Journal of Statistical Computation and Simulation
Vol. 75, No. 6, June 2005, 447–458

Comparison of exact tests for association in unordered contingency tables using standard, mid-*p*, and randomized test versions

STIAN LYDERSEN*†, VIVEK PRADHAN‡, PRALAY SENCHAUDHURI‡ and PETTER LAAKE§

†NTNU, Unit for Applied Clinical Research, The Cancer Building, 5th Floor, St Olavs Hospital, Norwegian University of Science and Technology, N-7006 Trondheim, Norway
‡Cytel Software Corporation, Cambridge, Massachusetts, USA
§Section of Medical Statistics, University of Oslo, Norway

(Received 16 May 2003; in final form 20 May 2004)

Pearson's chi-square (Pe), likelihood ratio (LR), and Fisher (Fi)–Freeman–Halton test statistics are commonly used to test the association of an unordered $r \times c$ contingency table. Asymptotically, these test statistics follow a chi-square distribution. For small sample cases, the asymptotic chi-square approximations are unreliable. Therefore, the exact p-value is frequently computed conditional on the row- and column-sums. One drawback of the exact p-value is that it is conservative. Different adjustments have been suggested, such as Lancaster's mid-p version and randomized tests. In this paper, we have considered 3×2, 2×3, and 3×3 tables and compared the exact power and significance level of these test's standard, mid-p, and randomized versions. The mid-p and randomized test versions have approximately the same power and higher power than that of the standard test versions. The mid-p type-I error probability seldom exceeds the nominal level. For a given set of parameters, the power of Pe, LR, and Fi differs approximately the same way for standard, mid-p, and randomized test versions. Although there is no general ranking of these tests, in some situations, especially when averaged over the parameter space, Pe and Fi have the same power and slightly higher power than LR. When the sample sizes (*i.e.*, the row sums) are equal, the differences are small, otherwise the observed differences can be 10% or more. In some cases, perhaps characterized by poorly balanced designs, LR has the highest power.

Keywords: Exact tests; p-value; Mid-p-value; Power; $r \times c$ table

1. Introduction

The three usual test statistics for association in $r \times c$ contingency tables are Pearson's chi-square (Pe), likelihood ratio (LR), and Fisher's (Fi) exact tests [1, 2]. These tests are asymptotically equivalent. For small sample situations, exact tests have been developed and are available in commercial software. For a given set of observations, the p-values can be different, and the tests can lead to different conclusions. But which test is preferable in small

*Corresponding author. Tel.: +47-73867270; Fax: +47-73867289; Email: stian.lydersen@medisin.ntnu.no samples? The research to answer this question is scant. It has focused primarily on the question of which of the three asymptotic tests matches its exact counterpart best; see StatXact 6 User Manual [3] and references therein.

When appropriately scaled, these test statistics asymptotically follow a chi-square distribution. In small sample situations, the chi-square approximation is not reliable and, usually, exact p-values are used. Exact tests use the distribution of the test statistics conditional on the row and column sums of the observed data. For 2×2 tables with equal row (or column) sums, Pe, LR, and Fi tests always give the same p-values and conclusions [4]. Otherwise, they may give different p-values leading to different conclusions. Lydersen and Laake [5] have computed the exact power of these tests for different situations in 2×2 tables. They concluded that these three tests may have different power, but there is no general ranking among these tests. In many cases, Pe chi-square and Fi exact tests have almost equal power and higher power than LR. In a few cases, perhaps characterized by poorly balanced designs, LR performs best.

The present paper deals with 3×2, 2×3, and 3×3 tables. The rows and columns are unordered. In the case of ordered categories, one should use more powerful tests like Kruskal–Wallis test or a score test.

Consider two factors with r and c levels, respectively. For example, $r = 3$ treatments may be compared with cure or no cure as the $c = 2$ possible outcomes. In general, the observations form an $r \times c$ table:

	1	$\cdots$	c	Sum
1	n_{11}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	$\cdots$	n_{rc}	n_{r+}
Sum	n_{+1}	$\cdots$	n_{+c}	N

We consider independent, multinomially distributed rows. The counts in row number i are multinomially distributed with parameters $n_{i+}, \pi_{i1}, \dots, \pi_{ic}$, where $\pi_{i1} + \cdots + \pi_{ic} = 1$. The null hypothesis is that each row has the same set of probability parameters $\pi_{i1}, \dots, \pi_{ic}$. We want to test:

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

versus

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

Alternative sampling schemes are table counts that follow a multinomial distribution with parameters $(N, \pi_{11}, \dots, \pi_{rc} | \pi_{11} + \cdots + \pi_{rc} = 1)$ or a Poisson distribution with expectations $(\lambda_{11}, \dots, \lambda_{rc})$. In each of these sampling schemes, the distribution of the test statistics, conditional on the row sums, will be the same as for independent multinomially distributed rows. With these alternative schemes, the power of a test may be computed using the law of total probability and the power for independent multinomial sampling; see Lydersen and Laake [5].

