Statistical Methods in Medical Research 2003; **12**: 91–108

Exact unconditional tests for a 2 × 2 matched-pairs design

RL Berger Statistics Department, North Carolina State University, Raleigh, NC, USA and **K Sidik** Biometrics Research, Wyeth Research, Princeton, NJ, USA

The problem of comparing two proportions in a 2 × 2 matched-pairs design with binary responses is considered. We consider one-sided null and alternative hypotheses. The problem has two nuisance parameters. Using the monotonicity of the multinomial distribution, four exact unconditional tests based on p-values are proposed by reducing the dimension of the nuisance parameter space from two to one in computation. The size and power of the four exact tests and two other tests, the exact conditional binomial test and the asymptotic McNemar's test, are considered. It is shown that the tests based on the confidence interval p-value are more powerful than the tests based on the standard p-value. In addition, it is found that the exact conditional binomial test is conservative and not powerful for testing the hypothesis. Moreover, the asymptotic McNemar's test is shown to have incorrect size; that is, its size is larger than the nominal level of the test. Overall, the test based on McNemar's statistic and the confidence interval p-value is found to be the most powerful test with the correct size among the tests in this comparison.

1 Introduction

The problem of comparing two proportions in a 2 × 2 matched-pairs sample with binary responses has been studied for many years. The null hypothesis in past studies is often set as a 'zero difference', that is, formulated to be equal proportions. For this null hypothesis, a test involves only one nuisance parameter, the common value of the two proportions in the null distribution of the data. For testing equality, the most commonly used test is McNemar's test,[1] an asymptotic test. Cochran[2] derives the same asymptotic test by conditioning on the number of discordant pairs observed. The exact conditional version of this test is obtained if it is based on the null conditional binomial distribution of the data. Suissa and Shuster[3] proposed an exact unconditional test for the null hypothesis of equality based on the standard definition of a p-value.[4] In this paper, we consider the problem of testing one-sided null and alternative hypotheses comparing two matched-pairs proportions. A null hypothesis of this study is 'one larger than or equal to the other,' not the zero difference between proportions. In this formulation, the problem involves testing in the presence of two nuisance parameters because of the unspecified two proportions under the null hypothesis. If exact unconditional p-values are considered, computation seems to be much more intensive and difficult. However, we introduce a method of reducing the dimension of the nuisance parameters from two

Address for correspondence: Roger L Berger, Statistics Department, Box 8203, North Carolina State University, Raleigh, NC 27695-8203, USA. E-mail: rberger@nsf.gov

 10.1191/0962280203sm312ra

to one and propose four exact unconditional tests based on the standard[4] and the confidence interval p-values.[5]

Let Y_1 and Y_2 be two binary random variables with the joint distribution $P(Y_1 = i, Y_2 = j) = p_{ij}$ for $i, j = 0, 1$. Consider a random sample of size n matched-pairs data from this distribution. This type of data is frequently displayed in a 2×2 table as shown in Table 1.

Table 1 A 2×2 table

Y_1	Y_2		Total
	0	1	
0	x_{00}	x_{01}	
1	x_{10}	x_{11}	
Total			n

where x_{ij} for $i, j = 0, 1$ is the ijth observed cell count. The distribution of the vector of random cell counts $(X_{00}, X_{01}, X_{10}, X_{11})$ is multinomial with $\sum_{i,j} p_{ij} = 1$. The multinomial probability mass function (pmf) will be denoted by

$$m(x_{00}, x_{01}, x_{10}, x_{11}; n, p_{00}, p_{01}, p_{10}, p_{11}) = \frac{n!}{x_{00}!\,x_{01}!\,x_{10}!\,x_{11}!} p_{00}^{x_{00}} p_{01}^{x_{01}} p_{10}^{x_{10}} p_{11}^{x_{11}}$$

where $\sum_{i,j} x_{ij} = n$.

Consider the problem of testing

$$H_0 : p_1 \geq p_2 \text{ versus } H_1 : p_1 < p_2$$

where $p_1 = P(Y_1 = 0) = p_{00} + p_{01}$ and $p_2 = P(Y_2 = 0) = p_{00} + p_{10}$, or equivalently

$$H_0 : p_{01} \geq p_{10} \text{ versus } H_1 : p_{01} < p_{10}. \tag{1}$$

A challenge in the construction of tests for this hypothesis is the presence of two nuisance parameters under H_0, that is, the discordant cell probabilities p_{01} and p_{10}. The discordant pair parameter space is

$$\Pi = \{(p_{01}, p_{10}) : 0 \leq p_{01} \leq 1, 0 \leq p_{10} \leq 1, \text{and } p_{01} + p_{10} \leq 1\}. \tag{2}$$

Exact tests for (1) will be considered. By showing that certain probabilities are maximized on the boundary of H_0 and H_1, we will propose exact unconditional tests using two different concepts of a p-value, the standard and the confidence interval p-values.[4,5] The p-value, size, and power computations use the exact multinomial distributions of the data. Because of the discrete nature of the data, the exact tests do not have sizes exactly equal to the specified α. Rather they are level-α tests with exact sizes less than or equal to α. In this problem exact size α tests must be randomized tests, but randomized tests are seldom used in practice. We consider only nonrandomized tests.

Hsueh *et al.*[6] considered exact unconditional tests for (1). Their RMLE test corresponds to the test we define in (8). However, they provided no specific information for the $\delta = 0$ case (their notation), which corresponds to (1). They did not provide size and power comparisons for the tests we will discuss.

The data in these types of problems are usually summarized as (X_{01}, X_{10}, $n - X_{01} - X_{10}$). That is, the cell counts X_{00} and X_{11} are summed. The heuristic idea is that the individual values X_{00} and X_{11} do not give information about the relative sizes of p_{01} and p_{10}. Tests could be defined that depend on the individual values X_{00} and X_{11}, but we do not know of any such tests that have been proposed. So we too will summarize the data in this way and simply denote the data as (X_{01}, X_{10}), because the third count is a function of the first two. The trinomial pmf of (X_{01}, X_{10}) will be denoted by

$$m(x_{01}, x_{10}; n, p_{01}, p_{10}) = \frac{n!}{x_{01}!\,x_{10}!\,(n - x_{01} - x_{10})!} p_{01}^{x_{01}} p_{10}^{x_{10}} (1 - p_{01} - p_{10})^{n - x_{01} - x_{10}} \quad (3)$$

where $x_{01} \geq 0$, $x_{10} \geq 0$ and $x_{01} + x_{10} \leq n$. The parameter space is given in (2).

2 Monotonicity of a joint distribution

To maximize certain probabilities in the calculation of p-values, we will use a monotonicity property described in this section.

