BIOMETRICS 54, 1541–1550
December 1998

Convex Hull Test for Ordered Categorical Data

Vance W. Berger,[1,*] Thomas Permutt,[1] and Anastasia Ivanova[2]

[1]Food and Drug Administration, Center for Biologics Evaluation and Research, 1401 Rockville Pike 200S, HFM-215, Rockville, Maryland 20852-1448, U.S.A.

[2]Department of Mathematics and Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, Maryland 21250, U.S.A.

SUMMARY

When testing for stochastic order in ordered $2 \times J$ contingency tables, it is common to select the cutoff required to declare significance so as to ensure that the size of the test is exactly α conditionally on the margins. It is valid, however, to use the margins to select not only the cutoff but also the form of the test. Linear rank tests, which are locally most powerful and frequently used in practice, suffer from the drawback that they may have power as low as zero to detect some alternatives of interest when the margins satisfy certain conditions. The Smirnov and convex hull tests are shown, through exact conditional power calculations and simulations, to avoid this drawback. The convex hull test is also admissible and palindromic invariant and minimizes the required significance level to have limiting power of one as the alternative moves away from the null in any direction.

1. Introduction

We consider the problem of testing for a treatment effect in a two-arm randomized clinical trial with J ordered outcome levels. For example, active and placebo antineoplastic agents might be compared for their effects on the size of tumors, with shrinkage rated as complete, partial, or none ($J = 3$), the data displayed as an ordered 2×3 contingency table, $\mathbf{X} = (X_1, X_2, X_3; Y_1, Y_2, Y_3)$. Moses, Emerson, and Hosseini (1984) and Rahlfs and Zimmermann (1993) cited the common practice of combining categories to create a 2×2 table as inefficient and suggested instead linear rank tests, in which the three outcome levels are assigned (without loss of generality) scores of $(0, \nu, 1)$ and the treatments are compared by the t-test, here denoted as ϕ_ν. The choice of $\nu \in \Re^1$ (often $\nu \in [0, 1]$) was discussed by Chakraborti and Schaafsma (1996) and Gautam (1997). Some authors recommend that scores be assigned according to an *a priori* assessment of how far apart the outcomes are, while others base them on the marginal totals for the outcome levels, usually in a way that corresponds to standard rank tests for continuous data. Graubard and Korn (1987) suggested that, if the choice is not apparent, then $\phi_{0.5}$ (equally spaced scores) should be used. Kimeldorf, Sampson, and Whitaker (1992) considered $\nu_{(1)}$ and $\nu_{(2)}$ to maximize and minimize, respectively, the p-value. The results may straddle α, with $p_{\min} < \alpha < p_{\max}$, meaning that the choice of scores influences the results.

As with continuous data, linear rank tests have reasonable power to detect shifts but poor (possibly zero) power for alternatives in which the cumulative distribution functions (c.d.f.'s) are strongly separated over part of the range but nearly coincide elsewhere. Thus, Podgor, Gastwirth, and Mehta (1996) and Sharp and Koch (1996) proposed looking at several sets of scores, and Behnen and Neuhaus (1989) proposed nonlinear rank tests. The Smirnov test (Hilton, Mehta, and Patel, 1994; Nikiforov, 1994), which has been proposed for situations in which nonshift alternatives

*Corresponding author's email address: bergerv@cber.fda.cov

Key words: Data peeling; Exact conditional test; Margin-based test; Smirnov test; Stochastic order.

Disclaimer: The views, opinions and assertions expressed in this article are those of the authors. No endorsement by the Food and Drug Administration is intended or should be inferred.

are of importance, is a nonlinear rank test based on the largest difference between empirical c.d.f.'s, and is denoted as ϕ_S. Because the acceptance regions of t-tests and ϕ_S are linear and piecewise linear, respectively, these tests are typically inadmissible (Berger and Sackrowitz, 1997; Berger, 1998). Berger (1995) and Cohen and Sackrowitz (1996) proposed applying convex hull peeling (Eddy, 1982; Green, 1985; Petitjean and Saporta, 1992) to the sample space. Unlike linear rank tests and the Smirnov test, these convex hull tests do not have close analogs for continuous data. Rather, the finiteness of the sample space is exploited by recursively adding points to the critical region one set at a time until the desired size is achieved. Berger (1998) established the equivalence of the class of convex hull tests and the minimal complete class of admissible tests. We describe the construction of the simplest convex hull test ϕ_{CH} in Section 3. In Sections 4 and 5, we compare the exact conditional and simulated unconditional power, respectively, of ϕ_{CH} to that of $\phi_{0.5}$ and ϕ_S. We refer to $\phi_{0.5}$ as either the exact t-test or the linear rank test with equally-spaced scores.

2. Notation and Formulation

Let $\pi_{i,j}$ be the probability that a patient results in outcome j (1, 2, or 3) if given treatment i (1 or 2), with $\boldsymbol{\pi_1} = \{\pi_{1,1}, \pi_{1,2}, \pi_{1,3}\}$ and $\boldsymbol{\pi_2} = \{\pi_{2,1}, \pi_{2,2}, \pi_{2,3}\}$ each summing to one and $\boldsymbol{\pi} = (\boldsymbol{\pi_1}, \boldsymbol{\pi_2})$. For $j = 1, 2$, let $\lambda_j(\boldsymbol{\pi}) = (\pi_{1,j}\pi_{2,3})/(\pi_{2,j}\pi_{1,3})$, $\theta_j(\boldsymbol{\pi}) = \ln(\lambda_j)$, and $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The row margins are $\mathbf{n} = \mathbf{n}(\mathbf{X}) = (\mathbf{n_1}, \mathbf{n_2})$ and the column margins are $\mathbf{T} = \mathbf{T}(\mathbf{X}) = (T_1, T_2, T_3)$. We consider exact conditional (on $\mathbf{T}$) tests of H*: $\boldsymbol{\pi_1} = \boldsymbol{\pi_2}$ against the one-sided stochastic order alternative hypothesis K*: $\boldsymbol{\pi_1} \neq \boldsymbol{\pi_2}$ and $\Sigma_{j=1}^{k} \pi_{1,j} \geq \Sigma_{j=1}^{k} \pi_{2,j}, k = 1, 2, 3$. As the conditional distribution of $\mathbf{X}$ given $\mathbf{T}$ (and hence the conditional power) depends on $\boldsymbol{\pi}$ only through $\boldsymbol{\theta}$ (Berger, 1998), the (conditional) hypotheses must be formulated in terms of $\boldsymbol{\theta}$ to be identifiable. Because $\boldsymbol{\pi_1} = \boldsymbol{\pi_2}$ if and only if $\boldsymbol{\theta}(\boldsymbol{\pi}) = \mathbf{0}$, the null hypothesis is H: $\boldsymbol{\theta} = \mathbf{0}$. However, $\boldsymbol{\theta}$ is insufficient to classify each alternative as being stochastically ordered or not, and thus the remaining parameters are not nuisance parameters. As a result, no conditional alternative hypothesis can be equivalent to K*, but we consider testing H against K: $\boldsymbol{\theta} \in \Omega = \{\boldsymbol{\theta} \mid \theta_1 > 0\}$ as in Berger (1998). The sample space Γ is the set of 2×3 contingency tables with nonnegative integer-valued cell counts with row totals $\mathbf{n}(\mathbf{X})$ and column totals $\mathbf{T}(\mathbf{X})$. As one can reconstruct the entire table from (X_1, X_2), we let $\mathbf{X} = (X_1, X_2)$ denote a point of Γ.