The situation in the case of an $r \times c$ table is more complex than that of a 2×2 table and the results from 2×2 tables may not be generalized:

- Even if the row sums are equal, the three tests often give different results.
- The test power is a function of $r \times c$ parameters, and substantially more difficult to study than for a 2×2 table with only four parameters.
- The number of possible outcomes explodes. For example, if the row sums are 20, then the number of possible outcomes in a 2×2, 3×2, and 3×3 table are 441, 9261, and 53361, respectively. For this reason, we used Monte Carlo simulations in the present study, while Lydersen and Laake [5] could perform exact calculations for 2×2 tables.

In a standard version of a test, the null hypothesis is rejected if the p-value does not exceed the significance level α. This guarantees that the type-I error probability does not exceed α. However, it is often much lower than α, and this fact causes much of the ongoing controversy about exact test [6]. In a mid-p test version, only half the probability of the observed value is included in the mid-p-value. In a randomized test, the decision is based on the p-value supplemented by a randomization procedure to ensure that the probability of type-I error equals α.

The purpose of the present paper is to compare the standard, mid-p, and randomized versions of Pe, LR, and Fi exact tests for $r \times c$ tables with r or c values up to 3. Test power and type-I error probability are compared. The objectives are to compare the three test versions and to identify areas of the parameter space where one test statistic is superior to another.

2. Test statistics

Let

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

be an observed $r \times c$ table. Let

$$x = \begin{bmatrix} x_{11} & \cdots & x_{1c} \\ \vdots & & \vdots \\ x_{r1} & \cdots & x_{rc} \end{bmatrix} \tag{2}$$

be a possible $r \times c$ table with the given marginals $(n_{1+}, \ldots, n_{r+}, n_{+1}, \ldots, n_{+c})$. In our notation, n denotes the actually observed $r \times c$ table and x denotes a possible $r \times c$ table with the same marginals as n. Conditional on the marginals, x has a multiple hypergeometric distribution under H_0 [2, p. 97]

$$P(x) = \frac{\left(\prod_{i=1}^{r} n_{i+}!\right)\left(\prod_{j=1}^{c} n_{+j}!\right)}{N! \prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}!}. \tag{3}$$

The exact distribution of a test statistic $T(n)$ conditional on the marginals is given by

$$P(T(x) = T(n)) = \sum_{x \in S: T(x) = T(n)} P(x), \tag{4}$$

where S is the set of possible x with the non-negative integer counts and the given marginals. In the present paper, we define the test statistics $T(n)$ such that high values provide evidence against H_0. The p-values are thus defined as

$$P(T(x) \geq T(n)) = \sum_{x \in S: T(x) \geq T(n)} P(x). \tag{5}$$

Conditional on the row and column sums, the expected count in cell i, j under H_0 is $m_{ij} = n_{i+}n_{+j}/N$. Rows or columns with zero sums are deleted from the table before calculating the test statistics. If this results in $r < 2$ or $c < 2$, we do not compute any test statistic, and define the p-value to be 1 regardless of the other observed cell counts. Else, Pe chi-square test, LR test, and Fi exact test have the following test statistics:

$$T_{\mathrm{Pe}}(n) = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \tag{6}$$

$$T_{\mathrm{LR}}(n) = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{m_{ij}}\right), \quad \text{where a term is 0 if } n_{ij} = 0, \tag{7}$$

and

$$T_{\mathrm{Fi}}(n) = -2\log(\gamma P(n)), \tag{8}$$

where

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

is a normalizing constant to make $T_{\mathrm{Fi}}(n)$ asymptotically chi-square distributed, like Pe chi-square and LR statistics. Of course, any strictly decreasing transformation of $P(n)$, such as $T^{*}_{\mathrm{Fi}}(n) = -P(n)$, could be used as an equivalent test statistic for Fi exact test. The extension of Fi exact test to $r \times c$ tables was first proposed by Freeman and Halton [7] and is often named Fisher–Freeman–Halton's test.

The test versions are defined as follows. First, decide on a significance level α, for example, $\alpha = 0.05$. Then, compute the test statistic $T(n)$ for the observed data set. In a standard test version, reject H_0 if p-value $\leq \alpha$, where

$$p\text{-value} = P(T(x) \geq T(n)). \tag{9}$$

In a mid-p test version [8], reject H_0 if mid-p-value $\leq \alpha$, where

$$\begin{aligned} \text{mid-}p\text{-value} &= P(T(x) > T(n)) + \frac{1}{2}P(T(x) = T(n)) \\ &= P(T(x) \geq T(n)) - \frac{1}{2}P(T(x) = T(n)). \end{aligned} \tag{10}$$

In a randomized test version, compute the next possibly lower p-value than what was actually observed:

$$\begin{aligned} p\text{-next} &= P(T(x) > T(n)) \\ &= P(T(x) \geq T(n)) - P(T(x) = T(n)). \end{aligned} \tag{11}$$

