Research Article

J. D. Consiglio, G. Shan and G. E. Wilding*

A Comparison of Exact Tests for Trend with Binary Endpoints Using Bartholomew's Statistic

Abstract: Tests for trend are important in a number of scientific fields when trends associated with binary variables are of interest. Implementing the standard Cochran–Armitage trend test requires an arbitrary choice of scores assigned to represent the grouping variable. Bartholomew proposed a test for qualitatively ordered samples using asymptotic critical values, but type I error control can be problematic in finite samples. To our knowledge, use of the exact probability distribution has not been explored, and we study its use in the present paper. Specifically we consider an approach based on conditioning on both sets of marginal totals and three unconditional approaches where only the marginal totals corresponding to the group sample sizes are treated as fixed. While slightly conservative, all four tests are guaranteed to have actual type I error rates below the nominal level. The unconditional tests are found to exhibit far less conservatism than the conditional test and thereby gain a power advantage.

Keywords: trend test, exact test, isotonic regression

DOI 10.1515/ijb-2014-0013

1 Introduction

Suppose realizations of a binary variable are recorded in a k -group setting where the groups are defined by ordered categories. The data associated with such experiments can be organized as a $2 \times k$ table; see Table 1. Most commonly, x_i , the number of responses in group i , is assumed to follow a binomial distribution with parameters π_i and n_i , $i = 1, 2, \dots, k$. Furthermore, we restrict interest to scenarios where x_i is assumed independent of x_j for all $i \neq j$. The parameter π_i is often referred to as the response probability in situations where the quantity represents the likelihood of positive response when the subject is exposed to the stimuli or intervention associated with the group. Under these situations, it is often of interest to test the equality of the response probabilities, and the null hypothesis takes the form

$$H_0 : \pi_1 = \pi_2 = \dots = \pi_k.$$

Consider an experiment with four dose groups, each fixed by the design to have 10 subjects. There is expected to be an increasing probability of response with increasing dose level, and doses administered are 0, 2.5, 25, and 250 units. In such scenarios, use of tests of the null in favor of the general alternative may result in an inefficient procedure in that they ignore the ordinal information which defines the groups. Rather, a test for the ordered alternative (a.k.a. “simple order hypothesis” [1]) is appropriate, and the alternative may be written as

$$H_a : \pi_1 \leq \pi_2 \leq \dots \leq \pi_k, \quad \pi_1 < \pi_k.$$

*Corresponding author: **G. E. Wilding**, Department of Biostatistics, The State University of New York at Buffalo, 3435 Main Street, Buffalo, NY 14214, USA, E-mail: gwilding@buffalo.edu

J. D. Consiglio, Department of Biostatistics, The State University of New York at Buffalo, 3435 Main Street, Buffalo, NY 14214, USA, E-mail: jdc46@buffalo.edu

G. Shan, Department of Environmental and Occupational Health, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA, E-mail: guogen.shan@unlv.edu

Table 1 Data structure

Response	Dose 1	Dose 2	...	Dose k
1	x_1	x_2	...	x_k
0	$n_1 - x_1$	$n_2 - x_2$	...	$n_k - x_k$
	n_1	n_2	...	n_k

Although the standard test due to Cochran and Armitage [2, 3] may be applied, the inherent weakness of this approach is the dependence on a choice of scores, which can be somewhat arbitrary and present the researcher with a dilemma in that the operating characteristics are dependent on the choice. Given that the procedure is based upon an assumed linear relationship between the chosen score and the probability of response, an incorrect choice in scores can result in losses in efficiency and vastly different conclusions for a given data set. In the aforementioned example, say responders totaled 0, 1, 4, and 3 in the four groups receiving the lowest to highest doses, respectively; see Table 2. Using dosage amount as the score, the Cochran–Armitage procedure does not detect any trend in response probabilities, with p -value 0.2615. However, if the transformation $\log\{\text{dose} + 0.01\}$ is taken as a score, the resulting p -value is 0.0332.