Definition 1 *In two dimensions, a set R is a Barnard convex set if $(x, y) \in R$, $x' \leq x$, and $y' \geq y$ imply $(x', y') \in R$.*

A Barnard convex set R contains all those points that lie above and to the left of a point (x, y) if $(x, y) \in R$. Note that a Barnard convex set is not necessarily convex in the sense of the usual mathematical definition of a convex set. The word convex is adopted because the shape property of the set is vaguely related to a convex set as described by Barnard.[7]

Sidik and Berger[8] proved the following theorem about the monotonicity of the joint distribution of random variables X and Y over a Barnard convex set.

Theorem 1 *Let $P_{\theta_1, \theta_2}(x, y)$ be a joint probability model for random variables X and Y indexed by parameters θ_1 and θ_2. Suppose that the marginal distribution of X depends only on θ_1 and the marginal distribution of Y depends only on θ_2. If for each y the family of conditional distributions of X given $Y = y$ is stochastically increasing in θ_1, and for each x the family of conditional distributions of Y given $X = x$ is stochastically* *increasing in* θ_2*, then, for any Barnard convex set R, when* $\theta_1' \leq \theta_1$ *and* $\theta_2' \geq \theta_2$

$$P_{\theta_1,\theta_2}((X, Y) \in R) \leq P_{\theta_1',\theta_2'}((X, Y) \in R). \tag{4}$$

This theorem presents sufficient conditions for achieving the distributional monotonicity (4) over a Barnard convex set. The monotonicity may also be seen as a type of multivariate stochastic order in parameters over a two-dimensional set. For discussion of multivariate stochastic orders, see Shaked and Shanthikumar.[9]

3 Four exact unconditional tests

In this section we define four exact unconditional tests of (1). The tests are defined by their p-values. The first two tests use the standard definition of a p-value, namely,

$$p(\mathbf{x}) = \sup_{\theta \in H_0} P_\theta(T(\mathbf{X}) \geq T(\mathbf{x})), \tag{5}$$

where $T(\mathbf{x})$ is the observed value of the test statistic. By Theorem 8.3.27 in Casella and Berger,[10] this defines a *valid p-value* in that the test that rejects H_0 if and only if $p(\mathbf{x}) \leq \alpha$ is a level-α test of H_0. The two tests we define use two different statistics: McNemar's Z statistic and the likelihood ratio (LR) statistic. The second two tests are defined by confidence interval p-values, namely,

$$p_C(\mathbf{x}) = \sup_{\theta \in C} P_\theta(T(\mathbf{X}) \geq T(\mathbf{x})) + \beta, \tag{6}$$

where $C = C(\mathbf{x})$ is a $100(1-\beta)\%$ confidence set for θ under H_0. By the Lemma in Berger and Boos,[5] this also defines a *valid p-value*. Again, the two tests use the Z and LR statistics. The suprema in these definitions are over two-dimensional sets for our matched pairs problem. However, we show in each case that the calculation can be reduced to a one-dimensional maximization, thereby greatly simplifying the numerical burden. The tests are exact unconditional tests because the exact trinomial distribution (3) of (X_{01}, X_{10}) is used in the calculation of the p-values, and, hence, the size of the tests is guaranteed to be less than or equal to the nominal value α.

Berger and Boos[5] noted that the standard p-value (5) may be very conservative if the supremum occurs at a point far from the true parameter value. To address this potential problem for matched pairs data, Hsueh *et al.*[6] mentioned the possibility of using the method proposed by Storer and Kim[11] and Kang and Chen,[12] namely, replace the supremum in (5) by the single probability calculation at the maximum likelihood estimate of the parameter under H_0. Unfortunately, as these authors noted, this method does not necessarily produce a valid p-value. The confidence interval p-value in (6) addresses the conservativeness of the standard p-value by considering only parameter values supported by the data, the values in the confidence set $C(\mathbf{x})$. However, it does this in such a way as to yield a valid p-value.

3.1 Standard p-value tests

3.1.1 The p-value using McNemar's test statistic

Consider the signed square root of McNemar's test statistic[1]

$$Z(x_{01}, x_{10}) = \frac{x_{10} - x_{01}}{\sqrt{x_{01} + x_{10}}}.$$

Because large values of Z give evidence against H_0, the p-value for testing (1) using $Z(x_{01}, x_{10})$ by the standard definition of p-value is

$$\begin{aligned} p_Z(x_{01}, x_{10}) &= \sup_{\{(p_{01}, p_{10}): p_{01} \geq p_{10}\}} P_{p_{01}, p_{10}}(Z(X_{01}, X_{10}) \geq Z(x_{01}, x_{10})) \\ &= \sup_{\{(p_{01}, p_{10}): p_{01} \geq p_{10}\}} \sum_{(u, v) \in R_Z(x_{01}, x_{10})} m(u, v; n, p_{01}, p_{10}) \end{aligned} \tag{7}$$

where $R_Z(x_{01}, x_{10}) = \{(u, v) : Z(u, v) \geq Z(x_{01}, x_{10})\}$. The supremum in this p-value is typically calculated numerically, and this may be difficult due to the maximization over a two-dimensional nuisance parameter space.

By finding the partial derivatives of $Z(x_{01}, x_{10})$ with respect to x_{01} and x_{10}, Sidik[13] showed that $R_Z(x_{01}, x_{10})$ is a Barnard convex set. (The partial derivative with respect to x_{10} is positive and the partial derivative with respect to x_{01} is negative.) In a 2×2 matched-pairs design, the conditional distribution of X_{01} given $X_{10} = x_{10}$ is binomial, that is,

$$b\left(x_{01}; n - x_{10}, \frac{p_{01}}{1 - p_{10}}\right) = \frac{(n - x_{10})!}{x_{01}!((n - x_{10}) - x_{01})!} \left(\frac{p_{01}}{1 - p_{10}}\right)^{x_{01}} \left(1 - \frac{p_{01}}{1 - p_{10}}\right)^{(n - x_{10}) - x_{01}}.$$

Similarly, the distribution of X_{10} given $X_{01} = x_{01}$ is binomial, $b(x_{10}; n - x_{01}, p_{10}/(1 - p_{01}))$. Using the result of Casella and Berger (Exercises 8.25 and 8.26),[10] it can be concluded that the family of conditional distributions of X_{01} given $X_{10} = x_{10}$ is stochastically increasing in p_{01} for any fixed p_{10}, and the family of conditional distributions of X_{10} given $X_{01} = x_{01}$ is stochastically increasing in p_{10} for any fixed p_{01}. Therefore, the joint distribution of X_{01} and X_{10} satisfies (4) of Theorem 1. For any p such that $p_{01} \geq p \geq p_{10}$ and $(p, p) \in \Pi$ (e.g., $p = p_{10}$),

$$P_{p_{01}, p_{10}}(Z(X_{01}, X_{10}) \geq Z(x_{01}, x_{10})) \leq P_{p,p}(Z(X_{01}, X_{10}) \geq Z(x_{01}, x_{10})),$$

and, hence,

$$\sup_{\{(p_{01}, p_{10}): p_{01} \geq p_{10}\}} P_{p_{01}, p_{10}}(Z(X_{01}, X_{10}) \geq Z(x_{01}, x_{10})) = \sup_{0 \leq p \leq \frac{1}{2}} P_{p,p}(Z(X_{01}, X_{10}) \geq Z(x_{01}, x_{10})).$$

