STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:4328–4343
Published online 20 February 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2839

Choice of test for association in small sample unordered $r \times c$ tables

S. Lydersen[1,*,†], V. Pradhan[2], P. Senchaudhuri[2] and P. Laake[3]

[1]*Unit for Applied Clinical Research, Department of Cancer Research and Molecular Medicine, The Norwegian University of Science and Technology, N-7006 Trondheim, Norway*
[2]*Cytel Inc., Cambridge, MA, U.S.A.*
[3]*Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Norway*

SUMMARY

Pearson's chi-squared, the likelihood-ratio, and Fisher–Freeman–Halton's test statistics are often used to test the association of unordered $r \times c$ tables. Asymptotical, exact conditional, or exact conditional with *mid-p* adjustment methods are commonly used to compute the *p-value*. We have compared test power and significance level for these test statistics and *p-value* calculations in small sample $r \times c$ tables, mostly 3×2 and some with both r and c are greater than 2. After extensive simulations, in general we recommend using an exact conditional *mid-p* test with Pearson's chi-squared or Fisher–Freeman–Halton's statistic, which usually is the most powerful test yet preserve the approximate significance level. Moreover, we recommend that the asymptotic Pearson's chi-squared or other asymptotic tests not be used for small sample $r \times c$ tables. Copyright 2007 John Wiley & Sons, Ltd.

KEY WORDS: unordered contingency tables; exact tests; *mid-p*; small sample; Fisher–Freeman–Halton's test; Pearson's chi-squared test

1. INTRODUCTION

An unordered $r \times c$ table summarizes counts for two nominal categorical variables, at least one of which has more than two levels. The asymptotic Pearson's chi-squared test, the asymptotic likelihood ratio test and Fisher–Freeman–Halton's exact conditional test (shortened to Fisher's exact test in this paper) are much used for testing for association. The Pearson's chi-squared test is not recommended in tables with few counts, because the difference between the true significance level (the type I error rate) and the nominal significance levels (usually $\alpha = 0.05$) may be substantial. In contrast, an exact conditional test preserves test size, so the true significance level does not exceed

*Correspondence to: S. Lydersen, Unit for Applied Clinical Research, The Norwegian University for Science and Technology, N-7006 Trondheim, Norway.
†E-mail: stian.lydersen@ntnu.no

Received 28 April 2006
Accepted 30 December 2006

Table I. Change in patient management by biopsy indication, from Pascual *et al.* [1].

	Patient management after biopsy		
Biopsy indication	Change	No change	Sum
Delayed graft function	0	6	6
Acute allograft dysfunction	20	25	45
Chronic allograft dysfunction	2	0	2
Sum	22	31	53

Table II. *P-values* calculated from renal allograft biopsy study data.

	Pearson	Likelihood-ratio	Fisher
Asymptotic	0.0268	0.0064	0.0334
Exact	0.0146	0.0101	0.0146
Exact *mid-p*	0.0112	0.0067	0.0112

the nominal level, though the test may be conservative. However, Lancaster's *mid-p* adjustment may be applied to make the test less conservative.

As an illustrative example, consider the data in Table I, from a study of renal allograft biopsy [1] aimed to assess whether a change in patient management depends on biopsy indication. *P-values* obtained with different test statistics using asymptotic, exact and exact *mid-p* methods are listed in Table II.

We see that the *p-values* are quite different. The challenge at hand is to determine which tests have the highest power and still preserve tests size. That is, which tests have true significance level closer to the nominal level.

In this paper, we present a study of various test methods with that goal in mind. In conclusion, we recommend using exact *mid-p* tests, with Pearson's chi-squared or Fisher's test statistics, which in Table II gives $p = 0.0112$ for the test of association between the change in patient management and biopsy indication.

2. NOTATION, MODELS, AND HYPOTHESES

We use the following notation for the table elements and their sums:

$$\mathbf{n} = \begin{bmatrix} n_{11} & \cdots & n_{1c} \\ \vdots & & \vdots \\ n_{r1} & \cdots & n_{rc} \end{bmatrix} \tag{1}$$

$$n_{i+} = \sum_{j=1}^{c} n_{ij}, \quad n_{+j} = \sum_{i=1}^{r} n_{ij} \quad \text{and} \quad N = \sum_{i=1}^{r} n_{i+} = \sum_{j=1}^{c} n_{+j}$$

DOI: 10.1002/sim

The marginals are the collection of row sums and column sums

$$\mathbf{n}_+ = (n_{1+}, \ldots, n_{r+}, n_{+1}, \ldots, n_{+c}) \tag{2}$$

Various statistical sampling models may give rise to an unordered $r \times c$ table. The independent multinomial model and the full multinomial model are the most used.

In the independent multinomial model, the counts in the ith row are multinomially distributed with parameters $n_{i+}, \pi_{i1}, \ldots, \pi_{ic}$, where $\pi_{i1} + \cdots + \pi_{ic} = 1$. Under the null hypothesis H_0, the probability parameters $\pi_{i1}, \ldots, \pi_{ic}$ are the same in each row

$$H_0: \quad \pi_{1j} = \pi_{2j} = \cdots = \pi_{rj} \quad \text{for all } j = 1, 2, \ldots, c$$

while under the alternative hypothesis H_1 this is not the case

$$H_1: \quad \pi_{ij} \neq \pi_{kj} \quad \text{for at least one } i, j, k$$

In a full multinomial model, the table counts are multinomially distributed with parameters $N, \pi_{11}, \ldots, \pi_{rc}$, where $\pi_{11} + \cdots + \pi_{rc} = 1$. Let the marginals be $\pi_{i+} = \sum_{j=1}^{c} \pi_{ij}$ and $\pi_{+j} = \sum_{i=1}^{r} \pi_{ij}$. The null hypothesis H_0 then is

$$H_0: \quad \pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for all } i, j$$

in contrast to the alternative hypothesis H_1 that is

$$H_1: \quad \pi_{ij} \neq \pi_{i+}\pi_{+j} \quad \text{for at least one } i, j$$

In some cases, counts may be considered to be jointly Poisson distributed. Conditioning on the observed total N then yields a full multinomial model.

In these models, conditional on row and column marginals, the expected count in cell i, j under null hypothesis is $m_{ij} = n_{i+}n_{+j}/N$. Unconditionally, m_{ij} is the estimated expected number of counts under H_0.

