STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1159–1175
Published online 23 Januray 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3531

TUTORIAL IN BIOSTATISTICS

Recommended tests for association in 2×2 tables

Stian Lydersen[1,*,†], Morten W. Fagerland[2] and Petter Laake[3]

[1]*Unit for Applied Clinical Research, Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway*
[2]*Ullevål Department of Research Administration, Oslo University Hospital, Norway*
[3]*Department of Biostatistics, University of Oslo, Norway*

SUMMARY

The asymptotic Pearson's chi-squared test and Fisher's exact test have long been the most used for testing association in 2×2 tables. Unconditional tests preserve the significance level and generally are more powerful than Fisher's exact test for moderate to small samples, but previously were disadvantaged by being computationally demanding. This disadvantage is now moot, as software to facilitate unconditional tests has been available for years. Moreover, Fisher's exact test with mid-p adjustment gives about the same results as an unconditional test. Consequently, several better tests are available, and the choice of a test should depend only on its merits for the application involved. Unconditional tests and the mid-p approach ought to be used more than they now are. The traditional Fisher's exact test should practically never be used. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: 2×2 tables; Fisher's exact test; Pearson's chi-squared test; unconditional tests; mid-p-value

1. INTRODUCTION

A 2×2 table is a way of summarizing the observed cross-classification of two dichotomous random variables. An example shown in Table I comprises the results of a double blind trial of high dose versus standard dose of epinephrine in children with cardiac arrest [1]. One of the 34 children in the high dose group survived 24 h, while 7 of the 34 children in the standard dose group survived 24 h. Another example is shown in Table II, which comprises a classification of CHRNA4 genotypes and presence of exfoliation syndrome in the eyes [2].

Association in 2×2 tables traditionally has been tested using the asymptotic Pearson's chi-squared test for larger samples and Fisher–Irwins's exact test, usually called Fisher's exact test, for smaller samples. But these two tests have lesser-known drawbacks. The asymptotic test may

*Correspondence to: Stian Lydersen, Unit for Applied Clinical Research, Department of Cancer Research and Molecular Medicine, NTNU, N-7489 Trondheim, Norway.
†E-mail: stian.lydersen@ntnu.no

Received 20 February 2008
Accepted 8 December 2008

Table I. Treatment of children with cardiac arrest. High dose versus standard dose epinephrine [1].

	Survival at 24 h		
Treatment	Yes	No	Sum
High dose	1	33	34*
Standard dose	7	27	34*
Sum	8	60	68*

An asterisk * denotes the sums fixed by design.

Table II. Genotype (CHRNA4-CC versus CHRNA4-CT or -TT) and presence of exfoliative syndrome in the eyes (XFS) [2].

	XFS		
	Yes	No	Sum
CHRNA4-CC	0	16	16
CHRNA4-TC/TT	15	57	72
Sum	15	73	88*

An asterisk * denotes the sum fixed by design.

not preserve the test size, that is, the actual significance level may be higher than the nominal significance level. Fisher's exact test is conservative, that is, other tests generally have higher power yet still preserve test size. In the examples of Tables I and II, neither of these traditional methods performs well.

Many significance tests are possible, depending on a variety of choices, including

- Level of conditioning in the sample space (possible tables): Should the p-value be computed unconditionally or conditionally on one or more of the marginal sums in the observed table?
- Choice of test statistic, such as Pearson's or Fisher's.
- Exact or asymptotic calculation of the p-value.
- Further adjustments, such as the mid-p-value.

The number of conceivable tests is large. For example, Martín Andrés and Silva Mato [3] studied 60 asymptotic tests for comparing binomial proportions. Exact analysis of discrete data, including 2×2 tables, is described in a more general framework in a recent book by Hirji [4].

In the present article, we describe the main principles underpinning various tests in 2×2 tables and recommend when each of them might best be used. Estimation and confidence intervals for effect size are outside the scope of the article. The common experimental designs underlying 2×2 tables are defined in Section 2. The most common test statistics are defined in Section 3. Various ways of defining and computing the p-value are described in Section 4. Section 5 summarizes what is known about the tests. Some readers may wish to skip Sections 3–5 and go straight to Section 6, where we summarize the definitions and properties of the most common tests, and the recommended tests. Examples illustrating the tests are given in Section 7. Power and sample size calculations are briefly covered in Section 8. Our recommendations for choice of tests are given in Section 9.

DOI: 10.1002/sim

2. EXPERIMENTAL DESIGNS AND HYPOTHESES

The counts of a 2×2 table may be summarized as illustrated in Table III. Such a table may result from different designs or sampling models, as described below.

2.1. Both margins fixed design

The classic example is 'a lady tasting a cup of tea' [5]. Fisher's colleague Muriel Bristol claims she can taste whether milk or tea was added first to her cup. Four tea first and four milk first cups are presented to her in randomized order. She is told that there are four of each kind and is asked to identify which is which. A possible result is given in Table IV. In this design, the row sums as well as the column sums are fixed beforehand. Such a design is hardly ever used in practice [6]. But understanding this design is important in understanding the nature of Fisher's exact test discussed in Section 3 below.

2.2. One margin fixed design

In clinical trials, one set of marginal sums usually is fixed beforehand, typically the number of patients in each treatment group, such as in the epinephrine example of Table I. This is also the basic design in case-control studies in epidemiology. The one margin fixed design is usually used for comparing two binomial proportions. We assume, without loss of generality, that the fixed margin comprises the row sums.

2.3. Total number fixed design

In cross-sectional studies of association, usually only the total sum N is fixed beforehand. This is the case for the exfoliation example in Table II, where only the total number of patients, $N = 88$, was fixed before genotype determination and eye examination.

Table III. The general counts of a 2×2 table.

		j		
		1	2	Sum
i	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
	Sum	n_{+1}	n_{+2}	N

Table IV. Fisher's tea-drinker.