Reject H_0 with probability

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

where

$$g(t) = \begin{cases} 0 & \text{if } t < 0 \\ t & \text{if } 0 \leq t \leq 1 \\ 1 & \text{if } t > 1. \end{cases} \tag{13}$$

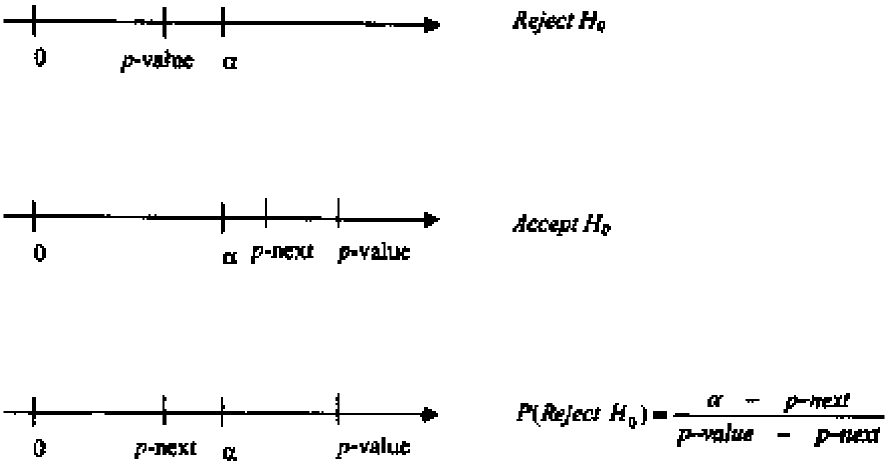
0
p-value
α
Reject H_0
0
α
p-next
p-value
Accept H_0
0
p-next
α
p-value
$P(\text{Reject } H_0) = \frac{\alpha - p\text{-next}}{p\text{-value} - p\text{-next}}$

Figure 1. The decision rule in randomized test version.

In other words, reject H_0 if p-value $\leq \alpha$, accept H_0 if p-next $\geq \alpha$, and reject H_0 with the above probability if p-next $< \alpha <$ p-value. This is illustrated in figure 1.

In order to compute the sums in equations (4) and (5), we could, in principle, compute $T(x)$ for all possible $r \times c$ tables x given the marginals. This may be quite computer time intensive unless efficient algorithms are used. We used the SAS procedure StatXact PROCs to compute $P(T(x) \geq T(n))$ and $P(T(x) = T(n))$ for given n, from which the p-value, mid-p-value and p-next are readily obtained from equations (9) to (11).

For a randomized test, the type-I error probability always equals the nominal significance level α. In practice, randomized tests are hardly ever used. One does not want to 'throw a dice' to decide on whether to reject a hypothesis. However, we believe that a randomized test version is a valuable tool in assessing the performance of different tests. If we only looked at the standard or mid-p version, we could risk unfair comparisons due to different obtained significance levels.

Some further remarks and references on the test versions are given in Lydersen and Laake [5].

3. Method for calculating power and type-I error probability

We consider independent, multinomially distributed rows where the parameters may be listed as

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

The probability of the $r \times c$ table n (1) is given by a product of r multinomial densities:

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

The probability $P(\text{Reject } H_0|\theta)$ is the power or type-I error probability when the parameter θ belongs to H_1 or H_0, respectively. In principle, these may be computed as

$$\beta(\theta) = P(\text{Reject } H_0|\theta) = \sum_{n} P(\text{Reject } H_0|n) \quad P(n;\theta), \tag{16}$$

where $P(\text{Reject } H_0|n)$ is 0 or 1 in standard or mid-p test version and may be between 0 and 1 in randomized test version. This was done for 2×2 tables by Lydersen and Laake [5]. However, the number of possible $r \times c$ tables is much larger than the number of possible 2×2 tables. Even with an efficient algorithm, we were confined to estimate $\beta(\theta)$ by Monte Carlo simulations. In the present paper, we did this by SAS and StatXact PROCs. For each given set of parameters θ, a total of M simulations were performed. Except where otherwise noted, $M = 100{,}000$. The SAS code using StatXact PROCs implementing the algorithm is available from the authors.

In simulation number m, draw an $r \times c$ table n_m from the probability distribution (15). The probabilities $P(T(x) \geq T(n_m))$ (5) and $P(T(x) = T(n_m))$ (4) are found by StatXact PROCs. Next, compute p-value, mid-p-value, and p-next. In the standard or mid-p test versions, set $P(\text{Reject } H_0|n_m) = 1$ (0) if p-value or mid-p-value is $\leq \alpha$ ($> \alpha$). In a randomized test, $P(\text{Reject } H_0|n_m)$ is given by equation (12). Finally, compute the estimate

$$\hat{\beta}(\theta) = \hat{P}(\text{Reject } H_0|\theta) = \frac{1}{M}\sum_{m=1}^{M} P(\text{Reject } H_0|n_m). \tag{17}$$

For the cases where we compute obtained significance level, we know that the true answer is $\beta_{\text{rnd}}(\theta) = \alpha$ for the randomized test versions. In these cases, we used the method of control variates [see ref. 9] to adjust the estimated values for the standard and the mid-p versions:

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

4. Results

We have performed simulations for 3×2, 2×3, and 3×3 tables. An overview of the performed simulation studies is given in table 1. In the table, we have also indicated the ranking of the powers of Pe, LR, and Fi tests. We have, to some extent, emphasized on 3×2 tables, since they occur more often, in practice, than 2×3 and 3×3 tables.

