Journal of Modern Applied Statistical Methods

Volume 1 | Issue 2 Article 36

11-1-2002

# Adaptive Tests for Ordered Categorical Data

Vance W. Berger
*Biometry Research Group, National Cancer Institute*

Anastasia Ivanova
*University of North Carolina*

Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 269-280


1538 – 9472/02/$30.00

# Adaptive Tests for Ordered Categorical Data

Vance W. Berger
Biometry Research Group
National Cancer Institute

Anastasia Ivanova
Department of Biostatistics
University of North Carolina

Consider testing for independence against stochastic order in an ordered $2xJ$ contingency table, under product multinomial sampling. In applications one may wish to exploit prior information concerning the direction of the treatment effect, yet ultimately end up with a testing procedure with good frequentist properties. As such, a reasonable objective may be to simultaneously maximize power at a specified alternative and ensure reasonable power for all other alternatives of interest. For this objective, none of the available testing approaches are completely satisfactory. A new class of admissible adaptive tests is derived. Each test in this class strictly preserves the Type I error rate and strikes a balance between good global power and nearly optimal (envelope) power to detect a specific alternative of most interest. Prior knowledge of the direction of the treatment effect, the level of confidence in this prior information, and possibly the marginal totals might be used to select a specific test from this class.

Key words: Contingency table; exact conditional test; linear rank test; omnibus test; permutation test.

## Introduction

When comparing two treatments on the basis of an ordinal endpoint, the data can be summarized as a $2xJ$ contingency table. The objective tumor response data, e.g., from 35 ovarian cancer patients treated with cisplatin-based combination chemotherapy and salvage platinum-based therapy (Chiara et al., 1993) are (4,7,2,2) and (1,6,7,6) for patients with treatment-free intervals ≤ 12 months and > 12 months, respectively, with categories for 'progressive disease', 'stable disease', 'partial response', and 'complete response'. Combining the two 'non-response' categories, as is common, yields counts $\boldsymbol{C_1}$ = (11,2,2) and $\boldsymbol{C_2}$ = (7,7,6) in the two groups. For simplicity, the case $J$ = 3 is treated, but with modification the results apply more generally. It is common in practice to dispense with the specification of the alternative hypothesis, and proceed directly to the analysis.

Vance W. Berger is Mathematical Statistician at the NCI and Adjunct Professor at University of Maryland Baltimore County. E-mail: vb78c@nih.gov. Anastasia Ivanova is Assistant Professor, Dept. of Biostatistics, School of Public Health., University of North Carolina – Chapel Hill. E-mail: aivanova@bios.unc.edu.

This failure to make the specific alternative hypothesis explicit is unfortunate, because it should serve as the basis for selecting and evaluating the analysis. Linear rank tests, based on assigning numerical scores to the categories, are the most powerful tests to detect point alternatives. If one wishes to test for the superiority of one treatment to another, then stochastic order serves as a reasonable (composite) alternative hypothesis (Cohen and Sackrowitz, 1998). Unless the margins satisfy pathological conditions, there is no uniformly most powerful test or monotone likelihood ratio. When testing for stochastic order, nonlinear rank tests, including the Smirnov, improved (Berger and Sackrowitz, 1997), convex hull (Berger, Permutt, and Ivanova, 1998; henceforth BPI), and *COM*(L) Fisher tests, tend to have better overall power profiles than linear rank tests do.

Berger's (1998) adaptive nonlinear rank test can be generalized to provide an entire class of exact, admissible, adaptive nonlinear rank tests, each of which balances omnibus power for any stochastically ordered alternative against optimal power to detect a specific alternative of greatest interest. The margins may be used to suggest the selection of one particular test from this novel class of tests. The exact conditional powers of some of the aforementioned tests are compared.

Notation and Formulation

Consider product multinomial sampling, with $n_1$ and $n_2$ (each fixed by the design) patients treated with the control and active treatments, respectively. The vectors of cell probabilities (each summing to one) are $\boldsymbol{\pi}_1=(\pi_{11},\pi_{12},\pi_{13})$ and $\boldsymbol{\pi}_2=(\pi_{21},\pi_{22},\pi_{23})$, respectively, and the corresponding trinomial random vectors are $\boldsymbol{C}_1 = (C_{11},C_{12},C_{13})$ and $\boldsymbol{C}_2 = (C_{21},C_{22},C_{23})$, with $n_i = C_{i1} + C_{i2} + C_{i3}$, $i = 1, 2$. The log odds ratios, $\theta_1$ and $\theta_2$, are calculated from $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ as

$$\theta_1 = \log\{(\pi_{11}\pi_{23})/(\pi_{21}\pi_{13})\} \text{ and}$$

$$\theta_2 = \log\{(\pi_{12}\pi_{23})/(\pi_{22}\pi_{13})\}.$$

Let $T_j = C_{1j} + C_{2j}$, $j = 1,2,3$. Conditional on $\boldsymbol{T} = (T_1,T_2,T_3)$, the sample space $\Gamma$ is the set of $2 \times 3$ contingency tables with nonnegative integer cell counts, and row and column totals $\boldsymbol{n} = (n_1,n_2)$ and $\boldsymbol{T}$, respectively. Given $\boldsymbol{T}$, $\boldsymbol{n}$, and $\boldsymbol{c} = (C_{11},C_{12})$, the entire $2 \times 3$ contingency table can be reconstructed as $C_{13} = n_1 - C_{11} - C_{12}$ and $\boldsymbol{C}_2 = \boldsymbol{T} - \boldsymbol{C}_1$. Thus, $\boldsymbol{c}$ suffices to denote a point of $\Gamma$.

Figure 1. The permutation sample space $\Gamma$ for the data set {(11,2,2);(7,7,6)}, with $\boldsymbol{n}$=(15,20) and $\boldsymbol{T}$=(18,9,8).

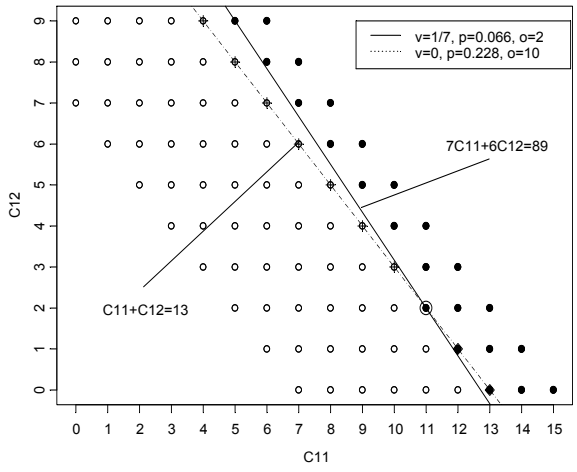


Figure 1 displays $C_{12}$ plotted against $C_{11}$ for all 87 tables of $\Gamma$ for the example, {(11,2,2);(7,7,6)}, with observed table (11,2) circled. With $K(\boldsymbol{T};\boldsymbol{\theta})=1/\sum_{\boldsymbol{c}\in\Gamma} H(\boldsymbol{c})\exp[\boldsymbol{\theta}'\boldsymbol{c}]$, $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, $\boldsymbol{\pi}=(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2)$, and $H(\boldsymbol{c})=n_1!n_2!/\Pi_{i=1}^{2}\Pi_{j=1}^{3} C_{ij}!$, the density follows the exponential family:

$$P_{\pi}\{c|T\} = P_{\boldsymbol{\theta}}\{\boldsymbol{c}|\boldsymbol{T}\} = K(\boldsymbol{T};\boldsymbol{\theta})H(\boldsymbol{c})\exp[\boldsymbol{\theta}'\boldsymbol{c}]. \quad (2.1)$$

Let $\Delta_1 = \pi_{11} - \pi_{21}$, and $\Delta_2 = (\pi_{11} + \pi_{12}) - (\pi_{21} + \pi_{22}) = \pi_{23} - \pi_{13}$. If $\Delta_1 \geq 0$, and $\Delta_2 \geq 0$, at least one strictly, then the active treatment is objectively superior to the control. One may wish to test $H$: $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ against the one-sided alternative hypothesis that the active response distribution is stochastically larger than the control response distribution, $H_A'$ : $\Delta_1 \geq 0$, $\Delta_2 \geq 0$, $\boldsymbol{\pi}_1 \neq \boldsymbol{\pi}_2$. As will be explained, this is not actually possible with a conditional test. By (2.1), $P_{\pi}\{\boldsymbol{c}|\boldsymbol{T}\}$ depends on $\boldsymbol{\pi}$ only through $\boldsymbol{\theta}(\boldsymbol{\pi})$, so if $\boldsymbol{\theta}(\boldsymbol{\pi}) = \boldsymbol{\theta}(\boldsymbol{\pi}^*)$, then $\boldsymbol{c}$ offers no information with which to distinguish $\boldsymbol{\pi}$ from $\boldsymbol{\pi}^*$. To be identifiable, then, the hypotheses must be formulated in terms of $\boldsymbol{\theta}$ (Berger, 1998).