	Guess poured first		
Poured first	Milk	Tea	Sum
Milk	3	1	4*
Tea	1	3	4*
Sum	4*	4*	8*

An asterisk * denotes the sums fixed by design.

DOI: 10.1002/sim

We consider the null hypothesis of no association between the variables defining rows and columns, for example, that the probability of success is independent of treatment. The present article focuses on the one margin fixed design and the total sum fixed design. The null hypothesis is formalized in Table V. If a row sum or column sum is zero, the table is uninformative about association. We assume that the row sums and column sums are nonzero. Most of the tests discussed are two-tailed, testing whether association is random or not.

3. COMMON TEST STATISTICS

A test statistic is a function of the observations, providing a measure of the observed table's compliance with the null hypothesis. Unless otherwise stated, we define a test statistic T in such a way that it is non-negative and tables with large values agree less with the null hypothesis than do tables with lower values.

Let $\mathbf{n}$ denote the observed table (Table III) with marginal sums $\mathbf{n}_+ = (n_{1+}, n_{2+}, n_{+1}, n_{+2})$. Under H_0, the estimated expected counts are

$$m_{ij} = n_{i+}n_{+j}/N \tag{1}$$

The most used test statistics are those defined for Pearson's chi-squared test, the likelihood ratio (LR) test, and Fisher's exact test. Pearson's chi-squared test statistic is

$$T_{\mathrm{Pe}}(\mathbf{n}) = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} \tag{2}$$

and the LR test statistic is

$$T_{\mathrm{LR}}(\mathbf{n}) = -2\log\frac{L_0}{L_1} = 2\sum_{i,j} n_{ij}\log\left(\frac{n_{ij}}{m_{ij}}\right) \quad \text{where a term is 0 if } n_{ij} = 0 \tag{3}$$

The maximum likelihood of the table under H_0 and H_1 are L_0 and L_1. Fisher's statistic is defined as the conditional probability

$$P(\mathbf{n}|\mathbf{n}_+) = \binom{n_{+1}}{n_{11}}\binom{n_{+2}}{n_{1+} - n_{11}} \Big/ \binom{N}{n_{1+}} \tag{4}$$

with small values providing evidence against H_0. All the sampling models described previously give the same results (1) to (4).

For testing equality between two proportions (one margin fixed design), another possible test statistic is the normalized difference between the observed proportions

$$z = \frac{\dfrac{n_{11}}{n_{1+}} - \dfrac{n_{21}}{n_{2+}}}{\sqrt{\dfrac{n_{+1}}{N}\cdot\dfrac{n_{+2}}{N}\left(\dfrac{1}{n_{1+}} + \dfrac{1}{n_{2+}}\right)}} \tag{5}$$

DOI: 10.1002/sim

Table V. Two common experimental designs for 2×2 tables.

Model	Fixed sums	Unknown parameters				Probability model	Null hypothesis
One	n_{1+} n_{2+} (and		Column 1	Column 2	Total	Two independent binomials	$\pi_1 = \pi_2$
margin	$N = n_{1+} + n_{2+}$)	Row 1	π_1	$1-\pi_1$	1		
fixed		Row 2	π_2	$1-\pi_2$	1	$P(\mathbf{n}) = \binom{n_{1+}}{n_{11}} \pi_1^{n_{11}} (1-\pi_1)^{n_{1+}-n_{11}} \binom{n_{2+}}{n_{21}} \pi_2^{n_{21}} (1-\pi_2)^{n_{2+}-n_{21}}$	
Total sum	N		Column 1	Column 2	Total	Multinomial	$\pi_{ij} = \pi_{i+}\pi_{+j}$
fixed		Row 1	π_{11}	π_{12}	π_{1+}	$P(\mathbf{n}) = \frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!} \pi_{11}^{n_{11}} \pi_{21}^{n_{21}} \pi_{12}^{n_{12}} \pi_{22}^{n_{22}}$	
		Row 2	π_{21}	π_{22}	π_{2+}		
		Total	π_{+1}	π_{+2}	1		

DOI: 10.1002/sim

However, $T_{\mathrm{Pe}} = z^2$, so this statistic is equivalent to Pearson's chi-squared statistic. It is sometimes denoted as z_{pooled}, as it is based on the two proportions being equal under H_0. If this is not the case, the corresponding normalized difference is

$$z_{\mathrm{unpooled}} = \frac{\dfrac{n_{11}}{n_{1+}} - \dfrac{n_{21}}{n_{2+}}}{\sqrt{\dfrac{n_{11}n_{12}}{n_{1+}^3} + \dfrac{n_{21}n_{22}}{n_{2+}^3}}} \tag{6}$$

The z_{unpooled} statistic is not recommended for testing the null hypothesis of no association [7]. Likewise, the Santner and Snell statistic [8], which is the unnormalized difference in proportions given by the numerator in (5), is not recommended for testing the null hypothesis of no association [7].

There are continuity corrections for some test statistics, aiming to improve the asymptotics in certain designs. Best known is Yates' correction for Pearson's statistic, described in Section 4 below.

4. DEFINING AND COMPUTING THE p-VALUE

In general, the p-value is defined as the probability of the test statistic T being equal to or more extreme than its value for the observed table (t_{obs})

$$p\text{-value} = P(T \geqslant t_{\mathrm{obs}} | H_0) \tag{7}$$

In general, H_0 is rejected if the p-value does not exceed α, the nominal significance level. If both marginal sums are fixed, the p-value (7) can be computed using the probabilities (4). Else, the calculated p-value depends on the design, as well as the value(s) of the unknown parameter(s), or nuisance parameter(s), under H_0. In the one margin fixed design, there is one nuisance parameter, $\pi = \pi_1 = \pi_2$, the common success probability in row 1 and 2. In the total sum fixed design, the row and column probabilities are unknown, so there are two nuisance parameters, π_{1+} and π_{+1}. A test is said to preserve test size if the actual significance level does not exceed the nominal significance level, for any value of the nuisance parameter(s). If the actual significance level is lower than α, the test is called conservative. If H_0 is rejected only if the p-value is less than α, the test may be unnecessarily conservative, since we have discrete data.