In all the simulations performed, we observed that the mid-p and the randomized test versions have approximately equal power. It follows from its definition that the standard version has lower power. In some cases, the power is notably lower, as shown for a 3×2 table in figure 2. In 3×3 tables, the three test versions have approximately the same power. This is understandable, since the number of outcomes is much larger in 3×3 tables, making the test statistic closer to a continuous stochastic variable.

The obtained significance level of the standard test is sometimes substantially less than α. The mid-p test version obtains a significance level closer to α, and seldom exceeds α. By definition, the obtained significance level of the randomized test equals α. This is illustrated in figure 3.

Figure 4 shows a comparison of the mid-p versions of Pe, LR, and Fi tests. In this particular case, Pe and Fi tests have approximately the same power and LR has somewhat lower power. Comparisons of the standard versions and comparisons of the randomized versions show approximately the same power differences (figures not included here). In fact, we have

Table 1. Summary of performed Monte Carlo simulations.

π	n_{i+}	Test power
3 × 2 Tables		

0.4 0.6
0.1 0.9
0.1 0.9
	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 100\}$	Pe > LR ≈ Fi

0.4 0.6
0.1 0.9
0.1 0.9
	$n_{1+} = 3k, n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 100\}$	LR > Pe ≈ Fi

0.4 0.6
0.1 0.9
0.1 0.9
	$n_{1+} = k, n_{2+} = n_{3+} = 3k, k \in \{5, 10, \ldots, 100\}$	Pe ≈ Fi > LR

0.1 0.9
0.4 0.6
0.4 0.6
	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 100\}$	LR ≈ Fi > Pe

0.1 0.9
0.4 0.6
0.4 0.6
	$n_{1+} = 3k, n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 100\}$	Pe ≈ Fi > LR

0.1 0.9
0.4 0.6
0.4 0.6
	$n_{1+} = k, n_{2+} = n_{3+} = 3k, k \in \{5, 10, \ldots, 100\}$	LR > Pe ≈ Fi

0.25 0.75
0.25 0.75
0.25 0.75
	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 100\}$	

$\pi_{11} \in \{0.01, 0.02, \ldots, 0.5\}$ $\pi_{31} = \pi_{21} = \pi_{11}$	$n_{1+} = n_{2+} = n_{3+} = 20$	

0.4 0.6
0.1 0.9
0.25 0.75
	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 100\}$	Pe ≈ Fi > LR

0.7 0.7
0.2 0.8
0.1 0.9
	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 25\}$	Pe > LR ≈ Fi

0.7 0.3 0.2 0.8 π_{31} $1 - \pi_{31}$ where $\pi_{31} \in \{0.05, 0.10, \ldots, 0.95\}$	$n_{1+} = n_{2+} = n_{3+} = 10$	Mainly Pe > LR ≈ Fi but for mid-*p* and π_{31} near 0.5: LR ≈ Fi > Pe
0.7 0.3 0.2 0.8 π_{31} $1 - \pi_{31}$ where $\pi_{31} \sim R(0, 0.3)$	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 100\}$ $M = 10{,}000$	Pe ≈ LR ≈ Fi
0.3 0.7 0.8 0.2 π_{31} $1 - \pi_{31}$ where $\pi_{31} \sim R(0, 0.3)$	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 30\}$ $M = 10{,}000$	Pe ≈ LR ≈ Fi
0.3 0.7 0.8 0.2 π_{31} $1 - \pi_{31}$ where $\pi_{31} \sim R(0, 0.7)$	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 75\}$ $M = 10{,}000$	Pe ≈ LR ≈ Fi

(*continued*)

Table 1. Continued.

π	n_{i+}	Test power
$(\pi_{11}, \pi_{21}, \pi_{31})$ consists of all triplets from $\{0.1, 0.2, \ldots, 0.9\}$	$n_{1+} = n_{2+} = n_{3+} = 25$	Small differences
$(\pi_{11}, \pi_{21}, \pi_{31})$ consists of all triplets from $\{0.1, 0.2, \ldots, 0.9\}$	$n_{1+} = 45, n_{2+} = 15, n_{3+} = 15$	Pe ≈ Fi. LR is larger or lower
2 × 3 Tables		
0.4 0.3 0.3 0.1 0.45 0.45	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 110\}$	Almost equal
0.25 0.375 0.375 0.25 0.375 0.375	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 110\}$	
0.1 0.2 0.7 0.5 0.3 0.2	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 25\}$	For $k \geq 10$: Pe ≈ Fi > LR For $k = 5$; LR > Pe ≈ Fi
0.1 0.2 0.7 π_{31} $\frac{1 - \pi_{31}}{2}$ $\frac{1 - \pi_{31}}{2}$ where $\pi_{31} \in \{0.05, 0.10, \ldots, 0.95\}$	$n_{1+} = n_{2+} = n_{3+} = 15$	For π_{31} near 0 or 1: LR > Pe ≈ Fi else: Pe ≈ Fi > LR
3 × 3 Tables		