3. TESTS OF INTEREST

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

$$T_{\mathrm{Pe}}(\mathbf{n}) = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{3}$$

and the likelihood ratio is defined as

$$T_{\mathrm{LR}}(\mathbf{n}) = -2 \log \frac{L_0}{L_1} = 2 \sum_{i,j} n_{ij} \log \left(\frac{n_{ij}}{m_{ij}} \right), 0 \quad \text{if } n_{ij} = 0 \tag{4}$$

where L_0 and L_1 are the respective maximum likelihoods of the observed table under H_0 and H_1. Both the independent multinomial model, the full multinomial model and the Poisson model yield

DOI: 10.1002/sim

the same result (4). Fisher's statistic may be defined in the following steps. First, let

$$\mathbf{x} = \begin{bmatrix} x_{11} & \cdots & x_{1c} \\ \vdots & & \vdots \\ x_{r1} & \cdots & x_{rc} \end{bmatrix} \tag{5}$$

be a possible $r \times c$ table with the same marginals $\mathbf{n}_+$ (2) as the observed table $\mathbf{n}$ (1). Conditional on the marginals, under H_0, $\mathbf{x}$ has a multiple hypergeometric distribution

$$P(\mathbf{x}|\mathbf{n}_+) = \frac{(\prod_{i=1}^{r} n_{i+}!)(\prod_{j=1}^{c} n_{+j}!)}{N! \prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}!} \tag{6}$$

for $\mathbf{x} \in S$, where S is the set of tables with non-negative integer counts and the given marginals. For an observed $\mathbf{n}$ with $\mathbf{n}_+$ marginals, Fisher's test statistic is defined as

$$T_{\mathrm{Fi}}(\mathbf{n}) = -2 \log(\gamma P(\mathbf{n}|\mathbf{n}_+)) \tag{7}$$

where

$$\gamma = \left[(2\pi)^{(r-1)(c-1)} N^{-(rc-1)} \prod_{i=1}^{r} (n_{i+})^{(c-1)} \prod_{j=1}^{c} (n_{+j})^{(r-1)} \right]^{1/2}$$

Fisher's test statistic is often defined as the equivalent $P(\mathbf{n}|\mathbf{n}_+)$, with small values providing evidence for rejecting H_0.

Equations (3), (4) and (7) are valid only when no marginals (2) are zero. Rows or columns with sum zero are deleted from the table before computing the test statistic. If this results in $r<2$ or $c<2$, the test statistic is defined as 0 and the *p-value* as 1, regardless of the observed cell counts.

The *p-value* or a corresponding test may be defined or computed in various ways. Five much used ways are the asymptotic *p-value*, the exact conditional *p-value*, the exact conditional *mid-p-value*, the randomized test and the exact unconditional *p-value*.

3.1. Asymptotic p-value

The test statistics of equations (3), (4) and (7) are asymptotically equivalent, and the asymptotic *p-value* is computed as

$$\text{asymp. } p\text{-}value = P(\chi^2_{(r-1)(c-1)} \geqslant T(\mathbf{n})) \tag{8}$$

where χ^2_v is chi-square distributed with v degrees of freedom. Pearson's asymptotic chi-square test is of this type.

Even under the null hypotheses, there are one or more unknown nuisance parameters. However, the conditional distribution (6) of a table given the marginals (2) has the remarkable property of not depending on unknown nuisance parameter(s).

3.2. Exact conditional p-value

The exact conditional *p-value* is given by

$$p\text{-}value = P(T(\mathbf{x}|\mathbf{n}_+) \geqslant T(\mathbf{n})) = \sum_{\mathbf{x} \in S: T(\mathbf{x}) \geqslant T(\mathbf{n})} P(\mathbf{x}|\mathbf{n}_+) \tag{9}$$

DOI: 10.1002/sim

For example, Fisher's exact test uses Fisher's statistic and the exact conditional *p-value*. It is a generalization of Fisher's exact test for 2×2 tables.

3.3. Exact conditional mid-p-value

An exact test preserves test size, that is, the true significance level does not exceed the nominal level. However, the significance level may be substantially lower than the nominal level, which makes the test conservative. This conservatism can be countered by using Lancaster's *mid-p* adjustment

$$\begin{aligned} \textit{mid-p-value} &= P(T(\mathbf{x}|\mathbf{n}_+) > T(\mathbf{n})) + \tfrac{1}{2} P(T(\mathbf{x}|\mathbf{n}_+) = T(\mathbf{n})) \\ &= P(T(\mathbf{x}|\mathbf{n}_+) \geqslant T(\mathbf{n})) - \tfrac{1}{2} P(T(\mathbf{x}|\mathbf{n}_+) = T(\mathbf{n})) \end{aligned} \tag{10}$$

3.4. Randomized test

A randomized test version entails computing the next possible lower *p-value* than the one actually observed

$$\begin{aligned} p_{\text{next}} &= P(T(\mathbf{x}|\mathbf{n}_+) > T(\mathbf{n})) \\ &= P(T(\mathbf{x}|\mathbf{n}_+) \geqslant T(\mathbf{n})) - P(T(\mathbf{x}|\mathbf{n}_+) \end{aligned} \tag{11}$$

Reject H_0 if the *p-value* $\leqslant \alpha$; reject H_0, and accept H_0 if $p_{\text{next}} \geqslant \alpha$. If $p_{\text{next}} < \alpha < p\text{-value}$, reject H_0 with the probability

$$P(\text{Reject } H_0|\mathbf{n}) = g\left(\frac{\alpha - p_{\text{next}}}{p\text{-value} - p_{\text{next}}}\right) \tag{12}$$

where

$$g(t) = \begin{cases} 0 & \text{if } t < 0 \\ t & \text{if } 0 \leqslant t \leqslant 1 \\ 1 & \text{if } t > 1 \end{cases}$$

In practice, randomized tests are seldom used, as 'throwing dice' arguably provides poor grounds for rejecting a hypotheses. But the significance level of a randomized test is exactly equal to α. Consequently, a randomized test is valuable in assessing the performance of various other tests and test versions.

The *mid-p-value* (or the decision in a randomized test) is calculated from the *p-value* and the point probability, which is the sum of probabilities of all possible tables with the same marginals and the same test statistic. In the example of Table I, this point probability is 0.00685 for all three test statistics. In this particular case it equals the probability of the observed table.

