

## The Analysis of Stratified $2 \times 2$ Contingency Tables

Vance W. Berger<sup>1</sup> and Catalina Stefanescu<sup>\*2</sup>

<sup>1</sup> University of Maryland Baltimore County and National Cancer Institute,  
Executive Plaza North, Suite 3131, Bethesda, MD 20892-7354

<sup>2</sup> London Business School, Regent's Park, London NW1 4SA, UK

### Summary

We consider the problem of testing for independence against the consistent superiority of one treatment over another when the response variable is binary and is compared across two treatments in each of several strata. Specifically, we consider the randomized clinical trial setting. A number of issues arise in this context. First, should tables be combined if there are small or zero margins? Second, should one assume a common odds ratio across strata? Third, if the odds ratios differ across strata, then how does the standard test (based on a common odds ratio) perform? Fourth, are there other analyzes that are more appropriate for handling a situation in which the odds ratios may differ across strata? In addressing these issues we find that the frequently used Cochran–Mantel–Haenszel test may have a poor power profile, despite being optimal when the odds ratios are common. We develop novel tests that are analogous to the Smirnov, modified Smirnov, convex hull, and adaptive tests that have been proposed for ordered categorical data.

*Key words:* Admissibility, binary data, clinical trial, common odds ratio, omnibus test.

## 1 Introduction and Motivating Examples

The problem we consider is to determine whether or not one treatment is superior to the other, on the basis of binary data, when the randomization is stratified, leading to stratified  $2 \times 2$  tables. Such stratified studies are frequently encountered in the medical literature, for the purpose of ensuring balance with respect to a prognostic factor, often gender or center in a multi-center randomized clinical trial (Kernan et al., 1999). Miller (1980) discussed several approaches for analyzing sets of  $2 \times 2$  contingency tables, including the Woolf logit approach, the Mantel–Haenszel (MH) approach, the sign test, and several methods of combining independent p-values from each table. It is also common to pool the data across the strata, especially when there is a zero margin in at least one table.

As an illustration, Table 1 adapted from Kuritz, Landis and Koch (1988, page 136) summarizes the results of a randomized placebo–controlled clinical trial comparing the effects of a combination drug labeled A& B with placebo, on the level of obstetrical related pain on women who recently had delivered a baby. The initial pain status immediately after delivery, but prior to treatment, was self-reported on two levels as either "some" or "lots", and this variable was used for stratification.

[Table 1 about here.]

As another example, Table 2 gives the results from a prospective study comparing several maintenance therapies for adults with reflux esophagitis. These results were reported in Vigneri et al. (1995). Participants were stratified according to their initial grade of esophagitis and randomly assigned to 12 months of treatment with cisapride or omeprazole. For each initial grade of esophagitis (grade 1 or grade 2), there were 15 patients in each treatment group (see the Table 2 footnote in Vigneri et al., 1995). After 12 months of treatment, the recurrence of endoscopic signs was recorded.

---

\* Corresponding author: e-mail: cstefanescu@london.edu, Phone: +44 (0)20 7262 5050, Fax: +44 (0)20 7724 7875

[Table 2 about here.]

Stratified  $2 \times 2$  contingency tables also have a natural connection with meta-analytic studies, in that the former combine strata and the latter combine studies. So stratified  $2 \times 2$  tables may result from meta-analyses with binary outcomes, in which the studies themselves serve as the strata. It is important to pre-specify an analysis, even prior to conducting a meta-analysis; however, at the outset the researchers would not have the data, and hence would not be in a position to determine the p-value for any given analysis. The basis for selecting one procedure must therefore be its general properties, rather than its performance for the given set of data. One sets out to find a difference, and so one would naturally wish to choose the analysis method with the best power. As we shall demonstrate, this turns out not to be the most commonly used analysis.

The issue of whether or not to condition on the margins of a contingency table is one of the oldest and most disputed controversies in all of statistics (Lloyd, 1988; Little, 1989; Routledge, 1992; Upton, 1992; Mehta and Hilton, 1993; Lehmann, 1993). When studying several  $2 \times 2$  contingency tables, we adopt the convention that conditioning on the margins refers to the margins within tables, but not to margins across tables. There is an argument against conditioning which is specific to the analysis of stratified tables, and does not typically come up in the analysis of a single contingency table of any dimensions. If, at one center, every patient received the active treatment, or every patient responded to each treatment, then any permutation will retain the zero margin. While tables with zero margins may contribute substantially for estimation, a stratified analysis with fixed effects for centers would make no use of such tables. Even so, there are good reasons to stratify the analysis, at least to match the stratification used in the design. First, there is no between-group information in tables with zero margins, and thus they are consistent with any hypothesized magnitude of treatment effect. Second, such a stratified analysis is the design-based analysis, which ensures validity (Matts and Lachin, 1988; Berger, 2000). A third reason to stratify the analysis, especially for a variable that was not used to stratify the randomization, is that table-wise superiority of the active treatment does not, generally speaking, imply superiority in the combined table. An illustration is, for example, Simpson's paradox (Berger, 2004) which shows how misleading results can be when Fisher's exact test, or any other test, is applied to data pooled across levels of a prognostic and unbalanced covariate (Berger, 2005). This issue is related to the concept of confounders discussed in epidemiology, or the collapsibility defined in Bishop, Fienberg, and Holland (1975). Stratification avoids these problems.

The rest of the paper is structured as follows. In Section 2 we discuss the background, introduce the notation, and formulate the hypothesis testing problem. Then we study the set of most powerful tests for detecting any specific simple alternative, including the one that specifies a common odds ratio across strata (this turns out to be the MH test). In Section 3 we show that certain constructions, such as the convex hull, adaptive, and modified Smirnov tests (analogous to those for ordered categorical data), result in admissible tests for all sets of margins and all  $\alpha$ -levels. In Section 3 we also consider another desirable property, specifically, proper monotonicity. In Section 4 we compare the power of the proposed tests on two data sets, and discuss the relative advantages of each test. The paper concludes with a discussion in Section 5.

## 2 Background and Notation

The model is as follows. With  $i \in \{1, 2, \dots, I\}$ ,  $j \in \{1, 2\}$ , and  $k \in \{1, 2\}$  identifying tables, rows, and columns, respectively, let  $X_{ijk}$  represent the number of patients in stratum  $i$  receiving treatment  $j$  and resulting in outcome  $k$ . Here  $j = 2$  corresponds to the active treatment and  $k = 2$  corresponds to positive response. The  $2I$  binomial variables  $\{X_{ij2}\}$  are independent, with parameters  $n_{ij} = X_{ij1} + X_{ij2}$  and  $\pi_{ij2}$  (the success probability). With  $\pi_{ij1} = 1 - \pi_{ij2}$  and  $\boldsymbol{\pi}_{ij} = (\pi_{ij1}; \pi_{ij2})$ , we wish to test the null hypothesis that the distributions  $\boldsymbol{\pi}_{i1}$  and  $\boldsymbol{\pi}_{i2}$  are the same for all  $i$  against the one-sided alternative hypothesis that the second treatment is consistently superior. That is, we test

$$H^* : \pi_{i11} = \pi_{i21}, \forall i, \quad \text{versus} \quad K^* : \exists i, \pi_{i11} \neq \pi_{i21}; \pi_{i11} \geq \pi_{i21}, \forall i. \quad (1)$$

We condition on  $n_{ij}$  and  $T_{ik} = X_{i1k} + X_{i2k}$ , but not  $\sum_{i=1}^I X_{ijk}$ . Each table then has one degree of freedom associated with it. Hence, the sample space  $\Gamma$  can be represented (Berger and Ivanova, 2001) as the set of  $I$ -tuples  $\mathbf{X} = (X_{111}, X_{211}, \dots, X_{I11})$  with row totals  $\mathbf{n} = \mathbf{n}(\mathbf{X}) = (n_{ij})$ , column totals  $\mathbf{T} = \mathbf{T}(\mathbf{X}) = (T_{ik})$ , and constraints

$$\Gamma = \{\mathbf{X} = (X_{111}, \dots, X_{I11}) \mid \max(0, T_{i1} - n_{i2}) \leq X_{i11} \leq \min(n_{i1}, T_{i1}), \forall i\}$$

to ensure that each cell count is nonnegative. For simplicity, we treat the case  $I = 2$ , which is an important problem in its own right (applying, e.g., when one stratifies by gender, or any other binary variable). However, our results apply broadly for all  $I$ . With  $\boldsymbol{\pi} = (\boldsymbol{\pi}_{11}, \boldsymbol{\pi}_{12}; \boldsymbol{\pi}_{21}, \boldsymbol{\pi}_{22})$ , let

$$\lambda_i = \frac{\pi_{i11}\pi_{i22}}{\pi_{i12}\pi_{i21}}, \theta_i = \ln \lambda_i, i \in \{1, 2\}; \boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\pi}) = (\theta_1, \theta_2).$$

The unconditional probability density for row  $j$  of table  $i$  is binomial:

$$P_{\boldsymbol{\pi}}(X_{ij1}, X_{ij2}) = \frac{n_{ij}!}{X_{ij1}!X_{ij2}!} \pi_{ij1}^{X_{ij1}} \pi_{ij2}^{X_{ij2}}.$$

As such, the density for both (independent) rows of table  $i$  is

$$\begin{aligned} P_{\boldsymbol{\pi}}(X_{i11}, X_{i12}, X_{i21}, X_{i22}) &= \frac{n_{i1}!n_{i2}!}{X_{i11}!X_{i12}!X_{i21}!X_{i22}!} \pi_{i11}^{X_{i11}} \pi_{i12}^{n_{i1}-X_{i11}} \pi_{i21}^{T_{i1}-X_{i11}} \pi_{i22}^{n_{i2}-T_{i1}+X_{i11}} \\ &= \frac{n_{i1}!n_{i2}!}{X_{i11}!(n_{i1}-X_{i11})!(T_{i1}-X_{i11})!(n_{i2}-T_{i1}+X_{i11})!} \pi_{i12}^{n_{i1}} \pi_{i21}^{T_{i1}} \pi_{i22}^{n_{i2}-T_{i1}} \exp(\theta_i X_{i11}). \end{aligned}$$