4.1. Exact conditional tests

The hindrance of unknown nuisance parameter(s) is overcome when the conditional p-value, given the marginal sums $\mathbf{n}_+ = (n_{1+}, n_{2+}, n_{+1}, n_{+2})$, is computed as

$$\text{Conditional } p\text{-value} = P(T \geqslant t_{\mathrm{obs}} | \mathbf{n}_+, H_0) \tag{8}$$

Then, H_0 is rejected if the conditional p-value does not exceed α. This is done in Fisher's exact test, which was first proposed by Irwin [9]. The widespread use of conditional tests for 2×2 tables may be ascribed to their independence of the nuisance parameter(s) and to their computational ease. In the present article, by 'conditional test' we mean conditional on row and column sums, unless otherwise stated.

DOI: 10.1002/sim

Any conditional test preserves test size, because

$$\begin{aligned} P(\text{Reject } H_0|H_0) &= \sum_{\mathbf{n}_+} P(\text{Reject } H_0|\mathbf{n}_+, H_0)\, P(\mathbf{n}_+|H_0) \\ &= \sum_{\mathbf{n}_+} P[t_{\text{obs}} \text{ is such that } P(T \geqslant t_{\text{obs}}|\mathbf{n}_+, H_0) \leqslant \alpha] P(\mathbf{n}_+|H_0) \\ &\leqslant \sum_{\mathbf{n}_+} \alpha P(\mathbf{n}_+|H_0) \\ &= \alpha \sum_{\mathbf{n}_+} P(\mathbf{n}_+|H_0) = \alpha \cdot 1 = \alpha \end{aligned} \tag{9}$$

A conditional test can be unnecessarily conservative, with actual significance level notably less than α. There are several approaches for reducing this conservatism, using

- An asymptotic method, like the asymptotic Pearson's chi-squared test. However, this may violate test size seriously for small samples.
- A mid-p-value with an exact conditional test. The test size may be violated, but typically not much [10]. This approach is called quasi-exact by Hirji *et al.* [11].
- An unconditional test that preserves test size and has high power. We regard this to be the best approach in small samples.

4.2. Asymptotic tests

The most used asymptotic tests are probably those using Pearson's chi-squared statistic (2) and the LR statistic (3), approximating the p-value as

$$\text{asymp } p\text{-value} = P(\chi_1^2 \geqslant t_{\text{obs}}) \tag{10}$$

where χ_1^2 is chi-squared distributed with one degree of freedom. These asymptotic tests can be used for all designs described above.

Yates' continuity correction for Pearson's statistic is given by

$$T_{\text{Pe,CC}}(\mathbf{n}) = \sum_{i,j} \frac{(|n_{ij} - m_{ij}| - 1/2)^2}{m_{ij}} = \frac{N\left(|n_{11}n_{22} - n_{12}n_{21}| - \dfrac{N}{2}\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}} \tag{11}$$

where m_{ij} is given by (1).

4.3. The mid-p-value

The mid-p-value is defined as

$$\text{mid-}p\text{-value} = P(T > t_{\text{obs}}) + \tfrac{1}{2} P(T = t_{\text{obs}}) \tag{12}$$

and the null hypothesis is rejected if mid-p-value $\leqslant \alpha$. The mid-p procedure was proposed by Lancaster [12] and recently has gained wider acceptance, see [4, 13, 14] and references therein. Barnard [15] recommends reporting both the p-value and the mid-p-value, arguing that the p-value measures the significance when the data are judged alone and the mid-p-value is suited for combining evidence from several studies. In fact, for a one-sided test with any discrete test statistic,

DOI: 10.1002/sim

we have $E(P\text{-value}|H_0) > \frac{1}{2}$ and $E(\text{mid-}p\text{-value}|H_0) = \frac{1}{2}$, see, for example, [4, 14]. A theoretical justification for mid-p-values in 2×2 contingency tables is provided by Hwang and Yang [16]. A mid-p test does not guarantee preservation of test size. However, a conditional mid-p test, as described above, approximately preserves test size *unconditionally* [4].

4.4. Unconditional tests

The exact tests considered this far are conditional on $\mathbf{n}_+ = (n_{1+}, n_{2+}, n_{+1}, n_{+2})$. Unconditional tests, on the other hand, assume no marginal sums fixed, save those fixed by design. A complication concerning the unconditional p-value is that $P(T \geqslant t_{\text{obs}})$ depends on the unknown nuisance parameter(s) under H_0. To ensure that the test size does not exceed α, the p-value is taken [17, 18] as

$$\max_{0 \leqslant \pi \leqslant 1} P(T \geqslant t_{\text{obs}}; \pi | H_0) \tag{13}$$

Equation (13) applies to a one margin fixed design, where the nuisance parameter is the success probability $\pi = \pi_1 = \pi_2$ under H_0. In a multinomial design (total number fixed design), maximization is performed over the two-dimensional area $[0, 1] \times [0, 1]$ for (π_{+1}, π_{1+}). The unconditional tests described here are for the one margin fixed design, unless otherwise specified.

The unconditional test described by Suissa and Shuster [19] is of this type, using Pearson's statistic. Barnard's unconditional test [20] uses a more computationally intensive algorithm for building a rejection region, and is, to our knowledge not included in any available software. StatXact provides the Suissa and Shuster test (somewhat misleadingly named Barnard's test in StatXact).