Table 2 Hypothetical data example

Response	Dose amount			
	0	2.5	25	250
1	0	1	4	3
0	10	9	6	7
	10	10	10	10

Bartholomew [4] proposed a test utilizing isotonic regression that addressed trend alternatives via a likelihood-ratio approach, the test statistic being a variant of the standard Pearson chi-square statistic. In that the null distribution depends on an unknown nuisance parameter in finite samples, Bartholomew published tables of asymptotic critical values for his statistic in the balanced sample size case [5] for use in practice. Poon [6] conducted a Monte Carlo study indicating that the trend test based on these critical values sometimes violated the nominal type I error rate; Collings et al. [7] corroborated these simulation results. Leuraud and Benichou [8, 9] also used simulation to assess the statistical properties of Bartholomew's test, finding the procedure to have good power properties under a number of alternatives.

Given the aforementioned type I error control problems, it is natural to alternatively consider use of the exact probability distribution. Complications arise in that, conditional on $n_1, n_2, \dots, n_k$, the distribution functions of $x_1, x_2, \dots, x_k$ are dependent upon the unknown nuisance parameter $\pi_1 = \pi_2 = \dots = \pi_k = \pi$. A number of classic solutions exist for the special case of $k = 2$. For testing equality of two binomial probabilities, Fisher's famous approach [10] conditions on the marginal totals of an observed 2×2 table. Eliminating dependence on the nuisance parameter π by restricting attention to 2×2 tables with the same marginal totals as the observed table, the p -value for Fisher's exact test is the sum of hypergeometric probabilities. Several exact tests for trend have been developed, many listed by Agresti and Coull [11], all of which used the exact conditional approach. The term conditional in this context refers to the fact that the corresponding null distributions are those conditioning on the total number of successes, and these procedures can be associated with a high degree of discreteness producing conservative tests.

For the 2×2 case, Barnard [12] proposed a more powerful test of $H_0 : \pi_1 = \pi_2 = \pi$ that called for conditioning only on n_1 and n_2 . An expression containing $\pi^{x_1+x_2}(1-\pi)^{n-x_1-x_2}$ is then maximized over the binomial parameter π to obtain the p -value. Berger and Boos [13] augment this procedure by maximizing

only over a $100(1 - \beta)\%$ confidence interval for π , then penalizing the resulting p -value by adding β . In the present context, we consider the more general problem of testing equality of $k \geq 2$ binomial parameters against a trend (simple order) alternative.

An approximate unconditional approach to eliminating the nuisance parameter is the notion of simply replacing it with an estimate, for instance one obtained via maximum likelihood [14]. This procedure provides an estimated (E) p -value. Lloyd [15] proposed using this E p -value as an ordering criterion and then maximizing the null likelihood, resulting in the $E + M$ (Estimation + Maximization) p -value. Shan et al. [16] investigated this approach in the context of detecting monotone trend using the Cochran-Armitage test statistic and subsequently using the Baumgartner-Weiss-Schindler statistic [17].

Due to the conservative nature of conditional tests, and the violations of type I error that accompany tests based on asymptotic critical values as observed by Poon [6], it is desirable to develop exact unconditional tests for trend for use with qualitatively ordered samples using the Bartholomew statistic.

Section 2 gives background information on Bartholomew's test for trend. Section 3 details the proposed exact testing procedures. Section 4 reports type I error and power properties of the examined procedures. Section 5 presents an example of the tests' application, and Section 6 contains concluding remarks.

2 Background

Bartholomew's [4] test of equality against the simple order alternative as applied to binomial counts is based on the statistic

$$T_B(\mathbf{x}) = \frac{\sum_{i=1}^k n_i (\hat{\pi}_i^* - \bar{\pi})^2}{\bar{\pi}(1 - \bar{\pi})},$$

where $\mathbf{x} = (x_1, x_2, \dots, x_k)$, $\bar{\pi}$ is the maximum likelihood estimator (MLE) of the nuisance parameter ($\bar{\pi} = \sum_{i=1}^k x_i / \sum_{i=1}^k n_i$), and the $\hat{\pi}_i^*$ s are the isotonic regression estimates of the π_i s. This specific application of Bartholomew's test, as described in Barlow et al. [1], begins with determination of the vector $\hat{\boldsymbol{\pi}}^* = (\hat{\pi}_1^*, \hat{\pi}_2^*, \dots, \hat{\pi}_k^*)$, the MLE of $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_k)$ subject to the order restriction [18],