3.5. Exact unconditional p-value

The exact tests discussed above are conditional on the marginals (2). On the other hand, unconditional tests assume no fixed marginals, except those fixed by design. The exact unconditional *p-value* has the drawback that $P(T \geqslant t_{\text{obs}})$ depends on the unknown nuisance parameter(s) under

DOI: 10.1002/sim

H_0. To keep the test size below α, the *p-value* is taken [2, 3] as

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

where π is the probability parameters of the multinomial distribution(s).

Unless otherwise stated, we refer to exact conditional tests as exact tests.

4. BACKGROUND

Cochran's [4] criterion for choice of test in $r \times c$ tables is extensively cited (more than 1800 times by 2006). It maintains that Pearson's asymptotic chi-square test is acceptable when all $m_{ij} > 1$ and no more than 20 per cent of the cells have $m_{ij} < 5$. Otherwise, Fisher's exact test is recommended. Cochran's criterion for 2×2 tables has been discussed and refined by Andres [5] and references therein, but we have found no study of Cochran's criterion for $r \times c$ tables. However, this criterion for $r \times c$ tables has endured extensive usage through many years.

Pearson's chi-squared, the likelihood ratio, and the Fisher's test statistics are asymptotically equivalent. Lydersen *et al.* [6] have compared the test statistics for exact conditional, exact conditional *mid-p*, and randomized tests in small to moderate samples and have concluded that in most practical designs, Pearson's chi-squared statistic and Fisher's statistic perform slightly better than does the LR statistic.

The test size is regarded as preserved if the significance level obtained does not exceed the nominal level α for any value(s) of the nuisance parameter(s). Among the methods of computing a *p-value*, only the exact conditional and exact unconditional methods guarantee preservation of the test size. The conservatism of exact methods has engendered much of the ongoing controversy about them [7]. For a detailed description of exact analysis of discrete data under a more general framework, see the recent book by Hirji [8].

A *mid-p* test does not always preserve the test size. For $r \times c$ tables, Lydersen *et al.* [6] found that the *mid-p* type I error typically is approximately equal to or slightly less than the nominal level.

Mehta and Hilton [9] studied unconditional tests for unordered 3×2 tables with fixed row sums. Computing time may be excessive in $r \times c$ tables when $r > 2$ or $c > 2$. There is no efficient algorithm for this computation, so we will not pursue it further. Instead of maximizing over all possible values of the nuisance parameter(s) under H_0, an approximate unconditional test may be formed by inserting their maximum likelihood estimates under H_0. Clearly, such a test does not automatically preserve the test size. Maiste and Weir [10] studied various tests for Hardy–Weinberg equilibrium and found that for them, approximate unconditional tests performed less well than conditional tests.

5. COMPUTATIONS IN THIS STUDY

For independent, multinomially distributed rows, the parameters may be listed as

$$\boldsymbol{\theta} = \begin{bmatrix} n_{1+}, & \pi_{11} & \cdots & \pi_{1c} \\ \vdots & \vdots & & \vdots \\ n_{r+}, & \pi_{r1} & \cdots & \pi_{rc} \end{bmatrix} \tag{14}$$

DOI: 10.1002/sim

The probability of an $r \times c$ table $\mathbf{n}$ is given by a product of r multinomial densities

$$P(\mathbf{n};\boldsymbol{\theta}) = \prod_{i=1}^{r}\left[\left(n_{i+}\Big/\prod_{j=1}^{c} n_{ij}\right)\prod_{j=1}^{c}\pi_{ij}^{n_{ij}}\right] \tag{15}$$

In principle, the power or type I error probability may be computed from

$$\beta(\boldsymbol{\theta};\alpha) = P(\text{Reject } H_0|\boldsymbol{\theta}) = \sum_{\mathbf{n}} I\,(p\text{-}value(\mathbf{n})\leqslant\alpha)\,P(\mathbf{n};\boldsymbol{\theta}) \tag{16}$$

where $I(\,)$ is the indicator function, taking the value 1 if the condition is true and 0 if it is false, as was done by Lydersen and Laake [11] for 2×2 tables. However, there are far more possible $r \times c$ tables than possible 2×2 tables. Even with an efficient algorithm, we were restricted to estimating $\beta(\boldsymbol{\theta})$ by using Monte Carlo simulations, which we carried out in SAS [12] and StatXact PROCs [13]. For each $\boldsymbol{\theta}$, a total of $M = 100\,000$ $r \times c$ tables, $\mathbf{n}(1), \ldots, \mathbf{n}(M)$, were drawn from the probability distribution (15). Then, for each such table, the *p-values* of each definition

Table III. Summary of simulations performed.

Simulation no.	$r \times c$	Fixed sums	Probability parameters
1	3×2	$n_{1+} = n_{2+} = n_{3+} = 10$	$\pi_{11}, \pi_{21}, \pi_{31}$ all the triplets with elements from $(0.1, 0.2, \ldots, 0.9)$. In total, $9^3 = 729$ triplets
2	3×2	$n_{1+} = 9, n_{2+} = 10, n_{3+} = 11$	As simulation no. 1
3	3×2	Equal row sums. $n_{1+} = n_{2+} = n_{3+}$ from 5 to 10 by 1	$\pi_{11} = \pi_{21} = \pi_{31}$ from 0.05 to 0.5 by 0.05
4	3×2	Slightly unequal row sums. n_{1+} from 4 to 9 by 1	As simulation no. 3
5	3×2	Large differences between row sums. $(n_{1+}, n_{2+}, n_{3+}) = (2, 3, 10)$ and $(3, 4, 17)$ and $(4, 8, 9)$	As simulation no. 3
6	3×2	$(n_{1+}, n_{2+}, n_{3+}) = (2, 3, 10)$	$\pi_{11}, \pi_{21}, \pi_{31}$ be all the triplets with elements from $(0.1, 0.2, \ldots, 0.9)$ such that $\pi_{11} \leqslant 0.5$
7	3×2	$N = 30$*	Cell probability $\pi_{ij} = \pi_{i+}\pi_{+j}$, $\pi_{+1} = 0.05, 0.10$ and 0.15, $\pi_{1+} = \pi_{2+} = \pi_{3+} = \frac{1}{3}$
8	3×9	$(n_{1+}, n_{2+}, n_{3+}) = (7, 10, 10)$	Let $\pi_{\cdot 1} = \frac{17}{27}$, $\pi_{\cdot 2} = \pi_{\cdot 3} = \frac{2}{27}$, $\pi_{\cdot 4} = \pi_{\cdot 5} = \cdots = \pi_{\cdot 9} = \frac{1}{27}$
9	4×3	$n_{1+} = n_{2+} = n_{3+} = n_{4+} = 5$	Let $\pi_{\cdot 1}$, $\pi_{\cdot 2}$, $\pi_{\cdot 3}$ be all the triplets with elements from $(0.05, 0.1, \ldots, 0.95)$ such that their sum is 1 and $\pi_{\cdot 1} \leqslant \pi_{\cdot 2} \leqslant \pi_{\cdot 3}$ (32 triplets)