Boschloo [21] and McDonald *et al.* [22] suggested a raised conditional level of significance to reduce the conservatism in an exact conditional test. That is, reject H_0 if the conditional p-value$\leqslant \delta$, where δ is the highest number such that $P(\text{Reject } H_0) \leqslant \alpha$ for all parameter values under H_0. This is a valid procedure, see, for example, [18]. An equivalent procedure is to use Fisher's exact conditional p-value as a test statistic in an unconditional test, with small values providing evidence against H_0. We call this the Fisher–Boschloo statistic. The resulting unconditional p-value for Boschloo's test is interpreted in the usual manner and is compared with the significance level α. This p-value can be computed using the Berger software [23]. Some authors modify Boschloo's test by using the one-sided Fisher's exact conditional p-value as test statistic in the two-sided test. This is done in the software by Martín Andrés [24].

It has been pointed out that the p-value defined by (13) is maximized over all values of π, including values highly unlikely in light of the observations (see, for example Agresti ([25], pp. 95–96)). This drawback is reduced in the Berger and Boos procedure [26], where the unconditional p-value is taken as

$$\max_{\pi \in C_\gamma} P(T \geqslant t_{\text{obs}}; \pi) + \gamma \tag{14}$$

where C_γ is a $100(1-\gamma)$ per cent confidence interval for π. Here, γ is taken to be very small, such as 0.001. This procedure also preserves the test size. It is fully implemented in [23, 27].

For an unconditional test in the one margin fixed design (two binomials), the Berger software [23] allows $N \leqslant 1000$, with optional Berger and Boos correction. Other relevant softwares are [24, 27, 28]. In the total sum fixed design, the softwares [23, 24], allow sample sizes $N \leqslant 400$ and $N \leqslant 40$, respectively.

DOI: 10.1002/sim

An approximate unconditional test may be formed by inserting the maximum likelihood estimate of π in (13) instead of maximizing over π. It seems similar to the conditional mid-p test, in terms of occasionally exceeding the nominal significance level [29].

4.5. *One-sided and two-sided tests*

The test statistics and p-values defined above are generally for two-sided tests. For a one-sided test, only outcomes in the direction of the one-sided hypothesis are included in computation of the p-value.

For two-sided exact tests, the two-sided p-value may alternatively be defined as twice the smallest tail (TST), that is, twice the smallest of the one-sided p-values, instead of a probability-based p-value (8). In exact tests, this can make the test slightly more conservative. On the other hand, with the TST one always is on the safe side: When rejecting a two-sided hypothesis at level α, the TST method implies that the corresponding one-sided hypothesis will be rejected at level $\alpha/2$. These issues are further discussed in Hirji ([4], pp. 206–210), Agresti ([25], p. 93), and Altman ([30], p. 256).

Note that many software programs compute the one-sided p-value only for the smallest tail.

5. WHAT IS KNOWN ABOUT THE TESTS

5.1. *Choice of test statistic*

For conditional p-values in 2×2 tables, common test statistics like Pearson's chi-squared, LR, and Fisher's statistic are equivalent in two cases [31]:

- If the row sums (or column sums) are equal.
- For one-sided tests, regardless of the marginal sums. Hence, the same applies to two-sided twice the smaller tail tests.

Else, the three test statistics may produce differing results. Power comparisons using conditional p-values for two binomials indicate some differences, though there is no general 'loser' or 'winner' among them [10, 32]. The statistics of Fisher, and particularly Pearson, tend to be slightly more powerful than LR. However, the Fisher statistic seems more robust to design and rarely performs poorly. The same conclusions may be made for rxc tables with one fixed margin [33].

These results for conditional p-values do not extend to unconditional p-values computed by (13). For example, the results in Table I give unconditional p-value $=0.0281$ with Pearson's and Fisher's statistics, and unconditional p-value $=0.0402$ with the LR statistic. For unconditional tests for the one margin fixed design, Mehrotra *et al.* [7] compared the test statistics Pearson (z_{pooled}), z_{unpooled}, Santner and Snell and Fisher–Boschloo. Pearson and Fisher–Boschloo are recommended, as they have the highest power. Andres *et al.* [34] compared 15 test statistics, including the original test by Barnard. Barnard's test and a simplified version of Barnard's test have highest power, but are considered too computer intensive for practical use. Among the others, Pearson's chi-squared and Fisher–Boschloo have power nearly as high as the optimal Barnard's test [34].

For unconditional tests in the total sum fixed design, we are not aware of comparisons of test statistics. Hence, we have no reason to recommend other test statistics than in the one margin fixed design.

DOI: 10.1002/sim

5.2. *p-value or mid-p-value*

The conditional p-value test is conservative. The conditional mid-p test is less conservative, but does not always preserve test size. However, in some cases, the conditional mid-p test preserves the nominal level for all values of the nuisance parameter, else it tends to slightly violate test size occasionally [10]. Hirji *et al.* [11] carried out a comprehensive comparison of Fisher's exact test, Fisher exact mid-p, the asymptotic Pearson's chi-squared and two other asymptotic tests. For both one- and two-sided tests, and for a wide range of sample sizes, they found that the actual significance levels of the mid-p tests tend to be closer to the nominal level as compared with the other tests. Empirical studies show that the performance of a conditional mid-p test resembles that of an unconditional test ([4], p. 219).

5.3. *Asymptotic tests*

The original rationale for using asymptotic expressions was their ease of computation. According to Cochran's criterion [35], Pearson's asymptotic chi-squared test is inaccurate in a 2×2 table if any of the expected counts are less than five ($m_{ij}<5$). Today, computations for other approaches are readily performed. A conditional mid-p test can easily be calculated and performs somewhat better than asymptotic tests also with counts slightly above Cochran's criterion [11]. Anyway, it should be noted that Yates' correction (11) for Pearson's chi-squared test assumes that all the marginal sums $\mathbf{n}_+=(n_{1+},n_{2+},n_{+1},n_{+2})$ are fixed, which makes it a valid approximation to an exact conditional test. This correction reduces the numerical value of the test statistic, and hence reduces the power and significance level of the test, making it overly conservative [6]. We agree with authors who state that Yates' correction should no longer be used [4, 6, 25].