The null hypothesis $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ is equivalent to $H$: $\boldsymbol{\theta}(\boldsymbol{\pi}) = \boldsymbol{0}$, but unless $0 \leq \theta_2 \leq \theta_1$, $\boldsymbol{\theta}(\boldsymbol{\pi})$ provides insufficient information with which to determine if $\boldsymbol{\pi}$ satisfies $H_A'$ because no conditional alternative hypothesis is equivalent to $H'_A$. Note, e.g., that {(3,3,4)/10;(2,4,4)/10} satisfies $H_A'$ and {(21,51,328)/400; (7,34,164)/205} does not, yet $\boldsymbol{\theta}$ = (log(3/2),log(3/4)) for both. The conditional power to detect $\boldsymbol{\pi}$ depends on $\boldsymbol{\theta}(\boldsymbol{\pi})$ only, so no conditional test that preserves the $\alpha$-level whenever $H'_A$ does not hold can be globally powerful whenever it does hold.

However, if $\boldsymbol{\pi}$ satisfies $H'_A$, then $\theta_1(\boldsymbol{\pi}) > 0$; and if $\theta_1 > 0$, then for any $\theta_2$ there exists (Berger and Sackrowitz, 1997) $\boldsymbol{\pi}$ satisfying $H'_A$ such that $\boldsymbol{\theta}(\boldsymbol{\pi}) = (\theta_1,\theta_2)$. As such, $\theta_1$ is the key parameter; the active treatment is superior on $\Omega_A = \{\boldsymbol{\theta} | \theta_1 >$

0}, no different on $\Omega_0 = \{\boldsymbol{\theta} \mid \theta_1 = 0\}$, and inferior on $\Omega_C = \{\boldsymbol{\theta} \mid \theta_1 < 0\}$. It is reasonable, then, to test $H$ against $H_A : \theta_1 > 0$. The large unconditional indifference region, where neither group stochastically dominates the other, has, by conditioning, been absorbed into $\Omega_0 \cup \Omega_A \cup \Omega_C$.

Let $\delta(\boldsymbol{\theta}) = 1 - \theta_2/\theta_1$ be the *direction* of the effect. As $\theta_1$ increases in both $\Delta_1$ and $\Delta_2$, while $\theta_2$ ($\theta_1 - \theta_2$) increases in $\Delta_2$ ($\Delta_1$), and decreases in $\Delta_1$ ($\Delta_2$), the superiority of the active treatment to the control is due primarily to a shift from the middle to the best outcome ($\Delta_2 > \Delta_1$) if $\delta(\boldsymbol{\theta})$ is small, or from the worst to the middle outcome ($\Delta_1 > \Delta_2$) if $\delta(\boldsymbol{\theta})$ is large. Let $\Omega_v = \{\boldsymbol{\theta} \mid \theta_1 > 0, \delta(\theta) = v\}$. As $\delta(\boldsymbol{\theta})$ is generally unknown *a priori*, omnibus tests that are sensitive to departures from $H_0$ in each direction of $\Omega_A = \cup_{v \in \Re^1} \Omega_v$ are preferred to tests that lack this desirable property.

If the $\varphi$ rejection region $R_\alpha(\varphi)$ contains $D[\Gamma]$, the set of directed extreme points of $\Gamma$ (BPI, 1998), then $\varphi$ is omnibus. The challenge is to exploit prior information about $\delta(\boldsymbol{\theta})$ to construct omnibus tests with especially good power in one preferred direction, $\Omega_v$. For reasons articulated by Berger (2000) and Berger et al. (2002), we consider only exact conditional tests in this formulation.

## A New Look at Linear Rank Tests

Linear rank tests are based on numerical scores $(v_1, v_2, v_3)$, $v_1 < v_3$, assigned to the three outcome levels. With $v = (v_2 - v_1)/(v_3 - v_1)$, $\varphi_v$ uses test statistic $z_v(\boldsymbol{c}) = C_{11} + (1 - v)C_{12}$. New notation allows for greater insight into linear rank tests. Let $M_v(\boldsymbol{c}) = \{\boldsymbol{c}^* \in \Gamma \mid z_v(\boldsymbol{c}^*) \geq z_v(\boldsymbol{c})\}$ be the $\varphi_v$ extreme region of $\boldsymbol{c}$, with boundary $B_v(\boldsymbol{c})$ and p-value $p_v(\boldsymbol{c}) = P_{\boldsymbol{0}}\{M_v(\boldsymbol{c}) | \boldsymbol{T}\}$. The level set (Frick, 2000, p. 719) of $z_v(\boldsymbol{c})$ is $B_v(\boldsymbol{c}) \cap \Gamma$, with $o_v(\boldsymbol{c})$ its order, or the number of points of $B_v(\boldsymbol{c}) \cap \Gamma$. If $\boldsymbol{c} = (C_{11}, C_{12}) \in \Gamma$ and $\boldsymbol{c}^* = (C^*_{11}, C^*_{12}) \in \Gamma - \boldsymbol{c}$, then $z_v(\boldsymbol{c}^*) = z_v(\boldsymbol{c})$ if and only if $v = 1 - (C_{11} - C^*_{11})/(C^*_{12} - C_{12})$, say $v = v_{\boldsymbol{c},\boldsymbol{c}^*}$ (vector valued for $J > 3$). Let $V(\boldsymbol{c}) = \{v_1(\boldsymbol{c}), v_2(\boldsymbol{c}), \ldots, v_{K_c}(\boldsymbol{c})\}$ be the ordered set $\{v_{\boldsymbol{c},\boldsymbol{c}^*} \mid |v_{\boldsymbol{c},\boldsymbol{c}^*}| < \infty, \boldsymbol{c}^* \in \Gamma - \boldsymbol{c}\}$, and let $v_0(\boldsymbol{c}) = -\infty$ and $v_{K_c+1}(\boldsymbol{c}) = \infty$. For finite $v$, $o_v(\boldsymbol{c}) > 1$ if and only if $v \in V(\boldsymbol{c})$.

Let $\varepsilon(\boldsymbol{c}) = \min_k[v_{k+1}(\boldsymbol{c}) - v_k(\boldsymbol{c})]/2$, $z^\perp_v(\boldsymbol{c}) = C_{12} + (v - 1)C_{11}$, $B^+_v(\boldsymbol{c}) = \{\boldsymbol{c}^* \in B_v(\boldsymbol{c}) \cap \Gamma \mid z^\perp_v(\boldsymbol{c}^*) > z^\perp_v(\boldsymbol{c})\}$, $B^-_v(\boldsymbol{c}) = \{\boldsymbol{c}^* \in B_v(\boldsymbol{c}) \cap \Gamma \mid z^\perp_v(\boldsymbol{c}^*) < z^\perp_v(\boldsymbol{c})\}$, $v^*(\boldsymbol{c}) = \{v^* \mid p_{v^*}(\boldsymbol{c}) \leq p_v(\boldsymbol{c})$ for all $v\}$.

By Lemma 1 (in the Appendix), $v^*(\boldsymbol{c})$ consists of the scores that minimize not just $p_v(\boldsymbol{c})$ but also $p_{\min(v)}(\boldsymbol{c}) = \min(\lim_{u \Downarrow v} p_u(\boldsymbol{c}), \lim_{u \Uparrow v} p_u(\boldsymbol{c})) = p_v(\boldsymbol{c}) - \max(P_0\{B^-_v(\boldsymbol{c})\}, P_0\{B^+_v(\boldsymbol{c})\})$. Hence, $p_{\min(v)}(\boldsymbol{c})$, which also equals $\min\{p_{v-\varepsilon(\boldsymbol{c})}(\boldsymbol{c}), p_{v+\varepsilon(\boldsymbol{c})}(\boldsymbol{c})\}$, is a true p-value. As $\Gamma$ has finitely many subsets, there can be only a finite number of values for $p_v(\boldsymbol{c})$, so the minimum p-value is attained, and $v^*(\boldsymbol{c}) \neq \varnothing$. If $v \in V(\boldsymbol{c})$, then $o_v(\boldsymbol{c}) > 1$, $B^-_v(\boldsymbol{c}) \cup B^+_v(\boldsymbol{c}) \neq \varnothing$, $p_{\min(v)}(\boldsymbol{c}) < p_v(\boldsymbol{c})$, and $v \notin v^*(\boldsymbol{c})$. Hence, $v^*(\boldsymbol{c}) \cap V(\boldsymbol{c}) = \varnothing$, and, by Lemma 1, $v^*(\boldsymbol{c})$ consists of one or more open intervals of the form $(v_k(\boldsymbol{c}), v_{k+1}(\boldsymbol{c}))$. For $\{(11,2,2);(7,7,6)\}$, $\boldsymbol{c} = (11,2)$, $K_{\boldsymbol{c}} = 42$, $\varepsilon(11,2) = 1/84$, and $V(\boldsymbol{c}) =$ {-6, -5, -4, -3, -5/2, -2, -5/3, -3/2, -4/3, -5/4, -6/5, -1, -5/6, -4/5, -3/4, -2/3, -3/5, -4/7, -1/2, -3/7, -2/5, -1/3, -2/7, -1/4, -1/5, -1/6, -1/7, 0, 1/7, 1/6, 1/5, ¼, 2/7, 1/3, 2/5, ½, 2/3, 1, 3/2, 2, 5/2, 3, 4, 5, 6}.