Thus, the standard p-value for testing (1) using McNemar's test statistic (given in (7)) can be simplified to

$$\begin{aligned} p_Z(x_{01}, x_{10}) &= \sup_{0 \le p \le \frac{1}{2}} P_{p,p}(Z(X_{01}, X_{10}) \ge Z(x_{01}, x_{10})) \\ &= \sup_{0 \le p \le \frac{1}{2}} \sum_{(u,v) \in R_Z(x_{01}, x_{10})} m(u, v; n, p, p). \end{aligned} \tag{8}$$

The p-value $p_Z(x_{01}, x_{10})$ can be computed as the maximum probability over one nuisance parameter rather than two under H_0. The p-value in (8) indicates that the supremum over the null parameter space occurs on the boundary of H_0 and H_1, that is $p_{01} = p_{10} = p$.

3.1.2 *The p-value using the likelihood ratio test (LRT) statistic*

In a 2×2 matched-pairs sample, the multinomial log-likelihood function is

$$\log m(p_{00}, p_{01}, p_{10}, p_{11}; x_{00}, x_{01}, x_{10}, x_{11}) = \log\left(\frac{n!}{\prod_{i=0}^{1} \prod_{j=0}^{1} x_{ij}!}\right) + \sum_{i=0}^{1} \sum_{j=0}^{1} x_{ij} \log(p_{ij})$$

where $\sum_{ij} x_{ij} = n$. Following Robertson *et al.*,[14] the order restricted maximum likelihood estimators (MLEs) of the multinomial parameters under the constraint of H_0 are[13]

$$\text{MLEs} = \begin{cases} \hat{p}_{ij} = \dfrac{x_{ij}}{n}, i, j = 0, 1, & \text{if } x_{01} \ge x_{10} \\ \hat{p}_{01} = \hat{p}_{10} = \dfrac{x_{01} + x_{10}}{2n}; \hat{p}_{ii} = \dfrac{x_{ii}}{n}, i = 0, 1, & \text{if } x_{01} < x_{10}. \end{cases}$$

Similarly, the MLEs under the constraint of H_1 are[13]

$$\text{MLEs} = \begin{cases} \hat{p}_{01} = \hat{p}_{10} = \dfrac{x_{01} + x_{10}}{2n}; \hat{p}_{ii} = \dfrac{x_{ii}}{n}, i = 0, 1 & \text{if } x_{01} \ge x_{10} \\ \hat{p}_{ij} = \dfrac{x_{ij}}{n}, i, j = 0, 1, & \text{if } x_{01} < x_{10}. \end{cases}$$

Consider the following form of the LRT statistic:

$$\lambda(x_{00}, x_{01}, x_{10}, x_{11}) = \frac{\sup_{H_0} m(p_{00}, p_{01}, p_{10}, p_{11}; x_{00}, x_{01}, x_{10}, x_{11})}{\sup_{H_1} m(p_{00}, p_{01}, p_{10}, p_{11}; x_{00}, x_{01}, x_{10}, x_{11})}.$$

The log of the LRT statistic can be expressed as[13]

$$L(x_{01}, x_{10}) = \begin{cases} -x_{01}\log\left(\dfrac{x_{01}+x_{10}}{2x_{01}}\right) - x_{10}\log\left(\dfrac{x_{01}+x_{10}}{2x_{10}}\right), & \text{if } x_{01} \geq x_{10} \\ x_{01}\log\left(\dfrac{x_{01}+x_{10}}{2x_{01}}\right) + x_{10}\log\left(\dfrac{x_{01}+x_{10}}{2x_{10}}\right), & \text{if } x_{01} < x_{10}. \end{cases}$$

Note, although we started with the full data likelihood, this LRT statistic depends only on X_{01} and X_{10}. This gives some justification to the summarization of the data that is usually made.

To define an exact unconditional test using $L(x_{01}, x_{10})$, because small values of $L(x_{01}, x_{10})$ support H_1, consider the set

$$R_L(x_{01}, x_{10}) = \{(u, v): L(u, v) \leq L(x_{01}, x_{10})\}.$$

The set contains all the data points whose test statistic is at most as large as the observed test statistic. By finding the partial derivatives with respect to x_{01} and x_{10}, it can be shown that $R_L(x_{01}, x_{10})$ is a Barnard convex set.[13] Therefore, arguing as we did for the Z statistic, an exact unconditional p-value for testing (1) using the LRT statistic is

$$\begin{aligned} p_L(x_{01}, x_{10}) &= \sup_{0 \leq p \leq \frac{1}{2}} P_{p,p}(L(X_{01}, X_{10}) \leq L(x_{01}, x_{10})) \\ &= \sup_{0 \leq p \leq \frac{1}{2}} \sum_{(u,v) \in R_L(x_{01}, x_{10})} m(u, v; n, p, p). \end{aligned} \tag{9}$$

3.2 Confidence interval p-value tests

In this section, we define two more exact unconditional tests, now using confidence interval p-values as defined in (6) and again using the statistics Z and L.

3.2.1 *The confidence interval p-value using McNemar's test statistic*

Suppose $C_\beta(x_{01}, x_{10})$ is a $100(1-\beta)\%$ confidence set for the parameters (p_{01}, p_{10}) calculated from the observed data under H_0. Then, the confidence set p-value using McNemar's test statistic is

$$\begin{aligned} p(x_{01}, x_{10}) &= \sup_{(p_{01}, p_{10}) \in C_\beta(x_{01}, x_{10})} P_{p_{01}, p_{10}}(Z(X_{01}, X_{10}) \geq Z(x_{01}, x_{10})) + \beta \\ &= \left(\sup_{(p_{01}, p_{10}) \in C_\beta(x_{01}, x_{10})} \sum_{(u,v) \in R_Z(x_{01}, x_{10})} m(u, v; n, p_{01}, p_{10}) \right) + \beta. \end{aligned}$$

$R_Z(x_{01}, x_{10})$ is the same as in $p_Z(x_{01}, x_{10})$. Although $C_\beta(x_{01}, x_{10})$ is a subset of H_0, it may still be computationally difficult to compute this supremum because $C_\beta(x_{01}, x_{10})$ is a two-dimensional set. To overcome this we introduce a specific confidence set $C_\beta(x_{01}, x_{10})$ and show that the maximization over this set can be reduced to a one-dimensional maximization over a confidence *interval* for p constructed under the assumption that $p_{01} = p_{10} = p$. To do this we use the following lemma, the proof of which follows from the results of Sidik and Berger.[8]