*In Simulation no. 7, only the total sum was fixed (full multinomial model). In the other simulations, the row sums were fixed (independent multinomial sampling model).

DOI: 10.1002/sim

were computed. Finally, the probability of rejecting H_0 is estimated as

$$\hat{\beta}(\boldsymbol{\theta};\alpha) = \frac{1}{M}\sum_{j=1}^{M} I\,(p\text{-}value(\mathbf{n}(j)) \leqslant \alpha) \tag{17}$$

For the full multinomial model, the same procedure is followed, with $M = 10\,000$ tables drawn from the probability distribution of the full multinomial model instead of the product of multinomial densities (15).

For the randomized test version, the indicator function in equation (17) is replaced by $P(\text{Reject } H_0|\mathbf{n})$.

In computing the significance level, which is $\beta_{\text{rnd}}(\boldsymbol{\theta}) = \alpha$ for the randomized test versions, we used the method of control variates (see, for instance, Ripley [14]) to adjust the estimated values for the asymptotic, exact and exact *mid-p* versions

$$\hat{\beta}_{\text{adj}}(\boldsymbol{\theta}) = \hat{\beta}(\boldsymbol{\theta}) - (\hat{\beta}_{\text{rnd}}(\boldsymbol{\theta}) - \alpha) \tag{18}$$

The approximate power of asymptotic Pearson and asymptotic LR tests can be computed as described by Agresti [15; pp. 243–244]. Under the alternative hypotheses, the test statistic is asymptotically distributed non-central chi-squared with $(r-1)(c-1)$ degrees of freedom. The non-centrality parameters of Pearson's chi-squared is $\lambda_{\text{Pe}} = N\sum_{i,j}(\pi_{ij} - \pi_{ij}^0)^2/\pi_{ij}^0$ and that of the likelihood ratio statistic is $\lambda_{\text{LR}} = 2N\sum_{i,j}\pi_{ij}\log(\pi_{ij}/\pi_{ij}^0)$, where π_{ij} is the true cell probability and π_{ij}^0 are the convergence limits of the ML estimates under the null hypothesis.

The simulations performed are listed in Table III. Simulation numbers 1, 2 and 6 are the computed powers, and the rest are the significance levels.

6. RESULTS

Figures 1–5 show significance levels for various 3×2 tables with fixed row sums (Simulations 1, 2 and 5) as function of the nuisance parameter π (probability of Column 1). The significance levels for a 3×2 table using full multinomial sampling (Simulation 7) and for a 3×9 table using independent multinomial sampling (Simulation 8) are shown in Table IV. Figure 6 shows the significance levels for Simulation 9 for a 4×3 table using Pearson's statistic as a function of $\frac{1}{3}\sum_{j=1}^{3}(\pi_{\cdot j} - 1/3)^2$. Similar results were obtained for the other two test statistics in Simulation 9, not included here.

In an ideal situation, the significance level obtained for a test is equal to the nominal level, as occurs in the randomized test version. However, this is not possible in practice, as the problem is discrete. Among the asymptotic, exact and exact *mid-p* tests, we find that the exact *mid-p* version usually has a significance level that is closer to this ideal, as shown in Figures 1–5 and in Table IV. In relatively few cases, the asymptotic test slightly outperforms the exact *mid-p* test (see Figure 3). In most cases, the *mid-p* test outperforms the asymptotic test. Moreover, the asymptotic tests are unreliable for small samples. For an example, see Table IV(b) with a 3×9 table for which the exact and exact *mid-p* tests perform well.

In general, we observe that the significance levels of the exact tests can be substantially lower than α, which demonstrates that exact tests are conservative. Typically, balanced designs are more conservative than unbalanced designs (Figure 1 *versus* Figure 4), as observed by

DOI: 10.1002/sim

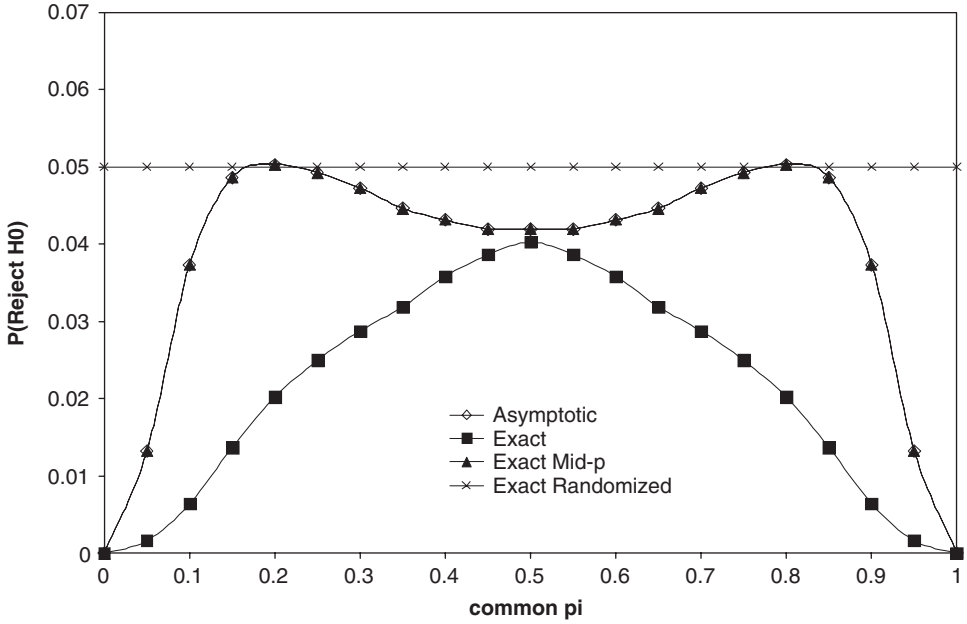

Figure 1. Significance level for a 3×2 table, Pearson's statistic, $n_{1+} = n_{2+} = n_{3+} = 10$, $\alpha = 0.05$. In this particular case, the asymptotic and exact *mid-p* versions are identical.