0.4 0.3 0.3
0.1 0.45 0.45
0.1 0.45 0.45
	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 45\}$	Pe > LR ≈ Fi

0.1 0.45 0.45
0.4 0.3 0.3
0.4 0.3 0.3
	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 45\}$	LR ≈ Fi > Pe

$\pi_{11} \in \{0.01, 0.02, \ldots, 0.5\}$ $\pi_{31} = \pi_{21} = \pi_{11}$ $\pi_{i2} = \pi_{i3}$	$n_{1+} = n_{2+} = n_{3+} = 20$	

0.4 0.3 0.3
0.1 0.45 0.45
0.25 0.375 0.375
	$n_{1+} = n_{2+} = n_{3+} = k, k \in \{5, 10, \ldots, 45\}$	Pe > LR ≈ Fi

observed that for almost any fixed set of parameters the power differences among Pe, LR, and Fi are approximately the same whether we look at the standard, mid-p, or randomized test versions. The exceptions are the cases where the randomized versions have approximately equal power and the standard versions have different power. However, in these cases, the standard versions with lowest power also had lowest obtained significance level. As the randomized versions always meet the target significance level, we mainly look at randomized versions for comparisons of Pe, LR, and Fi.

We have observed that in general the ranking among Pe, LR, and Fi does not depend on the row sum, as long as the π_{ij} and the ratio between the row sums are kept fixed. The only exception we have seen from this observation was for a certain 2 × 3 table in which LR has highest power for row sum less than 7, when the power was less than 0.45 and the test has limited practical application.

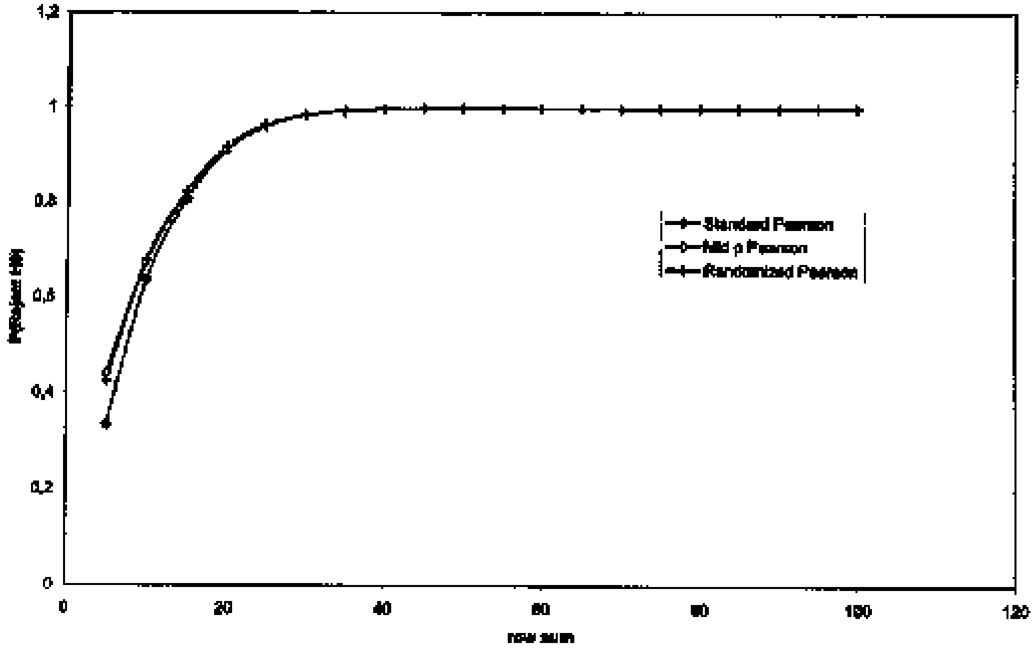

Figure 2. Power for standard test version, 3 × 2 table, equal row sums, $(\pi_{11}, \pi_{21}, \pi_{31}) = (0.4, 0.1, 0.1)$.