The likelihood ratio is $\Lambda_{\boldsymbol{\theta}}(\mathbf{X}) = P_{\boldsymbol{\theta}}\{\mathbf{X} \mid \mathbf{T}\}/P_{\mathbf{0}}\{\mathbf{X} \mid \mathbf{T}\} = k(\mathbf{T}; \boldsymbol{\theta}) \exp(\mathbf{X}'\boldsymbol{\theta})/k(\mathbf{T}; \mathbf{0})$, and the most powerful (MP) test to detect $\boldsymbol{\theta}$ is the linear rank test $\phi_{(\theta_1-\theta_2)/\theta_1}$, which depends on the direction (θ_2/θ_1), but not the magnitude $(\theta_1^2+\theta_2^2)^{1/2}$, of $\boldsymbol{\theta}$ (Berger, 1998). Because the quantity $(\theta_1-\theta_2)/\theta_1$ can assume any value in $\Re^1$ as $\boldsymbol{\theta}$ ranges over Ω, each linear rank test ϕ_ν, $\nu \in \Re^1$, is MP on a slice (with Lebesgue measure zero) $\Omega_\nu = \{\boldsymbol{\theta} \in \Omega \mid (\theta_1-\theta_2)/\theta_1 = \nu\}$ of Ω. Let $R(\alpha; \phi)$ be the critical region (including any randomization points) of ϕ at level α, and let $R(\alpha; \boldsymbol{\theta}) = R(\alpha; \phi_{(\theta_1-\theta_2)/\theta_1})$ be the critical region of the MP level-α test to detect $\boldsymbol{\theta}$. If $R(\alpha; \phi) \cap R(\alpha; \boldsymbol{\theta}) = \emptyset$, then ϕ will be seen to have poor power to detect $\boldsymbol{\theta}$. Let $I_\Gamma = \{R(\alpha; \phi) \mid \forall\, \alpha > 0,\ \boldsymbol{\theta} \in \Omega\ R(\alpha; \boldsymbol{\theta}) \cap R(\alpha; \boldsymbol{\theta}) \neq \emptyset\}$ be the class of critical regions $R(\alpha; \phi)$ that intersect with $R(\alpha; \boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Omega$. The corresponding tests will have good power to detect each $\boldsymbol{\theta} \in \Omega$. For $A \subset \Gamma$, let $D[A]$ be the set of (directed extreme) points of A that uniquely maximize $\Lambda_{\boldsymbol{\theta}}(\mathbf{X})$ for some $\boldsymbol{\theta} \in \Omega$. Berger and Ivanova (1997) showed $D[\Gamma]$ to consist of the points $W_1 = (\min(n_1, T_1), \max(0, n_1 - T_1 - T_3))$, $W_2 = (\min(n_1, T_1), \min(n_1, T_1 + T_2) - \min(n_1, T_1))$, and $W_3 = (\min(n_1, T_1 + T_2) - \min(n_1, T_2), \min(n_1, T_2))$. Because each $R(\alpha; \boldsymbol{\theta})$ must contain at least one of these three points, we see that $D[\Gamma] \in I_\Gamma$. In fact, $D[\Gamma] = \cap_{A \in I_\Gamma} A$, so $D[\Gamma]$ consists of precisely those points that are needed in the critical region to intersect with each $R(\alpha; \boldsymbol{\theta})$.

It would seem reasonable to construct the critical region by starting with $D[\Gamma]$. However, care needs to be exercised because there are sets of margins for which some directed extreme points are not suitable candidates for early entry into the critical region. If $n_1 \leq T_2$, then W_3 has a zero in the upper left cell and provides evidence that the control is superior to the treatment. If $n_1 > T_1 + T_3$, then $X_2 > 0$ for all $\mathbf{X} \in \Gamma$, so any alternative that specifies X_2 as a structural zero ($\pi_{1,2} = 0$) will not explain the observed data. If $\theta_2 = -\infty$, then necessarily $\pi_{1,2}\pi_{2,3} = 0$. Under H* or K*, $\pi_{2,3} = 0$ implies that $\pi_{1,3} = 0$, so $\pi_{2,3} = 0$ is inconsistent with T_3 being positive. This means that, if $n_1 > T_1 + T_3 > T_1$, we have no interest in detecting alternatives for which $\theta_2 = -\infty$, and W_1 (which derives its prominence as a directed extreme point from such alternatives) ought not be treated as a directed extreme point. Because the smaller of the lower ranges of Y_1 and X_3 and of Y_2 and X_3 are always structural zeros, we always wish to detect $\theta_1 = \infty$ and $\theta_2 = \infty$.

Like Ferron and Foster-Johnson (1996), we exploit the fact that, when conditioning on $\mathbf{T}$, margin-based tests, whose form depends on $\mathbf{T}$, are valid (preserve the nominal Type I error rate).

Specifically, we restrict attention to those margins for which it is reasonable to peel each directed extreme point into the critical region and which have positive marginal totals. We thus require that $\max(T_1, T_2) < n_1 < T_1 + \min(T_2, T_3 + 1) < N = n_1 + n_2$. As shown in Section 5, these conditions are not so restrictive, and data with such margins are frequently encountered (an example is presented in Section 4). With this requirement, the directed extreme points become $W_1 = (T_1, \mathbf{0}, n_1 - T_1; \mathbf{0}, T_2, n_2 - T_2)$, $W_2 = (T_1, n_1 - T_1, \mathbf{0}; \mathbf{0}, n_2 - T_3, T_3)$, and $W_3 = (n_1 - T_2, T_2, \mathbf{0}; n_2 - T_3, \mathbf{0}, T_3)$ and are all distinct.

