BIOMETRICS 43, 471–476
June 1987

Choice of Column Scores for Testing Independence in Ordered 2 × *K* Contingency Tables

Barry I. Graubard

National Institute of Child Health and Human Development, Biometry Branch,
Bethesda, Maryland 20892, U.S.A.

and

Edward L. Korn

Department of Biomathematics, University of California,
Los Angeles, California 90024, U.S.A.

SUMMARY

The numerous statistical methods for testing no association between a binary response (rows) and *K* ordered categories (columns) group naturally into two classes: those that require preassigned numerical column scores and those that do not. An example of the former would be a logistic regression analysis, and of the latter would be a Wilcoxon rank-sum test. In this paper we demonstrate that the perceived advantage of not preassigning scores is illusory. We do this by presenting an example from our consulting experience in which the midrank scores used by the rank tests that do not require preassigned scores are clearly inappropriate. Our recommendations are to assign reasonable column scores whenever possible, and to consider equally spaced scores when the choice is not apparent. Midranks as scores should always be examined for their appropriateness before a rank test is applied.

1. Introduction

The analysis of frequency data in 2 × *K* contingency tables with ordered categories appears frequently in the biomedical research literature—for example, in the analysis of a binary response to an increasing exposure dose. Investigators have been advised not to use a chi-square test for such data, but to use an appropriate analysis that incorporates the order of the columns (e.g., Moses, Emerson, and Hosseini, 1984). Many such appropriate analyses have been discussed in the statistical literature; see Agresti (1984) for a recent review. Proposed analyses may be divided into those that require a priori assignment of quantitative scores for each column, and those that do not. An example of the former type of analysis would be a logistic regression with the row variable as the binary response and with the column scores as the covariate. On the other hand, an analysis based on a rank statistic such as the Wilcoxon rank-sum test of row 1 versus row 2 does not require column scores, but uses only the column ordering.

A major drawback with using tests that require preassigned fixed scores is the need to specify the values for the scores. This is one of the considerations that has led some authors to recommend against the use of tests that require scores in favor of the rank tests (Kendall and Stuart, 1979, Chap. 33; Gross, 1981; Fleiss, 1981, §9.4). Additionally, allowing the investigator to select the scores leaves open the potential abuse of choosing scores that will

Key words: Exact conditional permutation test; Midrank scores; Ordered categorical data; Wilcoxon rank-sum test.

produce a desired result. However, even though analyses based on rank statistics are apparently objective, the investigator could easily choose from among many possible linear rank statistics (Hâjek and Šidák, 1967) to obtain a desired result. Several authors have stated that in practical cases the results from the two types of tests, those requiring scores and those not, are essentially the same (Armitage, 1955; Snedecor and Cochran, 1980, §11.8; Moses et al., 1984).

The purpose of this paper is to demonstrate that the rank statistics can be poor choices for testing independence when the column margin is far from uniformly distributed; see also Mantel (1963, 1979). This is because of the well-known correspondence between the rank tests and tests using scores with midranks as the preassigned scores. Therefore, the notion that rank tests avoid the arbitrary choice of column scores is misleading. Our recommendations for testing independence in an ordered 2 × K contingency table, which are similar to those given by Armitage (1955), are as follows:

(i) If possible, develop reasonable column scores based on the substantive meaning of the column categories, and use them in the analysis.
(ii) If no natural column scores are available, then consider using equally spaced column scores in the analysis.
(iii) Always examine the midranks as scores to make sure they are reasonable before using a rank test.

2. Exact Conditional Permutation Tests

We consider only exact permutation tests that are conditional on both the row and column margins in order to focus attention on the importance of the choice of column scores; see

Table 1
Column scores (x_i) for which the exact conditional permutation tests based on various test statistics are equivalent to the exact conditional permutation test based on $\sum x_i y_i$

	Test statistic	Column scores
(1)	Difference between the mean column scores for the two rows, i.e., a permutation t-test (Pitman, 1937a)	Any x_i
(2)	Pearson correlation coefficient of the y_i with the x_i, $i = 1, \ldots, N$ (Pitman, 1937b)	Any x_i
(3)	Slope in a weighted linear regression of the proportion positive in each column on the column scores, weights being inversely proportional to the binomial variances (Yates, 1948; Cochran, 1954)	Any x_i
(4)	Slope in linear regression of the y_i on the x_i, $i = 1, \ldots, N$ (Armitage, 1955)	Any x_i
(5)	Slope in a logistic regression of the y_i on the x_i, $i = 1, \ldots, N$ (Cox, 1970, §4.2)	Any x_i
(6)	Estimated row effect in the log-linear model considered by Haberman (1974)	Any x_i
(7)	Estimated row effect in the log-linear model considered by Simon (1974) provided the column scores are equally spaced	Equally spaced x_i
(8)	Estimated uniform association parameter in the uniform association model of Goodman (1979) provided the column scores are equally spaced (Agresti, 1984, p. 78)	Equally spaced x_i
(9)	Difference between the mean midranks for the two rows, i.e., a Wilcoxon rank-sum test using midranks for tied values (Lehmann, 1975, pp. 18–23)	x_i = midranks
(10)	Weighted least squares test of equal mean column percentiles (Williams and Grizzle, 1972)	x_i = midranks
(11)	Kendall's tau or Kendall's tau-B using midranks for tied values (Kendall, 1975; Hemelrijk, 1952)	x_i = midranks
(12)	Spearman's rho using midranks for the tied values (Lehmann, 1975, p. 301)	x_i = midranks
(13)	Difference between the mean ridits for the two rows (Bross, 1958; Selvin, 1977)	x_i = midranks
(14)	Efficient score statistic based on a proportional odds model (Snell, 1964; McCullagh, 1980)	x_i = midranks