Define

$$\begin{aligned} H_i(X) &= \frac{1}{X!(n_{i1}-X)!(T_{i1}-X)!(n_{i2}-T_{i1}+X)!}, \\ H(\mathbf{X}) &= H_1(X_{111}) \cdot H_2(X_{211}), \\ K(\mathbf{T}, \boldsymbol{\theta}) &= \frac{1}{\sum_{\mathbf{X} \in \Gamma} H(\mathbf{X}) \exp(\boldsymbol{\theta}'\mathbf{X})}. \end{aligned}$$

Then the conditional distribution of  $\mathbf{X}$  given  $\mathbf{T}$ ,

$$P_{\boldsymbol{\pi}}(\mathbf{X} | \mathbf{T}) = \frac{H(\mathbf{X}) \exp(\boldsymbol{\theta}'\mathbf{X})}{\sum_{\mathbf{X} \in \Gamma} H(\mathbf{X}) \exp(\boldsymbol{\theta}'\mathbf{X})} = K(\mathbf{T}, \boldsymbol{\theta}) H(\mathbf{X}) \exp(\boldsymbol{\theta}'\mathbf{X}) = P_{\boldsymbol{\theta}}(\mathbf{X} | \mathbf{T}),$$

depends on  $\boldsymbol{\pi}$  only through  $\boldsymbol{\theta}(\boldsymbol{\pi})$ . Therefore, we formulate the hypothesis testing problem in terms of  $\boldsymbol{\theta}$ , and observe that (1) is equivalent to testing  $H : \boldsymbol{\theta} = \mathbf{0}$  against  $K : \boldsymbol{\theta} \neq \mathbf{0}, \theta_1 \geq 0, \theta_2 \geq 0$ , so that the parameter space is  $\Omega = \{\boldsymbol{\theta} = (\theta_1, \theta_2) \mid \theta_1 \geq 0, \theta_2 \geq 0\}$ . A common odds ratio,  $\theta_1 = \theta_2$ , is often assumed. For example, StatXact (Mehta and Patel, 1999) allows for between-group analysis only when a common odds ratio is either assumed or inferred from failure to reject it as a preliminary hypothesis. In this case, the MH test is most powerful, as we show below. While we find it difficult to justify proceeding as though  $\theta_1 = \theta_2$  just because this null hypothesis was not rejected, we find even less merit in simply assuming that  $\theta_1 = \theta_2$ . As such, we find it instructive to evaluate the MH test in the context of the more general (and applicable) formulation that allows for  $\theta_1 \neq \theta_2$ . The power of test  $\phi : \Gamma \rightarrow [0, 1]$  to detect  $\boldsymbol{\theta} \in \Omega$  is  $E_{\boldsymbol{\theta}}(\phi | \mathbf{T}) = \sum_{\mathbf{X} \in \Gamma} \phi(\mathbf{X}) P_{\boldsymbol{\theta}}\{\mathbf{X} | \mathbf{T}\}$ , with significance level  $\alpha(\phi) = E_{\mathbf{0}}(\phi | \mathbf{T})$ . The most powerful (MP) test to detect  $\boldsymbol{\theta}$  is the likelihood ratio test based on

$$\Lambda_{\boldsymbol{\theta}}(\mathbf{X}) = \frac{P_{\boldsymbol{\theta}}(\mathbf{X} | \mathbf{T})}{P_{\mathbf{0}}(\mathbf{X} | \mathbf{T})} = \frac{K(\mathbf{T}, \boldsymbol{\theta})}{K(\mathbf{T}, \mathbf{0})} \exp(\boldsymbol{\theta}'\mathbf{X}), \quad (2)$$

requiring rejection of  $H$  when  $\theta' \mathbf{X}$  is large. The MH test statistic (Kuritz et al., 1988) is

$$\frac{\left(\sum_{i=1}^2 X_{i11} - E\left(\sum_{i=1}^2 X_{i11}\right)\right)^2}{\sum_{i=1}^2 V(X_{i11})}.$$

But  $E(X_{i11})$  and  $V(X_{i11})$  depend on  $\mathbf{X}$  only through  $\mathbf{n}$  and  $\mathbf{T}$ , and thus are constants within  $\Gamma$ . For a conditional analysis, then, this test statistic is equivalent to  $X_{111} + X_{211}$ , or  $\theta' \mathbf{X}$  when  $\theta_1 = \theta_2$ . That is, the MH test is MP only if we assume a common odds ratio across all tables. Otherwise, the MP test to detect  $\theta$ , or  $l\theta$ , if  $l > 0$ , depends on  $\theta$  through the maximal invariant  $\delta(\theta) = \frac{\theta_1}{\theta_2}$  to the transformation  $f(\theta) = l\theta$ , so there is no uniformly most powerful (UMP) test. We discuss the class of admissible tests in Appendix 1.

### 3 Desirable Properties of Specific Tests

In this section we investigate linear rank tests, as well as analogues of the Smirnov test and the convex hull test. We also study an analogue of Berger's (1998) adaptive test, and we introduce a novel class of polynomial tests that reject for large values of a linear function of the squares of the upper left cell counts. Finally, we consider modified versions of the Smirnov test. Specifically, we study the admissibility of these tests on the basis of the results of the previous section. We also consider another desirable property, specifically proper monotonicity, or directionality. Given the one-sided nature of the testing problem, a test that fails this condition would not be ideal.

**Definition 3.1** A test is properly monotonic if its test statistic is monotonically nondecreasing in  $X_{i,1,1}$  given  $X_{3-i,1,1}$ , for  $i = 1, 2$ .

#### 3.1 Linear rank tests

Linear rank tests require rejection for large values of  $\nu' \mathbf{X}$ , where  $\nu = (\nu_1, \nu_2)$  generally has nonnegative coordinates to ensure proper monotonicity. In fact, the test depends on  $\nu$  only through the ratio  $\nu = \nu_1/\nu_2$ , so we use the notation  $\phi_\nu$ . As in Section 2, the most powerful test to detect the simple (point) alternative  $l\theta = (l\theta_1, l\theta_2)$ ,  $l > 0$ , is  $\phi_\nu$ , where  $\nu = \delta(l\theta) = \theta_1/\theta_2$ . Notice that the MH test is  $\phi_1$ . For a comprehensive treatment of linear rank tests, including their drawbacks, see Ivanova and Berger (2001) and Berger and Ivanova (2002a, 2002b). The following theorem holds.

**Theorem 3.2** *If there are  $I$  tables  $2 \times 2$ , then for any  $\nu > \mathbf{0}$ ,  $\phi_\nu$  is admissible if and only if none of its randomization points is a convex combination of others (which implies that the test cannot randomize on more than  $I$  points of the permutation sample space  $\Gamma$ ). Randomization on no more than  $I$  points is sufficient for  $\phi_\nu$  to be admissible if and only if  $I = 2$ .*

#### 3.2 Convex hull type tests

Convex hull type tests were introduced by Berger (1998). The simplest convex hull test,  $\phi_{CH}$ , is constructed by placing  $D[\Gamma]$ , the set of all directed extreme points of  $\Gamma$ , in the critical region (assuming that  $\alpha \geq P_0\{D[\Gamma]\}$ ), and then iteratively using convex peeling to expand the critical region. That is, a reduced sample space,  $\Gamma_1 = \Gamma \setminus D[\Gamma]$ , is constructed,  $D[\Gamma_1]$  is placed in the critical region, and so on, until  $\alpha$  is spent. Berger (1998) showed the adaptive test,  $\phi_A$ , which uses as a test statistic the smallest  $p$ -value  $\mathbf{X}$  attains among the most powerful tests for all  $\theta \in \Omega$ , to be a convex hull type test. Entire classes of convex hull type tests can be constructed by adding some subset of  $D[\Gamma_m]$  to the rejection region at each step (Cohen and Sackrowitz, 1998), or modifying  $\phi_A$  to maximize power in a preferred direction. The admissibility of all convex hull type tests, including  $\phi_{CH}$  and  $\phi_A$ , follows from Theorems 5.12 and 5.13.

When applied to the analysis of  $2 \times 3$  contingency tables for which they were developed, convex hull tests need not be properly monotonic. However, the following theorem is true for stratified  $2 \times 2$  tables.

**Theorem 3.3** *All convex hull tests for stratified  $2 \times 2$  contingency tables are properly monotonic.*

### 3.3 Polynomial tests

The geometry of the sample space and the results above suggest that using a nonlinear test statistic can lead to a test with a good power profile. Polynomial tests generalize linear rank tests.

**Definition 3.4** The polynomial test,  $\phi_{\nu, k_1 k_2}$ , rejects  $H$  for large values of

$$T_P(\mathbf{X}) = \nu X_{111}^{k_1} + X_{211}^{k_2}. \quad (3)$$

With this definition the following theorem holds.

**Theorem 3.5** *If  $\nu > 0$ ,  $k_1, k_2 > 0$ , then  $\phi_{\nu, k_1 k_2}$  is properly monotonic. If  $\nu > 0$ ,  $k_1, k_2 \geq 1$ , and  $(k_1, k_2) \neq (1, 1)$ , then  $\phi_{\nu, k_1 k_2}$  is admissible.*

### 3.4 Modified Smirnov tests

The Smirnov test rejects  $H$  for large values of  $\max(X_{111}/n_1, X_{211}/n_2)$ , so the boundary of the acceptance region is piecewise linear, and the angle between adjacent pieces is  $\gamma = 90^\circ$ . Its randomization region may have three or more points on one line segment, and, therefore, by Theorem 5.12, this test can be inadmissible. We define the modified Smirnov test  $\phi_s^{MS}$  statistic by  $\max(n_1^{-1}(1, s)\mathbf{X}, n_2^{-1}(s, 1)\mathbf{X}) = \max((X_{111} + sX_{211})/n_1, (X_{211} + sX_{111})/n_2)$ . Now the angle  $\gamma$  between adjacent pieces depends on  $s$ ,  $\gamma(s) = 90^\circ + 2 \tan^{-1}(s)$ . The modified Smirnov test reduces to the Smirnov test when  $s = 0$ , and to  $\phi_{n_2/n_1}$  (or the MH test  $\phi_1$  if  $n_1 = n_2$ ) when  $s = 1$ .