Figure 1 shows $M_{1/7}(11,2)$ by dark dots and $M_0(11,2)$-$M_{1/7}(11,2)$ by crosses. Because (11,2) minimizes $z^\perp_{1/7}(11,2) = 7C_{12} - 6C_{11}$ over $B_{1/7}(11,2) \cap \Gamma$ (Table 1), $B^-_{1/7}(11,2) = \varnothing$ and $p_{1/7}(11,2) = \lim_{u \Uparrow 1/7} p_u(11,2) = 0.066$. Also $p_v(11,2) = 0.020$ for $v \in (1.0, 1.5) = v^*(11,2)$. If $v$

$\in V(11,2)$, then $P_0\{B_v^-\} \le P_0\{B_v^+\}$ for $v > 1.5$, and $P_0\{B_v^-\} \ge P_0\{B_v^+\}$ for $v < 1.0$. The optimality of most powerful (MP) test $\varphi_{\delta(\theta)}$ to detect $l\theta$, for $l>0$ (BPI, 1998), is offset by its potentially poor power on $\Omega_A$ - $\Omega_{\delta(\theta)}$. In fact, $D[\Gamma]$ may not be contained in the $\varphi_v$ critical region $R_\alpha(\varphi_v)$ for any ν, so for

Table 1. All possible linear rank tests with scores $(0,v,1)$, with middle score $v\in[0,2]$, for the data set {(11,2,2);(7,7,6)}, along with the number of points in its level set, the endpoints and null probabilities of each segment of its level set, and various p-values. (null probabilities of various extreme regions).

| $v$ | $o_v(11,2)$ | Endpoints of: $B_v^+$ | Endpoints of: $B_v^-$ | $p_v$ (minimum is underlined) | $p_v^-$ | $p_v^+$ | $P_0\{B_v^+\}$ | $P_0\{B_v^-\}$ | $p_{v,\infty}$ | $M_v$-$M_{v,\infty}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $v \in (-1/7,0)$ | 1 | | | 0.2262 | 0.2262 | 0.2262 | | | 0.2262 | |
| $v = 0$ | 10 | (4,9) -(10,3) | (12,1) -(13,0) | 0.2277 | 0.2262 | <u>0.0661</u> | 0.1615 | 0.0015 | 0.0726 | (7,6)-(10,3) |
| $v \in (0,1/7)$ | 1 | | | 0.0661 | 0.0661 | 0.0661 | | | 0.0661 | |
| $v = 1/7$ | 2 | (5,9) | | 0.0661 | 0.0661 | <u>0.0661</u> | $2.1*10^{-5}$ | | 0.0661 | |
| $v \in (1/7,1/6)$ | 1 | | | 0.0661 | 0.0661 | 0.0661 | | | 0.0661 | |
| $v = 1/6$ | 2 | (6,8) | | 0.0661 | 0.0661 | <u>0.0657</u> | 0.0004 | | 0.0661 | |
| $v \in (1/6,1/5)$ | 1 | | | 0.0657 | 0.0657 | 0.0657 | | | 0.0657 | |
| $v = 1/5$ | 2 | (7,7) | | 0.0657 | 0.0657 | <u>0.0629</u> | 0.0028 | | 0.0657 | |
| $v \in (1/5,1/4)$ | 1 | | | 0.0629 | 0.0629 | 0.0629 | | | 0.0629 | |
| $v = 1/4$ | 2 | (8,6) | | 0.0629 | 0.0629 | <u>0.0538</u> | 0.0091 | | 0.0629 | |
| $v \in (1/4,2/7)$ | 1 | | | 0.0538 | 0.0538 | 0.0538 | | | 0.0538 | |
| $v = 2/7$ | 2 | (6,9) | | 0.0538 | 0.0538 | <u>0.0538</u> | $5.7*10^{-6}$ | | 0.0538 | |
| $v \in (2/7,1/3)$ | 1 | | | 0.0538 | 0.0538 | 0.0538 | | | 0.0538 | |
| $v = 1/3$ | 3 | (7,8) -(9,5) | | 0.0538 | 0.0538 | <u>0.0387</u> | 0.0152 | | 0.0387 | (9,5) |
| $v \in (1/3,2/5)$ | 1 | | | 0.0387 | 0.0387 | 0.0387 | | | 0.0387 | |
| $v = 2/5$ | 2 | (8,7) | | 0.0387 | 0.0387 | <u>0.0382</u> | 0.0005 | | 0.0387 | |
| $v \in (2/5,1/2)$ | 1 | | | 0.0382 | 0.0382 | 0.0382 | | | 0.0382 | |
| $v = 1/2$ | 4 | (9,6) -(10,4) | (12,0) | 0.0385 | 0.0382 | <u>0.0237</u> | 0.0148 | 0.0003 | 0.0249 | (10,4) |
| $v \in (1/2,2/3)$ | 1 | | | 0.0237 | 0.0237 | 0.0237 | | | 0.0237 | |
| $v = 2/3$ | 2 | (10,5) | | 0.0237 | 0.0237 | <u>0.0220</u> | 0.0017 | | 0.0237 | |
| $v \in (2/3,1)$ | 1 | | | 0.0220 | 0.0220 | 0.0220 | | | 0.0220 | |
| $v = 1$ | 5 | (11,4) -(11,3) | (11,1) -(11,0) | 0.0276 | 0.0220 | <u>**0.0198**</u> | 0.0078 | 0.0056 | 0.0276 | |
| $v \in (1,3/2)$ | 1 | | | **0.0198** | **0.0198** | **0.0198** | | | 0.0198 | |
| $v = 3/2$ | 2 | | (10,0) | 0.0205 | <u>**0.0198**</u> | 0.0205 | | 0.0008 | 0.0205 | |
| $v \in (3/2,2)$ | 1 | | | 0.0205 | 0.0205 | 0.0205 | | | 0.0205 | |
| $v = 2$ | 4 | (12,3) | (10,1) -(9,0) | 0.0294 | <u>0.0205</u> | 0.0289 | 0.0005 | 0.0089 | 0.0294 | |
| $v \in (2,5/2)$ | 1 | | | 0.0289 | 0.0289 | 0.0289 | | | 0.0289 | |

Note that all the values are calculated at the outcome (11,2); $p_{v,\infty}$ and $M_{v,\infty}$ are the p-value and extreme region, respectively, of the adaptive test based on ν and $\tau = \infty$.

each ν there will exist $\theta \in \Omega_A$ for which the power of $\varphi_\nu$ to detect $l\theta$ tends to zero as $l$ gets large (BPI, 1998). Podgor, Gastwirth, and Mehta (1996) proposed the maximin efficiency robust test (MERT) in hopes of providing better power than linear rank tests. Ironically, the MERT is itself a linear rank test; its rejection region may also fail to contain $D[\Gamma]$, leading to poor power on parts of $\Omega_A$ and no power in the limit in some directions. Berger and Ivanova (2002) showed that at certain α-levels the most stringent linear rank test is $\varphi_{\nu_S}$, where $\nu_S$ is such that the two points of $D[\Gamma]$ that are furthest (in Euclidean distance) from each other are equated by $z_{\nu_S}(\boldsymbol{c})$. For {(11,2,2),(7,7,6)}, this gives $\nu_S = 0$, because Γ has two directed extreme points, $D[\Gamma]$={(15,0);(6,9)}, and $z_0(15,0) = 15+(1-0)(0)=15=6+(1-0)(9)= z_0(6,9)$.

Nonlinear Rank Tests

By allowing the boundary of $R_\alpha(\varphi)$ to curve, nonlinear rank tests often require smaller α-levels to ensure that $D[\Gamma] \subset R_\alpha(\varphi)$ than linear rank tests would. However, this is not always the case. Berger and Ivanova (2002) provide an example in which the proportional odds and proportional hazards tests (McCullagh, 1980) are not nonlinear enough to be omnibus at reasonable α-levels. The Smirnov test, $\varphi_S$, uses as the test statistic the largest of three quantities, 0, $D_1 = C_{11}/n_1 - C_{21}/n_2$, and $D_2 = (C_{11} + C_{12})/n_1 - (C_{21} + C_{22})/n_2$. Among tests routinely available in standard statistical software packages ($\varphi_S$ is a standard feature of StatXact), $\varphi_S$ minimizes the α-level required for its rejection region to contain $D[\Gamma]$. However, $\varphi_S$ is not generally admissible (Berger, 1998).