Since the ranking among Pe, LR, and Fi hardly depends on the sample size, it is of interest to study the power of these tests when the probability parameters π vary. This was done by estimating the power for all $9^3 = 729$ triplets $(\pi_{11}, \pi_{21}, \pi_{31})$ from $\{0.1, 0.2, \ldots, 0.9\}$ in a 3 × 2 table. We did this for two cases, when $n_{1+} = n_{2+} = n_{3+} = 25$ and when $n_{1+} = 45$ and $n_{2+} = n_{3+} = 15$. Out of these 729, there are 720 triplets where at least two π_{i1} differ. For the

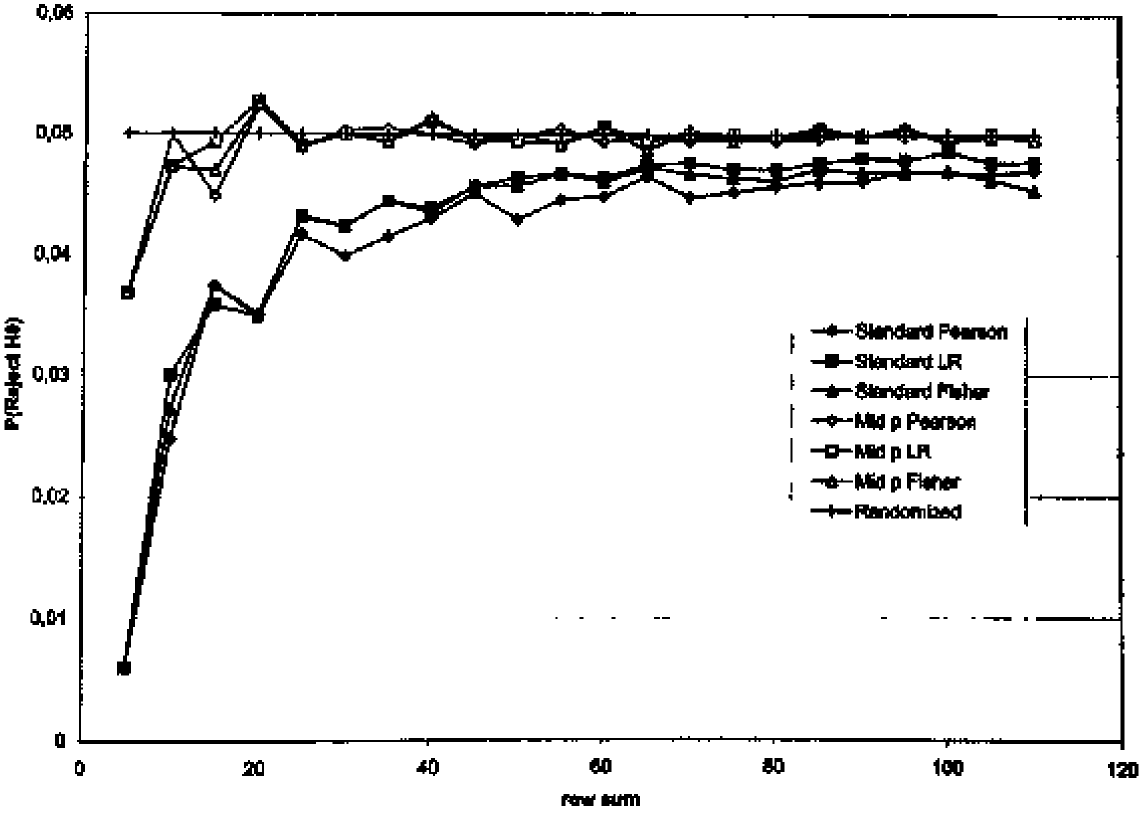

Figure 3. Obtained significance levels, 3 × 2 table, equal row sums, $\pi_{11} = \pi_{21} = \pi_{31} = 0.25$.