**Theorem 3.6** *If  $0 < s < 1/\max(n_1, n_2)$  or  $0 < s < 1$  and  $s$  is irrational, then  $\phi_s^{MS}$  is admissible. If  $0 < s < 1$ , then  $\phi_s^{MS}$  is properly monotonic.*

## 4 Power Comparisons

We compare the power of a variety of tests for the two pairs of tables,  $(20, 1; 16, 3) + (18, 4; 8, 8)$  and  $(3, 12; 0, 15) + (4, 11; 2, 13)$ , from the two applications we briefly presented in Section 1. The former pair of tables is taken from Kuritz et al. (1988), and the latter has been adapted from data reported by Vigneri et al. (1995). Note that the common Breslow–Day test does not fail to reject the null hypothesis of homogeneity of odds ratios for the latter set of tables, so one would usually proceed with the MH test. Table 3 shows the power of different non-randomized tests under a range of alternatives, with maximum nonrandomized power underlined for linear rank tests, modified Smirnov tests, and other convex hull type tests. The MH test performs poorly in both cases, while convex hull type tests exhibit much better power profiles, as can be seen in Table 4 which shows the shortcomings of the nonrandomized tests as compared to the power of the best randomized tests. The rejection regions for different tests are represented in Figures 1 and 2. Tables 5 and 6 summarize the test comparisons, showing for how many of the considered alternatives each test is superior to every other test. For tables  $(3, 12; 0, 15) + (4, 11; 2, 13)$ , the tests  $\phi_{0.1}$ ,  $\phi_{0.4}$ ,  $\phi_{0.2}^{MS}$ ,  $\phi_{\sqrt{2}/4}^{MS}$ ,  $\phi_{1.2, 2}$ , and  $\phi_A$  coincide, so there are only six different tests to compare.

Note that the MH test is not even best among the linear rank tests, in part because it is so conservative. This is predictable from a study of its ties (that is, the large number of tables to which it assigns the same value of its test statistic). See Ivanova and Berger (2001) for more details. In general, nonlinear rank tests tend to be better omnibus tests than any linear rank test can be (Berger et al., 1998; Berger and Ivanova, 2002a). The Smirnov test is a fairly simple nonlinear rank test, but it tends to be too conservative, and so it is usually less powerful than the modified Smirnov tests we have developed. The adaptive and convex

hull tests are excellent in terms of overall power, but they can be difficult to compute, especially with more than two  $2 \times 2$  tables. The polynomial test is almost as good, and it is quite simple to compute. While it is always difficult to rule on the tradeoff between simplicity and power, we would imagine that for many researchers the polynomial test would represent an ideal compromise between the two.

Our power calculations are all based on the conditional power, and some clarification may be in order. Conditioning on the margins, which are nearly ancillary, allows for an exact test, but there is a more fundamental reason for preferring the conditional power in the comparison of different hypothesis tests. Specifically, the relevant power is the one that applies to the study in question, complete with its margins. There is no consolation in using a test with unacceptable power for the observed set of margins by appealing to the compensation of its exceptional power for other sets of margins that did not occur. See Cox (1958) and Berger (2000) for further discussion of this point. On the other hand, the unconditional power is more useful when planning a study, because prior to the collection of the data one will not know the margins, and, hence, will not know the conditional power. So some studies have computed unconditional power. See, for example, Hilton and Mehta (1993) and Kang and Kim (2004). Unfortunately, the unconditional power is difficult to compute with stratified  $2 \times 2$  tables. Moreover, the time to compute unconditional power is when planning a study, perhaps to select the sample size. We still believe that the conditional power should inform the selection of the specific test, and our purpose is the comparison of tests, as opposed to the selection of a sample size. It is for these reasons that we consider only the conditional power.

## 5 Discussion

Our theoretical results and examples show that the most commonly used analysis for stratified  $2 \times 2$  tables, specifically the Mantel–Haenszel (MH) test, is often inadmissible. Inadmissibility may not by itself constitute sufficient reason not to use a procedure that otherwise has good properties, but in fact the MH test has power that can be substantially worse than that of other tests, some of which are admissible and have uniformly better power than the MH test over all considered alternatives for both examples we treated. The effort of having to switch from a known test to a more powerful novel one is more than offset by the savings that can result from the smaller studies that are allowed by a more powerful test. Moreover, these admissible and more powerful tests are not simply abstractions; many of them can be computed in closed form. We have proposed some such novel admissible and more powerful tests, including  $\phi_{CH}$ ,  $\phi_A$ ,  $\phi_{1,2,2}$ , and  $\phi_{\sqrt{2}/4}^{MS}$ . Each of these tests has uniformly better power than the MH test over all considered alternatives for both examples we treated. This finding is not surprising, because it confirms similar earlier findings (Berger et al., 1998) that the analogues of these tests for the analysis of  $2 \times 3$  contingency tables are more powerful, and potentially much more powerful, than the linear rank tests that are analogous to the MH test.

The primary purpose of this article is to clarify that for stratified  $2 \times 2$  tables, just as for  $2 \times 3$  contingency tables, tremendous increases in power can be realized by switching from the linear rank tests that are frequently used in practice to suitable nonlinear rank tests. It is not our purpose to identify one particular nonlinear rank test as being "best"; indeed, there is no unique best test by the usual standards (i.e., uniformly most powerful). However, we can offer some guidance. If power is the only concern, and computing difficulties are not an issue, then perhaps the convex hull test or one of the adaptive tests might be best. With no prior information, perhaps the convex hull test would be best, and with prior information, an appropriate adaptive test could be selected to ensure especially good power for the favored alternatives. See Berger and Ivanova (2002b) and Berger and Durkalski (2005) for details. If both power and simplicity are important, then a polynomial test might be the ideal analysis, because its test statistic is almost as simple as that of the MH test, requiring only that the cell counts be raised to a power before being summed. In addition, the power of the polynomial test is excellent.

## Acknowledgements

The authors would like to thank two anonymous referees and the associate editor for helpful comments and suggestions.

## Appendix 1: Minimal Complete Class. Directed Extremity

In this Appendix we provide a necessary and sufficient condition for admissibility, describing thus the minimal complete class of tests for the studied problem. The results obtained and the argument used are similar to those of Berger and Sackrowitz (1997), Berger (1998), and Berger et al. (1998). We start with definitions, then a lemma that states a useful property of certain convex sets.

**Definition 5.1** (Set dominance) Let  $B \subset \Gamma$  and  $\mathbf{X} \in \Gamma \setminus B$ . Then  $B$  dominates  $\mathbf{X}$  (denoted  $B \gg \mathbf{X}$ ) if there exists a non-negative weight function  $W$  on  $B$  satisfying  $\sum_{\mathbf{Y} \in B} W(\mathbf{Y}) = 1$  and  $\sum_{\mathbf{Y} \in B} W(\mathbf{Y})\Lambda_{\theta}(\mathbf{Y}) \geq \Lambda_{\theta}(\mathbf{X})$  for all  $\theta \in \Omega$ , strictly for some  $\theta \in \Omega$ .

**Definition 5.2** (Normal cones) The normal cone (in  $\mathbf{X}$ -space) of cone  $V$  in  $\theta$ -space is  $V^{-} \stackrel{def}{=} \{\mathbf{X} \mid \theta' \mathbf{X} \leq 0, \text{ for all } \theta \in V\}$ . Then

$$\Omega^{-} = \{\mathbf{X} : \forall i X_i \leq 0\}. \quad (4)$$

**Definition 5.3** (Directed extreme points) Let  $B \subset \Gamma$ . Then  $\mathbf{X} \in E_B$ , that is,  $\mathbf{X}$  is an extreme point of  $B$ , if  $\mathbf{X} \in B \setminus \text{conv}(B \setminus \mathbf{X})$ . The set of directed extreme points of  $B$  is

$$D[B] \stackrel{def}{=} \{\mathbf{Y} \in B \mid \mathbf{Y} \text{ uniquely maximizes } \Lambda_{\theta}(\mathbf{X}), \text{ or } \theta' \mathbf{X} \text{ by (2), over } B \text{ for some } \theta \in \Omega\}.$$

**Lemma 5.4** Let  $C \supset \Omega^{-}$  be a convex set in  $\mathbf{X}$ -space. If  $\mathbf{Y} \notin C \setminus E_C$ , then there exists  $\theta \in \Omega$  such that  $\theta' \mathbf{Y} > \theta' \mathbf{X}$  for all  $\mathbf{X} \in C \setminus E_C$ . If  $\mathbf{Y} \notin C$ , then  $\theta' \mathbf{Y} > \theta' \mathbf{X}$  for all  $\mathbf{X} \in C$ .