Permutt and Berger (2000) and Ivanova and Berger (2001) each proposed refinements of $\varphi_S$ that break its ties. Although such refinements are necessarily uniformly more powerful than $\varphi_S$ (Rohmel and Mansmann, 1999, p. 158), the term "improvement of φ" is reserved for a test whose exact (possibly randomized) version is uniformly more powerful than the exact (possibly randomized) version of φ. By this definition, refinements are rarely improvements. Berger and Sackrowitz (1997) developed methodology for constructing improvements of a given inadmissible test. In fact, by improving the "ignore-the-data" test, $\varphi_{ITD}(\boldsymbol{c}) = \alpha$ for all $\boldsymbol{c} \in \Gamma$, Berger and Sackrowitz (1997) constructed the first known test for this problem that is simultaneously admissible and unbiased. However, rejection regions at different α-levels need not be nested, so these improved tests may not yield unambiguous p-values, and thus are of somewhat limited value.

Berger (1998) established the one-to-one correspondence between the class of convex hull type tests and the minimal complete class of admissible tests. The convex hull test (BPI, 1998), $\varphi_{CH}$, is the simplest member of this convex hull class, and is qualitatively similar to the improvements of both $\varphi_S$ and $\varphi_{ITD}$, while minimizing, among all families of tests, the α-level required for its rejection region to contain $D[\Gamma]$.

In addition, $\varphi_{CH}$ is based on a test statistic, so rejection regions at different α-levels are nested, and p-values are provided. As such, $\varphi_{CH}$ is about as good a test as there is for testing $H$ against $H_A$, which is about as close as one can get to testing $H$ against $H'_A$ when dealing with $\boldsymbol{\theta}$ instead of $\boldsymbol{\pi}$. Specifically, admissible (unbiased) tests of $H$ against $H_A$ are conditionally admissible (unbiased) as tests of $H$ against $H'_A$ (Berger and Sackrowitz, 1997). However, $\boldsymbol{\theta}(\boldsymbol{\pi})$ is a nonlinear function, and maps small corners of $\boldsymbol{\pi}$-space (neighborhoods of structural zeros) into large regions of $\boldsymbol{\theta}$-space. By giving each direction $\delta(\boldsymbol{\theta})$ equal consideration, $\varphi_{CH}$ accommodates these small corners as much as it does the large regions of $\boldsymbol{\pi}$-space that are of greatest unconditional interest. As such, $\varphi_{CH}$ may not be ideal when viewed unconditionally. Cohen and Sackrowitz (1998) proposed another member of the convex hull class, called the *COM*(L) Fisher test, or

$\varphi_{COM}(L)$, based on repeatedly adding to the critical region those directed extreme points of the current acceptance region that are least likely under $H_0$. Because the test statistics of $\varphi_{COM}(L)$ and $\varphi_{CH}$ are defined not algebraically but relationally, by the relative position of $\boldsymbol{c}$ within $\Gamma$, the rejection regions need to be constructed recursively. This feature is a barrier to their use.

Adaptive Tests

Gross (1981, Section 5) suggested that an "analysis based on ... data-dependent scores may yield procedures that compare favorably to fixed-score procedures ...". Distinct from another definition used, e.g., by Rukhin and Mak (1992), Hogg (1974, p. 917) and Edgington (1995, pp. 371-373) defined adaptive tests as tests with data-based test statistics. This allows $\Gamma$ to be partitioned into regions sharing a common test statistic. Because the region need not be even nearly ancillary, conditioning on the region (as suggested by Donegani, 1991, and Good, 1994, p. 122) may entail a loss of power. Comparing the value of the test statistics across regions avoids this loss of power. The intuitive objection to "comparing apples to oranges" notwithstanding, such an approach is "good" or "bad" only to the extent to which it produces a "good" or "bad" test. This approach results in tests with excellent power properties. In fact, Gastwirth (1985) stated that "when the MERT for a particular problem has a low $r^2$, adaptive procedures are needed".

Without knowing $\theta$ *a priori*, it is unclear where to maximize the power. One could estimate $\delta(\boldsymbol{\theta})$ from $\boldsymbol{c}$, say as $\hat{\delta}_{p(\boldsymbol{c})}$, perhaps using maximum likelihood, and use the MP test $\varphi_{\hat{\delta}_P}$. The p-value of $\varphi_{\hat{\delta}_P}$ evaluated at observed outcome $\boldsymbol{c}$, $p_{\hat{\delta}_P}(\boldsymbol{c})$, is stochastically too small to serve as a valid p-value, but $p_{\hat{\delta}_P}(\boldsymbol{c})$ can be used as a *test statistic,* to be compared to its null distribution (Rohmel and Mansmann, 1999, p. 165). Variation in $\boldsymbol{c}$ is reflected in $p_{\hat{\delta}_P}(\boldsymbol{c})$ through *both* the argument and the subscript. Using either $p_{\hat{\delta}_P}(\boldsymbol{c})$ or $z_{\hat{\delta}_P}(\boldsymbol{c})$, suitably normalized, as a test statistic, any estimator $\hat{\delta}_{p(\boldsymbol{c})}$ of $\delta(\theta)$ induces an adaptive test, with regions $\Gamma_v = \hat{\delta}^{-1}(v) = \{\boldsymbol{c} \in \Gamma | \hat{\delta}_{p(\boldsymbol{c})} = v\}$. If the regions are $\Gamma_0 = \{\boldsymbol{c} \in \Gamma \mid C_{12} > n_1 T_2/(n_1 + n_2)\}$, $\Gamma_1 = \Gamma - \Gamma_0$, and $\Gamma_v = \varnothing$ for $v \notin \{0,1\}$, and the $\varphi_v$ test statistic $z_v(\boldsymbol{c})$ is used on $\Gamma_v$, with $C_{11} + C_{12}$ ($v = 0$) and $C_{11}$ ($v = 1$) normalized to $D_2$ and $D_1$, respectively, to facilitate the comparison of points from $\Gamma_1$ ($D_1 > D_2$) to those from $\Gamma_0$ ($D_2 \geq D_1$), then $\varphi_S$ results. Similar binary adaptive tests might define $\Gamma_0$ and $\Gamma_1$ by whichever of $\varphi_0$ and $\varphi_1$ yields a smaller p-value or a larger $\chi^2$.

Berger (1998) proposed judging outcome $\boldsymbol{c}$ by how small a p-value it can yield with an MP test; that is, $\varphi_A$ uses $p_{v^*(\boldsymbol{c})}(\boldsymbol{c}) = \min_{-\infty \leq v \leq \infty} p_v(\boldsymbol{c})$ as the test statistic. This is a continuous version of the adaptive test based on $\min(p_0(\boldsymbol{c}), p_1(\boldsymbol{c}))$, and estimates $\delta(\theta)$ non-uniquely as $\hat{\delta}_{\boldsymbol{c}} = v$ for any value $v \in v^*(\boldsymbol{c})$. The induced regions are $\Gamma_v = \{\boldsymbol{c} \in \Gamma | v \in v^*(\boldsymbol{c})\}$. The $\varphi_A$ critical region is $R_\alpha(\varphi_A) = \cup_{v \in R^1} R_{\alpha^*(v)}(\varphi_v)$ for some set of $\alpha^*(v) < \alpha$, so $\varphi_A$ is intuitively similar to union-intersection tests (Roy, 1953; Marden, 1991). Despite being constructed non-recursively, $\varphi_A$ is a convex hull type test (Berger, 1998); hence $\varphi_A$ is always admissible. Also, $\varphi_A$ tends to be omnibus, as $D[\Gamma] \subset R_\alpha(\varphi_A)$ for reasonable $\alpha$-levels.