Lemma 1 *In a* 2×2 *matched-pairs model, let*

$$I_\beta(X_{01}, X_{10}) = \{p: l(X_{01}, X_{10}) \le p \le u(X_{01}, X_{10})\}$$

be a $100(1-\beta)\%$ *equal-tailed confidence interval for p assuming that* $p_{01} = p_{10} = p$. *Suppose* $l(X_{01}, X_{10})$ *and* $u(X_{01}, X_{10})$ *are nondecreasing functions of* X_{01} *and* X_{10}. *Then, for* $p_{01} \ge p_{10}$

$$C_\beta(X_{01}, X_{10}) = \{(p_{01}, p_{10}): l(X_{01}, X_{10}) \le p_{01}, u(X_{01}, X_{10}) \ge p_{10}\}$$

is a $100(1-\beta)\%$ *confidence set for* (p_{01}, p_{10}).

Let $I_\beta(x_{01}, x_{10})$ be a $100(1-\beta)\%$ Clopper and Pearson[15] interval for p calculated from the data based on the variable $X_{01} + X_{10}$, where $X_{01} + X_{10} \sim \text{binomial}(n, 2p)$ if $p_{01} = p_{10} = p$. The lower and upper limits of $I_\beta(x_{01}, x_{10})$ are nondecreasing functions of x_{01} and x_{10} because this interval is based on the method in Theorem 9.2.14 of Casella and Berger,[10] and the binomial distribution function is decreasing in p. The interval is easily computed from the formula

$$\frac{t}{2[t + (n-t+1)F_{2(n-t+1),2t,\beta/2}]} \le p \le \frac{(t+1)F_{2(t+1),2(n-t),\beta/2}}{2[n - t + (t+1)F_{2(t+1),2(n-t),\beta/2}]}, \tag{10}$$

where $t = x_{01} + x_{10}$ and $F_{v,\eta,\beta/2}$ is the upper $100\beta/2$ percentile of an F distribution with v and η degrees of freedom. By Lemma 1 under H_0,

$$C_\beta(x_{01}, x_{10}) = \{(p_{01}, p_{10}): l(x_{01}, x_{10}) \le p_{01} \text{ and } u(x_{01}, x_{10}) \ge p_{10}\} \tag{11}$$

is a $100(1-\beta)\%$ confidence set for (p_{01}, p_{10}) (see Figure 1). Note that $l(x_{01}, x_{10})$ and $u(x_{01}, x_{10})$ are respectively the lower and upper limits of the interval (10). In addition, for any $(p_{01}, p_{10}) \in C_\beta(x_{01}, x_{10})$ one can find a p such that

$$p_{01} \ge p \ge p_{10} \text{ and } p \in I_\beta(x_{01}, x_{10}) \tag{12}$$

(e.g., $p = \max\{p_{10}, l(x_{01}, x_{10})\}$ satisfies these conditions). Note that $(p, p) \in C_\beta(x_{01}, x_{10})$ if $p \in I_\beta(x_{01}, x_{10})$. Because the trinomial distribution of (X_{01}, X_{10}) satisfies (4) of Theorem 1, for $(p_{01}, p_{10}) \in C_\beta(x_{01}, x_{10})$ and p satisfying (12),

$$P_{p_{01},p_{10}}((X_{01}, X_{10}) \in R_Z(x_{01}, x_{10})) \le P_{p,p}((X_{01}, X_{10}) \in R_Z(x_{01}, x_{10})),$$

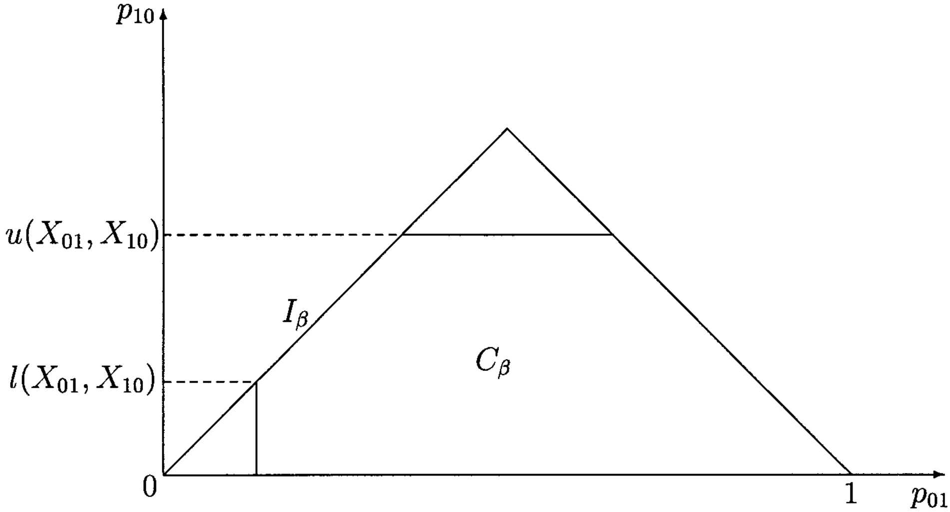

Figure 1 The confidence interval I_β and confidence set C_β under H_0

and, hence,

$$\sup_{(p_{01},p_{10})\in C_\beta} P_{p_{01},p_{10}}((X_{01}, X_{10}) \in R_Z(x_{01}, x_{10})) = \sup_{p\in I_\beta} P_{p,p}((X_{01}, X_{10}) \in R_Z(x_{01}, x_{10})).$$

Therefore, an exact unconditional confidence interval p-value for testing (1) using McNemar's test statistic is

$$\begin{aligned} p_{Z_C}(x_{01}, x_{10}) &= \sup_{p\in I_\beta(x_{01},x_{10})} P_{p,p}(Z(X_{01}, X_{10}) \geq Z(x_{01}, x_{10})) + \beta \\ &= \left(\sup_{p\in I_\beta(x_{01},x_{10})} \sum_{(u,v)\in R_Z(x_{01},x_{10})} m(u, v; n, p, p) \right) + \beta, \end{aligned} \tag{13}$$

where $I_\beta(x_{01}, x_{10})$ is the Clopper and Pearson interval for p calculated from (10). The supremum in (13) is taking over a one-dimensional interval $I_\beta(x_{01}, x_{10})$ rather than over a two-dimensional set $C_\beta(x_{01}, x_{10})$.

