Drawbacks of Integer Scoring of Ordered Categorical Data

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

[1] Department of Biostatistics, CB #7400

The University of North Carolina at Chapel Hill,

Chapel Hill, NC 27599-740, U.S.A.

[2] Biometry Research Group, DCP, NCI, Executive Plaza North, Suite 344

Bethesda, MD 20892-7354. U.S.A.

SUMMARY

Linear rank tests are widely used when testing for independence against stochastic order in a $2 \times J$ contingency table with two treatments and J ordered outcome levels. For this purpose, numerical scores are assigned to the outcome levels. When the choice of scores is not apparent, integer (equally-spaced) scores are often considered. We show that this practice leads to unnecessarily conservative tests. The use of slightly perturbed scores will result in a less conservative and more powerful test.

Corresponding author's e-mail address: aivanova@bios.unc.edu

Key words: Conservatism; Contingency Table; Permutation Test.

1. Introduction

Contingency tables with ordered categories are common in biostatistics. For example, Moses, Emerson, and Hosseini (1984) reported that ordered categorical data occurred in 32 out of 168 articles in Volume 36 (1982) of the *New England Journal of Medicine*. It is known that collapsing categories will result in a loss of power (Emerson and Moses, 1985). Tests which ignore the ordering among the categories may also lack adequate statistical power. To make use of the ordering, numerical scores are often assigned to the outcome levels, sometimes based on subject matter considerations (e.g., midpoints of the category intervals). When there is no indication of what scores to assign, Graubard and Korn (1987) argued for considering equally-spaced scores. In this paper we show that this choice makes the test excessively conservative, and slightly perturbed scores often lead to a uniformly more powerful test. In Section 2 we start with an example of a 2×3 table with ordered categories. In Section 3 we show that certain sets of scores result in excessive conservatism. In particular, the test is generally conservative when equally-spaced scores are chosen. We illustrate the point using the example, and then present power comparisons. In Section 4 we discuss the Wilcoxon rank-sum test, and other tests whose reliance on assigning scores is rarely made explicit. In Section 5 we generalize to $R \times C$ tables. In Section 6 we propose a new test that is an improvement of the Smirnov test in the same sense that the linear rank test with slightly perturbed scores is an improvement of the linear rank test with integer scores.

2. Example

Consider a response variable with three ordered outcome levels. Pneumonia status following treatment in a two-arm randomized clinical trial, e.g., may be classified as cured, improved, or failed, and summarized as a 2×3 contingency table with two treatments and three ordered outcome levels. Table 1 presents a hypothetical data set of this form.

Table 1

Hypothetical pneumonia status data.

	Failed	Improved	Cured	Total
Control	$C_{11} = 5$	$C_{12} = 3$	$C_{13} = 2$	$n_1 = 10$
Treatment	$C_{21} = 1$	$C_{22} = 4$	$C_{23} = 5$	$n_2 = 10$
Total	$T_1 = 6$	$T_2 = 7$	$T_3 = 7$	$N = 20$

Consider the problem of testing the null hypothesis of independence between rows and columns against the one-sided alternative of stochastic order. Let the vectors of cell probabilities (each summing to one) be $\boldsymbol{\pi_1} = (\pi_{11}, \pi_{12}, \pi_{13})$ and $\boldsymbol{\pi_2} = (\pi_{21}, \pi_{22}, \pi_{23})$, respectively. The corresponding trinomial random vectors (summing to n_1 and n_2, respectively) are $\boldsymbol{C_1} = (C_{11}, C_{12}, C_{13})$ and $\boldsymbol{C_2} = (C_{21}, C_{22}, C_{23})$. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$. The row margins $\boldsymbol{n} = (n_1, n_2)$ are fixed by design (product multinomial sampling). We condition on $\boldsymbol{T} = (T_1, T_2, T_3)$. The sample space Γ is the set of 2×3 contingency tables with non-negative integer-valued cell counts with row totals $\boldsymbol{n}$ and column totals $\boldsymbol{T}$. Given $\boldsymbol{T}, \boldsymbol{n}$, and $\boldsymbol{c} = (C_{11}, C_{12})$, we reconstruct the entire 2×3 contingency table as $C_{13} = n_1 - C_{11} - C_{12}$ and $\boldsymbol{C}_2 = \boldsymbol{T} - \boldsymbol{C}_1$, so we let $\boldsymbol{c}$ denote a point of Γ. For the data in our example, Γ consists of 44 tables. Figure 1 displays C_{12} plotted against C_{11} for each of these 44 tables. The conditional null probability of each table can be calculated using the hypergeometric distribution. The exact conditional linear rank test with scores (v_1, v_2, v_3), $v_1 \leq v_2 \leq v_3$, will order tables in the reference set according to the difference, $A_1 - A_2$, between two weighted sums:

$$A_1 = C_{11}v_1 + C_{12}v_2 + C_{13}v_3 \text{ and } A_2 = C_{21}v_1 + C_{22}v_2 + C_{23}v_3,$$

rejecting H_0 for tables with large values of $A_2 - A_1$. Without loss of generality the scores can be chosen as $(0, v, 1)$, with $v = (v_2 - v_1)/(v_3 - v_1)$, $0 \leq v \leq 1$. It can be shown that the linear rank test above is equivalent to test φ_v, that rejects H_0 for large values of $z_v(\boldsymbol{c}) = C_{11} + (1 - v)C_{12}$. In the sequel we consider the class of exact level-α linear rank tests with different scores: $\{\varphi_v : 0 \leq v \leq 1\}$, or even $\{\varphi_v : v \in R^1\}$. For example, if the three categories are assumed to be equally-spaced, then $\varphi_{0.5}$, the test with equally-spaced (integer) scores $(1,2,3)$, or $(0.0, 0.5, 1.0)$, is considered. The test statistic is $C_{11} + 0.5C_{12}$.

[Figure 1]

3. The conservatism of the linear rank test with integer scores

Let $M_v(\boldsymbol{c}) = \{\boldsymbol{c}^* \in \Gamma \mid z_v(\boldsymbol{c}^*) \geq z_v(\boldsymbol{c})\}$ denote the φ_v extreme region of $\boldsymbol{c}$, with $p_v(\boldsymbol{c}) = P_0\{M_v(\boldsymbol{c})|\boldsymbol{T}\}$ the corresponding p-value, where P_0 is the probability under the null hypothesis. Clearly $p_v(\bullet)$ is a monotonic set function for any v, so if $M_v(\boldsymbol{c}) \subset M_v(\boldsymbol{c}^*)$, then $p_v(\boldsymbol{c}) \leq p_v(\boldsymbol{c}^*)$. In the first panel of Figure 1, $M_{0.5}(5,3)$ is shown by dark dots. Fix $\boldsymbol{c} = (C_{11}, C_{12})$, and consider $\boldsymbol{c}^* = (C_{11}^*, C_{12}^*) \in \Gamma - \boldsymbol{c}$. Then $z_v(\boldsymbol{c}^*) = z_v(\boldsymbol{c})$ if and only if $v = v_{c,c^*} = 1 - (C_{11} - C_{11}^*)/(C_{12}^* - C_{12})$. Let $V(\boldsymbol{c}) = \{v_1(\boldsymbol{c}), v_2(\boldsymbol{c}), \ldots, v_{K_c}(\boldsymbol{c})\}$ be the ordered set of values v_{c,c^*}, as $\boldsymbol{c}^*$ ranges over $\Gamma - \boldsymbol{c}$. In our example, if $\boldsymbol{c} = (5,3)$, then $k_{(5,3)} = 25$ and

$$V(\boldsymbol{c}) = \{-\infty, -4, -3, -2, -\frac{3}{2}, -1, -\frac{2}{3}, -\frac{1}{2}, -\frac{1}{3}, -\frac{1}{4},$$
$$0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1, \frac{4}{3}, \frac{3}{2}, \frac{5}{3}, 2, \frac{5}{2}, 3, 4, 5, \infty\}.$$

If $v \in V(\boldsymbol{c})$, then there exists $\boldsymbol{c}^* \in \Gamma - \boldsymbol{c}$ such that $z_v(\boldsymbol{c}) = z_v(\boldsymbol{c}^*)$, so $\boldsymbol{c}^* \in M_v(\boldsymbol{c})$. If v were changed slightly to $v^* = v + \epsilon$ or $v^* = v - \epsilon$ so that $z_{v^*}(\boldsymbol{c}) > z_{v^*}(\boldsymbol{c}^*)$, then $\boldsymbol{c}^*$

would not be in $M_{v^*}(\boldsymbol{c})$, and would not inflate $p_{v^*}(\boldsymbol{c})$. We see that $p_v(\boldsymbol{c})$ is maximized locally when $v \in V(\boldsymbol{c})$.