Accommodating a Favored Alternative

Suppose that one believes *a priori* that $\delta(\theta) = \delta_P$. Let $\tau \geq 0$ be a measure of the strength in the belief that $\delta(\theta) = \delta_P$. The dual objectives are ensuring nearly MP power on $\Omega_{\delta_P}$ and reasonable power on $\Omega_A - \Omega_{\delta_P}$, with relative importance dictated by $\tau$. One might use $\varphi_{\delta_P}$ (which is MP on $\Omega_{\delta_P}$) for large $\tau$, or $\varphi_A$ (which is a good omnibus test) for small $\tau$, but none of the aforementioned test suffices for intermediate values of $\tau$. Linear

combinations such as $(\tau\varphi_{\delta_P} + \varphi_A)/(\tau + 1)$ would not suffice either, because they have large randomization regions and small critical regions, consisting only of the intersection $R_\alpha(\varphi_{\delta_P}) \cap R_\alpha(\varphi_A)$. Of course, these inadmissible tests could be improved to admissibility, but then the procedure would be complicated, and p-values may not be defined. There is another approach to bridge the gap between $\varphi_{\delta_P}$ and $\varphi_A$. Specifically, start with $\varphi_A$, but penalize those $\boldsymbol{c}$ whose minimizing MP p-value is obtained by $v$ far from $\delta_P$. To this end, let $\varphi_{\delta_P,\tau,\alpha}$ (or $\varphi_{\delta_P,\tau}$) be the level-$\alpha$ adaptive test based on the test statistic

$$A(\delta_P,\boldsymbol{\tau},\boldsymbol{c}) = \min_{-\infty\le v\le\infty} [\rho_{\min(v)}(\boldsymbol{c})(1 + |\delta_P - v|)^\tau].$$

Let $v_{[\delta_P,\tau]}(\boldsymbol{c}) = \{v \mid p_{\min(v)}(\boldsymbol{c})(1 + |\delta_P - v|)^\tau = A(\delta_P,\tau,\boldsymbol{c})\}$. Clearly, $\varphi_{\delta_P,0} = \varphi_A$ for any $\delta_P$ and $p_{\min(v)}(\boldsymbol{c})(1 + |\delta_P - v|)^\tau \le 1$ if $v \in v_{[\delta_P,\tau]}(\boldsymbol{c})$. Lemmas 2-4 confine $v_{[\delta_P,\tau]}(\boldsymbol{c})$ to a finite subset of an interval that shrinks, as $\tau$ gets large, to $\{\delta_P\}$. By Lemma 4, $\varphi_{\delta_P,\infty}$ induces the same ordering on $\Gamma$ as $\varphi_{\delta_P}$ does, thereby optimizing power on $\Omega_{\delta_P}$. Yet because the $\varphi_{\delta_P,\infty}$ test statistic is $p_{\min(\delta_P)}(\boldsymbol{c})$, and not necessarily $p_{\delta_P}(\boldsymbol{c})$, $\varphi_{\delta_P,\infty}$ is a refinement of $\varphi_{\delta_P}$, and $p_{\min(v)}(\boldsymbol{c}) \le p_{v,\infty}(\boldsymbol{c}) \le p_v(\boldsymbol{c})$ for all $v$ and $\boldsymbol{c}$. From Table 1, e.g., $p_{0.5}(11,2)=0.0385$, but $p_{0.5,\infty}(11,2)=0.0385-P_0\{(10,4)|T\}=0.0249$. Each test in the class of adaptive tests is admissible.

*Theorem 1.* For any triple $\delta_P \in \Re^1$, $\tau \ge 0$, and $\alpha \in [0,1]$, $\varphi_{\delta_P,\tau,\alpha}$ is admissible. Graubard and Korn (1987) suggested that without a reason to use a different $\delta_P$, $\varphi_{0.5}$ should be used. The desire to focus power on the "central" direction, $\Omega_{0.5}$, is understandable, but the use of linear rank tests in general (BPI, 1998; Berger and Ivanova, 2002), and $\varphi_{0.5}$ in particular (Ivanova and Berger, 2001), have been criticized. Now $\varphi_{0.5,\tau}$ offers good central power without sacrificing global power (unless $\tau = \infty$). But even if $\tau = \infty$, $\varphi_{0.5,\infty}$ is still more powerful than, and hence preferable to $\varphi_{0.5}$.

## Margin-Based Selection of $\delta_P$ and $\tau$

Recall that $v_S$ can be determined from the margins ($\boldsymbol{n}$ and $\boldsymbol{T}$, summarized by $\Gamma$). In some cases, it may be reasonable to use $v_S$ as $\delta_P$. In others, it may be reasonable to use the margins to find the largest $\tau$ that allows $R_\alpha(\varphi_{\delta_P,\tau,\alpha})$ to contain $D[\Gamma]$. Unless $|\delta_P - v_S|$ is small, the larger $\tau$ is, the less $\varphi_{\delta_P,\tau}$ focuses on omnibus power. Hence, the $\alpha$-level required for $R_\alpha(\varphi_{\delta_P,\tau,\alpha})$ to contain $D[\Gamma]$ tends to increase in $\tau$. If a range of $\alpha$-levels would be considered, say $0.01 \le \alpha \le 0.1$, then use the smallest $\alpha$-level in selecting $\tau$. Restricting attention to the integer values of $\tau$, and using $\delta_P = 0.5$, note that for $\{(11,2,2),(7,7,6)\}$, $D[\Gamma] = \{(6,9);(15,0)\}$ is contained by $R_{0.01}(\varphi_{0.5,18})$, $R_{0.025}(\varphi_{0.5,20})$, $R_{0.05}(\varphi_{0.5,22})$, and $R_{0.1}(\varphi_{0.5,24})$; but none of $R_{0.01}(\varphi_{0.5,19})$, $R_{0.025}(\varphi_{0.5,21})$, $R_{0.05}(\varphi_{0.5,23})$, or $R_{0.1}(\varphi_{0.5,25})$ contain (6,9). Consequently, $\varphi_{0.5,18}$ would be used by this approach.

## Comparisons of Tests

The exact conditional power of the one-sided nonrandomized versions of $\varphi_{0.0}$, $\varphi_{0.5}$, $\varphi_{1.0}$, $\varphi_S$, $\varphi_{CH}$, $\varphi_{COM}(\mathrm{L})$, and some adaptive tests, at $\alpha \le 0.05$, are compared considering all 87 $2 \times 3$ tables with row and column margins as in the example, $T$ = (18,9,8), $\boldsymbol{n}$ = (15,20). Figure 2 illustrates extreme regions. The exact conditional power of $\varphi$ to detect $\theta$ is calculated as $P_\theta\{R_{0.05}(\varphi)|\boldsymbol{T}\}$. Here $4 \times 7 = 28$ alternatives, with $\theta_1 \in \{0.5,1.0,1.5,2.0\}$ and $\theta_2 = \{-1.5,-1.0,-0.5,0.0,0.5,1.0,1.5\}$, are considered, along with the null case, $\theta_1 = \theta_2 = 0$. Bold entries represent the best power, for given $\theta$, among the six targeted tests in columns 4-9 and among five omnibus tests in columns 10-13. Because the linear rank tests $\varphi_{0.0}$ ($\alpha = 0.005$), $\varphi_{0.5}$ ($\alpha = 0.038$), and $\varphi_{1.0}$ ($\alpha = 0.028$) are excessively conservative, per the top