Any pair of cells not in the same column can be used, instead of (X_1, X_2), to uniquely identify a table (and this transformation preserves directed extremity). Thus, each directed extreme point is uniquely determined by its pattern of zero cell counts and has probability one under any alternative that specifies structural zeroes for its zero cell counts. This implies that a nonrandomized test will have power one or zero to detect alternatives of the form $\pi = (\pi_{1,1}, 0, \pi_{1,3}; 0, \pi_{2,2}, \pi_{2,3})$, $\pi = (\pi_{1,1}, \pi_{1,2}, 0; 0, \pi_{2,2}, \pi_{2,3})$, or $\pi = (\pi_{1,1}, \pi_{1,2}, 0; \pi_{2,1}, 0, \pi_{2,3})$, respectively, as its critical region does or does not contain W_1, W_2, or W_3. If θ_2/θ_1 is held constant and θ_1 gets large, $\boldsymbol{\theta}$ must tend to a limit that corresponds to a set of cell probabilities of one of these three types, so having $D[\Gamma]$ in the critical region is essential for having positive limiting power as $\boldsymbol{\theta}$ moves away from $\mathbf{0}$ in various directions. Linear rank tests have convex rejection regions in (X_1, X_2), so if $D[\Gamma] \subset R(\alpha; \phi_\nu)$, $CH(D[\Gamma]) \subset R(\alpha; \phi_\nu)$, where $CH(D[\Gamma])$ is the convex hull of $D[\Gamma]$. When $P_0\{CH(D[\Gamma]) \mid \mathbf{T}\} > \alpha$, no α-level linear rank test has positive limiting power as $\boldsymbol{\theta}$ moves away from $\mathbf{0}$ in all directions.

3. Derivative of the Convex Hull Test

Convex hull tests (Berger, 1998) are derived iteratively by constructing the critical region layer by layer with a recursive algorithm. The simplest convex hull test, proposed here as ϕ_{CH}, is constructed as follows. First, $D[\Gamma]$ is placed in the critical region. Then the reduced sample space $\Gamma_1 = \Gamma - D[\Gamma]$ is defined, and $D[\Gamma_1]$ is added to the critical region. This directed convex peeling process is continued until the desired size is achieved. The formal algorithm (arbitrarily selecting the sample size N as the value of the test statistic for the first peel) is as follows:

Step 0: Initialize each point in Γ by setting $Z_0(\mathbf{X}) = 0$ for all $\mathbf{X} \in \Gamma$.

Step 1: Let $Z_1(\mathbf{X}) = N$ (the sample size) if $\mathbf{X} \in D[\Gamma]$ and $Z_1(\mathbf{X}) = Z_0(\mathbf{X}) = 0$ if $\mathbf{X} \in \Gamma_1 = \Gamma - D[\Gamma]$.

Step $k+1$: $Z_{k+1}(\mathbf{X}) = N - k$ if $\mathbf{X} \in D[\Gamma_k] = D\{\mathbf{X} \mid Z_k(\mathbf{X}) = 0\}$ and $Z_{k+1}(\mathbf{X}) = Z_k(\mathbf{X}) = 0$ if $\mathbf{X} \in \Gamma_{k+1} = \Gamma_k - D[\Gamma_k]$.

Final step: Letting $Z(\mathbf{X}) = \max_{1 \le k \le N} Z_k(\mathbf{X})$, the level-$\alpha$ critical region is $\{\mathbf{X} \in \Gamma \mid Z(\mathbf{X}) > C(\alpha)\}$.

4. Exact Conditional Power Calculations

The exact conditional power of ϕ_{CH}, $\phi_{0.5}$ (the t-test with equally-spaced scores), and the Smirnov test (ϕ_S) are compared by enumerating the points of Γ and finding the probability of the critical region for each test under a variety of alternatives. Example J of Table 1 of Emerson and Moses (1985, p. 306) is $\mathbf{X} = (8, 6, 4; 1, 7, 10)$, with row margins $\mathbf{n} = (18, 18)$ and column margins $\mathbf{T} = (9, 13, 14)$ satisfying our margin conditions. We find that $D[\Gamma]$ consists of $W_1 = (9, \mathbf{0}, 9; \mathbf{0}, 13, 5)$, $W_2 = (9, 9, \mathbf{0}; \mathbf{0}, 4, 14)$, and $W_3 = (5, 13, \mathbf{0}; 4, \mathbf{0}, 14)$, so $Z(9, 0) = Z(9, 9) = Z(5, 13) = N = 36$. We place $D[\Gamma]$ in the critical region and define the reduced sample space, $\Gamma_1 = \{\mathbf{X} \mid Z(\mathbf{X}) < 36\}$. Now $Z(\mathbf{X}) = 35$ on $D[\Gamma_1]$, which consists of the points $\{(8, \mathbf{0}, 10; \mathbf{1}, 13, 4), (9, \mathbf{1}, 8; \mathbf{0}, 12, 6), (8, 10, \mathbf{0}; \mathbf{1}, 3, 14), (9, 8, \mathbf{1}; \mathbf{0}, 5, 13), (6, 12, \mathbf{0}; 3, \mathbf{1}, 14), (4, 13, \mathbf{1}; 5, \mathbf{0}, 13)\}$, and $Z(\mathbf{X}) = 34$ on $D[\Gamma_2] = \{(7, 0), (9, 2), (7, 11), (9, 7), (3, 13)\}$. Notice that $D[\Gamma_1]$ can be found as the set of points that start with a point of $D[\Gamma]$ and replace a zero with a one, $D[\Gamma_2]$ is the set of points that replace a zero from a point of $D[\Gamma]$ with a two, $D[\Gamma_3]$ is the set of points that either replace a zero with a three or replace both zeroes with ones, and so on. Figure 1 shows the derivation of ϕ_{CH} with $Z(\mathbf{X}) - 27$, instead of $Z(\mathbf{X})$, plotted on each point, so the points labeled 9 are peeled first, then 8, and so on. At significance level $\alpha = 0.025$, ϕ_{CH} rejects H when $Z(\mathbf{X}) - 27 > 1$ and has to randomize, to attain the exact significance level, when $Z(\mathbf{X}) - 27 = 1$.