5.4. *Conditional or unconditional tests*

Unconditional tests are generally more powerful than conditional tests, and have been recommended recently by many authors, see [7] and references therein. One outstanding comparison is that Fisher–Boschloo's test is uniformly more powerful than Fisher's exact test, because its rejection region always includes that of Fisher's exact test [21]. This is true for one-sided tests, and hence also for TST two-sided tests.

Mehrotra *et al.* [7] studied the Berger and Boos procedure (14) with $\gamma=0.001$ for unconditional tests with Pearson's and Fisher–Boschloo's statistics and found that the procedure gives a slight improvement in test power. Kang and Ahn [36] pointed out that the Berger and Boos procedure is particularly useful in extremely unbalanced designs, comparing two binomial proportions where one sample size is, say, 20 times the other sample size. This seems sensible: The situation is almost like a one sample test for a binomial probability, where you test if the probability in the small group equals the empirical probability in the large group. To our knowledge, no research has been conducted to find optimal values of γ. Most authors use 0.001 or 0.0001. An exception is StatXact 8 using the default value 0.000001.

As an illustration, Figure 1 shows the actual significance level obtained in comparing two groups with fixed row sums $n_{1+}=n_{2+}=34$, as in the epinephrine study. Fisher's exact test is far more conservative than Fisher–Boschloo's unconditional test, which also by definition preserves test size. Fisher's conditional mid-p test performs about as well as Fisher–Boschloo's unconditional test. The mid-p test does not always preserve test size, although, in this case, it does. The Pearson's asymptotic test is not conservative, but neither does it preserve test size.

DOI: 10.1002/sim

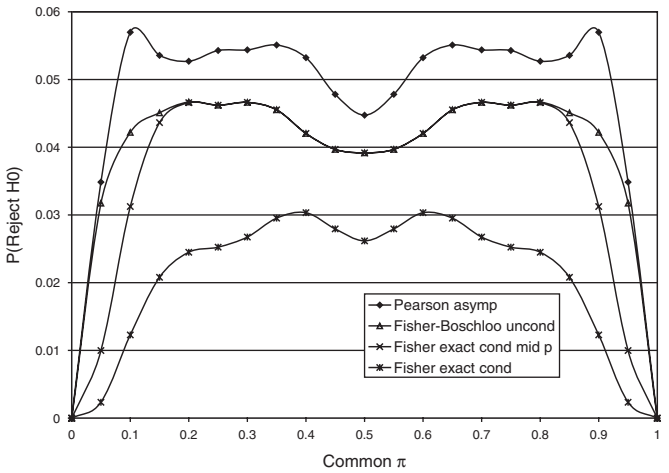

Figure 1. Actual significance level, two binomials (one margin fixed design), row sums 34, $\alpha = 0.05$.

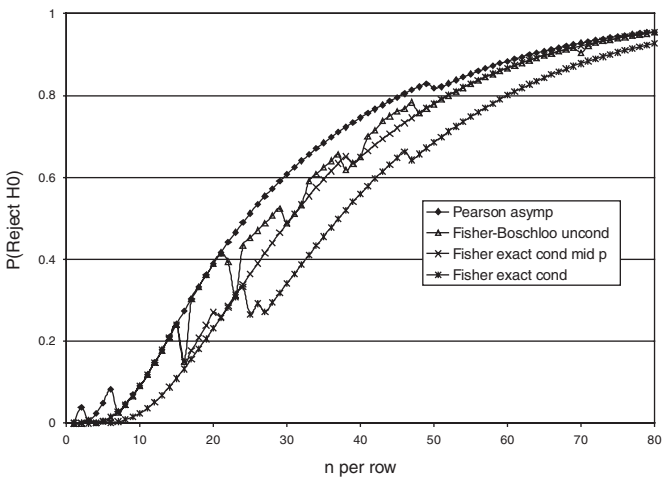

Figure 2. Power, two binomials (one margin fixed design), equal row sums, $\pi_1 = 0.03$, $\pi_2 = 0.2$, $\alpha = 0.05$.

Figure 2 shows an example of test power as a function of sample size for two equal groups with success probabilities $\pi_1 = 0.03$ and $\pi_2 = 0.2$, similar to the observed proportions in the epinephrine study, for a nominal significance level of $\alpha = 0.05$. The conditional test has lowest power. The mid-p test and the unconditional test have about the same power.

The conservatism of conditional tests is more pronounced in balanced designs than in unbalanced designs. This causes the paradox that a conditional test for a balanced design often gains power when the sample size is reduced by one in one group [37]. A conditional mid-p test, as well as an unconditional test, usually looses power with such a sample size reduction.

DOI: 10.1002/sim

Unconditional tests have been disputed, as by [38], see also page 191 of [39]. Such dispute is chiefly philosophical. Clearly, unconditional tests are legitimate when the relevant marginals are not fixed by design [17, 18]. Unconditional tests have higher power than conditional tests. This may lead to markedly lower sample sizes, even in moderately large samples. For example, to test for a difference in binomial proportions when $\pi_1 = 0.03$, $\pi_2 = 0.2$, and $\alpha = 0.05$, the sample size needed for 80 per cent power is 60 per group with Fisher's exact test and 52 per group with Fisher–Boschloo's unconditional tests, even though Martín Andrés *et al.* [40] report the conditional test to be acceptable in this case based on an average power consideration. In general, there should be no reason to condition on quantities not fixed by the design.

6. SUMMARY OF MUCH USED TESTS AND RECOMMENDED TESTS

6.1. Pearson's chi-squared test

This test uses the test statistic (2), with high values providing evidence against the null hypothesis. It is an asymptotic test that approximates the p-value by the upper tail probability from the chi-squared distributed with one degree and freedom. Yates' correction (11) for Pearson's chi-squared statistic was originally introduced to make it mimic Fisher's exact test. Pearson's chi-squared test with Yates' correction should no longer be used.