3.2.2 The confidence interval p-value using log LRT statistic

Similarly, we can derive a confidence interval p-value using $L(x_{01}, x_{10})$ that requires maximization only on the boundary of H_0 and H_1. An exact unconditional confidence interval p-value for testing (1) using the log LRT statistic is

$$p_{L_C}(x_{01}, x_{10}) = \sup_{p \in I_\beta(x_{01}, x_{10})} P_{p,p}(L(X_{01}, X_{10}) \leq L(x_{01}, x_{10})) + \beta$$
$$= \left(\sup_{p \in I_\beta(x_{01}, x_{10})} \sum_{(u,v) \in R_L(x_{01}, x_{10})} m(u, v; n, p, p) \right) + \beta$$

where $I_\beta(x_{01}, x_{10})$ is the Clopper and Pearson interval calculated from (10) and $R_L(x_{01}, x_{10})$ is the same as in $p_L(x_{01}, x_{10})$.

The confidence interval p-values p_{Z_C} and p_{L_C} depend on the error probability β which must be chosen *a priori*. For testing (1), Sidik[13] investigated several values of β with different values of α and concluded that $\beta = 0.0005$ yields good level-$\alpha = 0.05$ tests. The value $\beta = 0.0005$ is used in the remainder of this paper. Note, from definitions such as (13), a confidence interval p-value is always greater than or equal to β. Thus, in this context, one never uses a value of β such as 0.05. Rather one uses a very small value of β such as the $\beta = 0.0005$ used in this paper.

All four of the exact unconditional p-values defined in this section are valid p-values. The tests that reject H_0 if and only if the respective p-values are less than or equal to a specified α are level-α tests for (1). The exact unconditional p-values must be calculated numerically. However, all four p-values have been expressed in terms of a one-dimensional maximization of a polynomial in p, which is not difficult.

In this paper, we denote these four tests for testing (1) discussed in this section by Z, Z_C, L, and L_C corresponding to the p-values p_Z, p_{Z_C}, p_L, and p_{L_C}, respectively. In the following sections, we will also consider two more common tests, McNemar's asymptotic test, which we denote by M, and the exact conditional binomial test, which is defined by conditioning on the total number of discordant cell counts and which we denote by CB. For testing (1) the p-value of CB is

$$p_{CB}(x_{01}, x_{10}) = P(X_{01} \leq x_{01} | X_{01} + X_{10} = t).$$

The distribution of X_{01} given $X_{01} + X_{10} = t$ used to calculate p_{CB} is binomial$(t, 1/2)$, the conditional distribution assuming $p_{01} = p_{10}$. The p-value of M is

$$p_M(x_{01}, x_{10}) = P(Z^* \geq Z(x_{01}, x_{10})),$$

where Z^* has a standard normal distribution. CB is both conditionally and unconditionally a level-α test of (1), but M is only approximately a level-α test.

4 Example

Biomedical examples of matched-pairs data often involve comparison of standard and innovative procedures or drugs in which testing of (1) might be of interest. For illustration of the exact unconditional tests of this paper, let us assume the matched-

Table 2 Matched-pairs data from study

Standard	New		Total
	Yes	No	
Yes	9	4	
No	11	6	
Total			30

pairs data shown in Table 2 are from a study for comparing the standard and the innovative or new diagnostic procedures for a particular disease.

We wish to test the one-sided hypothesis (1), where $H_1 : p_{01} < p_{10}$ says that the probability of diagnosing the disease is smaller for the standard procedure than for the new procedure.

Although the previous data are fictional, such examples of comparing two diagnostic procedures for a disease are common. Hsueh *et al.*[6] described actual data comparing two diagnostic procedures for liver lesions in which patients with suspected liver lesions received both the noninvasive postcontrast magnetic resonance imaging (MRI) and the invasive computer tomography arterio portography (CTAP) diagnostic procedures. In the liver lesions study, one could be interested in testing whether the noninvasive MRI as a new procedure is more accurate than invasive CTAP in detection of liver lesions.

We applied the four exact unconditional tests Z, Z_C, L, and L_C along with the two common tests M and CB to the previous data. The results of the six tests are summarized in Table 3. In Z_C and L_C, $\beta = 0.0005$ is used. Among the four exact unconditional tests, the three tests Z, Z_C, L_C yield p-values less than 0.05; they reject H_0 in favor of H_1 at the $\alpha = 0.05$ test level. The asymptotic test M has the smallest p-value, and the exact conditional binomial test CB has the largest p-value among the six tests. We will see in the following section that M is a liberal test for testing (1), and CB is a very conservative test. Thus, it is common that they have the smallest and largest p-values, respectively. Because of the small sample size, M may not be an appropriate test of (1) at level-α.

In Table 3, $p_Z(4, 11) = 0.0494$, where this p-value is the sup over the complete null parameter space [0, 0.5] in computing p_Z, and the sup occurs at the maximum point $p = 0.5$. On the other hand, $p_{Z_C}(4, 11) = 0.0382$, In calculating p_{Z_C} the sup occurs at the maximum point $p = 0.2479$. The smaller $p_{Z_C}(4, 11)$ results from searching for a

Table 3 Results of the tests for the example data

Method	Test statistic	p-value	$I_\beta(x_{01}, x_{10})$	Max point
CB		0.0592		
M	1.8074	0.0354		
Z	1.8074	0.0494		0.0500
Z_C	1.8074	0.0382	(0.1007, 0.3993)	0.2479
L	−1.6985	0.0501		0.0664
L_C	−1.6985	0.0442	(0.1007, 0.3993)	0.1007

$\beta = 0.0005$ in Z_C and L_C.

maximum over only the interval $I_\beta(4, 11) = [0.1007, 0.3993]$ rather than over the entire interval $[0, 0.5]$. This example clearly demonstrates the important property of a confidence interval p-value, that it achieves better power in general by concentrating on the 'plausible' values of p with the restriction of the sup over the observed data driving interval $I_\beta(x_{01}, x_{10})$ rather than the complete null interval $[0, 0.5]$ in calculating the exact unconditional p-values. In L and L_C, we note the similar behavior, that is, $p_L(4, 11) = 0.0501$ with the maximum point $p = 0.0664$, and $p_{L_C}(4, 11) = 0.0442$ with the maximum point $p = 0.1007$. The p-values of the two tests lead to different decisions in which L_C rejects H_0 but L does not do so at the 5% level.