In contrast, $\phi_{0.5}$ rejects H when $2X_1 + X_2 > 20$ and randomizes when $2X_1 + X_2 = 20$, and ϕ_S rejects H when either $X_1 \ge 8$ or $X_1 + X_2 > 14$ and randomizes when $X_1 + X_2 = 14$. The auxiliary randomization required to obtain exact significance levels is ancillary and thus is conditioned upon. The power of the resulting conservative (nonrandomized) versions of each test (randomization points are not placed in the critical region) is given in Table 1. While Ω requires $\theta_1 > 0$, alternatives for which $\theta_1 = 0$ are included as limits of points in Ω (power functions are continuous). Table 1 reveals that $\phi_{0.5}$ is severely biased (with power much less than $\alpha = 0.025$), at $\theta = (1, -4)$, e.g., as

Table 1

Exact conditional power of the conservative (nonrandomized) versions of the tests, $\alpha \leq 0.025$. Data from Emerson and Moses (1985).

Theta	Envelope	$\phi_{0.5}$	ϕ_S	ϕ_{CH}	Theta	Envelope	$\phi_{0.5}$	ϕ_S	ϕ_{CH}
$(0,-\infty)$	1.000	**0.000**	1.000	1.000	$(1,0)$	0.222	0.141	0.155	0.173
$(0,-4)$	0.999	**0.000**	0.260	0.966	$(1,1)$	0.287	0.199	0.191	0.121
$(0,-3)$	0.978	**0.001**	0.190	0.835	$(1,2)$	0.708	0.230	0.433	0.293
$(0,-2)$	0.784	**0.003**	0.104	0.494	$(1,3)$	0.956	0.236	0.679	0.663
$(0,-1)$	0.289	**0.009**	0.039	0.130	$(1,4)$	0.997	0.237	0.825	0.909
$(0,0)$	0.025	0.017	0.016	0.017	$(2,-4)$	1.000	**0.001**	0.908	1.000
$(0,1)$	0.289	0.026	0.057	0.037	$(2,-3)$	0.999	**0.012**	0.881	0.996
$(0,1.3)$	0.442	0.028	0.088	0.072	$(2,-2)$	0.986	0.079	0.822	0.965
$(0,1.5)$	0.550	0.030	0.114	0.108	$(2,-1)$	0.888	0.240	0.711	0.818
$(0,1.6)$	0.602	0.030	0.128	0.130	$(2,0)$	0.627	0.429	0.546	0.569
$(0,1.7)$	0.652	0.031	0.143	0.156	$(2,1)$	0.612	0.567	0.485	0.440
$(0,1.8)$	0.700	0.031	0.160	0.184	$(2,2)$	0.773	0.633	0.682	0.509
$(0,1.9)$	0.744	0.032	0.177	0.216	$(2,3)$	0.961	0.641	0.874	0.736
$(0,2)$	0.784	0.032	0.194	0.250[a]	$(2,4)$	0.997	0.621	0.955	0.929
$(0,2.1)$	0.820	0.033	0.212	0.286[a]	$(3,-4)$	1.000	**0.002**	0.983	1.000
$(0,2.2)$	0.851	0.033	0.231	0.325	$(3,-3)$	1.000	**0.019**	0.977	1.000
$(0,2.3)$	0.879	0.034	0.250	0.365	$(3,-2)$	0.998	0.130	0.963	0.994
$(0,2.4)$	0.902	0.034	0.270	0.406	$(3,-1)$	0.980	0.416	0.931	0.961
$(0,2.5)$	0.922	0.034	0.289[a]	0.448	$(3,0)$	0.898	0.696	0.867	0.876
$(0,2.7)$	0.952	0.035	0.328	0.531	$(3,1)$	0.890	0.848	0.807	0.799
$(0,3)$	0.978	0.036	0.384	0.648	$(3,2)$	0.934	0.906	0.869	0.810
$(0,4)$	0.999	0.041	0.534	0.900	$(3,3)$	0.975	0.918	0.958	0.876
$(0,\infty)$	1.000	0.056	0.673	1.000	$(3,4)$	0.998	0.906	0.990	0.956
$(1,-4)$	1.000	**0.000**[a]	0.655	0.996	$(\infty,0)$	1.000	0.934	1.000	1.000
$(1,-3)$	0.995	**0.005**	0.585	0.971	(∞,∞)[b]	1.000	1.000	1.000	1.000
$(1,-2)$	0.925	0.026	0.461	0.831					
$(1,-1)$	0.610	0.075	0.295	0.475					

Bold entries represent points at which the given test is biased or has power below 0.025.

[a] The point of maximum shortcoming or difference from the envelope power.

[b] The limit of $\theta = (k,k)$ as k gets large.