THEOREM: If $v_k(\boldsymbol{c}) < v < v_{k+1}(\boldsymbol{c})$ for some k, then $p_{v_k(c)}(\boldsymbol{c}) \geq p_v(\boldsymbol{c})$ and $p_{v_{k+1}(c)}(\boldsymbol{c}) \geq p_v(\boldsymbol{c})$.

Proof. We first show that if $v_k(\boldsymbol{c}) < v < v_{k+1}(\boldsymbol{c})$, then $M_v(\boldsymbol{c})$ is independent of v and $M_v(\boldsymbol{c})$ is a subset of both $M_{v_k(c)}(\boldsymbol{c})$ and $M_{v_{k+1}(c)}(\boldsymbol{c})$. By the monotonicity of p, this suffices. The line through $\boldsymbol{c}$ with slope $1/(v-1)$ separates $M_v(\boldsymbol{c}) - \boldsymbol{c}$ from $\Gamma - M_v(\boldsymbol{c})$, and intersects with neither (because $v \notin V(\boldsymbol{c})$). See the second panel of Figure 1. Decreasing v will not change the sets $M_v(\boldsymbol{c}) - \boldsymbol{c}$ or $\Gamma - M_v(\boldsymbol{c})$ until $v = v_k(\boldsymbol{c})$. If $w_k(\boldsymbol{c}) = \{\boldsymbol{c}^* \in \Gamma - \boldsymbol{c} \mid z_{v_k(c)}(\boldsymbol{c}^*) = z_{v_k(c)}(\boldsymbol{c})\}$, then $M_{v_k(c)}(\boldsymbol{c}) = M_v(\boldsymbol{c}) \cup w_{v_k(c)}(\boldsymbol{c})$, and $w_{v_k(c)}(\boldsymbol{c}) \cap \{\Gamma - M_v(\boldsymbol{c})\}$ represents the set of points that migrated into the extreme region when v became $v_{v_k(c)}(\boldsymbol{c})$, making $M_{v_k(c)}(\boldsymbol{c})$ strictly larger than $M_v(\boldsymbol{c})$. Clearly, then, $p_{v_k(c)}(\boldsymbol{c}) \geq p_v(\boldsymbol{c})$. The same argument applies as v increases to $v_{k+1}(\boldsymbol{c})$. □

Because $0.5 = v_{14}(5,3) \in V(5,3)$, while $0.49 \notin V(5,3)$ and $0.51 \notin V(5,3)$, $\varphi_{0.5}$ assigns the same value of the test statistic to several points in the reference set, so each counts in the calculation of the p-value of all others with which it is tied. This makes the $\varphi_{0.5}$ critical region, and p-value, larger than those of $\varphi_{0.49}$ and $\varphi_{0.51}$ (Figure 1). Notice that $M_{0.49}(5,3) \subset M_{0.50}(5,3)$ and $M_{0.51}(5,3) \subset M_{0.50}(5,3)$ (the points in $M_{0.50}(5,3) - M_{0.49}(5,3)$ are marked by crosses). In fact, $p_{0.50}(5,3) = 0.0503$, $p_{0.49}(5,3) = 0.0490$, and $p_{0.51}(5,3) = 0.0383$, so the conservatism of $\varphi_{0.5}$ is enough to prevent statistical significance from being reached at $\alpha = 0.05$. The (null) expected p-values of these tests are $E_0[p_{0.5}|\boldsymbol{T}] = 0.576$ and $E_0[p_{0.49}|\boldsymbol{T}] = E_0[p_{0.51}|\boldsymbol{T}] = 0.538$. The larger critical region of $\varphi_{0.5}$ means that there are fewer points for which the $\varphi_{0.5}$ p-value is below 0.05 (or any other α-level). This makes the $\varphi_{0.50}$ critical region a proper subset of the $\varphi_{0.49}$ and $\varphi_{0.51}$ critical regions. Consequently, $\varphi_{0.5}$ is more conservative and less powerful than $\varphi_{0.49}$ and $\varphi_{0.51}$. Table 2 shows exact unconditional power comparisons of $\varphi_{0.5}$ to $\varphi_{0.49}$ and $\varphi_{0.51}$. We considered all 4356 tables with $n_1 = n_2 = 10$, and let $\pi_1 = (0.3, 0.4, 0.3)$, while π_2 varies.