Emerson and Moses (1985) for a comparison of exact and asymptotic tests. Let y_i be a binary outcome corresponding to the row of observation i in a $2 \times K$ contingency table, and let x_i be a score corresponding to the column of observation i, $i = 1, \ldots, N$. We assume that the scores x_i, which realize at most K distinct values, are chosen to be in increasing column order. This notation for the data is more convenient in what follows than the more typical cell counts of the table. Equally spaced column scores are defined by $x_i = k$ for observation i in column k. For the rank tests, let $r_i = \text{midrank}(x_i)$. Note that r_i do not depend on the actual values of the x_i but only on the column margin of the table. We consider exact permutation tests of no association conditional on both margins of the table, i.e., with $\{r_i\}$ and $\sum y_i$ held fixed. In principle, a general procedure for generating such a test based on a test statistic S is as follows: (i) Construct every $2 \times K$ table with the same margins as the observed table; (ii) calculate the probability of observing each of the tables under the null hypothesis of independence; (iii) calculate S for each of the tables; and (iv) add up the probabilities of the tables for which S is greater than or equal to S based on the observed data to yield the one-sided P-value. In practice, one can use the algorithms of Soms (1977) or Mehta, Patel, and Tsiatis (1984) to compute the one-sided P-values.

Consider the exact conditional permutation test derived from the following test statistic:

$$X = \sum x_i y_i.$$

As is well known, most of the commonly applied tests of no association for an ordered $2 \times K$ table are equivalent to this test statistic when applied as exact conditional tests; see Table 1. For example, the Wilcoxon rank-sum test is equivalent to X when the x_i equal the midranks (Table 1, line 9).

3. Example

Table 2 contains a subset of data from a prospective study of maternal drinking and congenital malformations. Women completed a questionnaire early in their pregnancy concerning alcohol use in the first trimester; complete data and details are available elsewhere (Mills and Graubard, 1987). Specifically, women were asked "During the first three months of this pregnancy ... Did you take any alcoholic beverages? If yes, did you average six or more, three to five, one or two, or less than one drink a day?" Data were later recorded on their pregnancy outcomes. Table 2 contains data on congenital sex organ malformations cross-classified by maternal alcohol consumption.

Three choices of column scores with the computed one-sided P-values are given in Table 3. To facilitate comparisons, the scores standardized to have means 0 and standard deviations 1 are also displayed. The midpoint scores correspond to the midpoints of the category intervals; the choice of 7.0 for "six or more" is somewhat arbitrary. Table 3 demonstrates two points. First, there can be large differences in the results of an analysis depending on the column scores chosen. Second, the midrank scores can be unreasonable in applications when the column margin is far from uniform. In the present application,

Table 2
Presence or absence of congenital sex organ malformation categorized by alcohol consumption of the mother

	Alcohol consumption (average # drinks/day)				
Malformation	0	<1	1–2	3–5	≥6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1
Total	17,114	14,502	793	127	38

Table 3
Alternative scoring systems for column categories with exact one-sided P-values

	Alcohol consumption (average # drinks/day)				
	0	<1	1–2	3–5	≥6
Midpoints	0	.5	1.5	4.0	7.0
Standardized	−.90	−.72	−.38	.48	1.52
			P-value = .0167		
Midranks	8557.5	24,365.5	32,013.0	32,473.0	32,555.5
Standardized	−1.69	−.16	.58	.63	.63
			P-value = .2860		
Equally spaced	1.0	2.0	3.0	4.0	5.0
Standardized	−1.26	−.63	.00	.63	1.26
			P-value = .1045		

both the consulting statisticians and the scientific investigator agreed that the closeness of the midrank scores at the higher levels of alcohol consumption was inappropriate. Thus, we decided that a rank test would not be a powerful test of independence for these data. These decisions can be made, and should be made, solely on the basis of the column margin of the table and not after computing *P*-values. Although the column margin of this example is somewhat extreme, we have seen similar margins in analyzing data on side effects where a high proportion of individuals have no or mild side effects.