$$0 < \pi_1 \leq \pi_2 \leq \dots \leq \pi_k < 1.$$

If the usual MLEs $\hat{\pi}_i = x_i/n_i$ are observed to be isotonic (i.e. $\hat{\pi}_1 \leq \hat{\pi}_2 \leq \dots \leq \hat{\pi}_k$), then $\hat{\boldsymbol{\pi}}^* = \hat{\boldsymbol{\pi}}$, for $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_k)$. If the $\hat{\pi}_i$ s are not in the order corresponding to H_a , then the order-restricted MLE $\hat{\boldsymbol{\pi}}^*$ may be obtained using the pool-adjacent-violators algorithm (PAVA) [19]. For more on the estimates computed by PAVA for independent binomial samples and the MLE of $\boldsymbol{\pi}$ under the simple order restriction, see Silvapulle and Sen [20]. Recall the hypothetical example presented earlier, with $n_i = 10$, $i = 1, 2, 3, 4$, and $\hat{\boldsymbol{\pi}} = (0, 0.1, 0.4, 0.3)$. These $\hat{\pi}_i$ s do not concur with the order restriction shared by the parameter space, since $0.3 \not\geq 0.4$. Following the steps dictated by PAVA [4], the violating quantities, namely 0.4 and 0.3, are pooled together, giving $(4 + 3)/(10 + 10) = 0.35$. Replacing both 0.4 and 0.3 by the pooled value 0.35 gives an array of nondecreasing estimates, and hence $\hat{\boldsymbol{\pi}}^* = (0, 0.1, 0.35, 0.35)$.

Although the statistic T_B is essentially Pearson's chi-square statistic applied to the isotonic regression estimates $\hat{\boldsymbol{\pi}}^*$ [1, 7, 11], the distribution of T_B is not as convenient as that of Pearson's chi-square. In the context of binary data, the sample proportions $\hat{\pi}_i$ would have only an asymptotic normal distribution, and the isotonic modification is an additional complicating factor. The asymptotic null distribution of T_B may be expressed as $\sum_{a=1}^k p_a \chi_{a-1}^2$ (referred to as the chi-square-bar distribution [4]) where χ_{a-1}^2 is a chi-squared random variable with $a - 1$ degrees of freedom ($\chi_0^2 \equiv 0$), and p_a is the probability that the vector $\hat{\boldsymbol{\pi}}^*$ has a distinct ordered values. While Barlow et al. [1, pp. 134–149] present formulas and recursions to calculate p_a , the expression is a function of the number of groups k , and

even for moderate k this may require several numerical integrations. [21]

Determining values for p_a can be highly prohibitive, necessitating the use of simulation to obtain approximate values [22]. Critical values are presented by Bartholomew [5] for the equal sample size case after a precise determination of the required p_a values.

Suppose the number of observed responders in the k groups of a particular data set is $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0k})$, and we define $s = \sum_{i=1}^k x_{0i}$, $n = \sum_{i=1}^k n_i$. By conditioning on s , which is a sufficient statistic, we obtain the exact conditional p -value as,

$$P_{\text{cond}}(\mathbf{x}_0) = \Pr\left(T_B(\mathbf{x}) \geq T_B(\mathbf{x}_0) \mid \sum \mathbf{x} = s, H_0\right) = \sum_{\mathbf{x} \in \Omega_{\text{cond}}(\mathbf{x}_0)} \frac{\prod_{i=1}^k \binom{n_i}{x_i}}{\binom{n}{s}},$$

where $\Omega_{\text{cond}}(\mathbf{x}_0) = \{\mathbf{x} : T_B(\mathbf{x}) \geq T_B(\mathbf{x}_0) \text{ and } \sum \mathbf{x} = s\}$.

Poon [6] generated tables of binomial counts from the conditional distribution and approximated the type I error and power of the exact test using the Bartholomew statistic under a variety of parameter settings, each with 500 simulations. Conditioning on the number of responses provides a p -value expression that is free of the nuisance parameter π . The exact type I error rate and power for this approach are computed according to the prescription of Mehta et al., who evaluated exact unconditional power of the conditional Cochran–Armitage test [23].

3 Methods of unconditional exact testing

Three unconditional exact testing procedures are now described in detail and distinguished from the conditional version.