[Table 2]

The last line of Table 2 shows that the actual sizes of $\varphi_{0.49}$ (0.034) and $\varphi_{0.51}$ (0.034) are closer to the nominal size of 0.05 than the actual size of $\varphi_{0.5}$ (0.026). This excessive conservatism of $\varphi_{0.5}$ is reflected in the power calculations: Specifically, both $\varphi_{0.49}$ and $\varphi_{0.51}$ are uniformly more powerful than $\varphi_{0.5}$. Note that $0 = v_{11}(5,3)$ and $1 = v_{16}(5,3)$ are also in $V(5,3)$, making binary tests on collapsed categories (φ_0 is the analysis of cure rates and φ_1 is the analysis of failure rates) overly conservative. Specifically, $E_0[p_{0.00}|\boldsymbol{T}] = 0.628$, $E_0[p_{-0.01}|\boldsymbol{T}] = E_0[p_{0.01}|\boldsymbol{T}] = 0.538$, and $p_{0.00}(5,3) = 0.175$, $p_{-0.01}(5,3) = 0.173$, $p_{0.00}(5,3) = 0.055$; $E_0[p_{1.00}|\boldsymbol{T}] = 0.633$, $E_0[p_{0.99}|\boldsymbol{T}] = E_0[p_{1.01}|\boldsymbol{T}] = 0.538$, and $p_{1.00}(5,3) = 0.070$, $p_{0.99}(5,3) = 0.038$, $p_{1.01}(5,3) = 0.062$. In practice, when linear rank tests are used, v is almost always chosen from $V = \cup_{c \in \Gamma} V(\boldsymbol{c})$, making φ_v overly conservative. Now $0.5 \in V$ for most, but not all sets of margins. If, e.g., $T_1 T_2 T_3 = 0$, then at least one column margin is zero, and there exists k such that $C_{11} + kC_{12}$ is constant on Γ. In this case, φ_v is the same test as φ_{v^*} provided that $v - k$ and $v^* - k$ have the same sign.

4. Tests that surreptitiously use scores

For the Cochran-Mantel-Haenszel test, scores are defined to be the row numbers and column numbers by default and hence are overly conservative integer scores. The uncritical data analyst may not even be aware of the fact that scores have been assigned by default. Other tests that rely on assignment of scores (that the user is rarely prompted to supply) include those based on correlation coefficients or ridits. The Wilcoxon rank-sum test (Emerson and Moses, 1985), when applied to this problem, is equivalent to a linear rank test with scores equal to midranks. In our example, the scores will be $v_1 = 3.5$ (since six observations are tied in the 'failed' category), $v_2 = 10$ (the midrank of ranks 7 – 13 for the 'improved' category), and $v_3 = 17$. The standardized midscore is $v = (10 - 3.5)/(17 - 3.5) = 0.482$. Graubard and Korn (1987) did not recommend the use of midrank scores for this problem because "midrank scores can be unreasonable in applications when the column margin is far from uniform". Of greater concern to us is that when the column margin is exactly uniform, i.e. $T_1 = T_2 = T_3$, the midrank scores will be exactly equivalent to the integer scores, and hence the test will be overly conservative.