6.2. Fisher's exact test

An exact p-value is the exact probability of observing a table at least as extreme as the observed one, under the null hypothesis. However, in 2×2 tables this probability typically depends on one or more unknown parameters, such as the common success probability in comparing two binomials in the one margin fixed design. This obstacle vanishes if we condition on the marginals (observed row and column sums)$\mathbf{n}_+ = (n_{1+}, n_{2+}, n_{+1}, n_{+2})$, as if these were fixed by design like in Fisher's tea drinker example. The conditional probability (4) of a table given the marginals does not depend on any unknown parameters. The p-value from the resulting Fisher's exact conditional test equals the probability of the observed table plus the sum of the probabilities (4) equal to or smaller than the probability of the observed table. The resulting test preserves test size, but is, however, unnecessarily conservative with lower power than conditional mid-p tests and unconditional tests. We do not recommend the use of Fisher's exact test.

6.3. Fisher's exact mid-p test

In the mid-p version of Fisher's exact conditional test, only half the probability of the observed outcome is included in the mid-p-value. The resulting test is less conservative than Fisher's exact test, but the test size is not necessarily preserved. Its performance approximates that of an unconditional test.

6.4. Exact unconditional tests

In an exact unconditional test, the exact p-value is first computed as a function of the unknown parameter(s). Then, the p-value is taken as the maximum over all possible values, typically from 0 to 1. Recommended test statistics are Pearson's chi-squared statistic (the Suissa and Shuster test) or the conditional p-value from Fisher's exact test (Fisher–Boschloo's test). An unconditional test

DOI: 10.1002/sim

preserves test size and is usually more powerful than Fisher's exact test. In fact, the one-sided Fisher–Boschloo test is uniformly more powerful than the one-sided Fisher's exact test. If the Berger and Boos correction is used, the p-value is maximized over a confidence interval of values rather than the whole interval 0 to 1, and this generally improves the test further. We consider exact unconditional tests to be the gold standard for testing association in 2×2 tables.

7. EXAMPLES

The various methods may be illustrated by the following examples of computing the p-values for the data in Tables I and II. In these examples, unconditional tests are our prime recommendations. In addition, we compute the p-values for Fisher's exact test, traditionally used in tables with small counts. This is done partly to show how they differ from p-values for unconditional tests, partly to illustrate how Fisher–Bochloo's unconditional p-value is computed.

7.1. The epinephrine study (*Table I*)

In Fisher's exact test with the given marginal sums, the possible counts n_{11} are $0, 1, \ldots, 8$. The conditional sample space, which is a one-dimensional subspace of the two-dimensional unconditional sample space, consists of nine outcomes, as illustrated in Figure 3. The corresponding nine conditional probabilities are 0.0025, 0.0247, 0.1021, 0.2253,0.2910, 0.2253, 0.1021, 0.00247, 0.0025. In fact, since the sample sizes are equal, there are only five possible different values of the test statistic, and hence only five possible conditional probabilities and five possible p-values. The exact conditional p-value is the probability of the test statistic being equal to (white in Figure 3) or more extreme than (grey in Figure 3) its value for the observed table (white with asterisk in Figure 3)

$$p\text{-value} = 0.0025 + 0.0247 + 0.0247 + 0.0025 = 0.0544$$

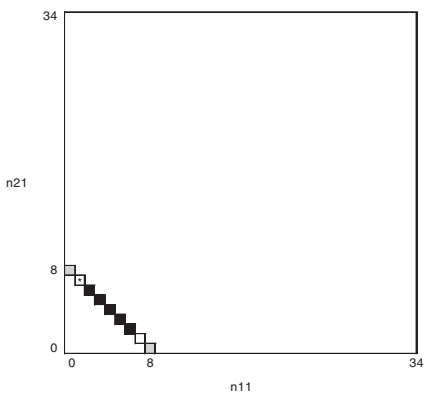

Figure 3. The conditional p-value for the epinephrine example (Table I) computed over the conditional sample space of the nine possible outcomes given the marginal sums. The observed outcome is marked *. Possible outcomes with less extreme, equal, and more extreme test statistic than the observed are marked black, white, and grey, respectively.

DOI: 10.1002/sim

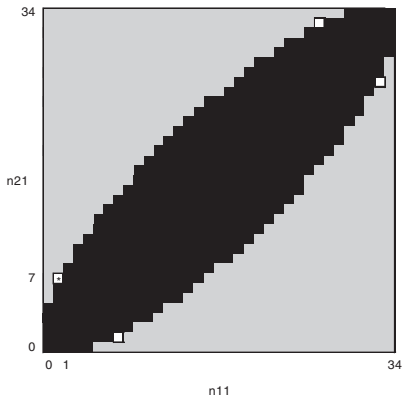

Figure 4. The unconditional p-value for the epinephrine example (Table I) in the one margin fixed design, computed over all $35 \times 35 = 1225$ possible outcomes in the unconditional sample space. The observed outcome is marked *. Possible outcomes with less extreme, equal, and more extreme test statistic than the observed are marked black, white, and grey, respectively.

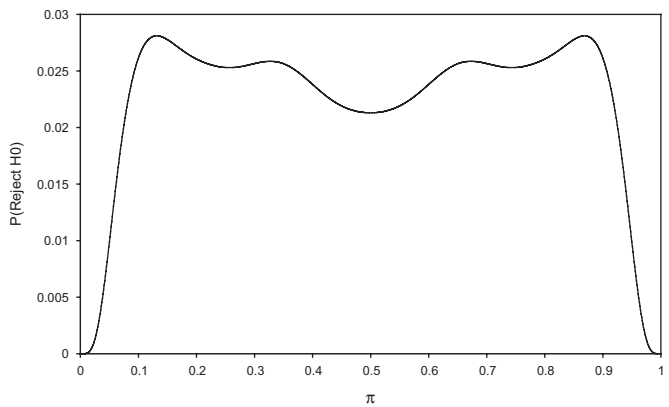

Figure 5. $P(T \geqslant t_{\text{obs}}; \pi)$ for the epinephrine example.

