	1	1	1	1	1	1	11	1	11	2	20
$10.5 < E \leqslant 12$	1559	1	0	0	0	1	0	1	0	6	0	9	2	18
$E > 12$	2857	1	0	0	0	1	0	1	0	6	1	9	2	17

Table 1 (*Continued*)

0.01–0.10	Two-tails									Methods				
$K = 1$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$2 < K \leqslant 3$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$E \leqslant 1.5$	1323	47	47	48	49	47	48	47	48	85	84	27	100	27
$1.5 < E \leqslant 2.5$	2184	11	13	14	17	12	16	12	16	81	83	37	64	23
$2.5 < E \leqslant 3.5$	3043	13	14	17	19	13	17	13	17	60	53	28	37	19
$3.5 < E \leqslant 4.5$	3386	2	2	2	2	2	2	2	2	48	28	27	21	24
$4.5 < E \leqslant 6.0$	5582	1	1	1	2	1	1	1	1	47	12	20	13	20
$6.0 < E \leqslant 7.5$	5537	1	1	0	1	1	0	1	0	38	4	15	6	20
$7.5 < E \leqslant 9.0$	5115	1	0	0	1	1	0	1	0	21	2	14	3	18
$9.0 < E \leqslant 10.5$	4296	1	1	0	1	1	0	1	0	11	1	12	3	19
$10.5 < E \leqslant 12$	3247	1	1	1	1	1	1	1	1	7	1	11	2	18
$E > 12$	2692	2	1	1	0	2	1	2	1	6	2	9	2	18
$K > 3$	N	UH	UM	PH	PM	UT	PT	UTA	PTA	A1	A2	FMH	FMA	FMM
$E \leqslant 1.5$	4888	51	52	52	53	51	52	51	52	92	97	1	100	1
$1.5 < E \leqslant 2.5$	4966	15	18	18	22	16	18	16	19	98	97	23	74	20
$2.5 < E \leqslant 3.5$	7629	8	10	11	15	8	11	8	11	69	59	24	41	19
$3.5 < E \leqslant 4.5$	8120	2	2	2	3	2	2	2	2	52	25	26	22	23
$4.5 < E \leqslant 6.0$	10 047	0	0	1	2	0	1	0	1	50	7	20	11	21
$6.0 < E \leqslant 7.5$	6391	0	0	0	0	0	0	0	0	37	2	15	5	21
$7.5 < E \leqslant 9.0$	3482	0	0	0	0	0	0	0	0	19	1	13	3	19
$9.0 < E \leqslant 10.5$	1452	0	0	0	0	0	0	0	0	10	1	13	2	19
$10.5 < E \leqslant 12$	327	0	0	0	0	0	0	0	0	14	0	9	0	24

Each sample space (n_1, n_2), where $10 \leqslant n_i \leqslant 50$, consists of $(n_1+1)\times(n_2+1)$ samples (x_1, x_2), where $0 \leqslant x_i \leqslant n/2$, which gives $N_1 = \sum_{n=10}^{n=50}(n+2)^2/4$ possible samples; (2) Each of the N_1 samples has a value $E = (\text{Min } a_i)/2$; the N_2 samples are selected with $E < 1.5$; (3) Each of the N_2 samples yields a one-tailed p-value (P_{E}) obtained by the optimal version of the Barnard test (the method of the minimum: Barnard, 1947); the N_3 samples are selected with $0.01 \leqslant P_{\text{E}} \leqslant 0.10$. The N_3 value is exactly $N = 82$. For a given asymptotic method, P_{A} is computed in each of these N samples and it is noted whether $|P_{\text{A}} - P_{\text{E}}| \leqslant \delta P_{\text{E}}$ or $|P_{\text{A}} - P_{\text{E}}| > \delta P_{\text{E}}$. In the first case, it is said that the asymptotic method does "not fail"; in the second, it is said that it "fails". If N^0 is the number of samples in which the method fails, then the value $H^0 = 100 \times N^0/N$ is the percentage of time it fails, which will be crucial for evaluating the method. To this end, the assumed values δ are those used by Martín et al. (1992) and proposed by Cochran, rather than the more liberal ones given in Martín and Silva (1996), that is

$$\delta = \begin{cases} 0.5 & \text{if } 0.001 \leqslant P_{\text{E}} \leqslant 0.01, \\ 0.575 - 7.5P_{\text{E}} & \text{if } 0.01 < P_{\text{E}} < 0.05, \\ 0.2 & \text{if } 0.05 \leqslant P_{\text{E}} \leqslant 0.10. \end{cases}$$