3.1 Maximization test

In the unconditional setting, the number of responses is not held fixed, and individual table probabilities depend on the nuisance parameter π . The p -value obtained by maximizing the null likelihood over the entire range of the nuisance parameter [24] is referred to as the M p -value. Since large values of the test statistic signify evidence against the null in our case, the rejection region for the M test is $\Omega_M(\mathbf{x}_0) = \{\mathbf{x} : T_B(\mathbf{x}) \geq T_B(\mathbf{x}_0)\}$, and the M p -value is

$$P_M(\mathbf{x}) = \sup_{\pi \in [0,1]} \left\{ \sum_{\mathbf{x} \in \Omega_M(\mathbf{x}_0)} \prod_{i=1}^k \binom{n_i}{x_i} \pi^{x_i} (1-\pi)^{n_i-x_i} \right\}.$$

This approach has been developed and applied in Suissa and Schuster [14].

3.2 Confidence interval test

An adjustment to the maximization approach was proposed by Berger and Boos [13] for the general testing problem involving a nuisance parameter and is referred to as the CI test. The approach is known elsewhere as partial maximization, e.g. Gunther et al. [25]. A confidence interval for the nuisance parameter is computed, and maximization is constrained to nuisance parameter values within this interval rather than the entire sample space (in this case $[0, 1]$). This approach has the potential to save on computational time and has the intuitive appeal of excluding parameter values that are inconsistent with the observed data. The usual $100(1 - \beta)\%$ interval $C(\mathbf{x}_0)$ for π is $\bar{\pi} \pm Z_{1-\beta/2} \sqrt{\bar{\pi}(1 - \bar{\pi})/n}$, where Z_γ is the γ th quantile from the standard normal distribution. We based our procedure on the Wald confidence interval for simplicity, but

we note that alternative methods of confidence interval estimation may be considered. As a penalty for not conducting the full maximization, β is added in the CI p -value expression, and one may take $\beta = 0.001$ as in Berger and Boos. The resulting p -value is

$$P_{CI}(\mathbf{x}_0) = \sup_{\pi \in C(\mathbf{x}_0)} \left\{ \sum_{\mathbf{x} \in \Omega_{CI}(\mathbf{x}_0)} \prod_{i=1}^k \binom{n_i}{x_i} \pi^{x_i} (1 - \pi)^{n_i - x_i} \right\} + \beta,$$

where $\Omega_{CI}(\mathbf{x}_0) = \Omega_M(\mathbf{x}_0)$.

In the case of the general alternative, software applying this approach to $2 \times k$ tables is available in *StatXact* [26], and for 2×2 tables via an internet application [27], which carries out the procedure based on a statistic of the user's choosing (choice is limited to z-pooled, z-unpooled, or Fisher–Boschloo statistic).

3.3 Estimation and maximization test

Lloyd's $E + M$ procedure [15, 28] for a general testing problem involving a nuisance parameter begins by computing an estimated (E) p -value. In this case, the nuisance parameter π is replaced by its MLE $\bar{\pi}$, and the E p -value for the problem at hand is computed as

$$P_E(\mathbf{x}_0) = \sum_{\mathbf{x} \in \Omega_E(\mathbf{x}_0)} \prod_{i=1}^k \binom{n_i}{x_i} \bar{\pi}^{x_i} (1 - \bar{\pi})^{n_i - x_i},$$

where once again $\Omega_E(\mathbf{x}_0) = \Omega_M(\mathbf{x}_0)$. Lloyd notes that this E p -value is typically slightly liberal. An exact test (the $E + M$ test) is obtained by defining the rejection region not in terms of $T_B(\mathbf{x}_0)$, but of $P_E(\mathbf{x}_0)$. In effect, the E p -value is used as a test statistic by which the space of possible $2 \times k$ tables is ordered. The ordering based on the estimated p -value is different than the ordering defined by the original statistic [25]. The rejection region for the $E + M$ test becomes $\Omega_{E+M}(\mathbf{x}_0) = \{\mathbf{x} : P_E(\mathbf{x}) \leq P_E(\mathbf{x}_0)\}$, and the $E + M$ p -value is

$$P_{E+M}(\mathbf{x}_0) = \sup_{\pi \in [0,1]} \left\{ \sum_{\mathbf{x} \in \Omega_{E+M}(\mathbf{x}_0)} \prod_{i=1}^k \binom{n_i}{x_i} \pi^{x_i} (1 - \pi)^{n_i - x_i} \right\}.$$

Applications of the $E + M$ approach are found in Lloyd [29, 30] and Shan et al. [16, 17].