In the previous example each of the confidence interval p-values was notably smaller than the corresponding exact unconditional standard p-value. This is not always the case, and one might ask whether the exact unconditional standard p-value can be notably smaller than the corresponding exact unconditional confidence interval p-value. The answer is 'no'. An exact unconditional confidence interval p-value will never exceed the corresponding exact unconditional standard p-value by more than the error probability β.[16] (Recall from section 3, when calculating confidence interval p-values β is always chosen to be a very small value, like the $\beta = 0.0005$ used in this paper.) For instance, $p_{Z_C}(x_{01}, x_{10}) \leq p_Z(x_{01}, x_{10}) + \beta$ for any (x_{01}, x_{10}). For a numerical illustration, consider the different set of data; $x_{00} = 5$, $x_{01} = 2$, $x_{10} = 10$, $x_{11} = 3$. With $\beta = 0.0005$, $p_Z(2, 10) = 0.01188$ and $p_{Z_C}(2, 10) = 0.01238$. Here, the sup over $[0, 0.5]$, which occurs at the maximum point $p = 0.3141$, in computing p_Z is equal to the sup over $I_\beta = [0.1093, 0.4536]$, which occurs at the same maximum point $p = 0.3141$, in computing p_{Z_C}. So we have $p_{Z_C}(2, 10) = p_Z(2, 10) + \beta = 0.01188 + 0.0005 = 0.01238$, and p_{Z_C} exceeds p_Z by the maximum possible amount for these data.

5 Exact size and power comparison

In this section, we compare the exact sizes and powers of the six tests of (1). We consider the four exact unconditional tests Z, Z_C, L, and L_C of this paper and the two common tests, the McNemar's asymptotic test M and the exact conditional binomial test CB. Exact sizes and powers of these six tests are computed using the trinomial distribution (3).

5.1 Size and power computations

Consider first the exact sizes of the tests Z, L, Z_C, and L_C. For a given value of α, the level-α rejection region of Z is

$$R_Z^\alpha = \{(u, v): p_Z(u, v) \leq \alpha\},$$

where p_Z is defined in (8). Define $Z' = \min\{Z(x_{01}, x_{10}) : p_Z(x_{01}, x_{10}) \leq \alpha\}$. Then $R_Z^\alpha = \{(u, v) : Z(u, v) \geq Z'\}$. By definition, the exact size of Z is

$$\text{size}(R_Z^\alpha) = \sup_{\{(p_{01}, p_{10}): p_{01} \geq p_{10}\}} P_{p_{01}, p_{10}}((X_{01}, X_{10}) \in R_Z^\alpha).$$

By the same argument as in section 3.1. R_Z^α is a Barnard convex set. Therefore, by Theorem 1 the exact size of Z is

$$\text{size}(R_Z^\alpha) = \sup_{0\leq p\leq \frac{1}{2}} P_{p,p}((X_{01}, X_{10}) \in R_Z^\alpha) = \sup_{0\leq p\leq \frac{1}{2}} \sum_{(u,v)\in R_Z^\alpha} m(u, v; n, p, p). \qquad (14)$$

Similarly, the size of L is calculated by replacing R_Z^α in (14) with the level-α rejection region $R_L^\alpha = \{(u, v) : p_L(u, v) \leq \alpha\}$.

For the confidence interval tests Z_C and L_C, it is not obvious that the level-α rejection regions of these tests are Barnard convex sets. For a given n and α we can examine the rejection regions of Z_C and L_C. If they are Barnard convex sets, then the tests' sizes can be calculated by following (14). For every one of the sample sizes in our comparisons and $\alpha = 0.05$, the rejection regions of Z_C and L_C are Barnard convex sets and the exact sizes were computed as in (14).

Consider computing the exact sizes of CB and M. For CB, $(x'_{01}, x'_{10}) \in R_{CB}^\alpha$ if $x'_{01} \leq x_{01}$ and $x'_{10} \geq x_{10}$ for $(x_{01}, x_{10}) \in R_{CB}^\alpha$, where $R_{CB}^\alpha = \{(u, v) : P(U \leq u|U + V = u + v) \leq \alpha\}$. This is because

$$P(X_{01} \leq x'_{01}|X_{01} + X_{10} = x'_{01} + x'_{10}) \leq P(X_{01} \leq x_{01}|X_{01} + X_{10} = x_{01} + x_{10}) \leq \alpha.$$

Hence, the level-α rejection region of CB is a Barnard convex set. The rejection region $R_Z^\alpha = \{(u, v) : Z(u, v) \geq z_\alpha\}$ of the asymptotic test M is also a Barnard convex set by the same argument as in section 3.1. (z_α is the $100(1 - \alpha)\%$ percentile from a standard normal distribution.) Therefore, the exact sizes of both CB and Z are computed similarly to (14).

Finally, the exact powers of these tests are calculated based on the trinomial distribution of the data. For example, the power of Z for $(p_{01}, p_{10}) \in H_1$ is

$$\text{power}(p_{01}, p_{10}; R_Z^\alpha) = \sum_{(u,v)\in R_Z^\alpha} m(u, v; n, p_{01}, p_{10}).$$

5.2 Size and power comparisons

The sizes and powers of the six tests, Z, L, Z_C, L_C, CB, and M, were computed as described in section 5.1. For Z_C and L_C, $\beta = 0.0005$ was used as the error probability for the confidence interval. In this study all comparisons were carried out using $\alpha = 0.05$. The first five tests are level-α tests. M is asymptotically level-α.

In Table 4 we list the exact sizes of the tests for 15 sample sizes, $n = 10(5)40(10)100(50)200$. First, consider the four exact unconditional tests, Z, L, Z_C, and L_C. For $n = 10$, the tests are identical and the sizes are equal. For $n = 15$, 20, and 25, the size of L is closest to $\alpha = 0.05$; in some cases the sizes of Z_C and L_C equal the size of L. In all but one case, for all $n \geq 30$ the sizes of all four tests are between 0.0484 and 0.05. So, all four exact unconditional tests do a good job of attaining a size close to but no more than the nominal level of $\alpha = 0.05$. Sidik[13] examined the sizes of the tests for 39 sample sizes and found that in cases when the sizes of Z and Z_C differed

Table 4 Exact sizes of the tests with $\alpha = 0.05$

	Tests					
n	CB	M	Z	Z_C	L	L_C
10	0.0208	0.0652	0.0265	0.0265	0.0265	0.0265
15	0.0304	0.0592	0.0369	0.0369	0.0498	0.0304
20	0.0339	0.0577	0.0393	0.0393	0.0485	0.0485
25	0.0342	0.0539	0.0382	0.0478	0.0478	0.0478
30	0.0498	0.0558	0.0494	0.0494	0.0450	0.0494
35	0.0448	0.0528	0.0499	0.0486	0.0499	0.0485
40	0.0404	0.0527	0.0499	0.0484	0.0497	0.0484
50	0.0373	0.0595	0.0495	0.0495	0.0495	0.0495
60	0.0462	0.0524	0.0493	0.0493	0.0493	0.0493
70	0.0397	0.0598	0.0492	0.0492	0.0492	0.0492
80	0.0465	0.0523	0.0492	0.0492	0.0492	0.0492
90	0.0401	0.0567	0.0491	0.0491	0.0491	0.0491
100	0.0443	0.0522	0.0495	0.0491	0.0491	0.0491
150	0.0439	0.0521	0.0489	0.0493	0.0489	0.0493
200	0.0441	0.0521	0.0488	0.0495	0.0489	0.0495

$\beta = 0.0005$ in Z_C and L_C.

greatly, Z_C had the larger size, closer to α. The same was true when comparing L and L_C; L_C had the size closer to α.