Table 1 contains the values of H^0 for all the methods indicated in the case of $1\% \leqslant P_{\text{E}} \leqslant 10\%$ (which are the most customary target significances). The results for low values of P_{E} may be obtained from the authors on request. From these results, the following general conclusions may be drawn:

(1) Generally speaking, all the methods improve with the increase of E, but are worse with the increase of the value of P_{E}. For large values of E (so that H^0 is small), all the methods generally improve with the increase in K, but perform worse as a two-tailed test than as a one-tailed test.

(2) All the asymptotic methods have some failure ($H^0 > 0\%$) in the two-tailed tests, high P_{E} values or balanced samples ($1 \leqslant K \leqslant 1.5$). This is an unfortunate circumstance given the usual situation. The problem persists even when $E > 18$ (these additional results may be requested from the authors).

(3) The minimum value of E for a type χ^2 statistic (with the relevant c.c.) not to fail is quite higher in the present case than when it is used as an approximation to Fisher's exact test (the results of Martín et al., 1992). Hence, it can be affirmed that the χ^2 test performs more erratically as an approximation to the unconditional test than to the conditional test.

(4) The A1 method performs very poorly, is worse than all the others and does not give reliable results even when $E > 12$. However, the slight modification to it which method A2 implies, makes its behaviour competitive.

(5) Of the type χ^2 methods, none behaves systematically better than the others. A good choice is the method UTA (PTA) for the low (high) P_{E}. Frequently, it is one of the best methods and has the advantage of being based on a formula that is easy to remember (and easy to apply).

(6) The mid-p method for one-tailed test (FM) performs quite well, especially with unbalanced samples or low P_{E} values. Of the two-tailed mid-p methods, the FMM (FMA) has the best performance in low (high) P_{E} values.

(7) Of the methods that can be applied "by hand" (arc sine and χ^2), the method A2 (PTA) is usually preferable in the low (high) P_{E} values.

(8) Overall, the selected mid-p methods are preferable to the others (of any type) in the two-tailed tests with low P_{E} values (FMM) and in those with one tail with high P_{E} values (FM).

A complementary aspect of the question is what happens when a method fails ($H^0 > 0$). If the failures are always conservative, that is, if $P_{\mathrm{A}} > (1 + \delta)P_{\mathrm{E}}$, then the test never gives false significances and is always reliable. This is what occurs almost always with the selected mid-p methods (see the data in Table 2). But the same does not hold for the other methods, as their failures are frequently liberal (these data may be requested from the authors).

3.3. Optimal method and validity conditions

In reality, it is not enough for a method to have the best general behaviour, rather, it is necessary to indicate in what conditions it is reliable (i.e. it does not fail). For a particular table in which the values of K and E are known, the authors advise applying any of those methods in Table 1 giving $H^0 = 0$ and, if there is none, the optimal mid-p (which is always conservative) or an exact method. The criterion of demanding that $H^0 = 0$ is useful for the researcher, for whom it is important to be sure that, in his/her

Table 2
Values of H^+/H^- (conservative/liberal failures) for the optimal mid- p methods