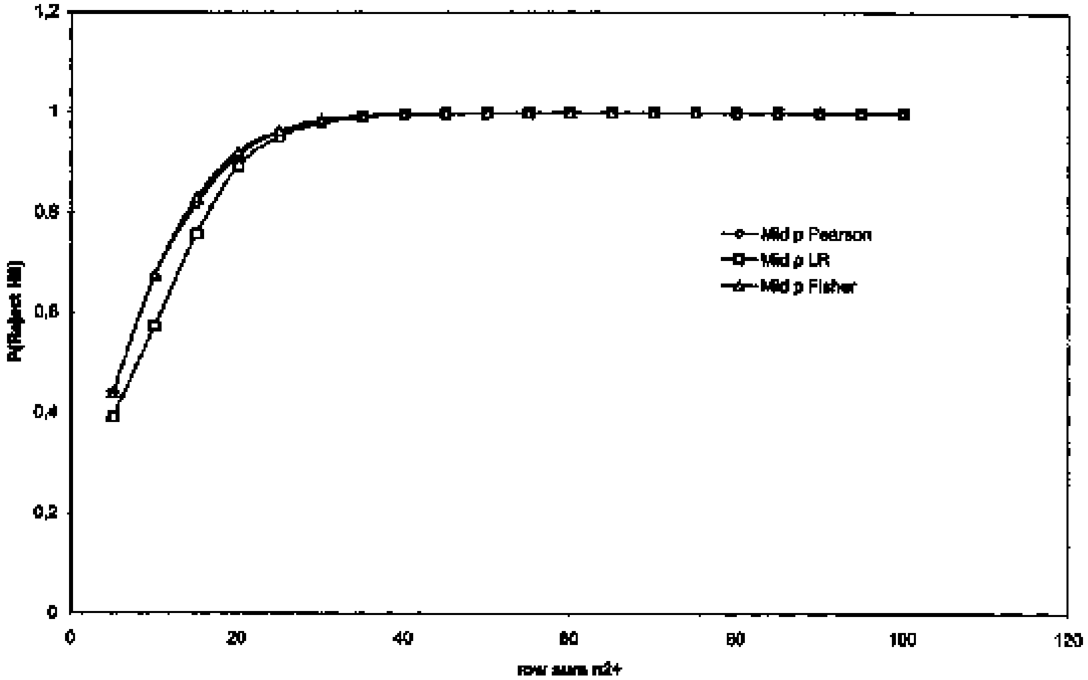

Figure 4. Power, 3 × 2 table. row sums $n_{1+} = 3n_{2+} = 3n_{3+}$, $(\pi_{11}, \pi_{21}, \pi_{31}) = (0.4, 0.1, 0.1)$.

case with equal row sums, the average power for these 720 triplets for Pe, LR, and Fi are shown in table 2. In this case, the ranges of power differences for the randomized version are found to be between Pe and LR (−0.0215, 0.0266), between Pe and Fi (−0.0208, 0.0224), and between LR and Fi (−0.0046, 0.0032). For the case row sums 45, 15, and 15, these ranges are between Pe and LR (−0.0711, 0.1019), between Pe and Fi (−0.0285, 0.0219), and between LR and Fi (−0.1159, 0.0842).

The power differences are larger for the case with unequal row sizes, and we explore these further. A point of interest would be to try to identify areas of the $(\pi_{11}, \pi_{21}, \pi_{31})$ space where a certain test is better than another. For the randomized test version (and the mid-p version), Pe and Fi have approximately the same power for all values of $(\pi_{11}, \pi_{21}, \pi_{31})$. In a closer investigation (not included here), we see that for π_{11} from 0.4 to 0.6, LR is most powerful and for π_{11} below 0.3 or above 0.7, Pe (and hence Fi) are most powerful. The areas where the test has a power closer to 1 than to 0 is of most practical interest. Figure 5 shows the power difference between Pe and LR as a function of the power of Pe. The figure shows that Pe, and hence also Fi, are most powerful mainly where the power is higher. Another point of interest is the area where there is an important difference between at least two π_{i1}. Table 3 shows the average test power for the 588 triplets where two of $(\pi_{11}, \pi_{21}, \pi_{31})$ differ by at least 0.3, when all row sums equal 25, and when $n_{1+} = 45$ and $n_{2+} = n_{3+} = 15$. The average

Table 2. Average power for a 3 × 2 table for the 720 triplets $(\pi_{11}, \pi_{21}, \pi_{31})$ from {0.1, 0.2, ..., 0.9} where at least two π_{i1} differ.

	$n_{1+} = n_{2+} = n_{3+} = 25$			$n_{1+} = 45, n_{2+} = 15, n_{3+} = 15$		
	Standard	Mid-p	Randomized	Standard	Mid-p	Randomized
Pe	0.749	0.760	0.762	0.706	0.712	0.721
LR	0.751	0.760	0.762	0.710	0.714	0.715
Fi	0.751	0.760	0.762	0.716	0.721	0.721

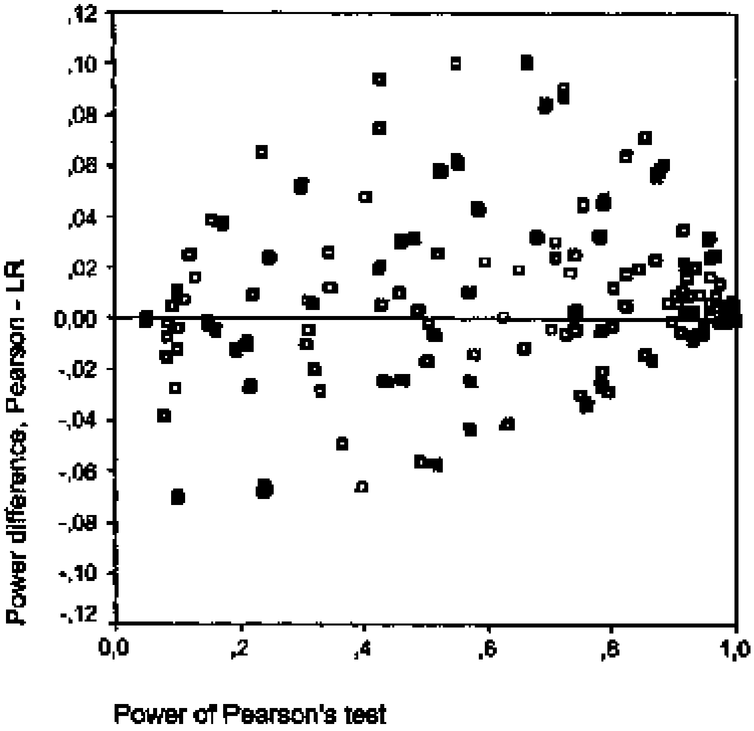

Figure 5. Difference in power for a 3 × 2 table, for $n_{1+} = 45$, $n_{2+} = 15$, $n_{3+} = 15$, for the randomized versions of Pe and LR, as a function of the power of Pearson's test.