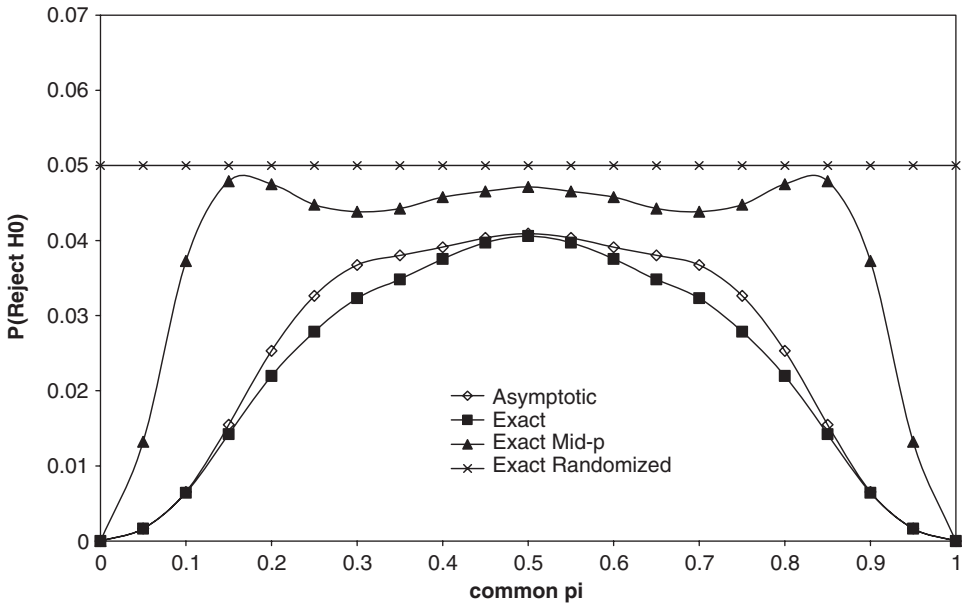

Figure 2. Significance level for a 3×2 table, Fisher's statistic, $n_{1+} = n_{2+} = n_{3+} = 10$, $\alpha = 0.05$.

DOI: 10.1002/sim

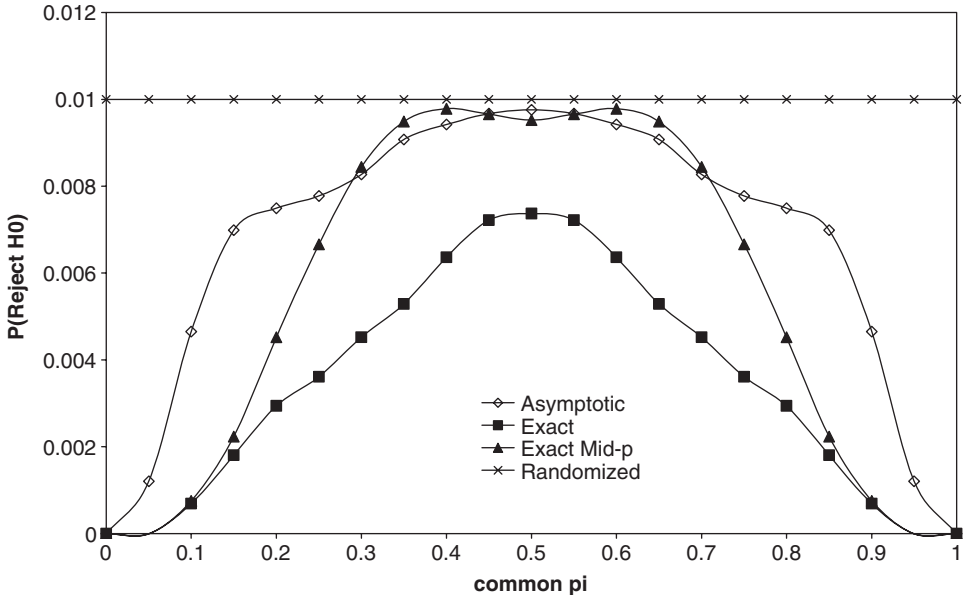

Figure 3. Significance level for a 3×2 table, Pearson's statistic, $n_{1+} = n_{2+} = n_{3+} = 10$, $\alpha = 0.01$.

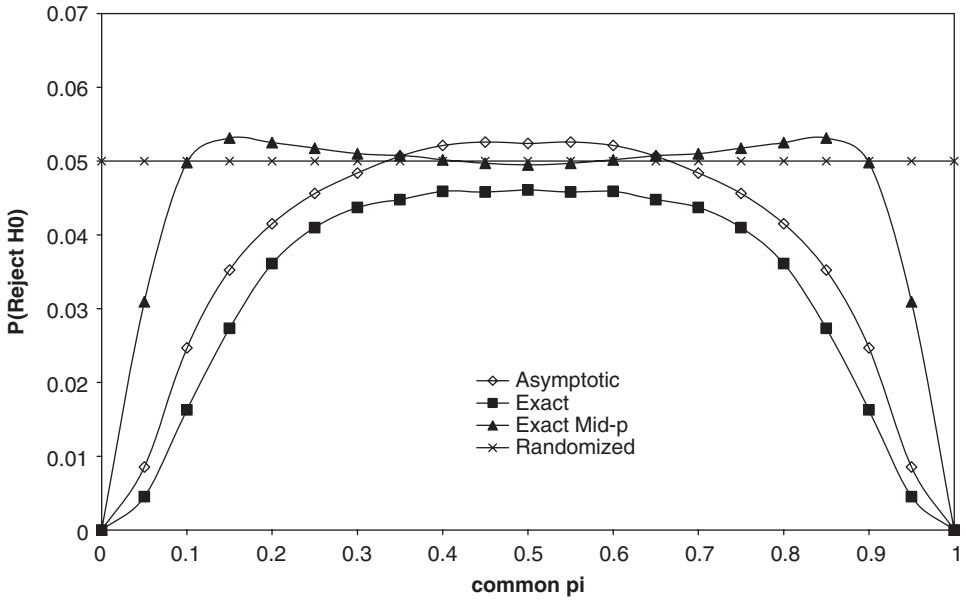

Figure 4. Significance level for a 3×2 table, Pearson's statistic, $n_{1+} = 9$, $n_{2+} = 10$, $n_{3+} = 11$, $\alpha = 0.05$.