*Proof.* By Theorem 2.14 of Valentine (1964), there exists a hyperplane  $H_{\theta, \mathbf{Y}} = \{\mathbf{X} \mid \theta' \mathbf{X} = \theta' \mathbf{Y}\}$  through  $\mathbf{Y}$  such that either  $C \setminus E_C$  or  $C$  itself, if  $\mathbf{Y} \notin C$ , is entirely on one side of  $H_{\theta, \mathbf{Y}}$ . Hence, the sign of  $\theta' \mathbf{Y} - \theta' \mathbf{X}$  is the same for all  $\mathbf{X} \in C$ . Since  $H_{\theta, \mathbf{Y}} = H_{-\theta, \mathbf{Y}}$ , we can pick  $\theta$  such that  $\theta' \mathbf{Y} > \theta' \mathbf{X}$  for all  $\mathbf{X} \in C$ . It remains to show that such  $\theta$  can be selected from  $\Omega$ . Because  $\Omega = (\Omega^{-})^{-}$ , it suffices (by Definition 5.2) to establish that the selected  $\theta$  satisfies  $\theta' \mathbf{X} \leq 0$  for all  $\mathbf{X} \in \Omega^{-}$ . Assume to the contrary that  $\theta' \mathbf{X} > 0$  for some  $\mathbf{X} \in \Omega^{-}$ . Because  $\Omega^{-}$  is a cone,  $k\mathbf{X} \in \Omega^{-}$  for all  $k > 0$ , and  $\theta' \mathbf{X}$  is unbounded in  $\Omega^{-}$ , hence in  $C$ . This contradicts  $\theta' \mathbf{X} < \theta' \mathbf{Y}$  for all  $\mathbf{X} \in C$ .  $\square$

We denote by  $NE(\mathbf{X}^*)$  the northeast quadrant of  $\mathbf{X}^*$  defined as  $NE(\mathbf{X}^*) = \{\mathbf{X} \mid X_{i11} \geq X_{i11}^* \text{ for } i = 1, 2\}$ . A necessary and sufficient condition for set dominance is given by the following lemma.

**Lemma 5.5** Let  $\mathbf{X}^*, \mathbf{Y} \in \Gamma$ . Then  $\mathbf{Y} \gg \mathbf{X}^*$  if and only if  $\mathbf{Y} \in NE(\mathbf{X}^*) \setminus \{\mathbf{X}^*\}$ . Let  $B \subset \Gamma$  consist of at least two points. If  $\mathbf{X}^* \in \Gamma \setminus B$ , then  $B \gg \mathbf{X}^*$  if and only if  $\text{conv}(B) \cap NE(\mathbf{X}^*) \neq \emptyset$ .

*Proof.* If  $\mathbf{Y} \in NE(\mathbf{X}^*) \setminus \{\mathbf{X}^*\}$ , then for all  $\theta \in \Omega$  we have  $\theta_1 Y_{111} \geq \theta_1 X_{111}^*$  and  $\theta_2 Y_{211} \geq \theta_2 X_{211}^*$ , at least one strictly, so  $\theta' \mathbf{Y} > \theta' \mathbf{X}^*$  and  $\Lambda_{\theta}(\mathbf{Y}) > \Lambda_{\theta}(\mathbf{X}^*)$ . Conversely, if  $\mathbf{Y} \notin NE(\mathbf{X}^*)$ , then either  $Y_{111} < X_{111}^*$  or  $Y_{211} < X_{211}^*$ . Using  $\theta = (1, 0)$  for the first case and  $\theta = (0, 1)$  for the second case shows that  $\Lambda_{\theta}(\mathbf{Y}) < \Lambda_{\theta}(\mathbf{X}^*)$ , and  $\mathbf{Y} \not\gg \mathbf{X}^*$ . Now if  $\text{conv}(B) \cap NE(\mathbf{X}^*)$  is not empty, then there exists a non-negative weight function  $W$  on  $B$  such that  $\mathbf{Y} = \sum_{\mathbf{X} \in B} W(\mathbf{X})\mathbf{X} \in NE(\mathbf{X}^*)$ , and

$$\sum_{\mathbf{X} \in B} W(\mathbf{X}) \exp(\theta' \mathbf{X}) \geq \exp\left(\sum_{\mathbf{X} \in B} W(\mathbf{X})\theta' \mathbf{X}\right) = \exp(\theta' \mathbf{Y}) \geq \exp(\theta' \mathbf{X}^*).$$

If  $\mathbf{Y} \neq \mathbf{X}^*$ , then the second inequality is strict. If  $\mathbf{Y} = \mathbf{X}^*$ , then, because  $\mathbf{X}^* \notin B$ , it must be the case that  $W(\mathbf{X}) > 0$  for more than one  $\mathbf{X} \in B$ , and by Jensen's inequality the first inequality is strict. This

yields  $B \gg \mathbf{X}^*$ . Conversely, if  $\text{conv}(B) \cap NE(\mathbf{X}^*) = \emptyset$ , then because  $\mathbf{X}$ -space is discrete, there exists an open neighborhood  $U$  of  $\text{conv}(B)$  satisfying  $U \cap NE(\mathbf{X}^*) = \emptyset$ . Theorem 2.9 of Valentine (1964) establishes the existence of a hyperplane,  $H_\theta = \{\mathbf{X} | \theta' \mathbf{X} = c\}$ , separating  $U$  and  $NE(\mathbf{X}^*)$ , so that  $\theta' \mathbf{Y} < c$  for all  $\mathbf{Y} \in U$ , and  $\theta' \mathbf{Y} \geq c$  for all  $\mathbf{Y} \in NE(\mathbf{X}^*)$ . An argument similar to that in Lemma 5.4, but this time based on the fact that  $NE(\mathbf{X}^*)$  is unbounded, shows that  $\theta \in \Omega$ . Because  $\mathbf{X}^* \in NE(\mathbf{X}^*)$ , we have  $\theta' \mathbf{X}^* \geq c$ , and hence

$$\sum_{\mathbf{Y} \in B} W(\mathbf{Y}) \exp(\theta' \mathbf{Y}) < \exp(\theta' \mathbf{X}^*)$$

for any weight function  $W$ . The last inequality implies that  $B \gg \mathbf{X}^*$ .  $\square$

Notice that unlike Diestel (2001), we do not require  $B$  to contain a single element  $\mathbf{Y}$  such that  $\mathbf{Y} \gg \mathbf{X}^*$  in order for  $B \gg \mathbf{X}^*$ . For example, if  $B = \{(4, 0), (0, 4)\}$  and  $\mathbf{X}^* = (1, 1)$ , then  $B \gg \mathbf{X}^*$ .

**Definition 5.6** (Property A) Test  $\phi$  satisfies Property A if the condition  $B \gg \mathbf{X}$  implies that according to  $\phi$ , either  $\mathbf{X}$  is “empty”,  $\phi(\mathbf{X}) = 0$ , or  $B$  is “full”,  $\max_{\mathbf{Y} \in B} \phi(\mathbf{Y}) = 1$ .

Property A tests form a complete class, as stated by the next theorem.

**Theorem 5.7** *If the test  $\phi$  is admissible for testing  $H$  versus  $K$ , then it satisfies Property A.*

*Proof.* Assume that  $\phi$  does not satisfy Property A. For some pair  $\{B, \mathbf{X}^* \in \Gamma \setminus B\}$ , then,  $B \gg \mathbf{X}^*$ ,  $\phi(\mathbf{X}^*) > 0$ , and  $\phi(\mathbf{Y}) < 1$  for all  $\mathbf{Y} \in B$ . Using the  $W$  required by  $B \gg \mathbf{X}^*$ , define  $H$  and, for  $0 < h \leq H$ ,  $\phi_h^*$  as follows:

$$H = \min \left( P_0(\mathbf{X}^* | \mathbf{T}) \phi(\mathbf{X}^*), \min_{\mathbf{Y} \in B} \frac{P_0(\mathbf{Y} | \mathbf{T}) (1 - \phi(\mathbf{Y}))}{W(\mathbf{Y})} \right),$$

$$\phi_h^*(\mathbf{X}) = \begin{cases} \phi(\mathbf{X}^*) - h/P_0(\mathbf{X}^* | \mathbf{T}), & \text{for } \mathbf{X} = \mathbf{X}^* \\ \phi(\mathbf{X}) + hW(\mathbf{X})/P_0(\mathbf{X} | \mathbf{T}), & \text{for } \mathbf{X} \in B, \\ \phi(\mathbf{X}), & \text{for } \mathbf{X} \in \Gamma \setminus (B \cup \{\mathbf{X}^*\}). \end{cases}$$

Then  $\phi_h^*$  is a test, satisfying  $0 \leq \phi_h^* \leq 1$ , and has size  $\alpha = E_0(\phi | \mathbf{T})$ , because

$$E_0[\phi_h^* - \phi | \mathbf{T}] = \sum_{\mathbf{Y} \in B} \frac{hW(\mathbf{Y})P_0(\mathbf{Y} | \mathbf{T})}{P_0(\mathbf{Y} | \mathbf{T})} - \frac{hP_0(\mathbf{X}^* | \mathbf{T})}{P_0(\mathbf{X}^* | \mathbf{T})} = 0.$$

Moreover, if  $\theta \in \Omega \setminus \mathbf{0}$ , then

$$E_\theta[\phi_h^* - \phi | \mathbf{T}] = \sum_{\mathbf{Y} \in B} \frac{hW(\mathbf{Y})P_\theta(\mathbf{Y} | \mathbf{T})}{P_0(\mathbf{Y} | \mathbf{T})} - \frac{hP_\theta(\mathbf{X}^* | \mathbf{T})}{P_0(\mathbf{X}^* | \mathbf{T})} = h \left[ \sum_{\mathbf{Y} \in B} W(\mathbf{Y}) \Lambda_\theta(\mathbf{Y}) - \Lambda_\theta(\mathbf{X}^*) \right] \geq 0,$$

strictly for some  $\theta \in \Omega \setminus \mathbf{0}$ , so  $\phi_h^*$  is uniformly more powerful than  $\phi$ . Hence,  $\phi$  is not admissible.  $\square$