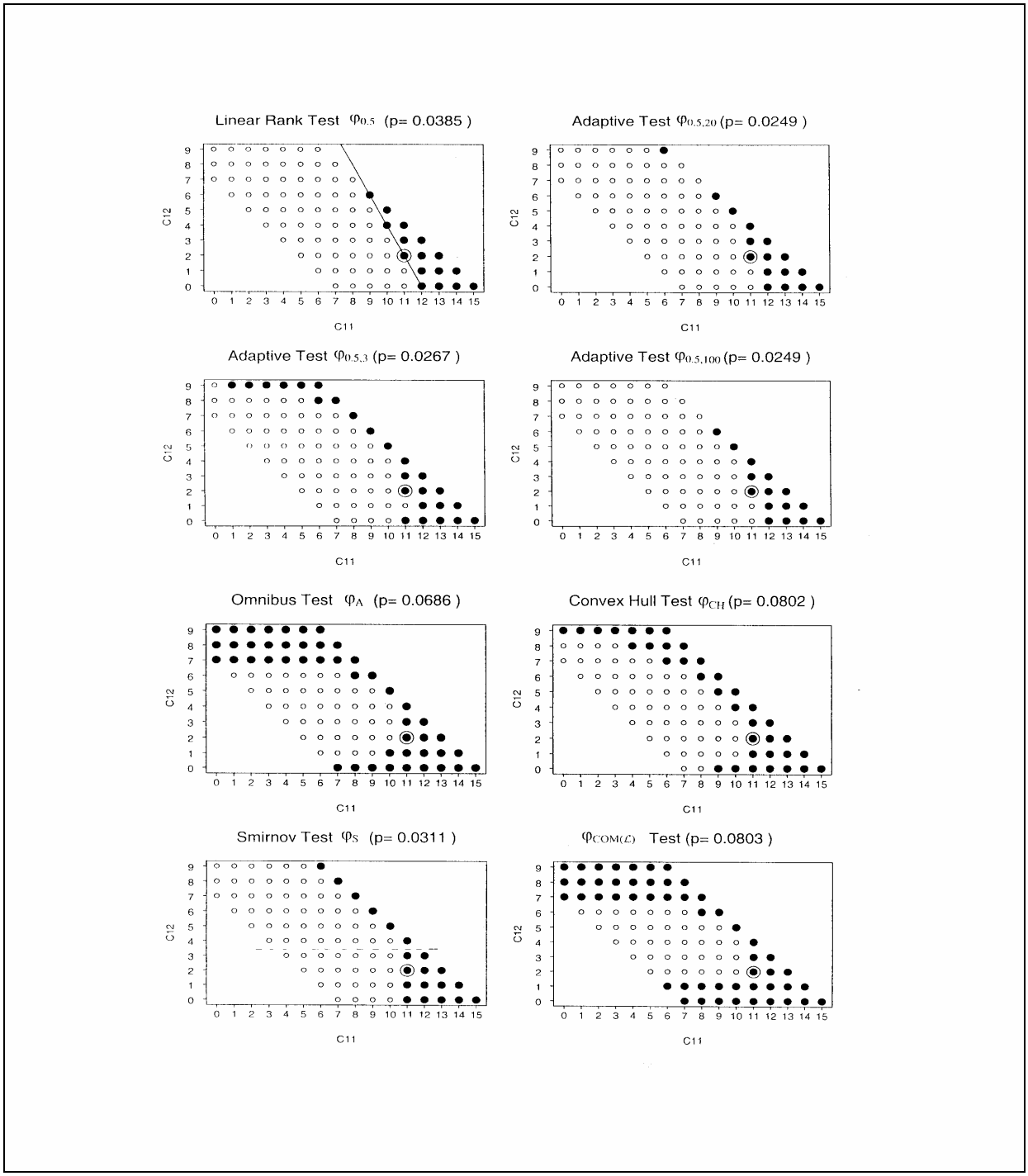


Figure 2. Extreme regions and p-values for {(11,2,2);(7,7,6)} and several tests including the linear rank test with equally-spaced scores $\varphi_{0.5}$, the adaptive tests with similar direction but varying second parameter $\varphi_{0.5,3}$, $\varphi_{0.5,20}$, $\varphi_{0.5,100}$, the omnibus adaptive test $\varphi_{A}$, the Smirnov test $\varphi_{S}$, the convex hull test $\varphi_{CH}$, and the $\varphi_{COM}$(L) test.

Table 2. Exact conditional power of the conservative (nonrandomized) versions of linear rank tests ($\varphi_0$, $\varphi_1$, $\varphi_{0.5}$), adaptive tests ($\varphi_{0,100}$, $\varphi_{1,100}$, $\varphi_{0.5,100}$, $\varphi_{0.5,1}$), omnibus adaptive test $\varphi_A$, the $\varphi_{COM}$(L) test, Smirnov test $\varphi_S$, and convex hull test $\varphi_{CH}$, with $\alpha \leq 0.05$, and table margins $\boldsymbol{T}$=(18,9,8), $\boldsymbol{n}$=(15,20). Bold entries represent the best power among the tests in each block (narrow and omnibus) for each given $\theta$.

| $\delta(\boldsymbol{\theta})$ | $\boldsymbol{\theta}$ | | $\varphi_0$ | $\varphi_{0,100}$ | $\varphi_{0.5}$ | $\varphi_{0.5,100}$ | $\varphi_1$ | $\varphi_{1,100}$ | $\varphi_{0.5,1}$ | $\varphi_A$ | $\varphi_{COM(L)}$ | $\varphi_S$ | $\varphi_{CH}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.0 | 0.0 | 0.005 | 0.040 | 0.038 | 0.044 | 0.028 | 0.039 | 0.046 | 0.047 | 0.050 | 0.031 | 0.035 |
| -2.000 | 0.5 | 1.5 | 0.054 | **0.232** | 0.046 | 0.063 | 0.006 | 0.015 | 0.258 | **0.375** | 0.316 | 0.058 | 0.255 |
| -1.000 | 0.5 | 1.0 | 0.038 | **0.163** | 0.071 | 0.080 | 0.021 | 0.032 | 0.150 | **0.198** | 0.152 | 0.053 | 0.145 |
| -0.500 | 1.0 | 1.5 | 0.107 | **0.325** | 0.151 | 0.174 | 0.039 | 0.067 | 0.290 | **0.332** | 0.244 | 0.131 | 0.285 |
| 0.000 | 0.5 | 0.5 | 0.025 | **0.120** | 0.103 | 0.110 | 0.057 | 0.070 | **0.109** | 0.108 | 0.090 | 0.073 | 0.093 |
| 0.000 | 1.0 | 1.0 | 0.079 | **0.264** | 0.212 | 0.223 | 0.099 | 0.126 | **0.219** | 0.215 | 0.169 | 0.151 | 0.200 |
| 0.000 | 1.5 | 1.5 | 0.184 | **0.447** | 0.352 | 0.371 | 0.149 | 0.208 | **0.366** | 0.361 | 0.292 | 0.270 | 0.349 |
| 0.250 | 2.0 | 1.5 | 0.280 | 0.606 | 0.603 | **0.615** | 0.370 | 0.445 | **0.524** | 0.491 | 0.460 | 0.485 | 0.489 |
| 0.333 | 1.5 | 1.0 | 0.143 | 0.417 | 0.442 | **0.455** | 0.288 | 0.328 | **0.379** | 0.333 | 0.310 | 0.346 | 0.330 |
| 0.500 | 1.0 | 0.5 | 0.055 | 0.231 | 0.274 | **0.291** | 0.200 | 0.225 | **0.244** | 0.196 | 0.189 | 0.223 | 0.193 |
| 0.500 | 2.0 | 1.0 | 0.231 | 0.593 | 0.689 | **0.704** | 0.560 | 0.597 | **0.615** | 0.543 | 0.537 | 0.605 | 0.542 |
| 0.667 | 1.5 | 0.5 | 0.109 | 0.390 | 0.521 | **0.550** | 0.454 | 0.481 | **0.483** | 0.391 | 0.395 | 0.475 | 0.390 |
| 0.750 | 2.0 | 0.5 | 0.188 | 0.560 | 0.754 | **0.785** | 0.723 | 0.741 | **0.738** | 0.634 | 0.640 | 0.735 | 0.634 |
| 1.000 | 0.5 | 0.0 | 0.015 | 0.096 | 0.137 | **0.157** | 0.121 | 0.147 | **0.140** | 0.104 | 0.116 | 0.127 | 0.100 |
| 1.000 | 1.0 | 0.0 | 0.038 | 0.201 | 0.333 | **0.378** | 0.332 | 0.368 | **0.347** | 0.258 | 0.283 | 0.339 | 0.257 |
| 1.000 | 1.5 | 0.0 | 0.082 | 0.349 | 0.585 | **0.646** | 0.612 | 0.642 | **0.621** | 0.499 | 0.521 | 0.617 | 0.499 |
| 1.000 | 2.0 | 0.0 | 0.153 | 0.514 | 0.799 | 0.851 | 0.836 | **0.852** | **0.841** | 0.736 | 0.748 | 0.839 | 0.736 |
| 1.250 | 2.0 | -0.5 | 0.126 | 0.467 | 0.830 | 0.896 | 0.906 | **0.924** | **0.908** | 0.828 | 0.844 | 0.906 | 0.828 |
| 1.333 | 1.5 | -0.5 | 0.062 | 0.302 | 0.634 | 0.729 | 0.736 | **0.779** | **0.744** | 0.628 | 0.665 | 0.737 | 0.628 |
| 1.500 | 1.0 | -0.5 | 0.026 | 0.167 | 0.384 | 0.472 | 0.471 | **0.536** | **0.483** | 0.377 | 0.432 | 0.472 | 0.375 |
| 1.500 | 2.0 | -1.0 | 0.106 | 0.429 | 0.854 | 0.920 | 0.944 | **0.963** | **0.948** | 0.899 | 0.915 | 0.944 | 0.899 |
| 1.667 | 1.5 | -1.0 | 0.048 | 0.262 | 0.671 | 0.784 | 0.822 | **0.876** | **0.834** | 0.752 | 0.795 | 0.823 | 0.750 |
| 1.750 | 2.0 | -1.5 | 0.093 | 0.401 | 0.874 | 0.930 | 0.965 | **0.983** | **0.970** | 0.945 | 0.958 | 0.965 | 0.945 |
| 2.000 | 0.5 | -0.5 | 0.010 | 0.077 | 0.171 | 0.221 | 0.212 | **0.273** | **0.227** | 0.167 | 0.217 | 0.214 | 0.163 |
| 2.000 | 1.0 | -1.0 | 0.018 | 0.136 | 0.426 | 0.552 | 0.593 | **0.692** | **0.617** | 0.524 | 0.604 | 0.593 | 0.520 |
| 2.000 | 1.5 | -1.5 | 0.038 | 0.231 | 0.703 | 0.811 | 0.877 | **0.934** | **0.895** | 0.848 | 0.889 | 0.877 | 0.845 |
| 2.500 | 1.0 | -1.5 | 0.013 | 0.111 | 0.463 | 0.602 | 0.687 | **0.810** | 0.730 | 0.668 | **0.756** | 0.687 | 0.660 |
| 3.000 | 0.5 | -1.0 | 0.006 | 0.059 | 0.203 | 0.292 | 0.318 | **0.433** | 0.350 | 0.284 | **0.377** | 0.318 | 0.275 |
| 4.000 | 0.5 | -1.5 | 0.004 | 0.044 | 0.231 | 0.348 | 0.419 | **0.591** | 0.487 | 0.435 | **0.558** | 0.419 | 0.416 |
| Mean power | | | 0.083 | 0.293 | 0.447 | 0.500 | 0.458 | **0.505** | **0.519** | 0.469 | 0.481 | 0.482 | 0.457 |
| p-value for (11,2,2;7,7,6) | | | 0.228 | 0.073 | 0.038 | 0.025 | 0.028 | 0.028 | 0.037 | 0.069 | 0.080 | 0.031 | 0.080 |