4. Discussion

We have shown for a particular 2 × *K* contingency table that different choices of column scores can lead to different conclusions concerning association of the rows and columns. The results of Gross (1981) suggest that this is not an unusual occurrence given the observed column margin of Table 2: If a table is generated by a logistic regression model based on a set of scores, then the asymptotic relative efficiency (ARE) for testing independence using a statistic based on an incorrect set of scores is given by the square of the Pearson correlation of the two sets of scores taken over the population. The correlation of the midpoint scores with the midrank scores over the 32,574 individuals in Table 2 is .74. Thus, if the midpoint scores were "correct," then use of the midrank scores would result in an ARE of 55%. Although this asymptotic result applies only to a sequence of tables approaching independence, it does suggest that the disparity in test results of Table 2 is not a chance finding; see also Simon (1978) and Tarone and Gart (1980) for related asymptotic results.

We suspect that if our recommendations given in the Introduction were followed in practice, there would be more use of tests involving preassigned scores for the analysis of ordered 2 × *K* contingency tables.

Acknowledgements

The authors wish to thank M. H. Gail for his helpful comments. This work was supported in part by National Cancer Institute Grant CA16042.

Résumé

On peut regrouper naturellement en deux classes les nombreuses méthodes statistiques pour tester l'absence d'association entre une réponse binaire (lignes) et *K* classes ordonnées (colonnes): celles qui demandent des scores numériques préalables sur les colonnes et celles qui n'en demandent pas. Un exemple des premières pourrait être une analyse de régression logistique et un exemple des secondes serait un test de Wilcoxon sur la somme des rangs. Dans ce papier, nous démontrons que l'avantage perçu en ne précissant pas de scores au préalable est illusoire. Nous présentons en effet un exemple tiré de notre expérience de consultant dans lequel les scores des rangs moyens utilisés par les tests de rang ne demandant pas de scores préalables sont clairement inadéquates. Nous recommandons d'assigner des scores raisonnables aux colonnes si possible, et d'utiliser des scores régulièrement espacés quand le choix n'est pas évident. Les scores des rangs moyens devront toujours être examinés pour leur adéquation, avant d'appliquer un test de rang.

References

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York: Wiley.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11,** 375–386.

Bross, I. D. J. (1958). How to use ridit analysis. *Biometrics* **14,** 18–38.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.

Cox, D. R. (1970). *Analysis of Binary Data.* London: Methuen.

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.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions,* 2nd edition. New York: Wiley.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* **74,** 537–552.

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

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics* **30,** 589–600.

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests.* New York: Academic Press.

Hemelrijk, J. (1952). Note on Wilcoxon's two-sample test when ties are present. *Annals of Mathematical Statistics* **23,** 133–135.

Kendall, M. G. (1975). *Rank Correlation Methods,* 4th edition. London: Griffin.

Kendall, M. G. and Stuart, A. (1979). *The Advanced Theory of Statistics, Volume* 2, 4th edition. London: Griffin.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association* **58,** 690–700.

Mantel, N. (1979). Ridit analysis and related ranking procedures—use at your own risk. *American Journal of Epidemiology* **109,** 25–29.

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

Mehta, C. R., Patel, N. R., and Tsiatis, A. A. (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* **40,** 819–825.

Mills, J. L. and Graubard, B. I. (1987). Is moderate drinking during pregnancy associated with an increased risk for malformations? *Pediatrics* (in press).

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

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *Supplement to Journal of the Royal Statistical Society* **4,** 119–130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Supplement to Journal of the Royal Statistical Society* **4,** 225–232.

Selvin, S. (1977). A further note on the interpretation of ridit analysis. *American Journal of Epidemiology* **105,** 16–20.

Simon, G. (1974). Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association* **69,** 971–976.

Simon, G. A. (1978). Efficacies of measures of association for ordinal contingency tables. *Journal of the American Statistical Association* **73,** 545–551.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods,* 7th edition. Ames, Iowa: Iowa State University Press.

Snell, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics* **20,** 592–607.

Soms, A. P. (1977). An algorithm for the discrete Fisher's permutation test. *Journal of the American Statistical Association* **72,** 662–664.

Tarone, R. E. and Gart, J. J. (1980). On the robustness of combined tests for trends in proportions. *Journal of the American Statistical Association* **75**, 110–116.

Williams, O. D. and Grizzle, J. E. (1972). Analysis of contingency tables having ordered response categories. *Journal of the American Statistical Association* **67**, 55–63.

Yates, F. (1948). The analysis of contingency tables with groupings based on quantitative characters. *Biometrika* **35**, 176–181.

Received April 1986; *revised December* 1986.