Table 3. Average power for a 3 × 2 table for the 588 triplets $(\pi_{11}, \pi_{21}, \pi_{31})$ from $\{0.1, 0.2, \ldots, 0.9\}$ where two π_{i1} differ by at least 0.3.

	$n_{1+} = n_{2+} = n_{3+} = 25$			$n_{1+} = 45, n_{2+} = 15, n_{3+} = 15$		
	Standard	Mid-p	Randomized	Standard	Mid-p	Randomized
Pe	0.869	0.878	0.879	0.823	0.835	0.836
LR	0.871	0.878	0.879	0.825	0.829	0.829
Fi	0.871	0.878	0.879	0.831	0.835	0.836

obtained significance levels are given in table 4. For equal row sums, there are practically no differences between the average test powers. For $n_{1+} = 45$ and $n_{2+} = n_{3+} = 15$, Pe and Fi are slightly better than LR for the mid-p and randomized version. For the standard version, Pe has lowest power and lowest obtained significance level, whereas Fi has highest power and highest obtained significance level.

Table 4. Average significance level obtained for a 3 × 2 table for the nine triplets $(\pi_{11}, \pi_{21}, \pi_{31})$ from $\{0.1, 0.2, \ldots, 0.9\}$ with equal π_{i1}.

	$n_{1+} = n_{2+} = n_{3+} = 25$			$n_{1+} = 45, n_{2+} = 15, n_{3+} = 15$		
	Standard	Mid-p	Randomized	Standard	Mid-p	Randomized
Pe	0.038	0.046	0.050	0.039	0.049	0.050
LR	0.039	0.046	0.050	0.043	0.051	0.050
Fi	0.040	0.045	0.050	0.044	0.051	0.050

5. Conclusions and discussion

The mid-p and randomized versions have approximately the same power. The standard version has lower power, especially in smaller tables (small N or small $r \times c$). The obtained significance level is, of course, smallest for the standard test versions. The mid-p and randomized test versions have nearly the same obtained significance level. The obtained significance level of the mid-p version seldom exceeds that of the randomized version, and never much.

There are cases where the power differences among Pe, LR, and Fi are of practical importance, with power differences up to about 0.10. If there is a certain ranking of the tests for a given set of parameter values, this ranking tends to be the same if the row sums are multiplied by the same constant. Further, for any given set of parameter values, the ranking among Pe, LR, and Fi tends to be the same for the standard, mid-p, and randomized versions.

If the row sums are equal, the power does not differ much. This is in accordance with the 2×2 tables, where the power is equal for equal row sums [4]. For the cases with unequal row sums, the 'winner' depends on the parameter values. There is no uniformly best test among the three. For the cases we have studied, averaged over the H_1 values of $(\pi_{11}, \pi_{21}, \pi_{31})$, the power of Pe and Fi are approximately equal and is slightly higher than LR. For some particular parameter values, we have seen power differences of about 0.10 in favor of Pe and Fi. In a few cases, perhaps characterized by poorly balanced designs, LR performs slightly better than the other. From the trends we have seen, this conclusion ought to be valid also for $r \times c$ tables with r or c greater than 3.

We have only carried out simulations for $\alpha = 0.05$. For the 2×2 tables studied by Lydersen and Laake [5], the mutual ranking of the tests did not depend on the significance level. We found no reason to believe this fact to be different for $r \times c$ tables.

References

[1] Mehta, C. and Patel, N., 1998, Exact inference for categorical data. *Encyclopaedia of Biostatistics* (Wiley).

[2] Agresti, A., 2002, *Categorical Data Analysis* (2nd edn) (Wiley).

[3] *StatXact 6 User Manual*, 2003, *A Statistical Software for Exact Nonparametric Inference* (Cambridge, MA: Cytel Software Corporation).

[4] Davis, L.J., 1986, Exact tests for 2 × 2 contingency tables. *The American Statistician*, **40**(2), 139–141.

[5] Lydersen, S. and Laake, P., 2003, 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*, **22**, 3859–3871.

[6] Agresti, A., 2001, Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine*, **20**, 2709–2722.

[7] Freeman, G.H. and Halton, J.H., 1951, Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, **38**, 141–149.

[8] Lancaster, H.O., 1961, Significance tests in discrete distributions. *Journal of the American Statistical Association*, **56**, 223–234.

[9] Ripley, B.D., 1987, *Stochastic Simulation* (Wiley).