We shall further show that this Property A complete class is the minimal complete class, so that if  $\phi$  is an inadmissible test, then  $\phi$  does not satisfy Property A. To this end, let us first denote  $A_{\phi, \phi'} \stackrel{\text{def}}{=} \{\mathbf{X} \in \Gamma | \phi(\mathbf{X}) > \phi'(\mathbf{X})\}$  for any two tests  $\phi$  and  $\phi'$ . Notice that if  $\phi'$  is a size  $\alpha$  test uniformly more powerful than  $\phi$ , which is of the same size  $\alpha$ , then clearly neither  $A_{\phi, \phi'}$  nor  $A_{\phi', \phi}$  is empty. Also, if  $W(\mathbf{X}) \stackrel{\text{def}}{=} [\phi'(\mathbf{X}) - \phi(\mathbf{X})]P_0(\mathbf{X})$ , then  $W(\mathbf{X}) > 0$  for  $\mathbf{X} \in A_{\phi', \phi}$ ,  $W(\mathbf{X}) < 0$  for  $\mathbf{X} \in A_{\phi, \phi'}$ , and  $W(\mathbf{X}) = 0$  for  $\mathbf{X} \in \Gamma \setminus (A_{\phi, \phi'} \cup A_{\phi', \phi})$ . Let

$$W_-(\mathbf{X}) \stackrel{\text{def}}{=} \frac{W(\mathbf{X})}{\sum_{\mathbf{Y} \in A_{\phi, \phi'}} W(\mathbf{Y})} \quad \text{and} \quad W_+(\mathbf{X}) \stackrel{\text{def}}{=} \frac{W(\mathbf{X})}{\sum_{\mathbf{Y} \in A_{\phi', \phi}} W(\mathbf{Y})}.$$

Then  $W_-(\mathbf{X}) > 0$  for  $\mathbf{X} \in A_{\phi, \phi'}$ ,  $W_+(\mathbf{X}) > 0$  for  $\mathbf{X} \in A_{\phi', \phi}$ , and

$$\sum_{\mathbf{Y} \in A_{\phi', \phi}} W_+(\mathbf{Y}) = \sum_{\mathbf{X} \in A_{\phi, \phi'}} W_-(\mathbf{X}) = 1.$$

So  $W_-$  and  $W_+$  are weight functions on  $A_{\phi, \phi'}$  and  $A_{\phi', \phi}$ , respectively.

**Lemma 5.8** For any inadmissible  $\phi$  and dominating  $\phi'$ , and for all  $\theta \in \Omega$ ,

$$\sum_{\mathbf{Y} \in A_{\phi', \phi}} W_+(\mathbf{Y}) \Lambda_{\theta}(\mathbf{Y}) \geq \sum_{\mathbf{X} \in A_{\phi, \phi'}} W_-(\mathbf{X}) \Lambda_{\theta}(\mathbf{X}). \quad (5)$$

*Proof.* Let  $\theta \in \Omega$ , and let  $\phi'$  dominate  $\phi$ . By definition,

$$\sum_{\mathbf{Y} \in A_{\phi', \phi}} W_+(\mathbf{Y}) \Lambda_{\theta}(\mathbf{Y}) = \frac{\sum_{\mathbf{Y} \in A_{\phi', \phi}} (\phi'(\mathbf{Y}) - \phi(\mathbf{Y})) P_{\theta}(\mathbf{Y})}{\sum_{\mathbf{Y} \in A_{\phi', \phi}} W(\mathbf{Y})} = \frac{\sum_{\mathbf{Y} \in A_{\phi', \phi}} (\phi'(\mathbf{Y}) - \phi(\mathbf{Y})) P_{\theta}(\mathbf{Y})}{\sum_{\mathbf{Y} \in A_{\phi', \phi}} (\phi'(\mathbf{Y}) - \phi(\mathbf{Y})) P_0(\mathbf{Y})}.$$

Because  $E_0[\phi'|\mathbf{T}] = E_0[\phi|\mathbf{T}]$ , the denominator is equal to  $\sum_{\mathbf{X} \in A_{\phi, \phi'}} (\phi(\mathbf{X}) - \phi'(\mathbf{X})) P_0(\mathbf{X})$ . Because  $E_{\theta}[\phi'|\mathbf{T}] \geq E_{\theta}[\phi|\mathbf{T}]$ ,  $\sum_{\mathbf{X} \in A_{\phi, \phi'}} (\phi(\mathbf{X}) - \phi'(\mathbf{X})) P_{\theta}(\mathbf{X})$  can never exceed the numerator. Hence,

$$\sum_{\mathbf{Y} \in A_{\phi', \phi}} W_+(\mathbf{Y}) \Lambda_{\theta}(\mathbf{Y}) \geq \frac{\sum_{\mathbf{X} \in A_{\phi, \phi'}} (\phi(\mathbf{X}) - \phi'(\mathbf{X})) P_{\theta}(\mathbf{X})}{\sum_{\mathbf{X} \in A_{\phi, \phi'}} (\phi(\mathbf{X}) - \phi'(\mathbf{X})) P_0(\mathbf{X})} = \sum_{\mathbf{X} \in A_{\phi, \phi'}} W_-(\mathbf{X}) \Lambda_{\theta}(\mathbf{X}). \quad \square$$

**Lemma 5.9** For any inadmissible  $\phi$  and dominating  $\phi'$ ,  $A_{\phi, \phi'} \subset \mathbf{conv}(A_{\phi', \phi} \cup \Omega^-)$ .

*Proof.* For  $\mathbf{X}^* \in \Gamma$ ,  $\mathbf{conv}(\mathbf{X}^* \cup \Omega^-) = SW(\mathbf{X}^*) = \{\mathbf{X} | X_{i11} \leq X^*_{i11}, \text{ for } i = 1, 2\}$  is the Southwest quadrant of  $\mathbf{X}^*$ . If  $\mathbf{X}^* \in A_{\phi, \phi'} \setminus \mathbf{conv}(A_{\phi', \phi} \cup \Omega^-)$ , then by Lemma 5.4 there exists  $\theta_1 \in \Omega$  such that  $\theta_1' \mathbf{X}^* > \theta_1' \mathbf{Y}$  for all  $\mathbf{Y} \in \mathbf{conv}(A_{\phi', \phi} \cup \Omega^-)$ , in particular, for all  $\mathbf{Y} \in A_{\phi', \phi}$ . Now for  $\lambda > 0$  we have

$$\begin{aligned} \frac{\sum_{\mathbf{Y} \in A_{\phi', \phi}} W_+(\mathbf{Y}) \exp(\lambda \theta_1' \mathbf{Y})}{\sum_{\mathbf{X} \in A_{\phi, \phi'}} W_-(\mathbf{X}) \exp(\lambda \theta_1' \mathbf{X})} &\leq \frac{\sum_{\mathbf{Y} \in A_{\phi', \phi}} W_+(\mathbf{Y}) \exp(\lambda \theta_1' \mathbf{Y})}{W_-(\mathbf{X}^*) \exp(\lambda \theta_1' \mathbf{X}^*)} \\ &= \sum_{\mathbf{Y} \in A_{\phi', \phi}} \frac{W_+(\mathbf{Y})}{W_-(\mathbf{X}^*)} \exp[\lambda \theta_1' (\mathbf{Y} - \mathbf{X}^*)]. \end{aligned}$$

The right hand side tends to 0 as  $\lambda$  tends to  $+\infty$ , yet Lemma 5.8, along with (2), establish that the left hand side is at least one. The contradiction leads to the conclusion that  $A_{\phi, \phi'} \subset \mathbf{conv}(A_{\phi', \phi} \cup \Omega^-)$ .  $\square$

With these observations we can now state the following theorem which shows that the Property A complete class is the minimal complete class.

**Theorem 5.10** If test  $\phi$  satisfies Property A, then it is admissible.

*Proof.* Let  $\phi'$  dominate  $\phi$ . If  $\mathbf{0} \in A_{\phi, \phi'}$ , then  $\phi(\mathbf{0}) > 0$ . By Lemma 5.5,  $\mathbf{0}$  is dominated by any  $\mathbf{Y} \in \Gamma$ , yet  $\phi(\mathbf{Y}) < 1$  for  $\mathbf{Y} \in A_{\phi', \phi}$ . This demonstrates the failure of Property A for  $\phi$ . If  $\mathbf{0} \notin A_{\phi, \phi'}$ , then select any  $\mathbf{X}^* \in A_{\phi, \phi'}$ . By Lemma 5.9,  $\mathbf{X}^* \in \mathbf{conv}(A_{\phi', \phi} \cup \Omega^-)$ , and we may write

$$\mathbf{X}^* = \sum_{\mathbf{Y} \in B \cup B^-} W''(\mathbf{Y}) \mathbf{Y},$$

where  $B \subset A_{\phi', \phi}$ ,  $B^- \subset \Omega^-$ , and  $W''(\cdot)$  is a positive weight function on  $B \cup B^-$ . Note that  $\Gamma \cap \Omega^- \setminus \{\mathbf{0}\} = \emptyset$ , so  $\mathbf{X}^* \notin \Omega^-$ , and therefore cannot be a convex combination of vectors from  $\Omega^-$  only. Hence  $B \neq \emptyset$ . Let  $\theta \in \Omega$ . Then  $\theta' \mathbf{Y} \leq 0$  for all  $\mathbf{Y} \in \Omega^-$ , in particular, for all  $\mathbf{Y} \in B^- \subset \Omega^-$ . Therefore

$$\theta' \mathbf{X}^* = \sum_{\mathbf{Y} \in B} W''(\mathbf{Y}) \theta' \mathbf{Y} + \sum_{\mathbf{Y} \in B^-} W''(\mathbf{Y}) \theta' \mathbf{Y} \leq \sum_{\mathbf{Y} \in B} W''(\mathbf{Y}) \theta' \mathbf{Y} \leq \sum_{\mathbf{Y} \in B} W'(\mathbf{Y}) \theta' \mathbf{Y},$$

where for  $\mathbf{Y} \in B$

$$W'(\mathbf{Y}) \stackrel{\text{def}}{=} \frac{W''(\mathbf{Y})}{\sum_{\mathbf{Y} \in B} W''(\mathbf{Y})} \geq W''(\mathbf{Y}).$$

Since  $\exp(\cdot)$  is a convex increasing function, Jensen's inequality yields

$$\sum_{\mathbf{Y} \in B} W'(\mathbf{Y}) \exp(\theta' \mathbf{Y}) \geq \exp\left(\sum_{\mathbf{Y} \in B} W'(\mathbf{Y}) \theta' \mathbf{Y}\right) \geq \exp(\theta' \mathbf{X}^*).$$

Because  $\mathbf{X}^* \notin B$ , there are two possibilities. Either  $B = \{\mathbf{Y}\}$ , in which case the second inequality above is strict (because  $\mathbf{X}^* \neq \mathbf{Y}$ ), or  $B$  consists of multiple points, in which case the first inequality above is strict. Either way,  $B \gg \mathbf{X}^*$ . Now  $\phi(\mathbf{Y}) < 1$  for all  $\mathbf{Y} \in A_{\phi', \phi}$ , in particular, for all  $\mathbf{Y} \in B$ . This, in conjunction with  $\phi(\mathbf{X}^*) > 0$ , demonstrates that Property A fails for  $\phi$ .  $\square$

It follows from Theorem 4.1 of Ledwina (1978) that  $\phi(\mathbf{X})$  is admissible if and only if there exists a closed convex set  $A^*(\phi) \subset \mathbb{R}^I$  satisfying  $\Omega^- \subset A^*(\phi)$  and the following two conditions:

Ledwina's  $\alpha$ .  $A^*(\phi)^c \cap \Gamma = \{\mathbf{X} \mid \phi(\mathbf{X}) = 1\}$ , where  $A^*(\phi)^c$  is the complement of  $A^*(\phi)$ .