One-tail	0.001–0.01 FM			0.01–0.10 FM			0.001–0.01 FMM			0.01–0.10 FMA			Two-tails
$K = 1$	N	H^+	H^-	N	H^+	H^-	N	H^+	H^-	N	H^+	H^-	
$E \leqslant 1.5$				82	80	0				76	87	0	
$1.5 < E \leqslant 2.5$	82	41	0	152	16	0	2	100	0	162	94	0	
$2.5 < E \leqslant 3.5$	152	70	0	166	1	1	162	72	0	98	4	0	
$3.5 < E \leqslant 4.5$	98	0	0	230	1	1	82	66	0	220	2	0	
$4.5 < E \leqslant 6.0$	239	1	0	350	2	0	234	3	0	300	1	0	
$6.0 < E \leqslant 7.5$	259	0	0	374	1	0	238	0	0	306	3	0	
$7.5 < E \leqslant 9.0$	249	1	0	363	0	1	233	1	0	312	5	0	
$9.0 < E \leqslant 10.5$	247	1	0	352	1	0	227	1	0	329	1	0	
$10.5 < E \leqslant 12$	255	0	0	336	2	0	219	0	0	289	2	0	
$E > 12$	1117	0	0	1553	0	1	1044	1	0	1376	4	0	
$1 < K \leqslant 1.25$	N	H^+	H^-	N	H^+	H^-	N	H^+	H^-	N	H^+	H^-	
$E \leqslant 1.5$				623	63	0				324	100	0	
$1.5 < E \leqslant 2.5$	283	78	0	993	21	0	181	2	0	845	76	0	
$2.5 < E \leqslant 3.5$	888	64	0	1104	0	1	762	13	0	916	18	0	
$3.5 < E \leqslant 4.5$	807	17	0	1508	2	0	721	7	0	1337	14	0	
$4.5 < E \leqslant 6.0$	1735	0	0	2399	2	0	1509	0	0	2186	9	0	
$6.0 < E \leqslant 7.5$	1797	1	0	2648	1	0	1673	0	0	2238	6	0	
$7.5 < E \leqslant 9.0$	1943	1	0	2782	2	0	1790	0	0	2453	4	0	
$9.0 < E \leqslant 10.5$	1965	0	0	2755	2	0	1839	0	0	2445	5	0	
$10.5 < E \leqslant 12$	1946	1	0	2753	1	0	1809	0	0	2424	4	0	
$E > 12$	9169	0	0	12 703	1	0	8515	0	0	11 273	3	0	
$1.25 < K \leqslant 1.5$	N	H^+	H^-	N	H^+	H^-	N	H^+	H^-	N	H^+	H^-	
$E \leqslant 1.5$				657	9	0				388	100	0	
$1.5 < E \leqslant 2.5$	367	60	0	1006	17	0	378	59	0	766	60	0	
$2.5 < E \leqslant 3.5$	940	48	0	1134	6	0	564	16	0	1077	30	0	
$3.5 < E \leqslant 4.5$	869	27	0	1576	6	0	933	0	0	1252	17	0	
$4.5 < E \leqslant 6.0$	1650	1	0	2358	4	1	1423	0	0	2084	13	0	
$6.0 < E \leqslant 7.5$	1846	1	0	2588	2	0	1660	0	0	2308	10	0	
$7.5 < E \leqslant 9.0$	1822	1	0	2650	1	1	1715	0	0	2316	6	0	
$9.0 < E \leqslant 10.5$	1818	1	0	2572	1	0	1709	0	0	2263	4	0	
$10.5 < E \leqslant 12$	1794	0	0	2529	1	0	1668	0	0	2242	2	0	
$E > 12$	6181	0	0	8561	0	0	5747	0	0	7592	2	0	
$1.5 < K \leqslant 1.75$	N	H^+	H^-	N	H^+	H^-	N	H^+	H^-	N	H^+	H^-	
$E \leqslant 1.5$	63	16	0	494	47	0	63	19	0	358	100	0	
$1.5 < E \leqslant 2.5$	312	24	0	944	16	0	292	56	0	675	54	0	
$2.5 < E \leqslant 3.5$	836	19	0	1097	9	0	542	0	0	981	28	0	
$3.5 < E \leqslant 4.5$	797	23	0	1360	5	0	824	0	0	1100	19	0	
$4.5 < E \leqslant 6.0$	1499	4	0	2138	0	0	1280	0	0	1900	13	0	
$6.0 < E \leqslant 7.5$	1557	2	0	2202	0	0	1424	0	0	1954	6	0	
$7.5 < E \leqslant 9.0$	1573	0	0	2222	0	0	1459	0	0	1954	4	0	
$9.0 < E \leqslant 10.5$	1527	0	0	2163	0	0	1441	0	0	1910	1	0	
$10.5 < E \leqslant 12$	1435	0	0	2046	0	0	1359	0	0	1793	2	0	
$E > 12$	3722	0	0	5145	0	0	3448	0	0	4576	2	0	

Table 2 (*Continued*)