On the other hand, in Table 4 the sizes of the asymptotic test M are larger than $\alpha = 0.05$ for all the sample sizes. Clearly, M is liberal for testing (1).

The sizes of CB are small, rising above 0.045 in only three cases in Table 4. The size of CB is smaller than all the other tests for all sample sizes except $n = 30$, for which its size is slightly larger than the sizes of the four exact unconditional tests. As expected, CB is very conservative because of the conditional nature of the test.

To better understand the sizes of the tests, we plotted the size functions of the six tests for $n = 50$ and $n = 100$ in Figure 2. The size function is the function that is maximized in computing the exact size of a test; for example, the size function of Z is

$$f(p; R_Z^\alpha) = \sum_{(u,v)\in R_Z^\alpha} m(u, v; n, p, p) \quad \text{for } 0 \le p \le \tfrac{1}{2}.$$

The size function of M exceeds the $\alpha = 0.05$ line over some regions of p for both sample sizes. The curve for CB is always much lower than the line $\alpha = 0.05$ over the complete region of p. On the other hand, the size function curves of Z, L, Z_C, and L_C are very close to and below the line $\alpha = 0.05$ over most of the region of p. In particular, the curves of Z, Z_C, and L_C for $n = 100$ are close to $\alpha = 0.05$. Note, the tests Z and L are identical, as are Z_C and L_C, for $n = 50$ and $\alpha = 0.05$ in Figure 2.

To compare the powers of the tests with $\alpha = 0.05$, we considered the nine sample sizes, $n = 10, 25, 35, 50, 60, 80, 100, 150, 200$. The exact powers were calculated for the grid of 100 pairs of p_{01} and p_{10} under H_1, which are determined by $p_{01} = 0.025(0.05)0.475$ and $p_{10} = [p_{01} + 0.05](0.05)[1 - p_{01}]$. The average powers are given in Table 5, and these relationships can be noted. For all nine sample sizes, the average power for M is the highest; but, of course, this is because M is a liberal test,

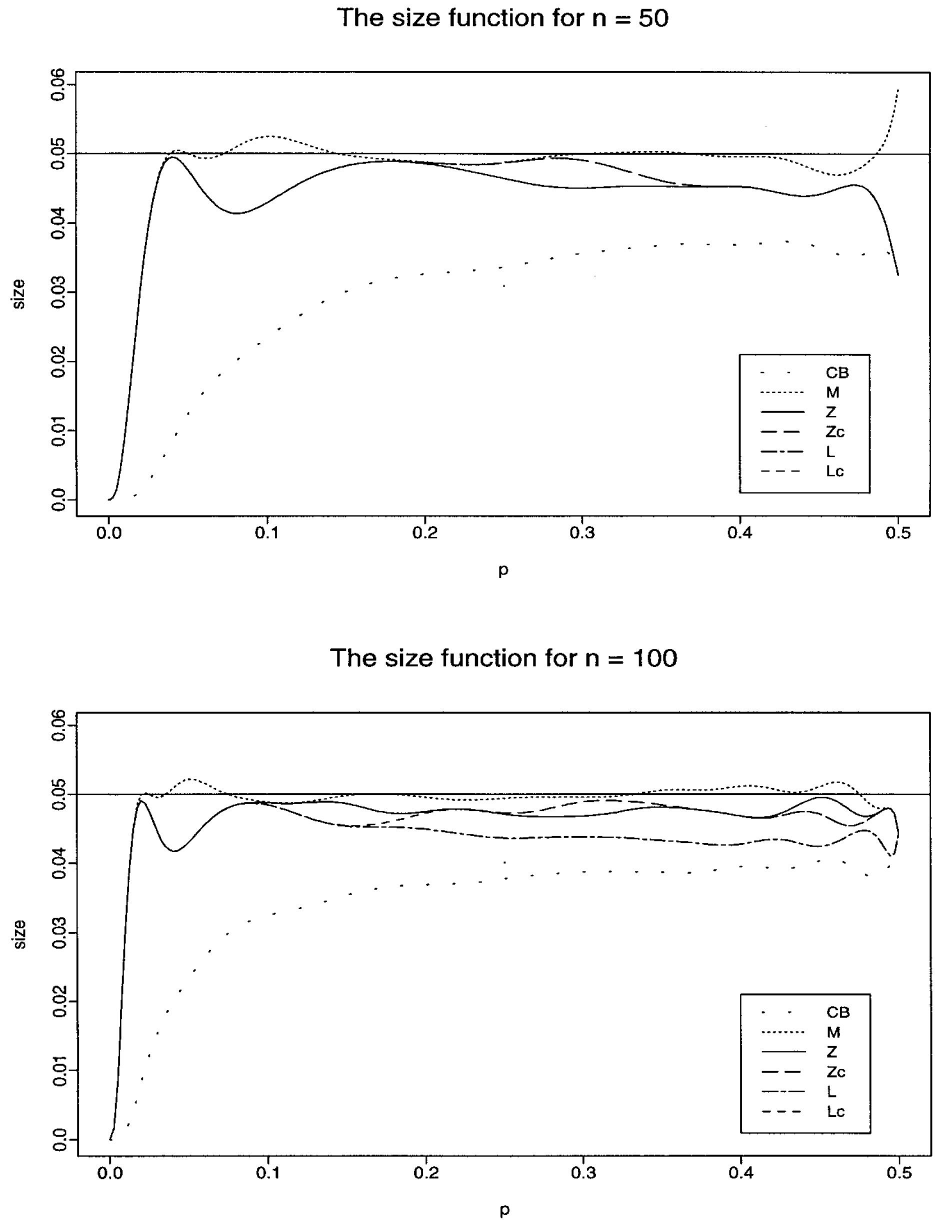

Figure 2 The size functions of the tests with $\alpha = 0.05$ ($\beta = 0.0005$ in Z_C and L_C)

and its size exceeds $\alpha = 0.05$. Among the five level-α tests, Z_C always has the highest average power or is tied for the highest. For all cases except one ($n = 25$ compared to L) *CB* has the lowest average power, confirming the conservativeness of the conditional test. In most cases in Table 5, the average powers of the four exact unconditional tests, Z, Z_C, L, and L_C, are very close, but Z_C has a slight advantage.