5. *RxJ* contingency tables

We showed, in the case of a 2×3 contingency table with ordered responses, that the linear rank test with integer scores, or any middle score v from V, is overly conservative. The test based on a score close to v, but not exactly v, has a larger critical region, and is less conservative. Similar results hold for $R \times J$ contingency tables, for which the reference set Γ is $(R - 1)(J - 1)$ dimensional (to match the degrees of freedom), and the standardized v is a vector with $J - 2$ components. The reason is the same: Tests with integer scores assign the same value of the test statistic to several points in the reference set, so each counts in the calculation of the p-value of all others with which it is tied. The StatXact manual (1995, page 602) has, as an example, a comparison of five chemotherapy regiments in a small clinical trial. Tumor response was measured with three ordered response categories, 'None', 'Partial', and 'Complete'. The data are arranged as a 5×3 contingency table, with rows $(2,0,0)$, $(1,1,0)$, $(3,0,0)$, $(2,2,0)$, and $(1,1,4)$. The manual suggests assigning scores $(0,100,150)$, or equivalently $(0,\frac{2}{3},1)$, because these scores are "reasonable estimates of the number of weeks in remission following a response of the specified type". But $\frac{2}{3} \in V(\boldsymbol{c})$, and a slight change in the middle score (in either direction) can reduce the p-value slightly: $p_{\frac{2}{3}}(\boldsymbol{c}) = 0.0450$, $p_{0.66}(\boldsymbol{c}) = 0.0436$, and $p_{0.67}(\boldsymbol{c}) = 0.0446$. Note that the choice of scores should be made in advance and not on the basis of which set of scores will yield the smallest p-value. In this example we argue that either $(0,0.66,1)$ or $(0,0.67,1)$ should be chosen but not $(0,\frac{2}{3},1)$.

6. The Smirnov test

Because of some general deficiencies of linear rank tests, Berger, Permutt, and Ivanova (1998) suggested using nonlinear rank tests for this problem. The Smirnov test is the simplest nonlinear rank test, and uses as the test statistic the larger of $D_1 = C_{11}/n_1 - C_{21}/n_2$ and $D_2 = (C_{11} + C_{12})/n_1 - (C_{21} + C_{22})/n_2$, or equivalently $D_1 = C_{11} - T_1 n_1/n$ and $D_2 = C_{11} + C_{12} - (T_1 + T_2)n_1/n$. The test is a combination of φ_1 and φ_0, and retains the excessive conservatism of both of these components. A combination of $\varphi_{0.99}$ and $\varphi_{0.01}$ instead, with the test statistic the larger of $D_1 = C_{11} + 0.01C_{12} - T_1 n_1/n$ and $D_2 = C_{11} + 0.99C_{12} - (T_1 + T_2)n_1/n$, will be less conservative, and therefore more powerful.

7. Discussion

Berger, Permutt, and Ivanova (1998) provided conditions on the margins for which suitable nonlinear rank tests are preferable to linear rank tests. Nevertheless, linear rank tests remain popular and are often used. We confine attention to ordinal, and not interval-scaled, data so by definition there is no basis for the selection of scores and any set of scores must inherently be arbitrary. The column scores are neither data (observable from the sample) nor parameters (observable from the population), yet are said to be "correct" when they reflect the subject matter (Graubard and Korn, 1987). For example, suppose that the response to the question "How much would you pay, out of your own pocket, to be improved instead of experiencing a treatment failure?" would meet with an unqualified response of $\$M_1$. Likewise, suppose that one could assign a possibly different monetary value ($\$M_2$) to shifting from improved to cured. If $M_1 > 0$ and $M_2 > 0$ were known, then they would provide a clear basis for spacing the three outcome levels relative to each other, with column scores of $(0, M_1, M_1 + M_2)$, or, equivalently, $(0, M_1/[M_1 + M_2], 1)$. It would then be logical to select as preferable whichever treatment provides a larger mean score. The problem is that the monetary values, and consequently the sets of column scores, would vary both across individuals and within each individual over time. So the only reason to choose integer (equally-spaced) scores is that they "look good". We argue that, though somewhat less attractive, slightly perturbed integer scores will lead to better results. The scores should be chosen prospectively, but not as integer-scores. If a nonlinear rank test is to be used, then the Smirnov test with slightly perturbed scores, discussed in Section 6, might be considered, on the basis that is more powerful than the Smirnov test, but easier to compute than the uniformly improved Smirnov test (Berger and Sackrowitz, 1997), adaptive test (Berger, 1998), or convex hull test (Berger, Permutt, and Ivanova, 1998).

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., Permutt, T. and Ivanova A. (1998). The convex hull test for ordered categorical data. *Biometrics* **54**, 157-166.

Berger, V.W. and Sackrowitz, H. (1997). Improving tests for superior treatments in contingency tables. *Journal of American Statistical Association* **92**, 438, 700-705.