Fisher's exact conditional mid-p-value is

$$\text{mid-}p\text{-value} = 0.0544 - \tfrac{1}{2}(0.0247 + 0.0247) = 0.0297$$

In principle, the unconditional p-value is found by computing the test statistic for all the $(n_{1+}+1)(n_{2+}+1) = 35 \times 35 = 1225$ possible outcomes. The p-value for Fisher–Boschloo's unconditional test is the probability of the test statistic being equal to (white in Figure 4) or more extreme than (grey in Figure 4) than its value for the observed table (white with asterisk in Figure 4). This p-value depends on the value of the common success probability π under H_0, as illustrated in Figure 5. The unconditional p-value is taken as the maximum of this function, which is

$$\begin{aligned}\text{unconditional } p\text{-value} &= \max_{0 \leqslant \pi \leqslant 1} P(T \geqslant t_{\text{obs}}; \pi | H_0) = P(T \geqslant t_{\text{obs}}; 0.128 | H_0) \\ &= P(T \geqslant t_{\text{obs}}; 0.872 | H_0) = 0.0281\end{aligned}$$

DOI: 10.1002/sim

For the Berger and Boos correction, we first compute a confidence interval for the common success probability under H_0. Since we are performing exact tests, we should use an exact confidence interval. The Clopper–Pearson interval guarantees that the coverage probability is at least $1-\gamma$. It is given by (π_l, π_h) where

$$\pi_l = \left[1+\frac{N-n_{+1}+1}{n_{+1}F^{-1}_{2n_{+1},2(N-n_{+1}+1)}(1-\gamma/2)}\right]^{-1}$$
$$\pi_h = \left[1+\frac{N-n_{+1}}{(n_{+1}+1)F^{-1}_{2(n_{+1}+1),2(N-n_{+1})}(\gamma/2)}\right]^{-1} \tag{15}$$

and $F^{-1}_{a,b}(c)$ is the quantile of the Fisher distribution with a and b degrees of freedom, given by the inverse distribution function of c. With $n_{+1}=8$ successes out of $N=68$ trials, a $1-10^{-3}$ confidence interval becomes (0.0271, 0.2939), and the corresponding p-value is

$$p\text{-value} = \max_{0.0271\leqslant\pi\leqslant 0.2939} P(T\geqslant t_{\text{obs}};\pi|H_0)+10^{-3} = P(T\geqslant t_{\text{obs}};0.128|H_0)+10^{-3}$$
$$= 0.0281+10^{-3} = 0.0291$$

7.2. The genotype—exfoliation example (*Table II*)

Only the total sum N is fixed beforehand. p-values conditional on row and column sums are computed as before, and, using Fisher's test statistic, are obtained as p-value$=0.0629$ and mid-p-value$=0.0447$. The p-value for an unconditional test using the Fisher–Boschloo statistic is computed to be 0.0514 or 0.0486 without or with Berger–Boos correction ($\gamma=10^{-3}$) by the Berger software [23].

In the above examples, we see that the p-values from the recommended unconditional tests can be substantially lower than those from the traditional Fisher's exact test. Also, the conditional mid-p-value is approximately equal to the unconditional p-value, as expected.

8. POWER AND SAMPLE SIZE CALCULATIONS

Tests may be conditional or unconditional. However, power calculations must, by their nature, always be performed unconditionally on the marginal sums not fixed by design. Exact power or sample size calculations for the recommended tests cannot always be performed with existing software. StatXact provides power calculations for conditional as well as unconditional tests. The software [28] performs power calculations for all recommended tests of the one margin fixed design, save for unconditional tests with the Berger and Boos correction. Power calculations for unconditional tests with the Berger and Boos correction can be performed, slightly conservatively, as power calculations for unconditional tests without the correction. Computing time may be excessive for moderate to large sample sizes with unconditional tests. Most commercial softwares that provide sample size calculations provide asymptotic calculations for asymptotic tests. In cases

DOI: 10.1002/sim

for which exact calculations cannot be made, computing power or sample size asymptotically is preferable to no computation at all.

9. RECOMMENDATIONS

Exact tests have the important property of always preserving test size. Our general recommendation is not to condition on any marginals not fixed by design. In practice, this means that an exact unconditional test is ideal. Pearson's chi-squared (z_{pooled}) statistic or Fisher–Boschloo's statistic works well with an exact unconditional test. Further, such a test can be approximated by an exact conditional mid-p test or, in large samples, see, for example, the traditional asymptotic Pearson's chi-squared test. However, when an exact test is chosen, an unconditional test is clearly recommended. The traditional Fisher's exact test should practically never be used.