DOI: 10.1002/sim

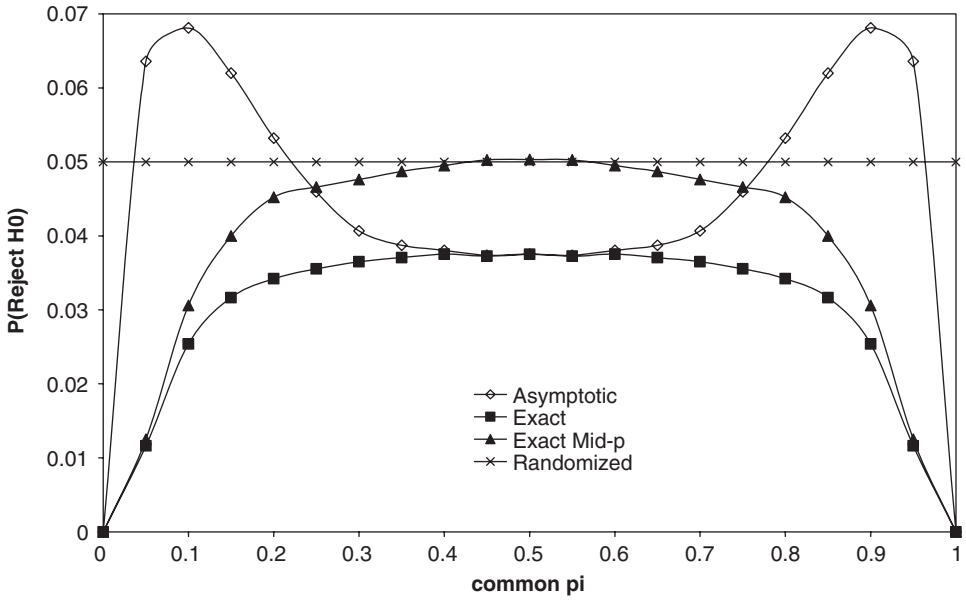

Figure 5. Significance level for a 3×2 table, Pearson's statistic, $n_{1+} = 3$, $n_{2+} = 4$, $n_{3+} = 17$, $\alpha = 0.05$.

Table IV. (a) Significance level for the 3×2 table, full multinomial model, $N = 30$, (Simulation 7) and (b) significance level for a small sample 3×9 table (Simulation 8).

		(a) Simulation 7			
		Column 1 probability, π_{+1}			
Statistic	Test version	0.05	0.10	0.15	(b) Simulation 8
Pearson	Asymptotic	0.014	0.029	0.039	0.012
	Exact	0.011	0.024	0.033	0.048
	Exact *mid-p*	0.021	0.038	0.046	0.050
	Exact rand	0.050	0.050	0.050	0.050
LR	Asymptotic	0.018	0.047	0.065	0.034
	Exact	0.010	0.025	0.033	0.047
	Exact *mid-p*	0.021	0.039	0.047	0.049
	Exact rand	0.050	0.050	0.050	0.050
Fisher	Asymptotic	0.003	0.010	0.019	0.003
	Exact	0.010	0.024	0.033	0.046
	Exact *mid-p*	0.022	0.040	0.047	0.049
	Exact rand	0.050	0.050	0.050	0.050

Duchateau and Janssen [16] for 2×2 tables. But the exact tests are less conservative for the 9×3 table (Table IV(b)). Exact *mid-p* tests are less conservative and have a level closer to the nominal level. Further, the significance level of the exact *mid-p* tests typically remains close to

DOI: 10.1002/sim

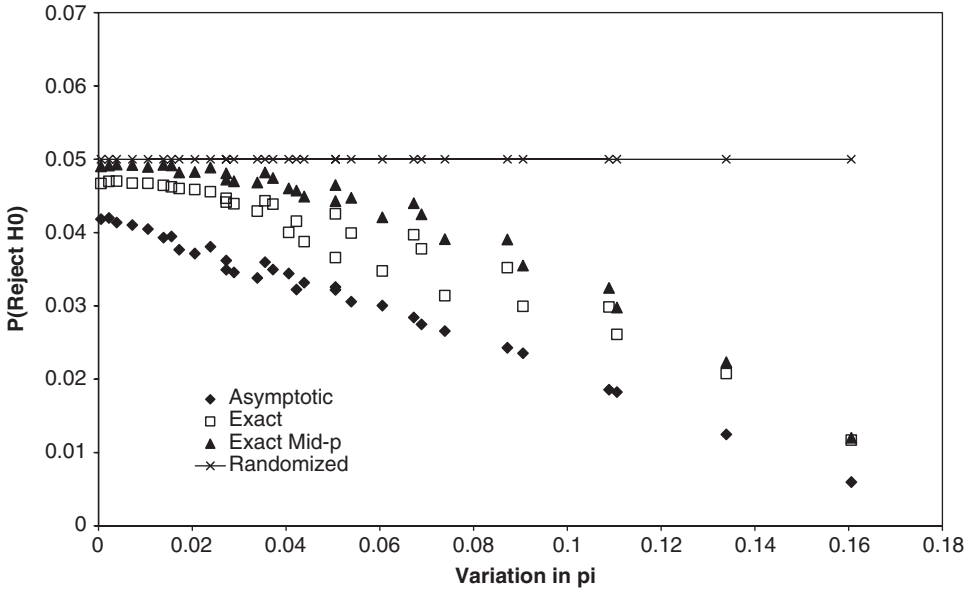

Figure 6. Significance level for a 4×3 table, Pearson's statistic, $n_{1+} = n_{2+} = n_{3+} = n_{4+} = 5$, $\alpha = 0.05$.