4 Test properties

The exact methods detailed in Section 3 and the exact conditional approach from Section 2 are compared with the test using the asymptotic critical values presented in Bartholomew [5]. For additional comparison, the Cochran–Armitage test with equally spaced scores based on the asymptotic and exact conditional null distributions are included. Type I and type II error rates were calculated through complete enumeration, i.e. presented rates are exact and not obtained through Monte Carlo simulation. In the case of the conditional procedure, the unconditional error rates are presented for comparability which are the rates across possible values of s . Complications arise in computing T_B when the observed $\bar{\pi}$ is equal to 0 or 1. In these instances, since such tables show no evidence of increasing trend, we set $T_B \equiv 0$.

Figure 1 displays exact type I error plotted over the range of the nuisance parameter for the four exact tests and the test based on asymptotic critical values. For scenarios defined by the number of groups (3 or 4) and the sample size within each group (10 or 20), nuisance parameter values between 0 and 1 by an increment of 0.01 were used in implementation of the exact tests. An alternate approach is a two-stage procedure involving an initial pass to identify possible regions for the global maximum, followed by a more refined search of the identified regions. The critical values of T_B used to evaluate the asymptotic test were

Figure 1 Exact type I error rates of the Bartholomew test based on the unconditional maximization (M), confidence interval (CI), expectation and maximization ($E + M$), conditional ($C(T_B)$), and asymptotic-based ($A(T_B)$) approaches, and the Cochran–Armitage test based on conditional ($C(T_{CA})$) and asymptotic-based ($A(T_{CA})$) approaches

3.820 and 4.528, respectively, for $k = 3$ and $k = 4$ [5]. The profiles corresponding to the exact conditional tests using Bartholomew's statistic and the Cochran–Armitage statistic are labeled, respectively, as $C(T_B)$ and $C(T_{CA})$; likewise, the tests based on asymptotic critical values of these same statistics are labeled $A(T_B)$ and $A(T_{CA})$. Profiles based on the unconditional maximization approach, confidence interval approach, and the application of Lloyd's approach are labeled M , CI , and $E + M$, respectively.

As can be seen in Figure 1, the unconditional tests preserve the nominal α -level in all four scenarios. In the first panel of Figure 1, the M , $E + M$, and $A(T_B)$ profiles coincide. While adequate in this situation ($k = 3$, $n_i = 10$), the test using the asymptotic critical value of T_B does not control type I error in the remaining panels, as the $A(T_B)$ profile is seen to exceed 0.05. The Cochran–Armitage test based on an asymptotic critical value violates the nominal level of type I error in all scenarios. Type I error of the unconditional tests is uniformly less conservative than that of the conditional tests. For extreme values of the nuisance parameter, the unconditional tests are seen to have significant advantage over the conditional test.

Power is first summarized in Figure 2 by choosing a parameter setting similar to those listed in Neuhauser [31], of the form $\pi_a = [0.1, 0.1 + \frac{\gamma}{2}, 0.1 + \gamma]$ for the case of $k = 3$, and $\pi_a = [0.1, 0.1 + \frac{\gamma}{3}, 0.1 + \frac{2\gamma}{3}, 0.1 + \gamma]$ for the case of $k = 4$, with $\gamma \in (0, 0.9)$. These settings are linear in nature, describing the “center” of the alternative space. It is difficult to distinguish the power curves in the first panel of Figure 2, except to note that in general, power is lowest for $C(T_{CA})$ and $A(T_{CA})$, both tests utilizing the Cochran–Armitage statistic. This holds for the remaining three panels of Figure 2. With some effort, power for $C(T_B)$ can be identified as having slightly worse power than $A(T_B)$, M , CI , and $E + M$,

Figure 2 Exact power for center of alternative space of the Bartholomew test based on the unconditional maximization (M), confidence interval (CI), expectation and maximization ($E + M$), conditional ($C(T_B)$), and asymptotic-based ($A(T_B)$) approaches, and the Cochran–Armitage test based on conditional ($C(T_{CA})$) and asymptotic-based ($A(T_{CA})$) approaches

which are all nearly identical. The third panel of Figure 2 indicates that for $k = 4$ and $n_i = 10$, $A(T_B)$ and CI enjoy a minor power advantage over the other tests, the top performer being dependent on the value of γ . We remind the reader of the type I error violations associated with the test based on T_B in conjunction with asymptotic critical values, thereby detracting from the power advantages.