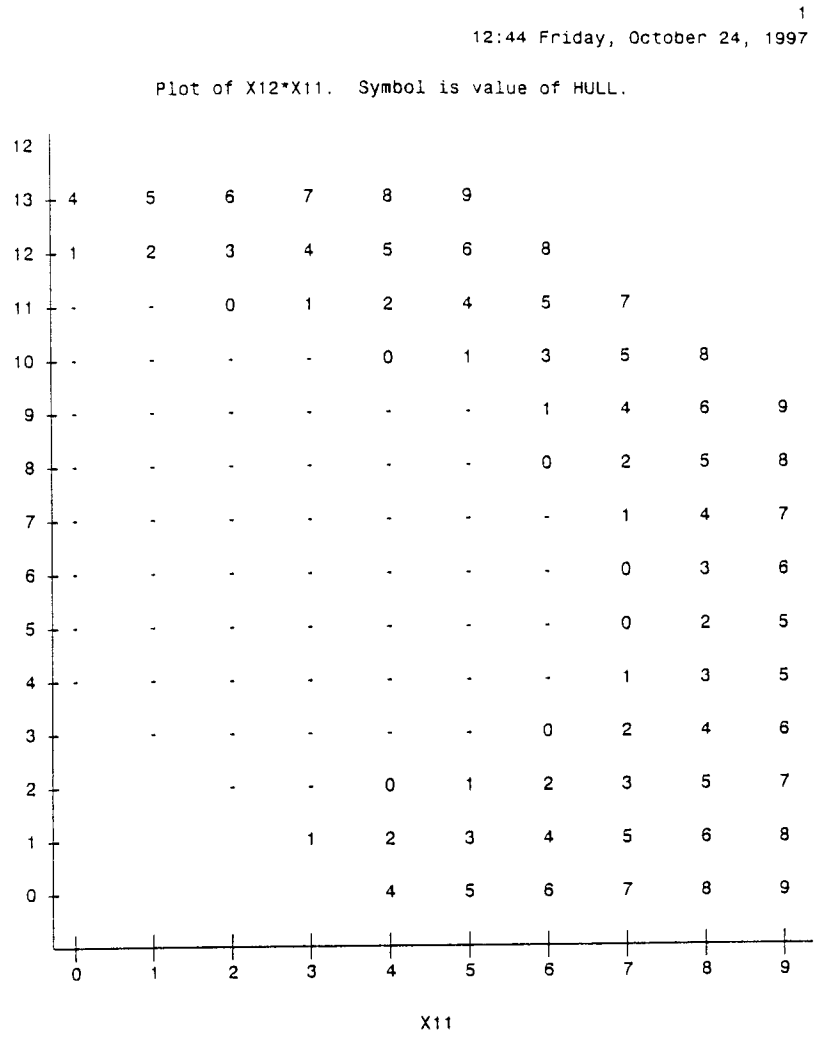

Figure 1. The derivation of the convex hull test on (8, 6, 4; 1, 7, 10). Data from Emerson and Moses (1985). The first peel is labeled 9, the second peel is labeled 8, and so on.

is predictable from the fact that it fails to reject H on W_1. As expected, $\phi_{0.5}$ is most powerful on and near $\Omega_{0.5}$ ($\theta_1 = 2\theta_2$) and fails to attain the envelope power function (maximal attainable power of any level-α test) on $\Omega_{0.5}$ only because of the conservatism induced by discreteness.

The maximum shortcomings (largest difference in power compared to the envelope power function) over the range of alternatives considered (including a grid search to find the maximum shortcomings of each test) are 0.633 for ϕ_S at $(0.0, 2.5)$, 0.534 for ϕ_{CH} at $(0.0, 2.0)$, and 1.00 for $\phi_{0.5}$ at $(1, -4)$, indicating that ϕ_{CH} appears to be most stringent (the desirable property of minimizing the maximum shortcoming) among these three tests and $\phi_{0.5}$ is least stringent (a most undesirable property) among all tests, as in Berger and Ivanova (1997). Notice that not only does $\phi_{0.5}$ have poor power when $\theta_2 < 0$ but also when $\theta_2 > \theta_1$ and that this poor power persists even in the limit as θ_1 gets large. The power of ϕ_S and ϕ_{CH} exceeds 0.025 for all alternatives considered, so these tests appear to be unbiased based on this analysis. In fact, these tests are not unbiased, by virtue of their power functions having negative directional derivatives at the null point, but having a power function dip slightly below α close to the null point seems less worrisome than having limiting power of 0.

When $\theta_1 = 0$, ϕ_{CH} is more powerful than ϕ_S for $\theta_2 \geq 1.6$ or $\theta_2 \leq 0$ but less powerful than ϕ_S for $1 \leq \theta_2 \leq 1.5$. When $\theta_1 = 1$, ϕ_{CH} is more powerful than ϕ_S for $\theta_2 \geq 4$ or $\theta_2 \leq 0$ but less powerful than ϕ_S for $1 \leq \theta_2 \leq 3$. When $\theta_1 = 2$ or 3, ϕ_{CH} is more powerful than ϕ_S for $\theta_2 \leq 0$ but less powerful than ϕ_S for $1 \leq \theta_2 \leq 4$. While ϕ_S tends to have somewhat better power than ϕ_{CH}, by as much as 0.173 at $(2, 2)$, when $\theta_1 > 0$ and $\theta_2 > 0$ are close, ϕ_{CH} may have much better

Table 2

Simulated unconditional power of the conservative (nonrandomized) versions of the equally spaced scores linear rank test, the Smirnov test, and the convex hull test, $\alpha \leq 0.025$, 20 observations per row, keeping only data satisfying the margin conditions

π_2	$\phi_{0.5}$	ϕ_S	ϕ_{CH}	Reps[a]	π_2	$\phi_{0.5}$	ϕ_S	ϕ_{CH}	Reps[a]
	$\pi_1 = (0.1, 0.8, 0.1)$					$\pi_1 = (0.3, 0.4, 0.3)$			
(0.0, 0.4, 0.6)	1.00	0.98	1.00	42	(0.0, 0.2, 0.8)	1.00	0.97	1.00	587
(0.1, 0.1, 0.8)	0.89	1.00	0.98	208	(0.0, 0.3, 0.7)	0.95	0.85	0.97	249
(0.1, 0.2, 0.7)	0.93	0.98	0.95	246	(0.0, 0.4, 0.6)	0.86	0.73	0.96	184
(0.1, 0.3, 0.6)	0.75	0.96	0.77	690	(0.0, 0.5, 0.5)	0.71	0.65	0.92	178
(0.1, 0.4, 0.5)	0.72	0.85	0.62	47	(0.0, 0.6, 0.4)	0.68	0.74	0.94	254
(0.1, 0.5, 0.4)	0.50	0.67	0.50	12	(0.0, 0.7, 0.3)	0.49	0.78	0.97	489
	$\pi_1 = (0.2, 0.6, 0.2)$				(0.1, 0.1, 0.8)	0.84	0.93	0.77	565
(0.0, 0.3, 0.7)	0.99	0.94	0.97	236	(0.1, 0.2, 0.7)	0.72	0.67	0.61	242
(0.0, 0.4, 0.6)	0.97	0.82	0.97	299	(0.1, 0.3, 0.6)	0.73	0.56	0.58	150
(0.0, 0.5, 0.5)	0.88	0.68	0.93	516	(0.1, 0.4, 0.5)	0.55	0.43	0.49	135
(0.0, 0.6, 0.4)	0.66	0.57	0.81	74	(0.1, 0.5, 0.4)	0.28	0.25	0.32	162
(0.0, 0.7, 0.3)	0.58	0.67	0.92	12	(0.1, 0.6, 0.3)	0.24	0.31	0.33	211
(0.1, 0.1, 0.8)	0.89	0.99	0.91	243	(0.2, 0.0, 0.8)	0.64	0.92	0.99	555
(0.1, 0.2, 0.7)	0.82	0.89	0.76	166	(0.2, 0.1, 0.7)	0.53	0.71	0.59	231
(0.1, 0.3, 0.6)	0.75	0.80	0.69	174	(0.2, 0.2, 0.6)	0.45	0.59	0.34	135
(0.1, 0.4, 0.5)	0.55	0.51	0.50	236	(0.2, 0.3, 0.5)	0.19	0.20	0.15	129
(0.1, 0.5, 0.4)	0.46	0.38	0.40	448	(0.2, 0.4, 0.4)	0.16	0.13	0.16	127
(0.1, 0.6, 0.3)	0.33	0.20	0.26	76	(0.2, 0.5, 0.3)	0.06	0.09	0.10	154
(0.1, 0.7, 0.2)	0.16	0.21	0.21	19	(0.3, 0.0, 0.7)	0.32	0.72	0.94	242
(0.2, 0.0, 0.8)	0.70	0.99	1.00	225	(0.3, 0.1, 0.6)	0.20	0.40	0.47	151
(0.2, 0.1, 0.7)	0.66	0.98	0.88	135	(0.3, 0.2, 0.5)	0.13	0.18	0.18	118
(0.2, 0.2, 0.6)	0.41	0.76	0.58	139	(0.3, 0.3, 0.4)	0.12	0.10	0.12	113
(0.2, 0.3, 0.5)	0.29	0.44	0.30	142	(0.3, 0.4, 0.3)	0.03	0.02	0.02	121
(0.2, 0.4, 0.4)	0.22	0.29	0.15	229					
(0.2, 0.5, 0.3)	0.03	0.10	0.07	458					
(0.2, 0.6, 0.2)	0.03	0.03	0.04	75					