Table 5 Average (over 100 points) power with $\alpha = 0.05$

	Tests					
n	CB	M	Z	Z_C	L	L_C
10	0.287	0.439	0.316	0.316	0.316	0.316
25	0.555	0.618	0.577	0.588	0.554	0.581
35	0.635	0.676	0.662	0.663	0.662	0.659
50	0.700	0.733	0.723	0.725	0.723	0.725
60	0.733	0.757	0.751	0.753	0.751	0.752
80	0.775	0.794	0.789	0.791	0.787	0.790
100	0.804	0.820	0.817	0.817	0.813	0.816
150	0.849	0.860	0.858	0.858	0.854	0.858
200	0.876	0.884	0.880	0.882	0.879	0.882

$\beta = 0.0005$ in Z_C and L_C.

Another summary of the pairwise comparisons for the same nine sample sizes is presented in Table 6. Each block of nine entries represents a comparison of the row test and the column test, and the nine positions in each block correspond to the nine sample sizes in this pattern:

$$\begin{array}{ccc} 10 & 25 & 35 \\ 50 & 60 & 80 \\ 100 & 150 & 200. \end{array}$$

The symbol '$=$' indicates the power function of the two tests are exactly equal because the rejection regions of the two tests are identical. Notation '$<$' means the column test is uniformly more powerful than the row test because the rejection region of the row test is a proper subset of the rejection region of the column test. Symbol '$>$' indicates the row test is uniformly more powerful than the column test because the rejection region of the column test is a proper subset of the rejection region of the row test. In cases where none of the uniform comparisons apply, the powers were computed for all the 100 paired points (p_{01}, p_{10}). The proportion of the points at which the column test's power exceeds the row test's power is listed as a percent. These comparisons show that M is uniformly more powerful than all the other tests because of the incorrect, liberal size of the test. The four exact unconditional tests are uniformly more powerful than CB in all cases except the comparison with L when $n = 25$. Z_C is identical to or uniformly more powerful than Z and L for seven of the nine sample sizes. L_C is identical to or uniformly more powerful than L for eight of the nine sample sizes. In addition, L_C is identical to or uniformly more powerful than Z for five of the nine sample sizes, and its power is higher than Z more frequently for another three sample sizes. As far as the comparisons of Z_C and L_C are concerned, Z_C is uniformly more powerful than L_C for five of the nine and identical to L_C for another three of the nine sample sizes. In the other case, Z is uniformly more powerful than L_C. In all five cases when Z_C is not the same as or uniformly more powerful than another level-α test, the power of Z_C exceeds the power of the other test over more than 50% of the alternative points. Thus, for the cases

Table 6 Pairwise power comparison of the tests for $\alpha = 0.05$

	M			Z			Z_C			L			L_C		
CB	<	<	<	<	<	<	<	<	<	<	25	<	<	<	<
	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<
	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<
L_C	<	<	<	=	33	32	=	78	=	=	>	32			
	<	<	<	>	>	>	=	<	<	>	>	>			
	<	<	<	<	41	>	<	<	<	>	>	>			
L	<	<	<	=	91	=	=	91	67						
	<	<	<	=	<	<	<	<	<						
	<	<	<	<	<	<	<	<	<						
Z_C	<	<	<	=	>	32									
	<	<	<	>	>	>									
	<	<	<	46	>	>									
Z	<	<	<												
	<	<	<												
	<	<	<												

sample size in each block		
10	25	35
50	60	80
100	150	200

Note: = means row and column tests are the same. < means column test is uniformly more powerful than row test. > means row test is uniformly more powerful than column test. Numeric value is percentage of H_1 on which column test's power exceeds row test's power.
$\beta = 0.0005$ in Z_C and L_C.

considered in Tables 5 and 6, Z_C appears to be the level-α test with the best power properties.

6 Conclusions

In this paper, we introduced four exact unconditional tests for the problem of testing the one-sided hypothesis about two paired proportions. By considering the monotonicity of the joint distribution, these tests can be defined by considering one nuisance parameter on the boundary of H_0 and H_1. This simplifies the computation of p-values for these tests.

The size and power of the four exact unconditional tests, an asymptotic test, and a conditional test were compared. We found that the exact unconditional tests, Z, L, Z_C, and L_C, have accurate size properties; their exact sizes are less than and very close to the level of the tests. For sample sizes like $n = 100$, the size function curves suggested that Z, Z_C, and L_C are approximately unbiased for testing (1) with the curves being very close to $\alpha = 0.05$ over almost the whole region of p on the boundary between H_0 and H_1. Also, we found that the exact size of the asymptotic test M is always larger than the nominal level of the test. Therefore, it is not appropriate to use M for testing (1). In addition, it has been shown that the exact conditional binomial test CB is conservative, and its size is usually much smaller than the level of the test. Furthermore, the results of the power comparisons indicate that the confidence interval tests Z_C and L_C are generally more powerful than the non-interval tests Z and L. Among the four tests, Z_C,

L_C, and Z generally have better power than L. The exact unconditional tests are almost always uniformly more powerful than the exact conditional binomial test CB. The asymptotic test M is uniformly more powerful than all the other five tests because of its incorrect, liberal size. Overall, in this comparison, Z_C appears to be the level-α test with the best power properties.

Acknowledgement

We thank Dennis D. Boos, David A. Dickey, and William H. Swallow for their helpful comments.

References

1 McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12**: 153–57.
2 Cochran WG. The comparisons of percentages in matched samples. *Biometrika* 1950; **37**: 256–66.
3 Suissa S, Shuster JJ. The 2 × 2 matched-pairs trials: exact unconditional design and analysis. *Biometrics* 1991; **47**: 361–72.
4 Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden-Day, 1977.
5 Berger RL, Boos DD. *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89**: 1012–16.
6 Hsueh HM, Liu JP, Chen JJ. Unconditional exact tests for equivalence or noninferiority for paired binary endpoints. *Biometrics* 2001; **57**: 478–83.
7 Barnard GA. Significance tests for 2 × 2 tables. *Biometrika* 1947; **34**: 123–38.
8 Sidik K, Berger RL. Theoretical consideration of exact unconditional tests for one-sided comparisons of two parameters in discrete data. Technical Report. Raleigh, NC: North Carolina State University, Statistics Department, 1997.
9 Shaked M, Shanthikumar JG. *Stochastic orders and their applications*. San Diego, CA: Academic Press, 1994.
10 Casella G, Berger RL. *Statistical inference*, second edition. Pacific Grove, CA: Duxbury, 2002.
11 Storer BE, Kim C. Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association* 1990; **85**: 146–55.
12 Kang SH, Chen JJ. An approximate unconditional test of non–inferiority between two proportions. *Statistics in Medicine* 2000; **19**: 2089–2100.
13 Sidik K. Exact unconditional tests for discrete data. Unpublished Ph.D. dissertation. Raleigh (NC): North Carolina State University, 1997.
14 Robertson T, Wright FT, Dykstra RL. *Order restricted statistical inference*. New York: John Wiley, 1988: 38–39.
15 Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**: 404–13.
16 Berger, RL. More powerful tests from confidence interval *p* values. *The American Statistician* 1996; **50**: 314–18.