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

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

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

Mehta, C. and Patel, N. (1995). *StatXact* 3 *for Windows*. User Manual. CYTEL Software Corporation.

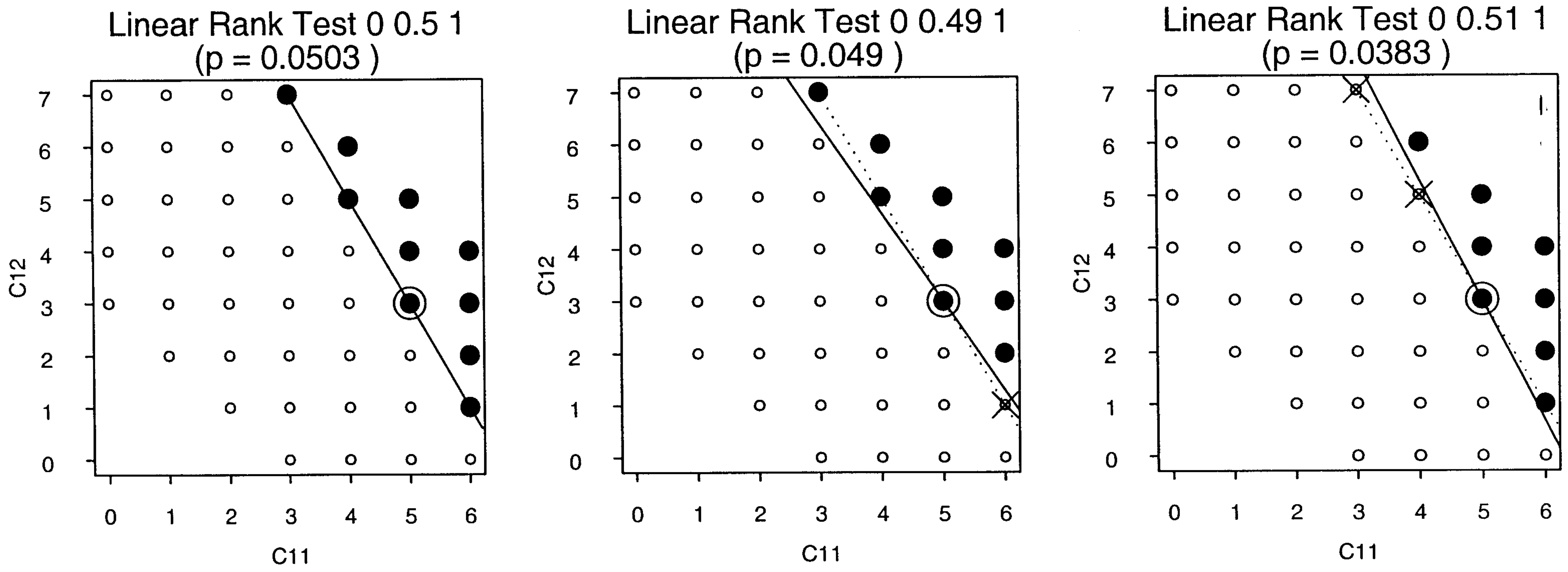
Table 2

Exact power comparisons of three linear rank tests with $n_1 = n_2 = 10$,

with alternative $\pi_1 = (0.3, 0.4, 0.3)$ *and* π_2 varying, at nominal $\alpha = 0.05$.

The last row of the table represents the actual size of the test.

π_2	$\varphi_{0.5}$	$\varphi_{0.49}$	$\varphi_{0.51}$
(0.1,0.0,0.9)	0.683	**0.797**	0.683
(0.1,0.1,0.8)	0.562	**0.641**	0.569
(0.1,0.2,0.7)	0.437	**0.489**	0.456
(0.1,0.3,0.6)	0.320	**0.354**	**0.354**
(0.1,0.4,0.5)	0.219	0.240	**0.267**
(0.1,0.5,0.4)	0.139	0.151	**0.195**
(0.1,0.6,0.3)	0.082	0.088	**0.139**
(0.3,0.4,0.3)	0.026	0.034	0.034



Linear Rank Test 0 0.5 1
(p = 0.0503)

Linear Rank Test 0 0.49 1
(p = 0.049)

Linear Rank Test 0 0.51 1
(p = 0.0383)

Figure 1. Three Linear Rank Tests