Table 3. Pairwise comparisons of 11 tests for 4x7=28 values of $\boldsymbol{\theta}$, where each entry is the number of parameter values (out of 28 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).

| | $\varphi_0$ | $\varphi_{0,100}$ | $\varphi_{0.5}$ | $\varphi_{0.5,100}$ | $\varphi_1$ | $\varphi_{1,100}$ | $\varphi_{0.5,1}$ | $\varphi_A$ | $\varphi_{COM(L)}$ | $\varphi_S$ | $\varphi_{CH}$ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi_0$ | - | **0** | 1 | **0** | 4 | 3 | **0** | **0** | **0** | **0** | **0** | 8 |
| $\varphi_{0,100}$ | **28** | - | 7 | 6 | 10 | 9 | 7 | 7 | 9 | 9 | 9 | 101 |
| $\varphi_{0.5}$ | 27 | 21 | - | **0** | 14 | 12 | 6 | 14 | 13 | 12 | 17 | 136 |
| $\varphi_{0.5,100}$ | **28** | 22 | **28** | - | 18 | 15 | 13 | 21 | 18 | 17 | 21 | 201 |
| $\varphi_1$ | 24 | 18 | 14 | 10 | - | **0** | **0** | 19 | 14 | **0** | 20 | 119 |
| $\varphi_{1,100}$ | 25 | 19 | 16 | 13 | **28** | - | 17 | 20 | 21 | 19 | 20 | 198 |
| $\varphi_{0.5,1}$ | **28** | 21 | 22 | 15 | **28** | 11 | - | 25 | 23 | **28** | **28** | **229** |
| $\varphi_A$ | **28** | 21 | 14 | 7 | 9 | 8 | 3 | - | 10 | 8 | **28** | 136 |
| $\varphi_{COM(L)}$ | **28** | 19 | 15 | 10 | 14 | 7 | 5 | 18 | - | 12 | 20 | 148 |
| $\varphi_S$ | **28** | 19 | 16 | 11 | **28** | 9 | **0** | 20 | 16 | - | 21 | 168 |
| $\varphi_{CH}$ | **28** | 19 | 11 | 7 | 8 | 8 | **0** | **0** | 8 | 7 | - | 96 |
| Total | 272 | 179 | 144 | 79 | 161 | 82 | **51** | 144 | 132 | 112 | 184 | |

row of Table 2, they are dominated at $\alpha = 0.05$ by their corresponding adaptive tests $\varphi_{0.0,100}$ ($\alpha = 0.040$), $\varphi_{0.5,100}$ ($\alpha = 0.044$), and $\varphi_{1.0,100}$ ($\alpha = 0.039$). This is not surprising, and will be the case quite generally. Note that $\varphi_{0.5,1}$ maximizes the average power, at 0.519, or the area under the power curve. The non-adaptive tests did not fare as well. Among the omnibus tests ($\varphi_A$, $\varphi_{COM}$(L), $\varphi_S$, and $\varphi_{CH}$), $\varphi_{0.5,1}$ maximizes the power for 22 of the 28 $\theta$ values ($\varphi_A$ and $\varphi_{COM}$(L) each maximize the power for three $\theta$ values). Also, $\varphi_{0.5,1}$ ($p = 0.037$) and $\varphi_S$ ($p = 0.031$) are the only omnibus tests to yield statistical significance at $\alpha = 0.05$ for {(11,2,2);(7,7,6)}. Table 3, above, shows that $\varphi_{0.5,1}$ dominates both $\varphi_S$ and $\varphi_{CH}$, and almost dominates $\varphi_A$ and $\varphi_{COM}$(L) too, and does dominate them when $\delta(\theta)$ is near the $\delta_P$ value of 0.5 used by $\varphi_{0.5,1}$. In fact, only where $\delta(\theta) \le -0.5$ or $\delta(\theta) \ge 2.5$ is $\varphi_A$ or $\varphi_{COM}$(L) more powerful than $\varphi_{0.5,1}$. Among pairwise comparisons, $\varphi_{0.5,1}$ has larger power than its competitor (each of the other ten tests are considered for each of 28 alternatives) for 229 out of 280 comparisons, and 104 of the 112 comparisons to omnibus tests. The non-adaptive tests did not fare as well, but $\varphi_S$ attained168/280 or 57/112, respectively, which is quite respectable.

## Conclusion

In an effort to improve the comparison of two treatments on the basis of ordinal data, a new class of adaptive tests was defined, and shown to be admissible, while providing unambiguous p-values and a non-iterative construction. If one is interested in testing for $\theta_1 > 0$, and has no particular preference for any subset of $\Omega_A$ relative to any other, then $\varphi_{CH}$ would be a fine test to use.

However, $\varphi_A$ and $\varphi_{0.5,1}$ are also excellent omnibus tests, and are easier to compute then $\varphi_{CH}$. If one is interested in testing for stochastic order, and uses $\theta_1 > 0$ only as a surrogate, then $\varphi_A$ and $\varphi_{0.5,1}$ are probably better tests than $\varphi_{CH}$. Certainly if one is in the situation treated in this article, with a preferred direction, then an appropriate adaptive test would be the test of choice. There is nothing particular about ordered trinomial distributions that makes this problem especially amenable to treatment with the adaptive

approach. For any hypothesis testing problem with a composite alternative hypothesis, one can enumerate the alternatives and the corresponding MP test for each. One can then apply each of these MP tests to a given outcome, and find the smallest of the resulting p-values. Using this minimized MP p-value as a test statistic produces a test analogous to $\varphi_A$, and reduces to the uniformly most powerful test if one exists. If not, then the adaptive tests that bridge the gap between $\varphi_A$ and the MP tests to detect a favored direction should have good properties in a variety of contexts.

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. *Statistics in Medicine*, *19*, 1319-1328.

Berger, V. W., & Ivanova, A. (2002). The bias of linear rank tests when testing for stochastic order in ordered categorical data. *Journal of Statistical Planning and Inference*, *107*, 237-247.

Berger, V. W., Lunneborg, C., Ernst, M. D., Levine, J. G. (2002), Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods*, *1*, 74-82.

Berger, V. W., Permutt, T., & Ivanova A. (1998). The Convex hull test for ordered categorical data. *Biometrics*, *54***,** 1541-1550.

Berger, V. W., & Sackrowitz, H. (1997). Improving tests for superior treatments in contingency tables. *Journal of American Statistical Association*, *92***,** 700-705.

Chiara, S., Compora, E., Merlini, L., et al. (1993). Recurrent ovarian carcinoma: salvage treatment with platinum in patients responding to first-line platinum-based regiments. *European Journal of Cancer*, *29A*, 652.

Cohen, A., & Sackrowitz, H. (1998). Directional tests for one-sided alternatives in multivariate models. *Annals of Statistics*, *26***,** 2321-2338.

Donegani, M. (1991). An adaptive and powerful randomization test. *Biometrika*, *78*, 930-933.

Edgington, E. S. (1995). *Randomization tests.* (3rd ed.). New York: Marcel Dekker.

Frick, H. (2000). Undominated p-values and property *c* for unconditional one-sided two-sample binomial tests. *Biometrical Journal*, *42*, 715-728.

Gastwirth, J. L. (1985). The use of maximum efficiency robust tests in combining contingency tables and survival analysis. *The Journal of the American Statistical Association, 80*, 380-384.

Good, P. (1994). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. New York: Springer-Verlag.

Gross, S. T. (1981). On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. *Journal of American Statistical Association*, *76***,** 935–941.

Graubard, B. I., & Korn, E. L.(1987). Choice of column scores for testing independence in ordered 2xk contingency tables. *Biometrics*, *43*, 471-476.

Hogg, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of American Statistical Association*, *69*, 909-923.

Ivanova, A., & Berger, V. W. (2001). Drawbacks of integer scoring of ordered categorical data. *Biometrics*, *57*, 567-570.

Marden, J. I. (1991). Sensitive and sturdy p-values. *Annals of statistics, 19*, 918-934.

McCullagh, P. (1980). Regression methods for ordinal data. *Journal of Royal Statistical Society*, *B 42*, 109-142.

Permutt, T., & Berger, V. W. (2000). Rank tests in ordered 2xk contingency tables. *Communications in Statistics, Theory and Methods*, *29*, 989-1003.