REFERENCES

1. Perondi MBM, Reis AG, Paiva EF, Nadkarni VM, Berg RA. A comparison of high-dose and standard-dose epinephrine in children with cardiac arrest. *New England Journal of Medicine* 2004; **350**(17):1722–1730.
2. Ritland JS, Utheim TP, Utheim OA, Espeseth T, Lydersen S, Semb SO, Rootwelt H, Elsås T. Effects of APOE and CHRNA4 genotypes on retinal nerve fibre layer thickness at the optic disc and on risk for developing exfoliation syndrome. *Acta Ophthalmologica Scandinavica* 2007; **85**(3):257–261.
3. Martín Andrés A, Silva Mato A. Choosing the optimal unconditioned test for comparing 2 independent proportions. *Computational Statistics and Data Analysis* 1994; **17**(5):555–574.
4. Hirji KF. *Exact Analysis of Discrete Data*. Chapman & Hall: Boca Raton, 2006.
5. Fisher RA. *The Design of Experiments*. London Oliver and Boyd: Edinburgh, 1966.
6. Haviland MG. Yates's correction for continuity and the analysis of 2×2 contingency-tables. *Statistics in Medicine* 1990; **9**(4):363–367.
7. Mehrotra DV, Chan ISF, Berger RL. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* 2003; **59**(2):441–450.
8. Santner TJ, Snell MK. Small-sample confidence-intervals for P1–P2 and P1/P2 in 2×2 contingency-tables. *Journal of the American Statistical Association* 1980; **75**(370):386–394.
9. Irwin JO. Tests of significance for differences between percentages based on small numbers. *Metron* 1935; **12**(2):83–94.
10. Lydersen S, Laake P. Power comparison of two-sided exact tests for association in 2×2 contingency tables using standard, mid p, and randomized test versions. *Statistics in Medicine* 2003; **22**(24):3859–3871.
11. Hirji KF, Tan SJ, Elashoff RM. A quasi-exact test for comparing 2 binomial proportions. *Statistics in Medicine* 1991; **10**(7):1137–1153.
12. Lancaster H. Significance tests in discrete-distributions. *Journal of the American Statistical Association* 1961; **56**(294):223–234.
13. Agresti A, Gottard A. Nonconservative exact small-sample inference for discrete data. *Computational Statistics and Data Analysis* 2007; **51**(12):6447–6458.
14. Berry G, Armitage P. Mid-P confidence intervals: a brief review. *The Statistician* 1995; **44**(4):417–423.
15. Barnard GA. On alleged gains in power from lower P-values. *Statistics in Medicine* 1989; **8**(12):1469–1477.
16. Hwang JTG, Yang MC. An optimality theory for mid p-values in 2×2 contingency tables. *Statistica Sinica* 2001; **11**(3):807–826.
17. Casella G, Berger RL. *Statistical Inference* (2nd edn). Duxbury: Pacific Grove, CA, 2002.
18. Lehmann EL, Romano JP. *Testing Statistical Hypotheses* (3rd edn). Springer: New York, 2008.
19. Suissa S, Shuster JJ. Exact unconditional sample sizes for the 2×2 binomial trial. *Journal of the Royal Statistical Society, Series A—Statistics in Society* 1985; **148**:317–327.
20. Barnard GA. Significance tests for 2×2 tables. *Biometrika* 1947; **34**(1–2):123–138.
21. Boschloo RD. Raised conditional level of significance for the 2×2-table when testing the equality of two probabilities. *Statistica Neerlandica* 1970; **24**(1):1–35.

DOI: 10.1002/sim

22. McDonald LL, Davis BM, Milliken GA. Non-randomized unconditional test for comparing 2 proportions in 2×2 contingency-tables. *Technometrics* 1977; **19**(2):145–157.
23. Berger RL. *Exact Unconditional Homogeneity/Independence Tests for* 2×2 *Tables*. 29 April 2005. Available at: http://www.stat.ncsu.edu/exact/.
24. Martín Andrés A. *Software* 2×2 *Tables*. 2000. Available at: http://www.ugr.es/∼bioest/software.htm.
25. Agresti A. *Categorical Data Analysis* (2nd edn). Wiley: Hoboken, NJ, 2002.
26. Berger RL, Boos DD. *P*-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89**(427):1012–1016.
27. StatXact. Cytel Software Inc., 2007.
28. Fagerland M, Lydersen S, Laake P. *Two-by-Two. Software for Testing Association in* 2×2 *Tables*. 1 October 2004. Available at: http://www.med.uio.no/imb/stat/two-by-two/manual.html.
29. Storer BE, Kim C. Exact properties of some exact test statistics for comparing 2 binomial proportions. *Journal of the American Statistical Association* 1990; **85**(409):146–155.
30. Altman DG. *Practical Statistics for Medical Research*. Chapman & Hall: London, 1991.
31. Davis LJ. Exact tests for 2×2 contingency-tables. *American Statistician* 1986; **40**(2):139–141.
32. Kang SH, Kim SJ. A comparison of the three conditional exact tests in two-way contingency tables using the unconditional exact power. *Biometrical Journal* 2004; **46**(3):320–330.
33. Lydersen S, Pradhan V, Senchaudhuri P, Laake P. Comparison of exact tests for association in unordered contingency tables using standard, mid-*p*, and randomized test versions. *Journal of Statistical Computation and Simulation* 2005; **75**(6):447–458.
34. Andres AM, Quevedo MJS, Mato AS. Fisher's mid-*P*-value arrangement in 2×2 comparative trials. *Computational Statistics and Data Analysis* 1998; **29**(1):107–115.
35. Cochran WG. Some methods for strengthening the common chi squared tests. *Biometrics* 1954; **10**(4):417–451.
36. Kang S-H, Ahn CW. Tests for homogeneity of two binomial proportions in extremely unbalanced 2×2 contingency tables. *Statistics in Medicine* 2008; **27**:2524–2535.
37. Duchateau L, Janssen P. Small vaccination experiments with binary outcome: the paradox of increasing power with decreasing sample size and/or increasing imbalance. *Biometrical Journal* 1999; **41**(5):583–600.
38. Yates F. Tests of significance for 2×2 contingency-tables. *Journal of the Royal Statistical Society*, *Series A—Statistics in Society* 1984; **147**:426–463.
39. Senn S. *Statistical Issues in Drug Development* (2nd edn). Wiley: Chichester, England, 2007.
40. Martín Andrés A, Silva Mato A, Tapia Garcia JM, Sanches Quevedo MJ. Comparing the asymptotic power of exact tests in 2×2 tables. *Computational Statistics and Data Analysis* 2004; **47**(4):745–756.

DOI: 10.1002/sim