α when the nuisance parameter varies. We also observe that the nominal level of a *mid-p* test is seldom exceeded, and if so, only slightly. On the other hand, for small samples, the significance level of an asymptotic test may be substantially less than or greater than α. This is especially true for the unbalanced design shown in Figure 5, and in Table IV(b) for the small sample 9×3 table. Even in these cases, the significance level of the *mid-p* test remains close to α. In some cases, asymptotic tests are extremely conservative, even more so than the *mid-p* tests (Figures 2 and 4, and Table IV(b)). In particular, the asymptotic version of Fisher's test seems very conservative, as seen in Figure 2 and in Table IV(b).

The choice of test statistic for the exact or *mid-p* tests has little influence on the significance level (results are not shown here). But Pearson's and Fisher's statistics tend to perform slightly better than the LR statistic, as noted by Lydersen *et al.* [6]. However, in the asymptotic versions of the tests, Fisher's asymptotic test tends to be more conservative than Pearson's asymptotic test (Figure 1 *versus* Figure 2). Maiste and Weir [10] studied various tests for Hardy–Weinberg equilibrium, which is similar to $r \times c$ tables. For conditional tests, they found that Fisher's statistic performs better than Pearson's chi-squared or the LR statistic. Moreover, they found that Pearson's asymptotic test outperforms the asymptotic LR test. Parshall *et al.* [17] studied various test statistics for asymptotic tests in small sample $r \times c$ tables. They found that the Pearson statistic and the Read and Cressie statistic give conservative tests, while the LR statistic is quite liberal.

Table V shows the average powers of various tests for a 3×2 tables in which the total sum is 30. Typically, the power of the *mid-p* test tends to be close to the power of the randomized tests. The asymptotic tests have a power close to that of the *mid-p* tests (Pearson) or somewhat lower (Fisher), but as noted earlier, the asymptotic tests preserve test size to a lesser extent. The exact tests have a lower power, as expected.

DOI: 10.1002/sim

Table V. Average test power, averaged over all triplets $\pi_{11}, \pi_{21}, \pi_{31}$ with elements from $(0.1, 0.2, \ldots, 0.9)$, where at least two elements differ (a) and where the asymptotic power is at least 0.6 (b).

		Row sums = 10		Row sums 9,10,11	
Statistic	Test version	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Pearson (a)	Asymptotic	0.507	0.338	0.521	0.329
	Exact	0.475	0.303	0.505	0.332
	Exact *mid-p*	0.507	0.337	0.521	0.342
	Exact rand	0.515	0.339	0.520	0.343
	No. of triplets	720	720	720	720
Pearson (b)	Asymptotic	0.836	0.812	0.849	0.808
	Exact	0.817	0.772	0.835	0.808
	Exact *mid-p*	0.836	0.807	0.845	0.817
	Exact rand	0.842	0.804	0.845	0.817
	No. of triplets	276	144	278	134
LR (a)	Asymptotic	0.558	0.386	0.561	0.388
	Exact	0.483	0.306	0.506	0.328
	Exact *mid-p*	0.507	0.326	0.519	0.336
	Exact rand	0.514	0.337	0.516	0.338
	No. of triplets	720	720	720	720
LR (b)	Asymptotic	0.864	0.837	0.870	0.850
	Exact	0.817	0.775	0.832	0.795
	Exact *mid-p*	0.834	0.785	0.840	0.803
	Exact rand	0.839	0.797	0.839	0.803
	No. of triplets	276	144	278	134
Fisher (a)	Asymptotic	0.488	0.325	0.499	0.318
	Exact	0.481	0.306	0.506	0.331
	Exact *mid-p*	0.507	0.326	0.520	0.336
	Exact rand	0.515	0.337	519	0.341
	No. of triplets	720	720	720	720
Fisher (b)	Asymptotic	0.823	0.797	0.831	0.793
	Exact	0.820	0.775	0.834	0.802
	Exact *mid-p*	0.834	0.785	0.843	0.806
	Exact rand	0.840	0.797	0.842	0.811
	No. of triplets	276	144	278	134

Figure 7 shows the actual power of four versions of Pearson's test, as function of the asymptotically computed power for Simulation 1. We observe that the true power of the asymptotic and *mid-p* tests tends to be higher than the asymptotic power computed if the power is above 0.5 and the other way round if the power is less than 0.5. In practice, experiments are designed with a sample size to achieve a test power greater than 0.5, typically 0.8 or 0.9. This implies that the asymptotic powers are conservative and consequently useful for the powers in the range of practical interest. This result is promising, because exact power calculations for $r \times c$ tables require considerable computing time. That said, we have not performed a sufficiently large number of computations to verify the generality of this result.

DOI: 10.1002/sim

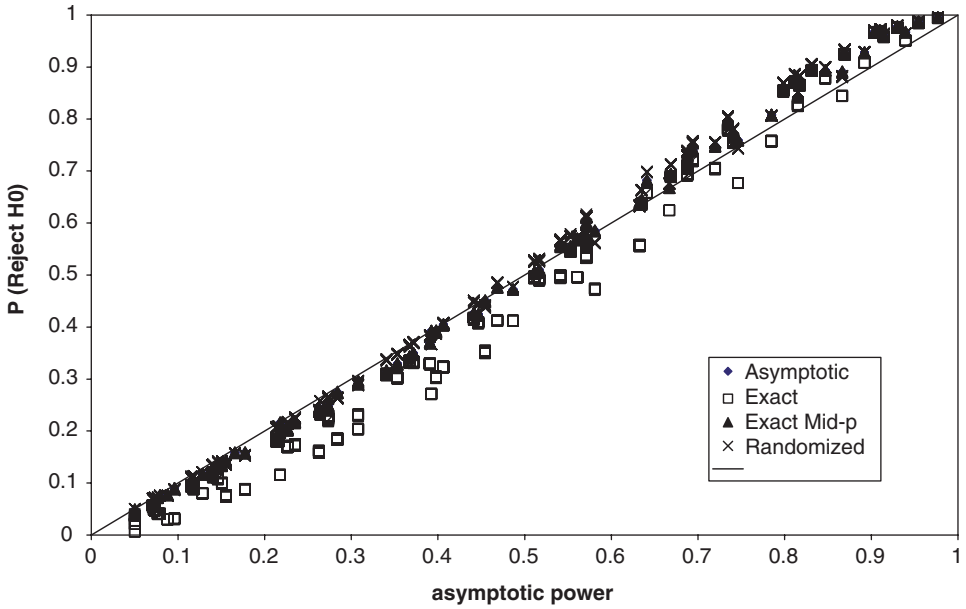

Figure 7. Power for a 3×2 table, Pearson's statistic, $n_{1+} = n_{2+} = n_{3+} = 10$, $\alpha = 0.05$, as function of the asymptotically computed power.