To investigate power further, in Figure 3 we consider a parameter setting closer to the edge of the alternative space, choosing $\pi_a = [0.1, 0.1, 0.1 + \gamma]$ when $k = 3$, and $\pi_a = [0.1, 0.1, 0.1, 0.1 + \gamma]$ when $k = 4$, again with $\gamma \in (0, 0.9)$. The power properties are largely unchanged from the earlier parameter setting, with the tests based on T_{CA} suffering a power disadvantage, and power curves for the remaining tests grouping tightly together. The CI test shows slightly superior power properties in the third panel of Figure 3. Power for $C(T_B)$ can be seen below the main cluster of curves in all four panels.

5 Example

We consider the hypothetical example given in Fleiss et al. [32], tabulating one-month release rates as a function of initial severity measured on an ordinal scale. The observed counts appear in Table 3, and testing results are found in Table 4. For the unconditional exact approaches, all 2×4 tables with the identical column totals are enumerated, and the Bartholomew statistic corresponding to each table is computed. p -Values for the M and $E + M$ tests are calculated by maximizing the expressions in Section 3.3 over $\pi \in (0, 1)$ by increments of 0.01. The p -values for the conditional tests are calculated by enumerating all 2×4 tables with the same row and column totals as the original data.

Figure 3 Exact power for edge of alternative space of the Bartholomew test based on the unconditional maximization (M), confidence interval (CI), expectation and maximization ($E + M$), conditional ($C(T_B)$), and asymptotic-based ($A(T_B)$) approaches, and the Cochran–Armitage test based on conditional ($C(T_{CA})$) and asymptotic-based ($A(T_{CA})$) approaches

Table 3 Example data from Fleiss et al.

Released	Initial severity			
	Mild	Moderate	Serious	Severe
Yes	25	22	12	6
No	5	3	8	19
	30	25	20	25

Table 4 Hypothesis test results for analysis of data in Fleiss et al.

Test	p -Value
M	0.00000149
CI	0.00100149
$E + M$	0.00000046
$C(T_B)$	0.00000020
$C(T_{CA})$	0.00000072
$A(T_B)$	0.00000101
$A(T_{CA})$	<0.005

Figure 4 depicts the p -value profiles of the M and $E + M$ tests. As the height of these curves depends on π , the M and $E + M$ p -values are the respective maxima of these profiles. The limits of the 99.9% Wald confidence interval for π obtained from the observed data are (0.493, 0.807), represented in the figure by vertical lines. The CI p -value is obtained by adding $\beta = 0.001$ to the maximum of the M test profile between

Figure 4 p -Value profiles for example in Fleiss et al.

these confidence limits. The horizontal dashed lines in the figure represent p -values for the conditional exact tests based on statistics T_B and T_{CA} . The values of these statistics obtained from the example data are 28.81 and 23.91, respectively. Significance can be assigned to these values using standard asymptotic methods by appealing to the limiting chi-square and chi-square-bar distributions. As the group sample sizes in this example are not balanced, the p -value from the asymptotic chi-square-bar distribution can be identified as <0.005 using the Table A.2 of Barlow et al. [1]. All tests detect a highly significant decreasing trend in the probability of release as initial severity increases at a 0.05 nominal significance level.

6 Conclusion and recommendation

The test using the asymptotic critical values of Bartholomew does not have acceptable type I error control. The exact conditional test is guaranteed to control type I error, but is conservative in all instances, sometimes severely so. The maximization, confidence interval, and Estimation + Maximization tests all maintain type I error control and show markedly less conservatism than the conditional test. Furthermore, each unconditional test exhibits a modest gain in power over conditional tests.

Based on the type I error and power properties, no single unconditional test is uniformly preferable to the other two. The $E + M$ test requires far more computational effort than the M and CI tests. The CI test involves a penalty term that can affect the p -value by orders of magnitude, in addition to the minor task of requiring calculation of a confidence interval for the nuisance parameter. By way of simplicity and computational convenience, we recommend Maximization as the unconditional method of choice.

References

1. Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. Statistical inference under order restrictions. New York: Wiley, 1972.
2. Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955;11:375–86.
3. Cochran WG. Some methods for strengthening the common χ^2 tests. *Biometrics* 1954;10:417–51.