Table 2. *Continued*

π_2	$\phi_{0.5}$	ϕ_S	ϕ_{CH}	Reps[a]
	$\pi_1 = (0.4, 0.2, 0.4)$			
(0.0, 0.2, 0.8)	1.00	0.98	1.00	61
(0.0, 0.3, 0.7)	0.93	0.97	1.00	478
(0.0, 0.4, 0.6)	0.79	0.93	1.00	242
(0.0, 0.5, 0.5)	0.68	0.91	1.00	162
(0.0, 0.6, 0.4)	0.41	0.85	1.00	138
(0.1, 0.1, 0.8)	0.83	0.80	0.69	64
(0.1, 0.2, 0.7)	0.67	0.62	0.55	439
(0.1, 0.3, 0.6)	0.56	0.58	0.65	219
(0.1, 0.4, 0.5)	0.46	0.58	0.71	164
(0.1, 0.5, 0.4)	0.24	0.42	0.65	116
(0.2, 0.0, 0.8)	0.63	0.85	0.92	52
(0.2, 0.1, 0.7)	0.49	0.48	0.33	513
(0.2, 0.2, 0.6)	0.38	0.36	0.24	225
(0.2, 0.3, 0.5)	0.22	0.20	0.22	140
(0.2, 0.4, 0.4)	0.07	0.11	0.26	108
(0.3, 0.0, 0.7)	0.31	0.56	0.79	751
(0.3, 0.1, 0.6)	0.19	0.19	0.19	235
(0.3, 0.2, 0.5)	0.14	0.16	0.09	154
(0.3, 0.3, 0.4)	0.00	0.05	0.08	113
(0.4, 0.0, 0.6)	0.12	0.21	0.51	338
(0.4, 0.1, 0.5)	0.05	0.06	0.14	173
(0.4, 0.2, 0.4)	0.02	0.02	0.06	134
	$\pi_1 = (0.5, 0.5, 0.0)$			
(0.1, 0.4, 0.5)	1.00	0.99	1.00	154
(0.1, 0.5, 0.4)	1.00	0.95	1.00	226
(0.1, 0.6, 0.3)	1.00	0.98	1.00	454
(0.1, 0.7, 0.2)	1.00	0.97	1.00	70
(0.1, 0.8, 0.1)	1.00	0.91	1.00	11
(0.2, 0.3, 0.5)	0.98	0.98	1.00	124
(0.2, 0.4, 0.4)	0.96	0.90	1.00	148
(0.2, 0.5, 0.3)	0.97	0.85	1.00	247
(0.2, 0.6, 0.2)	0.92	0.70	0.94	464
(0.2, 0.7, 0.1)	0.75	0.52	0.78	40
(0.3, 0.2, 0.5)	0.84	0.99	1.00	118
(0.3, 0.3, 0.4)	0.90	0.94	1.00	132
(0.3, 0.4, 0.3)	0.78	0.70	0.95	180
(0.3, 0.5, 0.2)	0.67	0.48	0.82	380
(0.3, 0.6, 0.1)	0.53	0.30	0.55	94
(0.4, 0.2, 0.4)	0.63	0.90	0.99	151
(0.4, 0.3, 0.3)	0.57	0.70	0.97	170
(0.4, 0.4, 0.2)	0.53	0.43	0.77	263
(0.4, 0.5, 0.1)	0.24	0.06	0.46	722
(0.5, 0.1, 0.4)	0.52	0.94	1.00	208
(0.5, 0.2, 0.3)	0.31	0.64	0.99	275
(0.5, 0.3, 0.2)	0.19	0.32	0.77	324
(0.5, 0.4, 0.1)	0.14	0.08	0.40	862
	$\pi_1 = (0.5, 0.0, 0.5)$			
(0.0, 0.5, 0.5)	0.62	0.99	1.00	210
(0.1, 0.2, 0.7)	0.78	0.87	1.00	55
(0.1, 0.3, 0.6)	0.61	0.89	1.00	448
(0.1, 0.4, 0.5)	0.42	0.89	1.00	235
(0.2, 0.1, 0.7)	0.57	0.71	0.80	51
(0.2, 0.2, 0.6)	0.39	0.59	0.88	535
(0.2, 0.3, 0.5)	0.25	0.60	0.94	232
(0.3, 0.1, 0.6)	0.19	0.26	0.50	90
(0.3, 0.2, 0.5)	0.14	0.30	0.72	338
(0.4, 0.1, 0.5)	0.06	0.11	0.26	643

[a] The number presented in the column entitled "Reps" gives either the number of replications required to obtain 100 replications satisfying the margin conditions (if it is at least 100) or the number of replications satisfying the margin conditions out of 1000 total replications (if it is under 100).

power, by as much as 0.706 at $(0, -4)$, than ϕ_S otherwise. Notice that, when $\theta_1 = \theta_2$, the power advantage of ϕ_S over ϕ_{CH} is not monotonic in this common value. Note also that, when $\theta_1 = 0$, the envelope power is symmetric about $\theta_2 = 0$, but, when $\theta_1 > 0$, no such symmetry exists and the envelope power is actually larger when $\theta_2 < 0$. Also, for fixed values of θ_2, the envelope power is not monotonic in θ_1.