Ledwina's  $\gamma$ .  $\{\mathbf{X} \mid 0 < \phi(\mathbf{X}) < 1\} \subset E_{A^*(\phi)} \cap \Gamma$ .

If  $\mathbf{X} \in [A^*(\phi) \setminus E_{A^*(\phi)}] \cap \Gamma$ , then  $\mathbf{X} \notin A^*(\phi)^c$ , and, by Ledwina's  $\alpha$ ,  $\phi(\mathbf{X}) < 1$ . Also,  $\mathbf{X}$  is not an extreme point of  $A^*(\phi)$ , so by Ledwina's  $\gamma$  it cannot be true that  $0 < \phi(\mathbf{X}) < 1$ . Therefore,  $\phi(\mathbf{X}) = 0$ . This is stronger than Ledwina's  $\beta$ ,  $\text{Int}(A^*(\phi)) \cap \Gamma \subset \{\mathbf{X} \mid \phi(\mathbf{X}) = 0\}$ , because  $\text{Int}(A^*(\phi)) \subset A^*(\phi) \setminus E_{A^*(\phi)}$ , but they are not necessarily equal. Another way to characterize admissible tests is based on directed extremity (Definition 5.3). First we formulate a lemma regarding directed extreme points.

**Lemma 5.11** *If  $A \subset B$  and  $\mathbf{Y} \in D[B]$ , then  $\mathbf{Y} \in D[\{\mathbf{Y}\} \cup A]$ .*

*Proof.* If  $\theta' \mathbf{Y} \geq \theta' \mathbf{X}$  for all  $\mathbf{X} \in B$ , then this is true for all  $\mathbf{X} \in A$ .  $\square$

Then the following two theorems hold, analogous to Theorems 3.1 and 3.3 of Berger (1998) for  $2 \times J$  contingency tables.

**Theorem 5.12** *Let  $A(\phi) = \{\mathbf{X} \in \Gamma \mid \phi(\mathbf{X}) < 1\}$  be the acceptance region of  $\phi$ . Then test  $\phi$  is admissible if and only if  $\phi(\mathbf{Y}) > 0$  implies  $\mathbf{Y} \in D[\{\mathbf{Y}\} \cup A(\phi)]$ .*

*Proof.* If  $\phi$  is admissible and  $\phi(\mathbf{Y}) > 0$ , then it follows from Ledwina's  $\alpha$  and  $\gamma$  that  $\mathbf{Y} \notin A^*(\phi) \setminus E_{A^*(\phi)}$ . But  $\Omega^- \subset A^*(\phi)$ , so by Lemma 5.4  $\mathbf{Y} \in D[\{\mathbf{Y}\} \cup A^*(\phi)]$ . Then  $A(\phi) = A^*(\phi) \cap \Gamma \subset A^*(\phi)$ , so by Lemma 5.11  $\mathbf{Y} \in D[\{\mathbf{Y}\} \cup A(\phi)]$ . Conversely, suppose that  $\phi(\mathbf{Y}) > 0$  implies that  $\mathbf{Y} \in D[\{\mathbf{Y}\} \cup A(\phi)]$ . Suppose also that  $\phi(\mathbf{Y}) > 0$ , and  $\max_{\mathbf{X} \in B} \phi(\mathbf{X}) < 1$ . Then  $B \subset A(\phi)$ , and, by Lemma 5.11,  $\mathbf{Y} \in D[\{\mathbf{Y}\} \cup B]$ . But  $\mathbf{Y}$  uniquely maximizing  $\Lambda_{\theta}(\mathbf{X})$  over  $D[\{\mathbf{Y}\} \cup B]$  implies that  $B \gg \mathbf{Y}$  cannot be true. Therefore, Property A is satisfied and by Theorem 5.10  $\phi$  is admissible.  $\square$

**Theorem 5.13** *Let  $E_0(\phi_1) = \alpha_1 < \alpha_2 = E_0(\phi_2)$ ,  $\phi_1$  be admissible at level  $\alpha_1$ , and  $\phi_2(\mathbf{X}) \geq \phi_1(\mathbf{X})$  for all  $\mathbf{X} \in \Gamma$ . If  $\phi_2(\mathbf{X}) > \phi_1(\mathbf{X}) \Rightarrow \mathbf{X} \in D[\mathbf{X} \cup A(\phi_1)]$ , then  $\phi_2$  is admissible at level  $\alpha_2$ .*

*Proof.* The proof of Theorem 3.3 of Berger (1998) applies verbatim.  $\square$

## Appendix 2: Proofs of Results in Section 3

**Proof of Theorem 3.2** The acceptance region for  $\phi_\nu$  is  $A(\phi_\nu) = \{\mathbf{X} \in \Gamma \mid \nu' \mathbf{X} \leq \mathbf{a}\}$ , with randomization nowhere except possibly on  $\partial A(\phi_\nu) \stackrel{def}{=} \{\mathbf{X} \in \Gamma \mid \nu' \mathbf{X} = \mathbf{a}\}$ . If one of the points from  $\partial A(\phi_\nu)$  is a convex combination of the others, then it is not a directed extreme point, so if randomization takes place on this point, then the necessary condition for admissibility given by Theorem 5.12 is not satisfied. Conversely, if none of the randomization points is a convex combination of the others, then they are all directed extreme points, uniquely maximizing  $\theta' \mathbf{X}$  for some  $\theta \in \Omega$ .  $\square$

**Proof of Theorem 3.3** For any  $A \subset \Gamma$ , if  $(X_{111}, X_{211}) \in D[A]$  and  $(X_{111}^*, X_{211}^*) \in A \setminus D[A]$ , then either  $X_{111} > X_{111}^*$  or  $X_{211} > X_{211}^*$  (or both). Clearly, then,  $X_{3-i11} = X_{3-i11}^* \Rightarrow X_{i11} > X_{i11}^*$ ,  $i = 1, 2$ .  $\square$

**Proof of Theorem 3.5** The proper monotonicity follows from (3). On the contour  $\nu x_1^{k_1} + x_2^{k_2} = c$ , one can solve for  $x_2$  as a function of  $x_1$ . Implicit differentiation shows that the second derivative of this function is

$$x_2''(x_1) = -\frac{\nu k_1 x_1^{k_1-2}}{k_2^2 x_2^{2k_2-1}} \left( k_2(k_1 - 1)x_2^{k_2} + \nu k_1(k_2 - 1)x_1^{k_1} \right).$$

So  $x_2''(x_1) = 0$  if  $(k_1, k_2) = (1, 1)$ , but  $x_2''(x_1) < 0$  if  $x_1, x_2 > 0$ ,  $k_1, k_2 \geq 1$ , and  $(k_1, k_2) \neq (1, 1)$ . Therefore,  $x_2(x_1)$  is strictly concave, and all possible randomization points  $\{\mathbf{X} \in \Gamma \mid T_P(\mathbf{X}) = c\}$  are directed extreme points of  $A(\phi) = \{\mathbf{X} \in \Gamma \mid T_P(\mathbf{X}) \leq c\}$ . Admissibility follows from Theorem 5.12.  $\square$

**Proof of Theorem 3.6** The boundary of  $A^*(\phi_s)$  consists of two straight lines,  $sX_{111} + X_{211} = n_2c(s)$  and  $X_{111} + sX_{211} = n_1c(s)$ . If  $0 < s < 1/\max(n_1, n_2)$  or  $0 < s < 1$  and  $s$  is irrational, then neither line intersects  $\Gamma$  at more than one point. If  $0 < s < 1$ , then the acceptance region is convex, and  $\phi_s^{MS}$  is properly monotonic. So, if  $\mathbf{X} \in \Gamma \cap (A^*(\phi_s^{MS}) \setminus \text{Int}(A^*(\phi_s^{MS})))$ , then  $\mathbf{X} \in E_{A^*(\phi_s^{MS})}$ . Ledwina's conditions  $\alpha$  and  $\gamma$  are thus satisfied and  $\phi_s^{MS}$  is admissible.  $\square$

## References

- Berger, V.W. (1998). Admissibility of exact conditional tests of stochastic order. *Journal of Statistical Planning and Inference* **66**, 39–50.
- Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine* **19**, 1319–1328.
- Berger, V.W. (2004). Valid adjustment for binary covariates of randomized binary comparisons. *Biometrical Journal* **46**, 589–594.
- Berger, V.W. (2005). Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials (with discussion). *Biometrical Journal* **47**, 119–127.
- Berger, V.W. and Durkalski, V.L. (2005). Analysis of trichotomous pharmaceutical endpoints. *Journal of Biopharmaceutical Statistics* **15**, 739–745.
- Berger, V.W. and Ivanova, A. (2001). Permutation tests for Phase III clinical trials. In *Applied Statistics in the Pharmaceutical Industry With Case Studies Using S-Plus*. S.P. Millard and A. Krause, eds. Springer, New York.
- Berger, V.W. and Ivanova, A. (2002a). Bias of linear rank tests for stochastic order in ordered categorical data. *Journal of Statistical Planning and Inference* **107**, 237–247.
- Berger, V.W. and Ivanova, A. (2002b). Adaptive tests for ordinal data. *Journal of Modern Applied Statistical Methods* **1**, 269–280.
- Berger, V.W., Permutt, T. and Ivanova, A. (1998). Convex hull set for ordered categorical data. *Biometrics* **54**, 1541–1550.