One-tail	0.001–0.01	FM	0.01–0.10	FM	0.001–0.01	FMM	0.01–0.10	FMA	Two-tails
$1.75 < K \leqslant 2$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	
$E \leqslant 1.5$	157	45 0	608	31 0	157	50 0	446	100 0	
$1.5 < E \leqslant 2.5$	385	19 0	915	6 0	391	20 0	688	59 0	
$2.5 < E \leqslant 3.5$	752	29 0	1035	4 0	406	1 0	963	34 0	
$3.5 < E \leqslant 4.5$	824	33 0	1379	1 0	846	0 0	1106	21 0	
$4.5 < E \leqslant 6.0$	1458	2 0	2095	0 1	1274	0 0	1852	11 0	
$6.0 < E \leqslant 7.5$	1521	0 0	2137	0 0	1397	0 0	1912	3 0	
$7.5 < E \leqslant 9.0$	1526	0 0	2154	0 0	1400	0 0	1895	2 0	
$9.0 < E \leqslant 10.5$	1387	0 0	1961	0 0	1323	0 0	1733	2 0	
$10.5 < E \leqslant 12$	1251	0 0	1799	0 0	1188	0 0	1559	2 0	
$E > 12$	2320	0 0	3207	0 0	2134	0 0	2857	2 0	
$2 < K \leqslant 3$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	
$E \leqslant 1.5$	796	78 0	1672	18 0	801	79 0	1323	100 0	
$1.5 < E \leqslant 2.5$	1182	2 0	3111	5 0	1162	2 0	2184	63 1	
$2.5 < E \leqslant 3.5$	2448	12 0	3354	0 0	1470	3 0	3043	36 0	
$3.5 < E \leqslant 4.5$	2612	29 0	4188	0 0	2545	0 0	3386	21 0	
$4.5 < E \leqslant 6.0$	4362	0 0	6341	0 0	3929	0 0	5582	13 0	
$6.0 < E \leqslant 7.5$	4415	0 0	6181	0 0	3990	0 0	5537	6 0	
$7.5 < E \leqslant 9.0$	4138	0 0	5792	0 0	3800	0 0	5115	3 0	
$9.0 < E \leqslant 10.5$	3404	0 0	4884	0 0	3256	0 0	4296	3 0	
$10.5 < E \leqslant 12$	2594	0 0	3685	0 0	2447	0 0	3247	2 0	
$E > 12$	2199	0 0	3019	0 0	1984	0 0	2692	2 0	
$K > 3$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	N	H$^+$ H$^-$	
$E \leqslant 1.5$	3094	6 0	5354	1 0	3090	6 0	4888	100 0	
$1.5 < E \leqslant 2.5$	3385	0 0	7518	0 0	3380	0 0	4966	71 3	
$2.5 < E \leqslant 3.5$	5726	1 0	8656	0 0	3672	0 0	7629	41 0	
$3.5 < E \leqslant 4.5$	6431	9 0	9999	0 0	5924	0 0	8120	22 0	
$4.5 < E \leqslant 6.0$	7818	0 0	11 338	0 0	7023	0 0	10 047	11 0	
$6.0 < E \leqslant 7.5$	5042	0 0	7144	0 0	4568	0 0	6391	5 0	
$7.5 < E \leqslant 9.0$	2751	0 0	3977	0 0	2612	0 0	3482	3 0	
$9.0 < E \leqslant 10.5$	1154	0 0	1663	0 0	1107	0 0	1452	2 0	
$10.5 < E \leqslant 12$	262	0 0	371	0 0	241	0 0	327	0 0	

particular table, the asymptotic method gives a reliable p-value, it is not enough to know that the method "generally" works well.

If one wishes to select the optimal method in terms of the objective error α, it may be concluded in a very general way that

(I) For researchers interested only in the very low significances (because they have to apply Bonferroni's method, for example), the optimal is the method FMM for the two-tailed tests. In the one-tailed test, the selection depends on the sample inbalance: methods UTA, A2 or FM for low, moderate and high inbalances.

(II) For researchers interested in the usual significances, the optimal is the method FM (PTA) in one-(two) tailed tests.