5. Simulated Unconditional Power Calculations

To explore the unconditional power of the conservative nonrandomized version of each test, we used various sets of cell probability vectors to generate simulated data. The row margins were chosen to be 20 each, allowing for 861 possible sets of column margins (T_1 ranges from 0 to 40 and T_2 ranges from 0 to $40 - T_1$), of which only the 171 (T_1 ranges from 2 to 19 and T_2 ranges from $21 - T_1$ to 19) meeting our conditions are considered. For some sets of cell probability vectors, such as $(0.4, 0.2, 0.4; 0.2, 0.4, 0.4)$, the probability of obtaining one of these 171 sets of margins is close to one. There were either 1000 total replications (in which case the number of replications with the required margin structure is shown and is no larger than 100) or 100 replications with the required margin structure (in which case the total number of replications is shown and is between 100 and 1000). For π_1 as each of $(0.1, 0.8, 0.1)$, $(0.2, 0.6, 0.2)$, $(0.3, 0.4, 0.3)$, $(0.4, 0.2, 0.4)$, $(0.5, 0.5, 0.0)$, $(0.5, 0.0, 0.5)$, and $(0.0, 0.5, 0.5)$, we let π_2 range freely subject to being an integer divided by 10, provided that $\pi \in \mathrm{H}^* \cup \mathrm{K}^*$ and that there was at least 1 replication out of 1000 with the required margin structure.

For brevity, cases for which ϕ_S and ϕ_{CH} each had power 1.00 were deleted (we note that, for $\pi = (0.5, 0.0, 0.5; 0.0, 0.4, 0.6)$, 552 replications were required to obtain 100 with the required margin structure, and the power of ϕ_S and ϕ_{CH} each were 1.00, while the power of $\phi_{0.5}$ was only 0.73). The simulation shows that ϕ_{CH} tends to have better unconditional power than ϕ_S, especially when π_1 is roughly uniform or $\theta_2 < 0$, with uniform superiority when there is a structural zero. Also, as with the exact conditional power calculations, when ϕ_S was more powerful than ϕ_{CH}, it tended not to be so by much, whereas when ϕ_{CH} was more powerful than ϕ_S, it tended to be much more powerful, as shown in Table 2.

6. Discussion

By declaring significance only when the observed outcome is one of the $100\alpha\%$ most extreme of all outcomes obtained by permuting the treatment assignments, permutation tests have size exactly (or, without auxiliary randomization, no greater than) α for all null parameter values. What constitutes an extreme observation may depend on both the margins and the direction, θ_2/θ_1. It is reasonable, therefore, to select the form of the test based on the margins. For margins that satisfy $\max(T_1, T_2) < n_1 < T_1 + \min(T_2, T_3 + 1) < N = n_1 + n_2$, there is a different most powerful test for each direction, each of which may have poor (or no) power to detect alternatives in other directions. The failure of any one linear test statistic to capture extremity for all directions transcends the realm of hypothesis testing and has implications even for data description and estimation. The proportional hazards and proportional odds tests (McCullagh, 1980) are admissible (Berger, 1998), but the boundaries of their critical regions are well approximated by linear functions, and thus they may still have zero power to detect certain alternatives of interest. The Smirnov test always has positive power, and nonrandomized Smirnov tests are admissible at their actual (conservative) level. However, like linear rank tests, randomized Smirnov tests tend to be inadmissible, indicating that the nonrandomized versions of both tests are overly conservative.

By repeatedly improving the trivially unbiased "ignore-the-data test" (which randomizes on each point of Γ with the same probability), Berger and Sackrowitz (1997) constructed the first test for this problem that is simultaneously admissible and unbiased. This is a rather general construct for obtaining tests with good power properties. However, it is difficult to obtain a nested family of critical regions, meaning that observed data may be significant at one significance level yet not at a larger significance level. A second approach to constructing tests with good power properties is to estimate the direction for each point in Γ. Then the p-value of the most powerful test for the estimated direction is computed for each point in Γ, and the p-value for the observed data is compared to its null distribution. Berger (1998) proved the admissibility of such an adaptive test when the direction is estimated by finding the linear rank test that minimizes the p-value for the given data, using the algorithm of Kimeldorf et al. (1992). However, the computations involved in this approach can be complicated.

Convex hull peeling is a third general construct, which is based on the directed extreme points, i.e., on those points that are most extreme for some direction. The convex hull test is admissible

(Berger, 1998), has power tending to one as θ moves away from 0 in any direction for α as small as $P_0\{D[\Gamma] \mid \mathbf{T}\}$ (which is the minimum α-level required for any test to have this property), and palindromic invariant (McCullagh, 1980) to reversals of both rows and columns. Palindromic invariance is seen as follows. The reversal maps $\boldsymbol{\theta} = (\theta_1, \theta_2)$ into $R(\boldsymbol{\theta}) = (\theta_1, \theta_1 - \theta_2)$ and $\mathbf{X} = (X_1, X_2)$ into $R(\mathbf{X}) = (Y_3, Y_2) = (T_3 - n_1 + X_1 + X_2, T_2 - X_2)$. Either both $\boldsymbol{\theta}$ and $R(\boldsymbol{\theta})$ or neither are in Ω. Further, $\Lambda_{R(\boldsymbol{\theta})}(R(\mathbf{X})) = \Lambda_{\boldsymbol{\theta}}(\mathbf{X})$, and the points of $A \subset \Gamma$ that uniquely maximize one of these likelihood ratios will also maximize the other, so $D[A]$ can be defined in an unambiguous manner. Linear rank tests require the significance level to be at least $P_0\{CH(D[\Gamma]) \mid \mathbf{T}\}$ to have limiting power of one in all directions. The convex hull test is qualitatively similar to the test that is uniformly more powerful than the Smirnov test (Berger and Sackrowitz, 1997) because the improvements to the Smirnov test are based on transfer of rejection mass to the directed extreme points.