- Berger, V.W. and Sackrowitz, H. (1997). Improving tests for superior treatment in contingency tables. *Journal of the American Statistical Association* **92**, 700–705.
- Bishop, Y.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press: Cambridge, MA.
- Brand, R. and Kragt, H. (1992). Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* **11**, 2077–2082.
- Cohen, A. and Sacrowitz, H. (1998). Directional tests for one-sided alternatives in multivariate models. *Annals of Statistics* **26**, 2321–2338.
- Cox, D.R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics* **29**, 357–372.
- Diestel, R. (2001). Relating subsets of a poset, and a partition theorem for WQOS. *ORDER* **18**, 275–279.
- Hartung, J. and Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* **20**, 3875–3889.
- Hilton, J. and Mehta, C.R. (1993). Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics* **49**, 609–616.
- Ivanova, A. and Berger, V.W. (2001). Drawbacks to integer scoring for ordered categorical data. *Biometrics* **57**, 567–570.
- Kang, S.H. and Kim, S. (2004). A comparison of the three conditional exact tests in two-way contingency tables using the unconditional exact power. *Biometrical Journal* **46**, 320–330.
- Kernan, W.N., Viscoli, C.M., Makuch, R.W., Brass, L.M., and Horwitz, R.I. (1999). Stratified randomization for clinical trials. *Journal of Clinical Epidemiology* **52**, 19–26.
- Kuritz, S.J., Landis, J.R. and Koch, G.G. (1988). A general overview of Mantel–Haenszel methods. *Annual Review of Public Health* **9**, 123–160.
- Ledwina, T. (1978). Admissible tests for exponential families with finite support. *Statistics* **1**, 105–118.
- Lehmann, E.L. (1993). The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association* **88**, 1242–1249.
- Little, R.J.A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician* **43**, 283–288.
- Lloyd, C.J. (1988). Some issues arising from the analysis of  $2 \times 2$  contingency tables. *Australian Journal of Statistics* **30**, 35–46.
- Matts, J.P. and Lachin, J.M. (1988). Properties of permuted-block randomization in clinical trials. *Controlled Clinical Trials* **9**, 327–344.
- Mehta, C. and Patel, N. (1999). *StatXact for Windows: User Manual*. CYTEL Software Corporation, Cambridge, MA.
- Mehta, C.R. and Hilton, J.F. (1993). Exact power of conditional and unconditional tests: Going beyond the  $2 \times 2$  contingency table. *The American Statistician* **47**, 91–98.
- Miller, R.G. (1980). Combining  $2 \times 2$  contingency tables. In *Biostatistics Casebook*, R. Miller, B. Efron, B. Brown and L. Moses, eds., John Wiley and Sons, New York.
- Routledge, R.D. (1992). Resolving the conflict over Fisher’s exact test. *The Canadian Journal of Statistics* **20**, 201–209.
- Upton, G.J.G. (1992). Fisher’s exact test. *Journal of the Royal Statistical Society* **155**, 395–402.
- Valentine, A.V. (1964). *Convex Sets*. McGraw–Hill: New York.
- Vigneri, S., Termini, R., Leandro, G., Badalamenti, S., Pantalena, M., Savarino, V., Di Mario, F., Battaglia, G., Mela, G.S., Pilotto, A., Plebani, M. and Davi, G. (1995). Comparison of five maintenance therapies for reflux esophagitis. *New England Journal of Medicine* **333**, 1106–1110.

**Table 1** Distribution of total number of hours of at least some post-partum pain by treatments and strata. Adapted from Kuritz et al. (1988).

Initial pain	Treatment	Hours with at least some post-partum pain	
		0–4	5–8
Some	A&B	20	1
	Placebo	16	3
Lots	A&B	18	4
	Placebo	8	8

**Table 2** Recurrence of endoscopic signs at 12 months in two treatment groups, according to the initial grade of esophagitis. Adapted from Vigneri et al. (1995).

Initial grade	Treatment	Number of patients	
		Recurrence	No recurrence
Grade 1	Cisapride	3	12
	Omeprazole	0	15
Grade 2	Cisapride	4	11
	Omeprazole	2	13

**Table 3** Conditional non-randomized power calculations and envelope exact (randomized) power for the tables (20, 1; 16, 3) and (18, 4; 8, 8) from Kuritz et al. (1988), and tables (3, 12; 0, 15) and (4, 11; 2, 13) from Vigneri et al. (1995). Nominal size  $\alpha = 0.05$  (one-sided, i.e., reject for large values of the test statistic). Columns 5–9 are linear rank tests, columns 10–12 are piecewise linear tests, and columns 13–15 are nonlinear tests. The maximal power in each class of tests is underlined.

$\delta(\theta) = \frac{\theta_1}{\theta_2}$	$\theta_1$	$\theta_2$	$\phi_{\delta(\theta)}$	$\phi_{0.1}$	$\phi_{0.4}$	$\phi_1$	$\phi_{2.5}$	$\phi_{10}$	$\phi^S$	$\phi_{0.02}^{MS}$	$\phi_{\sqrt{2}/4}^{MS}$	$\phi_{1,2,2}$	$\phi_{CH}$	$\phi_A$
Tables (20, 1; 16, 3) and (18, 4; 8, 8)														
1	0.0	0.0	0.050	0.047	0.044	0.019	0.037	0.044	0.043	0.049	0.042	0.028	0.040	0.030
$\infty$	1.0	0.0	0.235	0.067	0.066	0.054	0.161	<u>0.223</u>	0.221	<u>0.226</u>	0.163	0.102	<u>0.163</u>	0.103
$\infty$	2.0	0.0	0.540	0.101	0.101	0.097	0.358	<u>0.530</u>	0.528	<u>0.531</u>	0.359	0.214	<u>0.359</u>	0.214
0	0.0	1.0	0.415	<u>0.407</u>	0.391	0.190	0.144	0.078	0.080	0.187	<u>0.246</u>	0.199	0.205	<u>0.234</u>
1	1.0	1.0	0.512	<u>0.456</u>	0.455	0.363	0.392	0.288	0.251	0.344	<u>0.438</u>	0.408	<u>0.429</u>	0.417
2	2.0	1.0	0.716	0.541	0.541	0.517	<u>0.662</u>	0.587	0.546	0.602	<u>0.674</u>	0.624	<u>0.673</u>	0.625
0	0.0	2.0	0.880	<u>0.877</u>	0.862	0.593	0.372	0.187	0.371	0.645	<u>0.715</u>	0.594	0.610	<u>0.695</u>
0.5	1.0	2.0	0.902	<u>0.894</u>	0.893	0.790	0.687	0.497	0.488	0.726	<u>0.822</u>	0.796	0.798	<u>0.819</u>
1	2.0	2.0	0.944	<u>0.924</u>	<u>0.924</u>	0.899	0.885	0.771	0.689	0.835	<u>0.918</u>	0.913	0.915	<u>0.915</u>
mean power			0.577	<u>0.479</u>	0.475	0.391	0.411	0.356	0.357	0.461	<u>0.486</u>	0.431	<u>0.466</u>	0.450
observed p-value				0.009	0.008	0.004	0.018	0.044	0.043	0.049	0.022	0.008	0.018	0.009
Tables (3, 12; 0, 15) and (4, 11; 2, 13)														
1	0.0	0.0	0.050	<u>0.046</u>	<u>0.046</u>	0.013	0.040	0.036	0.008	<u>0.046</u>	<u>0.046</u>	<u>0.046</u>	0.043	<u>0.046</u>
$\infty$	1.0	0.0	0.160	0.070	0.070	0.034	<u>0.121</u>	0.117	0.008	<u>0.070</u>	<u>0.070</u>	0.070	<u>0.122</u>	0.070
$\infty$	2.0	0.0	0.291	0.081	0.081	0.058	<u>0.215</u>	0.212	0.008	<u>0.081</u>	<u>0.081</u>	0.081	<u>0.215</u>	0.081
0	0.0	1.0	0.260	<u>0.245</u>	<u>0.245</u>	0.080	0.120	0.085	0.092	<u>0.245</u>	<u>0.245</u>	<u>0.245</u>	0.156	<u>0.245</u>
1	1.0	1.0	0.361	<u>0.341</u>	<u>0.341</u>	0.185	0.313	0.271	0.092	<u>0.341</u>	<u>0.341</u>	<u>0.341</u>	0.328	<u>0.341</u>
2	2.0	1.0	0.553	0.385	0.385	0.288	<u>0.521</u>	0.492	0.092	<u>0.385</u>	<u>0.385</u>	0.385	<u>0.524</u>	0.385
0	0.0	2.0	0.584	<u>0.564</u>	<u>0.564</u>	0.223	0.244	0.108	0.350	<u>0.564</u>	<u>0.564</u>	<u>0.564</u>	0.379	<u>0.564</u>
0.5	1.0	2.0	0.707	<u>0.698</u>	<u>0.698</u>	0.438	0.504	0.344	0.350	<u>0.698</u>	<u>0.698</u>	<u>0.698</u>	0.563	<u>0.698</u>
1	2.0	2.0	0.778	<u>0.759</u>	<u>0.759</u>	0.613	0.733	0.626	0.350	<u>0.759</u>	<u>0.759</u>	<u>0.759</u>	0.747	<u>0.759</u>
mean power			0.416	<u>0.354</u>	<u>0.354</u>	0.215	0.312	0.255	0.150	<u>0.354</u>	<u>0.354</u>	<u>0.354</u>	0.342	<u>0.354</u>
observed p-value				0.084	0.076	0.013	0.009	0.009	0.084	0.084	0.076	0.076	0.013	0.046

**Table 4** Shortcomings of the above tests for tables (20, 1; 16, 3) and (18, 4; 8, 8) from Kuritz et al. (1988), and tables (3, 12; 0, 15) and (4, 11; 2, 13) from Vigneri et al. (1995), i.e., the difference between the envelope (randomized) test power and the power of the test in question. Columns 4–8 are linear rank tests, columns 9–11 are piecewise linear tests, and columns 12–14 are nonlinear tests.