7. DISCUSSION AND CONCLUSIONS

In an earlier study of the tests for association in unordered $r \times c$ tables, Lydersen *et al.* [6] found that *mid-p* and randomized test versions have approximately the same power, higher than that of the standard test version. The *mid-p* type-I error seldom exceeds the nominal level.

The present study comprises a more detailed examination of performance in small sample 3×2 tables, for balanced designs, for slightly unbalanced designs and for highly unbalanced designs. Further, we study the performance of the tests as the nuisance parameter varies. Some larger tables (4×3 and 3×9) also have been included in this study. In addition to the exact, exact *mid-p* and randomized test versions, asymptotic test versions also have been included.

We have confirmed that exact tests are conservative for small samples. The *mid-p* tests are less conservative, and the test size is approximately preserved. As seen for the 3×9 table in Table IV(b), exact tests tend to be less conservative as r or c increase, since the increasing number of possible values for the test statistic tends to increase [5; p. 98]. Moreover, we found that asymptotic tests are less reliable in small samples, as at times, their significance level may fluctuate substantially. In some cases, the asymptotic tests are the most conservative of all. Hence, asymptotic tests should not be used for small sample $r \times c$ tables, not even in attempts to reduce the conservatism of exact tests.

This agrees with results for similar situations. Tang [18] compared the asymptotic, conditional, and conditional *mid-p* score test for ordered $r \times c$ tables and found that the *mid-p* score test was the most powerful, with the actual significance level close to the nominal level. Similar results were reported by Tang [19] for multinomial goodness-of-fit tests. Our recommendation for the

DOI: 10.1002/sim

mid-p test version in unordered $r \times c$ tables also agrees with the general recommendation of *mid-p* by Hirji [8; pp. 218–219].

The choice of test statistic in exact or *mid-p* tests does not substantially influence the results, as earlier pointed out by Lydersen *et al.* [6]. We see no reason to change the common practice of using Pearson's or Fisher's statistic, as also recommended by Lydersen *et al.* [6].

Our simulations were performed for 3×2 tables, a 4×3 table and a 3×9 table. The above conclusions all imply similar trends, so we extend the recommendations to other $r \times c$ tables, such as 3×3, 4×2 and so forth. With increasing numbers of rows or columns, the number of possible outcomes conditional on the margin sums increases and diminishes the difference between the exact and *mid-p* tests.

Two matters have not been addressed in our study: we have not scrutinized Cohran's criteria for the validity of Pearson's asymptotic test, to possibly suggest refinements. Further, we have not searched criteria (of similar type as Cochran's criteria) for when exact tests are not substantially more conservative than exact *mid-p* tests. These two matters would require a much more comprehensive study and were outside our scope.

The *mid-p-value* may be computed using the *p-value* and the point probability. Most general statistical software programs provide the exact *p-value*s for Fisher's test in $r \times c$ tables. SAS and StatXact apparently are the only commercial software packages which report this point probability, even though it is no more difficult to compute than the *p-value* itself. We challenge the compilers statistical software to provide the point probability or the *mid-p-value* as well.

ACKNOWLEDGEMENTS

We thank the reviewers for their valuable suggestions.

REFERENCES

1. Pascual M, Vallhonrat H, Cosimi AB, Tolkoff-Rubin N, Colvin RB, Delmonico FL, Ko DSC, Schoenfeld DA, Williams WW. The clinical usefulness of the renal allograft biopsy in the cyclosporine era—a prospective study. *Transplantation* 1999; **67**(5):737–741.
2. Casella G, Berger RL. *Statistical Inference*. Duxbury: Belmont, CA, 2002.
3. Lehmann EL, Romano JP. *Testing Statistical Hypotheses*. Springer: New York, 2005.
4. Cochran WG. Some methods for strengthening the common chi squared tests. *Biometrics* 1954; **10**(4):417–451.
5. Andres AM, Quevedo MJS, Garcia JMT, Silva-Mato A. On the validity condition of the chi-squared test in 2×2 tables. *Test* 2005; **14**(1):99–128.
6. 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.
7. Agresti A. Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine* 2001; **20**(17–18):2709–2722.
8. Hirji KF. *Exact Analysis of Discrete Data*. Chapman & Hall: Boca Raton, FL, 2006.
9. Mehta CR, Hilton JF. Exact power of conditional and unconditional tests—going beyond the 2×2 contingency table. *American Statistician* 1993; **47**(2):91–98.
10. Maiste PJ, Weir BS. Optimal testing strategies for large, sparse multinomial models. *Computational Statistics and Data Analysis* 2004; **46**(3):605–620.
11. 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.
12. SAS Institute Inc. *SAS Language Reference* [*8*]. SAS Institute Inc.: Cary, NC, 1999.
13. Cytel Inc. *StatXact 6 PROCs for SAS Users. A Software Package for Exact Nonparametric Inferences*. Cytel Inc.: Cambridge, MA 02139, 2004.

DOI: 10.1002/sim

14. Ripley BD. *Stochastic Simulation*. Wiley: New York, 1987.
15. Agresti A. *Categorical Data Analysis*. Wiley: New York, 2002.
16. 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.
17. Parshall CG, Kromrey JD, Dailey R. Comparative performance of three statistical tests of homogeneity for sparse $I \times J$ contingency tables. *Communications in Statistics-Simulation and Computation* 1999; **28**(1):275–289.
18. Tang ML. Exact and asymptotic power determination for dose–response studies with ordinal response. *Biometrical Journal* 2002; **44**(3):273–287.
19. Tang ML. Small-sample study of the use of *mid-P* power divergence goodness-of-fit tests. *Journal of Statistical Computation and Simulation* 1998; **62**(1–2):137–180.

DOI: 10.1002/sim