Table 3
Overall values of $\Delta = (P_A - P_E)/P_E$ for $0.01 \leqslant P_E \leqslant 0.10$, $n = 20(1)100$, any K and the asymptotic methods indicated under the conditions shown (E refers to the minimum expected quantity)

Family	Test	Method	Condition	Δ minimum	Δ maximum	Δ median	$\lvert\Delta\rvert$ median
χ^2	One-tailed	PTA	$E \leqslant 9$	−0.978	0.874	−0.066	0.113
			$E > 9$	−0.338	0.383	−0.030	0.064
	Two-tailed	PTA	$E \leqslant 6$	−0.956	0.998	0.021	0.105
			$E > 6$	−0.392	0.522	−0.004	0.067
Arc sine	One-tailed	A2	$E \leqslant 10.5$	−0.549	0.234	−0.084	0.091
			$E > 10.5$	−0.370	0.337	−0.047	0.068
	Two-tailed	A2	$E \leqslant 9$	−0.778	1.428	−0.037	0.167
			$E > 9$	−0.434	0.362	−0.048	0.077
mid-p	One-tailed	FM	$E \leqslant 6$	−0.246	0.580	0.073	0.085
			$E > 6$	−0.306	0.539	0.029	0.060
	Two-tailed	FMA	$E \leqslant 10.5$	−0.291	2.093	0.133	0.139
			$E > 10.5$	−0.348	0.507	0.054	0.078

However, it is clear that the reader is interested in more generic rules of performance, which means having to allow between 1% and 2% failure in some circumstances. For the ordinary significances

(A) The method PTA may be applied for $E > 9$ (6) in the one(two)-tailed tests.

(B) The method A2 may be applied for $E > 10.5$ ($E > 9$ and $K \neq 1$) in the one(two)-tailed tests.

(C) The method FM (FMA) may always be applied in the one(two)-tailed tests, although for $E \leqslant 6$ ($E \leqslant 10.5$) it tends to give too many conservative results.

For the very low significances

(A′) The method UTA may be applied for $E > 7.5$ (9) in the one(two)-tailed tests.

(B′) The method A2 may be applied for $E > 7.5$ in the one- and two-tailed tests.

(C′) The method FM (FMM) may always be applied in the one- and two-tailed tests, although for $E \leqslant 4.5$ it tends to give too many conservative results.

An alternative way of evaluating the 3 previous methods (under the afore-mentioned v.c.) consists in determining the experimental value for the relative error $\Delta = (P_A - P_E)/P_E$ (note that $\delta = |\Delta|$) for all the tables studied here. The minimum, median and maximum values for Δ (as well as the median value for $|\Delta|$) are given in Table 3. It can be seen that when the v.c.'s are not verified, the tests are not reliable (particularly the two-tailed tests). For example (and for two-tailed tests), method FMA can give an approximate p-value of more than three times the real p-value (Δ maximum = 2.093); method PTA can give an approximate p-value of either nearly twice (Δ maximum = 0.998) or almost 4% (Δ minimum = −0.956) that of the real one.

4. Summary, conclusions and an example

The present article has evaluated some 60 asymptotic methods constituting an approximation to Barnard's unconditional test (comparison of two independent proportions).

Included in these are the 4 optimal methods selected by Martín and Silva (1996). The base criterion for the evaluation is that the asymptotic method should yield a p-value that is approximately equal to that of the exact method (using the criterion of Cochran). In this sense, it is said that an asymptotic method does not fail if it yields, for an exact p-value of 5% (for example), an approximate p-value which is between 4% and 6%. From the article, it can be deduced that for the ordinary significances and tests of one (two) tails, the competitive methods are PTA, A2 and FM (PTA, A2 y FMA). The method A1 (traditional arc sine) should not be applied.

The methods do not have a universal application, but generally will be valid when E (the minimum expected quantity) is sufficiently large. All of them behave badly when the samples are balanced. For the ordinary significances: (a) The method PTA, which is the simplest, can be applied when $E > 9$ (6) in the one(two)-tailed tests, but this implies the acceptance of failure between 1% and 2%, (b) The method A2 may be applied for $E > 10.5$ ($E > 9$ and $K \neq 1$) in the one(two)-tailed tests; (c) The method FM (FMA), which is the most complex of all, can always be applied as a one(two)-tailed test; its failures, which can be excessive for $E \leqslant 6$ ($E \leqslant 10.5$), are always conservative.