Our power calculations show that the convex hull test tends to be much more powerful than the Smirnov test when θ_1 or θ_2 is large or when $\theta_2 < 0$. When the Smirnov test is more powerful than the convex hull test, which tends to be the case when θ_1 and θ_2 are comparable in size and not too far from zero, the difference in power tends to be rather modest. Nonlinear rank tests can be exploited to allow for smaller studies (exposing fewer patients to a potentially harmful or ineffective treatment), with roughly the same power as one would obtain with linear rank tests. It must be borne in mind when interpreting this statement that the sample size for a nonlinear rank test (such as the Smirnov test) will generally be larger than that for a linear test because the latter is computed under the optimistic assumption that the direction of the effect is known. If it is not, the power of the linear rank test may be much less than expected.

When one can predict the direction of an effect, the convex hull test will not be much worse than the linear rank test, which is seen as follows. Recall that $R(\alpha; \phi)$ is the critical region of ϕ at level α. Let α_1 be the smallest number such that $R(\alpha; \phi_{CH}) \subset R(\alpha + \alpha_1; \phi_\nu)$ and let α_2 be the smallest number such that $R(\alpha - \alpha_2; \phi_\nu) \subset R(\alpha; \phi_{CH})$. Both α_1 and α_2 will depend on α, ν, and $\mathbf{X}$, but in general, α_1 will be large and α_2 will be small. By using the convex hull test even when one expects that the direction of the effect is known, one can pay a small premium (α_2 of the critical region of the linear rank test is taken away) to receive a large return (an additional α_1 would be required for the linear rank test to have the same sensitivity in all directions as the convex hull test) in case the direction did not turn out as planned. The convex hull test would still have good power properties when there are more than three columns except that Γ could then not be plotted as it was in this paper. It seems reasonable to expect that the convex hull peeling approach can offer globally powerful tests for other problems involving discrete data, composite alternatives, and no monotone likelihood ratio. However, for a convex hull test to be unbiased, it would need to account for both the null probabilities of the points being peeled into the critical region and the correlation between the cell counts X_1 and X_2. This would add an extra level of complexity to the construction.

Acknowledgements

This research was supported in part by grant RSR-96-004A from the Center for Drug Evaluation and Research of the Food and Drug Administration. The authors are very much obliged to Susan Ellenberg, Terry Neeman, Marvin Podgor, William Rosenberger, Knut Wittkowski, Dan Zelterman, and the anonymous referees and editors for useful comments and suggestions.

Résumé

Quand on fait le test d'un ordre stochastique pour un tableau de contingence ordonné $2 * J$, il y a quelques ensembles de marges pour lesquelles il existe un test optimal ou des tests de rang linéaire ayant une bonne puissance globale. Pour la majorité, néanmoins, les tests de rang linéaire ont une faible puissance pour détecter des alternatives éloignées de la région étroite dans laquelle ils sont optimaux. La surface de l'enveloppe convexe de l'espace de permutation échantillonné peut être utilisée pour déduire un test globalement puissant.

References

Behnen, K. and Neuhaus, G. (1989). *Rank Tests with Estimated Scores and Their Applications*. Stuttgart: Teubner.

Berger, V. (1998). Admissibility of tests of stochastic order. *The Journal of Statistical Planning and Inference* **66**, 39–50.

Berger, V. W. (1995). *Testing for Stochastic Order in Contingency Tables*. Ann Arbor, Michigan: UMI Dissertation Services.

Berger, V. and Ivanova, A. (1997). *The conditional t-test of stochastic order is biased and least stringent for the product multinomial distribution*. Technical Report TR97-05, Department of Mathematics and Statistics, University of Maryland–Baltimore County, Baltimore.

Berger, V. and Sackrowitz, H. (1997). Improving tests for superior treatments in contingency tables. *The Journal of the American Statistical Association* **92**, 700–705.

Chakraborti, S. and Schaafsma, W. (1996). On the choice of scores in contingency tables. *Proceedings of the Biometrics Section of the American Statistical Association*, 329–333.

Cohen, A. and Sackrowitz, H. B. (1996). *New testing methodology for one-sided alternatives in discrete multivariate models*. Technical Report 96-017, Department of Statistics, Rutgers University, New Brunswick, New Jersey.

Eddy, W. F. (1982). Convex hull peeling. *COMPSTAT* **82**, 42–47.

Emerson, J. D. and Moses, L. E. (1985). A note on the Wilcoxon–Mann–Whitney test for $2 \times k$ ordered tables. *Biometrics* **41**, 303–309.

Ferron, J. and Foster-Johnson, L. (1996). Randomization tests without predetermined test statistics. *XXVI International Congress of Psychology*.

Gautam, S. (1997). Test for linear trend in $2 \times K$ ordered tables with open-ended categories. *Biometrics* **53**, 1163–1169.

Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered $2 \times k$ contingency tables. *Biometrics* **43**, 471–476.

Green, P. J. (1985). Peeling data. In *The Encyclopedia of Statistical Sciences*, Volume 6, S. Kotz, N. L. Johnson, and C. B. Read (eds), 660–664. New York: Wiley.

Hilton, J. F., Mehta, C. R., and Patel, N. R. (1994). An algorithm for conducting exact Smirnov tests. *Computational Statistics and Data Analysis* **17**, 351–361.

Kimeldorf, G., Sampson, A. R., and Whitaker, L. R. (1992). Min and max scorings for two-sample ordinal data. *Journal of the American Statistical Association* **87**, 241–247.

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

Moses, L. E., Emerson, J. D., and Hosseini, H. (1984). Analyzing data from ordered categories. *New England Journal of Medicine* **311**, 442–448.

Nikiforov, A. M. (1994). Exact Smirnov two-sample tests for arbitrary distributions. *Applied Statistics* **43**, 265–284.

Petitjean, M. and Saporta, G. (1992). On the performance of peeling algorithms. *Applied Stochastic Models and Data Analysis* **8**, 91–98.

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

Rahlfs, V. W. and Zimmermann, H. (1993). Scores: Ordinal data with few categories—How should they be analyzed? *Drug Information Journal* **27**, 1227–1240.

Sharp, T. J. and Koch, G. G. (1996). Some bivariate strategies for extended Mantel–Haenszel tests. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 288–293.

Received December 1996; *revised September and November* 1997; *accepted December* 1997.