Podgor, M. J., Gastwirth, J. L., Mehta, C. R. (1996). Efficiency robust tests of independence in contingency tables with ordered categories. *Statistics in Medicine*, *15*, 2095-2105.

Rohmel, J. & Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal*, *41*, 149-170.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Statistics*, *24***,** 220-238.

Rukhin, AL, Mak, KS (1992). Adaptive Test Statistics and Bahadur Efficiency. *Statistica Sinica*, *2*, 541-552.

## Appendix
## Lemmas (with Proofs), and Proofs of Theorems

*Lemma 1*. Let $\boldsymbol{c} \in \Gamma$ and $k \in \{0,1,...,K_{\boldsymbol{c}}\}$. If $|v_k(\boldsymbol{c}) \pm \varepsilon(\boldsymbol{c})| < \infty$ then $v_k(\boldsymbol{c}) \pm \varepsilon(\boldsymbol{c}) \notin V(\boldsymbol{c})$. If $v \in (v_k(\boldsymbol{c}), v_{k+1}(\boldsymbol{c}))$, then $M_v(\boldsymbol{c}) = M_{v_{(k+1)}(\boldsymbol{c})}(\boldsymbol{c}) - B^-_{v_{(k+1)}(\boldsymbol{c})}(\boldsymbol{c}) = M_{v_{(k)}(\boldsymbol{c})}(\boldsymbol{c}) - B^+_{v_{(k)}(\boldsymbol{c})}(\boldsymbol{c})$.

*Proof*. Increasing (decreasing) $v$ by $\varepsilon(\boldsymbol{c})$ moves $B^-_v(\boldsymbol{c})$ ($B^+_v(\boldsymbol{c})$) into the interior of, and $B^+_v(\boldsymbol{c})$ ($B^-_v((\boldsymbol{c}))$) completely out of, the new critical region, but if $v \in V(\boldsymbol{c})$, then no points of $\Gamma - M_v(\boldsymbol{c})$ are moved into the new critical region (Table 1). Hence, $o_{v-\varepsilon(c)}(\boldsymbol{c}) = o_{v+\varepsilon(\boldsymbol{c})}(\boldsymbol{c}) = 1$, and neither $v_k(\boldsymbol{c}) - \varepsilon(\boldsymbol{c})$ nor $v_k(\boldsymbol{c}) + \varepsilon(\boldsymbol{c})$ is in $V(\boldsymbol{c})$. If $v \notin V(\boldsymbol{c})$, say $v_k(\boldsymbol{c}) < v < v_{k+1}(\boldsymbol{c})$, then $o_v(\boldsymbol{c}) = 1$, so $B^+_v(\boldsymbol{c}) = B^-_v(\boldsymbol{c}) = \varnothing$ and $M_v(\boldsymbol{c})$ will not change when $v$ varies within $(v_k(\boldsymbol{c}), v_{k+1}(\boldsymbol{c}))$.

*Lemma 2*. If $\delta_P \in \Re^1$, $\tau > 0$, $v_* \in v_{[\delta_P,\tau]}(\boldsymbol{c})$, and $v^* \in v^*(\boldsymbol{c})$, then $|\delta_P - v_*| \le |\delta_P - v^*|$.

*Proof*. If there exist $v^* \in v^*(\boldsymbol{c})$ and $v_* \in v_{[\delta P,\tau]}(\boldsymbol{c})$ such that $|\delta_P - v^*| < |\delta_P - v_*|$, then $p_{v^*}(\boldsymbol{c})(1 + |\delta_P - v^*|)^\tau < p_{\min(v^*)}(\boldsymbol{c})(1 + |\delta_P - v_*|)^\tau$, and $v_*$ cannot be in $v_{[\delta_P,\tau]}(\boldsymbol{c})$.

*Lemma 3*. For any $\delta_P$, $\tau > 0$, and $\boldsymbol{c} \in \Gamma$, $v_{[\delta_P,\tau]}(\boldsymbol{c}) \subset V(\boldsymbol{c}) \cup \delta_P$.

*Proof*. Assume there exists $v \ne \delta_P$ in $v_{[\delta_P,\tau]}(\boldsymbol{c}) - V(\boldsymbol{c})$, say $v_k(\boldsymbol{c}) < v < v_{k+1}(\boldsymbol{c})$. Let $v^* = v_k(\boldsymbol{c})$ if $\delta_P \le v_k(\boldsymbol{c})$, $v^* = \delta_P$ if $v_k(\boldsymbol{c}) < \delta_P < v_{k+1}(\boldsymbol{c})$, or $v^* = v_{k+1}(\boldsymbol{c})$ if $v_{k+1}(\boldsymbol{c}) \le \delta_P$. Now $v^* \subset V(\boldsymbol{c}) \cup \delta_P$ and $p_{\min(v)}(\boldsymbol{c})(1 + |\delta_P - v|)^\tau > p_{\min(v^*)}(\boldsymbol{c})(1 + |\delta_P - v^*|)^\tau$.

*Lemma 4*. For any $\delta_P$ and $\boldsymbol{c} \in \Gamma$, $v_{[\delta_P,\tau]}(\boldsymbol{c}) = \{\delta_P\}$ for sufficiently large $\tau$.

*Proof*. Let $D_{\boldsymbol{c}}(\delta_P) = \min_{v \in V(c) - \delta_P} |\delta_P - v| > 0$. For $\tau > 0$, let $v \in v_{[\delta_P,\tau]}(\boldsymbol{c}) - \delta_P$. By Lemma 3, $v \in V(\boldsymbol{c}) - \delta_P$, so $|\delta_P - v| \ge D_{\boldsymbol{c}}(\delta_P)$. If $\tau > -\ln(p_{\min(\delta_P)}(\boldsymbol{c}))/\ln(1 + D_{\boldsymbol{c}}(\delta_P))$, then $p_{\min(v)}(\boldsymbol{c})(1 + |\delta_P - v|)^\tau \ge p_{\min(v)}(\boldsymbol{c})(1 + |D_{\boldsymbol{c}}(\delta_P)|)^\tau > 1$, contradicting $v \in v_{[\delta_P,\tau]}(\boldsymbol{c})$.

*Proof of Theorem 1*. By Theorem 3.3 of Berger (1998), it suffices to show that for any $B \subset \Gamma$, if $\boldsymbol{c}^*$ minimizes $A(\delta_P,\tau,\boldsymbol{c})$ over $B$, then $\boldsymbol{c}^* \in D[B]$. If $\boldsymbol{c}^* \notin D[B]$, then $\boldsymbol{c}^*$ cannot, for any $v$, uniquely minimize $p_v$ over $B$, and for every $v$ there exists $\boldsymbol{c} \in B - \boldsymbol{c}^*$ such that $p_v(\boldsymbol{c}) \le p_v(\boldsymbol{c}^*)$. If $v \notin V(\boldsymbol{c}^*)$, then $o_v(\boldsymbol{c}^*) = 1$, so $p_v(\boldsymbol{c}) \ne p_v(\boldsymbol{c}^*)$, and $p_v(\boldsymbol{c}) \le p_v(\boldsymbol{c}^*) - \min_{\boldsymbol{c} \in \Gamma} P_0\{\boldsymbol{c}|\Gamma\}$. Let $v_1 \in v_{[\delta_P,\tau]}(\boldsymbol{c}^*)$. By the continuity in $v$ of the function $(1 + |\delta_P - v|)^\tau$, one can, for any $\varepsilon > 0$, choose $v_2 \notin V(\boldsymbol{c}^*)$ suitably close to $v_1$ to satisfy $p_{v_2}(\boldsymbol{c}^*) = p_{\min(v_1)}(\boldsymbol{c}^*)$, and, thus,

$$A(\delta_P,\tau,\boldsymbol{c}) = \min_{-\infty \le v \le \infty} [p_{\min(v)}(\boldsymbol{c})(1 + |(\delta_P - v|)^\tau] \le p_{v_2}(\boldsymbol{c})(1 + |(\delta_P - v_2|)^\tau$$

$$\le [p_{v_2}(\boldsymbol{c}^*) - \min_{c \in \Gamma} P_0\{\boldsymbol{c}|\Gamma\}](1 + |\delta_P - v_2|)^\tau$$

$$= [p_{\min(v_1)}(\boldsymbol{c}^*) - \min_{c \in \Gamma} P_0\{\boldsymbol{c}|\Gamma\}](1 + |\delta_P - v_2|)^\tau$$

$$< A(\delta_P,\tau,\boldsymbol{c}^*) - \min_{c \in \Gamma} P_0\{\boldsymbol{c}|\Gamma\}(1 + |\delta_P - v_2|)^\tau + \varepsilon < A(\delta_P,\tau,\boldsymbol{c}^*),$$

the last inequality holding for $\varepsilon < \min_{\boldsymbol{c} \in \Gamma} P_0\{\boldsymbol{c}|\Gamma\}$. This is a contradiction.