As an example, let the data be $x_1/n_1 = 14/40$ vs. $x_2/n_2 = 28/50$ with a two-tailed exact p-value $P_{\mathrm{E}} = 5.068\%$. Since $E = 18.7$, all the previous asymptotic methods are valid. For the PTA method, $P_{\mathrm{PTA}} = 4.900\%$ since $\chi^2_{\mathrm{PTA}} = \{|14 \times 22 - 26 \times 28| - 1\}^2 \times 89/\{42 \times 48 \times 40 \times 50\} = 3.875$. For the A2 method, $P_{\mathrm{A2}} = 4.783\%$ since $\chi^2_{\mathrm{A2}} = (81 \times 101/91)\{\sin^{-1}(14.375/40.75)^{0.5} - \sin^{-1}(28.375/50.75)^{0.5}\}^2 = 3.916$. For the FMA method, since $\mathrm{FMA}(x_1) = 2\mathrm{FM}(x_1)$, it is necessary first to apply expression (6); in this, $x_1 = 14$ is not larger than $E_{11} = 18.7$, for which reason one will have to permute the x_i values for the y_i values and compare the samples 26/40 vs. 22/50. In this way, $\mathrm{FM}(x_1 = 26) = 2.5736\%$, and thus $P_{\mathrm{FMA}} = 5.147\%$. As one can see (and this is quite usual), the three methods yield a p-value that is very near to the real one, but PTA and A2 are liberal, while FMA is conservative.

Acknowledgements

This research was supported by the Dirección General de Investigación, Spain, Grant No. BFM-2000-1472.

References

Agresti, A., 2001. Exact inference for categorical data: recent advances and continuing controversies. Statist. Med. 20, 2709–2722.

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

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

Cochran, W.G., 1954. Some methods for strengthening the common χ^2 tests. Biometrics 10, 417–451.

Conover, W.J., 1974. Some reasons for not using Yates' continuity corrections on 2 × 2 contingency tables. J. Amer. Statist. Assoc. 69, 374–376.

Cook, I.T., 1981. On the continuity correction for bivariate discrete distribution. Private communication to Upton, 1982.

Cox, D.R., 1970. The continuity correction. Biometrika 57, 217–219.

D'Agostino, R.B., Chase, W., Belanger, A., 1988. The appropriateness of some common procedures for testing the equality of two independent binomial populations. Amer. Statist. 42 (3), 198–202.

Davis, A.B., 1993. Power of testing proportions in small two-sample studies when sample sizes are equal. Statist. Med. 12, 777–787.

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

Haber, M., 1982. The continuity correction and statistical testing. Internat. Statist. Rev. 50, 135–144.

Haber, M., 1986. A modified exact test for 2 × 2 contingency tables. Biometrical J. 28 (4), 455–463.

Haber, M., 1987. A comparison of some conditional and unconditional exact tests for 2 × 2 contingency tables. Comm. Statist. - Simula. Comp. 16 (4), 999–1013.

Haber, M., 1992. On the expected significance probabilities and Bahadur efficiencies of tests for comparing two binomial proportions. J. Statist. Comput. Simula. 43, 243–251.

Hirji, K.F., Tan, S.J., Elashoff, R.M., 1991. A quasi-exact test for comparing two binomial proportions. Statist. Med. 10, 1137–1153.

Kendall, M.G., Stuart, A., 1973. The Advanced Theory of Statistics, Vol. 2, 2nd Edition. Hafner Pub. Co, New York.

Mantel, N., 1974. Comment and a suggestion. J. Amer. Statist. Assoc. 69, 378–380.

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

Martín Andrés, A., Herranz Tejedor, I., 1997. On condition for validity of the approximations to Fisher's exact test. Biometrical J. 39, 935–954.

Martín Andrés, A., Herranz Tejedor, I., 2000. On the minimum expected quantity for the chi-square test in 2 × 2 tables. J. Appl. Statist. 27 (7), 807–820.

Martín Andrés, A., Silva Mato, A., 1996. Optimal correction for continuity and conditions for validity in the unconditional chi-squared test. Comput. Statist. Data Anal. 21, 609–626. Erratum in 26, 1997, p. 235.

Martín Andrés, A., Herranz Tejedor, I., Luna del Castillo, J.D., 1992. Optimal correction for continuity in the chi-squared test in 2 × 2 tables (conditioned method). Comm. Statist. - Simula. Comp. 21 (4), 1077–1101.

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.

Routledge, R.D., 1992. Resolving the conflict over Fisher's exact test′. Canad. J. Statist. 20 (2), 201–209.

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

Silva Mato, A., Martín Andrés, A., 1997. Simplifying the calculation of the P-value for Barnard's test and its derivatives. Statist. Comput. 7, 137–143.

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

Yates, F., 1934. Contingency tables involving small numbers and the χ^2 test. J. Roy. Statist. Soc. Suppl. 1, 217–235.

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