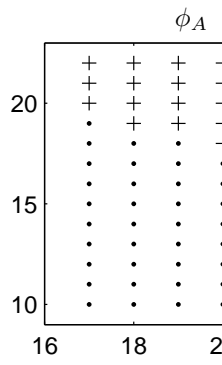
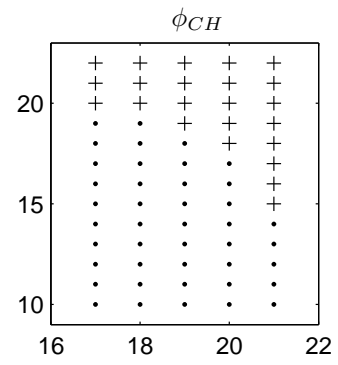
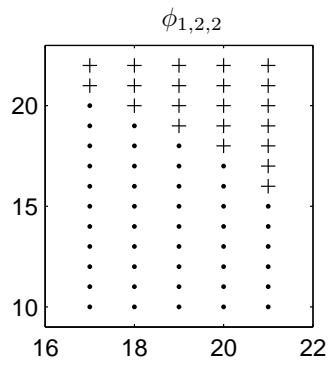
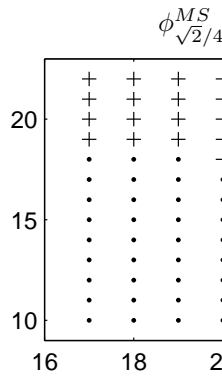
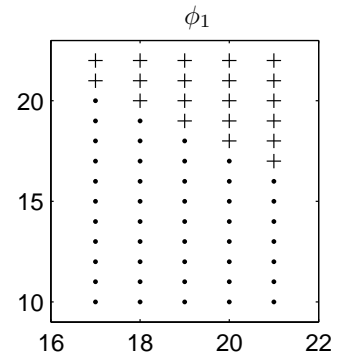
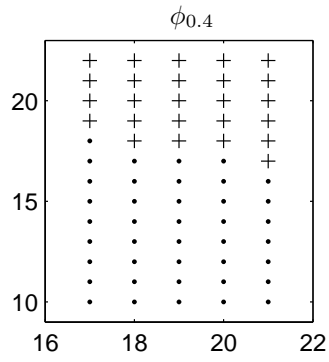
$\delta(\theta) = \frac{\theta_1}{\theta_2}$	$\theta_1$	$\theta_2$	$\phi_{0.1}$	$\phi_{0.4}$	$\phi_1$	$\phi_{2.5}$	$\phi_{10}$	$\phi^S$	$\phi_{0.02}^{MS}$	$\phi_{\sqrt{2}/4}^{MS}$	$\phi_{1,2,2}$	$\phi_{CH}$	$\phi_A$
Tables (20, 1; 16, 3) and (18, 4; 8, 8)													
1	0.0	0.0	0.003	0.006	0.031	0.013	0.006	0.007	0.001	0.008	0.022	0.010	0.020
$\infty$	1.0	0.0	0.168	0.169	0.181	0.074	0.011	0.014	0.009	0.071	0.133	0.072	0.132
$\infty$	2.0	0.0	0.439	0.439	0.443	0.182	0.010	0.012	0.009	0.181	0.326	0.181	0.326
0	0.0	1.0	0.008	0.024	0.225	0.271	0.337	0.335	0.228	0.169	0.216	0.210	0.181
1	1.0	1.0	0.055	0.057	0.148	0.119	0.223	0.261	0.168	0.074	0.104	0.083	0.095
2	2.0	1.0	0.174	0.174	0.199	0.054	0.129	0.170	0.113	0.042	0.092	0.043	0.091
0	0.0	2.0	0.003	0.017	0.287	0.508	0.693	0.509	0.235	0.165	0.286	0.270	0.185
0.5	1.0	2.0	0.008	0.009	0.112	0.215	0.405	0.414	0.176	0.080	0.106	0.104	0.083
1	2.0	2.0	0.021	0.021	0.045	0.060	0.174	0.255	0.110	0.027	0.032	0.029	0.029
average			0.098	0.102	0.186	0.166	0.221	0.220	0.117	0.091	0.146	0.111	0.127
maximum			0.439	0.439	0.443	0.508	0.693	0.509	0.235	0.181	0.326	0.270	0.326
Tables (3, 12; 0, 15) and (4, 11; 2, 13)													
1	0.0	0.0	0.004	0.004	0.037	0.010	0.014	0.042	0.004	0.004	0.004	0.004	0.004
$\infty$	1.0	0.0	0.090	0.090	0.126	0.039	0.043	0.151	0.090	0.090	0.090	0.038	0.090
$\infty$	2.0	0.0	0.210	0.210	0.233	0.076	0.079	0.282	0.210	0.210	0.210	0.076	0.210
0	0.0	1.0	0.015	0.015	0.179	0.139	0.175	0.167	0.015	0.015	0.015	0.103	0.015
1	1.0	1.0	0.020	0.020	0.177	0.048	0.091	0.269	0.020	0.020	0.020	0.033	0.020
2	2.0	1.0	0.168	0.168	0.266	0.032	0.061	0.461	0.168	0.168	0.168	0.029	0.168
0	0.0	2.0	0.020	0.020	0.361	0.341	0.476	0.234	0.020	0.020	0.020	0.205	0.020
0.5	1.0	2.0	0.009	0.009	0.268	0.203	0.362	0.356	0.009	0.009	0.009	0.144	0.009
1	2.0	2.0	0.020	0.020	0.165	0.045	0.152	0.428	0.020	0.020	0.020	0.031	0.020
average			0.062	0.062	0.201	0.104	0.161	0.266	0.062	0.062	0.062	0.074	0.062
maximum			0.210	0.210	0.361	0.341	0.476	0.461	0.210	0.210	0.210	0.205	0.210

**Table 5** Pairwise comparisons of 11 tests for 36 values of  $\theta$ , where each entry is the number of parameter values (out of 36 considered in the power calculations) for which the test to the left (defining the row) had greater power than the test above (defining the column) for tables (20, 1; 16, 3) and (18, 4; 8, 8).

	$\phi_{0.1}$	$\phi_{0.4}$	$\phi_1$	$\phi_{2.5}$	$\phi_{10}$	$\phi^S$	$\phi_{0.02}^{MS}$	$\phi_{\sqrt{2}/4}^{MS}$	$\phi_{1,2,2}$	$\phi_{CH}$	$\phi_A$
$\phi_{0.1}$	-	<b>36</b>	<b>36</b>	22	25	25	24	21	21	21	21
$\phi_{0.4}$	<b>0</b>	-	<b>36</b>	22	25	25	24	21	21	21	21
$\phi_1$	<b>0</b>	<b>0</b>	-	17	23	23	17	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>
$\phi_{2.5}$	14	14	19	-	28	27	16	<b>0</b>	15	<b>0</b>	15
$\phi_{10}$	11	11	13	8	-	28	<b>0</b>	8	11	8	11
$\phi^S$	11	11	13	9	8	-	<b>0</b>	8	9	8	9
$\phi_{0.02}^{MS}$	12	12	19	20	<b>36</b>	<b>36</b>	-	8	16	12	11
$\phi_{\sqrt{2}/4}^{MS}$	15	15	<b>36</b>	<b>36</b>	28	28	28	-	<b>36</b>	<b>36</b>	<b>36</b>
$\phi_{1,2,2}$	15	15	<b>36</b>	21	25	27	20	<b>0</b>	-	<b>0</b>	<b>0</b>
$\phi_{CH}$	15	15	<b>36</b>	<b>36</b>	28	28	24	<b>0</b>	<b>36</b>	-	19
$\phi_A$	15	15	<b>36</b>	21	25	27	25	<b>0</b>	<b>36</b>	17	-
Total	108	144	280	212	251	274	178	66	201	123	143
Dominate	2	1	0	0	0	0	2	5	1	3	2
Dominated by	0	1	6	2	1	1	0	0	3	1	1

**Table 6** Pairwise comparisons of 6 different tests for 36 values of  $\theta$ , where each entry is the number of parameter values (out of 36 considered in the power calculations) for which the test to the left (defining the row) had greater power than the test above (defining the column) for tables (3, 12, 0, 15) and (4, 11, 2, 13).

	$\phi_A$	$\phi_1$	$\phi_{2.5}$	$\phi_{10}$	$\phi^S$	$\phi_{CH}$
$\phi_A$	-	<b>36</b>	21	23	<b>36</b>	21
$\phi_1$	<b>0</b>	-	<b>0</b>	12	30	<b>0</b>
$\phi_{2.5}$	15	<b>36</b>	-	<b>36</b>	32	<b>0</b>
$\phi_{10}$	13	24	<b>0</b>	-	26	<b>0</b>
$\phi^S$	<b>0</b>	6	4	10	-	1
$\phi_{CH}$	15	<b>36</b>	<b>36</b>	<b>36</b>	35	-
Total	43	138	61	117	159	22
Dominate	2	0	2	0	0	3
Dominated by	0	3	1	2	1	0



**Fig. 2** Rejection regions for different size  $\alpha = 0.05$  tests. Tables  $(3, 12; 0, 15)$  and  $(4, 11; 2, 13)$ .