4. Bartholomew DJ. A test of homogeneity for ordered alternatives. *Biometrika* 1959;46:36–48.
5. Bartholomew DJ. A test of homogeneity for ordered alternatives II. *Biometrika* 1959;46:328–35.
6. Poon A. A Monte Carlo study of the power of some k-sample tests for ordered binomial alternatives. *J Stat Comput Simulation* 1980;11:251–9.
7. Collings BJ, Margolin BH, Oehlert GW. Analyses for binomial data, with application to the fluctuation tests for mutagenicity. *Biometrics* 1981;37:775–94.
8. Leuraud K, Benichou J. A comparison of several methods to test for the existence of a monotonic dose-response relationship in clinical and epidemiological studies. *Stat Med* 2001;20:3335–51.
9. Leuraud K, Benichou J. Tests for monotonic trend from case-control data: Cochran–Armitage–Mantel trend test, isotonic regression and single and multiple contrast tests. *Biom J* 2004;46:731–49.
10. Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd, 1970.
11. Agresti A, Coull BA. The analysis of contingency tables under inequality constraints. *J Stat Plan Inference* 2002;107:43–73.
12. Barnard GA. A new test for 2×2 tables. *Nature* 1945;156:177.
13. Berger RL, Boos DD. Values maximized over a confidence set for the nuisance parameter. *J Am Stat Assoc* 1994;89:1012–16.
14. Suissa S, Schuster JJ. Exact unconditional sample sizes for the 2×2 binomial trial. *J R Stat Soc Ser A Stat Soc* 1985;148:317–27.
15. Lloyd CJ. Exact p -values for discrete models obtained by estimation and maximization. *Austr N Z J Stat* 2008;50:329–45.
16. Shan G, Ma C, Hutson AD, Wilding GE. An efficient and exact approach for detecting trends with binary endpoints. *Stat Med* 2012;31:155–64.
17. Shan G, Ma C, Hutson AD, Wilding GE. Some tests for detecting trends based on the modified Baumgartner-Weiss-Schindler statistics. *Comput Stat Data Anal* 2013;57:246–61.
18. Robertson T, Wright F, Dykstra RL. *Order restricted statistical inference*. New York: Wiley, 1988.
19. Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E. An empirical distribution function for sampling with incomplete information. *Ann Math Stat* 1955;26:641–7.
20. Silvapulle MJ, Sen PK. *Constrained statistical inference: inequality, order, and shape restrictions*. Hillsdale, NJ: Wiley, 2005.
21. Robertson T, Wright FT. One-sided comparisons for treatments with a control. *Can J Stat* 1985;13:109–22.
22. Agresti A, Coull BA. Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics* 1996;52:1103–11.
23. Mehta CR, Patel NR, Senchaudhuri P. Exact power and sample-size computations for the Cochran–Armitage trend test. *Biometrics* 1998;54:1615–21.
24. Basu D. On the elimination of nuisance parameters. *J Am Stat Assoc* 1977;72:355–66.
25. Gunther CC, Bakke O, Havard R, Langaas M. Trondheim (Norway): Norwegian university of science and technology; 2009:4. Statistical hypothesis testing for categorical data using enumeration in the presence of nuisance parameters.
26. StatXact. Cytel Software Inc., 2007.
27. Berger RL. Exact unconditional homogeneity/independence tests for 2×2 tables. 1996 [cited 2013 March 19]. Available at: <https://urldefense.proofpoint.com/v1/url?u=http://www.stat.ncsu.edu/exact/>.
28. Lloyd CJ. Efficient and exact tests of the risk ratio in a correlated 2×2 table with structural zero. *Comput Stat Data Anal* 2007;51:3765–75.
29. Lloyd CJ. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics* 2008;64:716–23.
30. Lloyd CJ. Exact tests based on pre-estimation and second order pivotals: non-inferiority trials. *J Stat Comput Simul* 2008;80:841–51.
31. Neuhauser M. And exact test for trend among binomial proportions based on a modified Baumgartner-Weiss-Schindler statistic. *J Appl Stat* 2006;33:79–88.
32. Fleiss JL, Levin B, Paik MC. *Statistical methods for rates and proportions*. Hillsdale, NJ: Wiley, 2003.