NONPARAMETRIC ANALYSIS OF BLOCKED ORDERED CATEGORIES DATA: SOME EXAMPLES REVISITED

D. J. BEST, J. C. W. RAYNER, AND O. THAS

Nonparametric analysis for general block designs can be given by using the Cochran-Mantel-Haenszel (CMH) statistics. We demonstrate this with four examples and note that several well-known nonparametric statistics are special cases of CMH statistics.

1. Introduction

In this paper, we will use Cochran-Mantel-Haenszel (CMH) statistics to analyse four data sets which have appeared in the literature. It is well known that tests based on the CMH statistics are equivalent to certain standard rank tests but here we show how CMH statistics also apply in less standard situations. In particular, examples are given for randomized block designs both with and without missing values, for balanced incomplete block designs, and for supplemented balanced designs.

Recent descriptions of CMH statistics have been in Davis [6, Chapter 8] and Agresti [1, Chapter 7, Section 5]. We now give a very brief outline of CMH statistics, mainly from Davis [6, Chapter 8].

2. Cochran-Mantel-Haenszel statistics

The CMH statistics apply to counts N_{ijh} in which $i = 1,\ldots,r$, $j = 1,\ldots,c$, and $h = 1,\ldots,s$. Typically, the layer index h reflects the subjects or experimental units, usually referred to as the strata; the row index i reflects the levels of the factor of interest, and the column index j reflects the values of the response variable. The marginal totals $\{n_{\cdot jh}\}$ and $\{n_{i\cdot h}\}$ for each of the s strata are taken to be fixed. For each stratum, the vector of counts $N_h = (n_{11h},\ldots,n_{1ch},\ldots,n_{r1h},\ldots,n_{rch})^T$ has probability function

$$\left\{\prod_{i=1}^{r} n_{i\cdot h}!\right\}\left\{\prod_{j=1}^{c} n_{\cdot jh}!\right\} \Big/ \left\{n_{\cdot\cdot h}!\prod_{i=1}^{r}\prod_{j=1}^{c} n_{ijh}!\right\}. \tag{2.1}$$

Initially no assumption is made about the ordering of the row and column variables: both are taken to be nominal. The null hypothesis of interest, that there is no association between row and column variables in any of the s tables, is first tested against its negation.

Davis [6, Section 8.2.2] shows that for a table consisting of only a single stratum, the CMH statistic to test for randomness in a 2×2 table is $\{(n-1)/n\}X^2$, where X^2 is the familiar Pearson test statistic $\sum(\text{observed} - \text{expected})^2/\text{expected}$. A test statistic for testing no association between row and column variables across s 2×2 tables is due to Cochran [5] and Mantel and Haenszel [11]. For an arbitrary single stratum $r \times c$ table, a test for randomness may be based on $\{(n-1)/n\}X^2$. The test for no association between s 2×2 tables can be generalized to s $r \times c$ tables. The details follow.

2.1. CMH general association statistic. Suppose now we have counts in s independent $r \times c$ tables. The test statistic may be derived by considering the vector of counts for the hth stratum, N_h, modified by removing the redundant counts for the final row and column; these are known if the row and column totals and the other row or column entries are known. We also need the expected value under the null hypothesis of no association, $E[N_h]$, and the difference, $G_h = N_h - E[N_h]$. Now $G = \sum_h G_h$ is the aggregation over all strata of $(r-1)\,(c-1)$ differences between observation and expectation, and G has expectation zero and covariance matrix V_G, say under the null hypothesis, so that $Q_G = G^T V_G^{-1} G$ has asymptotic distribution $\chi^2_{(r-1)(c-1)}$ as the total sample size $n_{...} = \sum_h n_{..h}$ approaches infinity. This is known as the CMH *general association statistic*. The Anderson [3] and McNemar [12] statistics are particular cases of the CMH general association statistic.

2.2. CMH mean score statistic. Assume now that the column variable is ordinal or interval, and that every observation in the jth column of the hth stratum is scored as b_{hj}, $j = 1,\ldots,c$. The null hypothesis, that there is no association between row and column variables in any of the s tables, is now tested against the alternative that the r row mean scores differ, on average, across strata. First, define N_{jh} as the $r-1$ vector of counts N_{ijh}, $i = 1,\ldots,r-1$, and then define $M_h = (\sum_{j=1}^c b_{hj}(N_{jh} - E[N_{jh}]))$ as the vector containing the first $r-1$ row sums for the hth stratum. It is routine to show that under the null hypothesis of no association $M = \sum_h M_h$ has expectation zero and covariance matrix V_M say, so that $Q_M = M^T V_M^{-1} M$ has asymptotic distribution χ^2_{r-1} as the total sample size $n_{...} = \sum_h n_{..h}$ approaches infinity. The statistic Q_M is known as the CMH *mean score statistic*. If mid-rank scores are used, then if $s = 1, Q_M$ is the Wilcoxon-Mann-Whitney statistic for $r = 2$ and the Kruskal-Wallis [10] statistic for $r > 2$, while if $s > 1$ and all row totals for all strata are unity, Q_M is the Friedman [8] statistic. If the "natural" scores, $b_{hj} = j$, $j = 1,\ldots,c$, are used when $s = 1$ and $r > 2$, then a statistic due to Yates [19] is obtained.

2.3. CMH correlation statistic. Assume now that both the row and column variables are ordinal or interval, and that every observation in the ith row of the hth stratum is scored as a_{hi}, $i = 1,\ldots,r$, and that every observation in the jth column of the hth stratum is scored as b_{hj}, $j = 1,\ldots,c$. The null hypothesis, that there is no association between row

and column variables in any of the s tables, is now tested against the alternative that across strata there is a consistent association, positive or negative, between the row scores and column scores. Let C_h be a scalar given by $C_h = \sum_i \sum_j a_{hi} b_{hj} \{N_{ijh} - E[N_{ijh}]\}$. It is routine to show that under the null hypothesis of no association, $C = \sum_h C_h$ has expectation zero and variance V_C say, so that $Q_C = C^T V_C^{-1} C = C^2/V_C$ has asymptotic distribution χ_1^2 as the total sample size $n_{...} = \sum_h n_{..h}$ approaches infinity. The statistic Q_C is known as the CMH *correlation statistic*. If $s = 1$, then Q_C is $(n_{...} - 1)$ times the square of the Pearson correlation between the row and column variables; if $s = 1$ and natural scores $a_{hi} = i$ and $b_{hj} = j$ are used, then Q_C is $(n_{...} - 1)$ times the square of the Spearman correlation. The CMH correlation test is a detector of linear-linear association.

2.4. Generalized CMH statistics.

Suppose that the row variable is not ordered (nominal) while the column variable is ordinal or interval, with scores $\{b_{hj}\}$. Suppose that the scores satisfy $b_{hj} = b_v(j)$ for all h with $\sum_j b_r(j) b_s(j) N_{.j.}/n_{...} = \delta_{rs}$. Then M is an $(r-1)$ vector with typical element $\sum_j b_v(j)\{N_{ij.} - E[N_{ij.}]\}$. It follows from Rayner and Best [16, Section 4.4] that M standardised is the vth component of Pearson's X^2 in the sense that the sum of the squares of the $(c-1)$ components is X^2. This order v component detects departures of the data from the model of homogeneity of row means. As before, if natural linear scores are used, the resulting test is related to that of Yates [19]. However, if the scores are quadratic, the resulting test detects dispersion differences between rows. The set of p-values resulting from applying all $(c-1)$ component tests gives a detailed and informative scrutiny of the data, albeit an informal one.

Suppose that both row and column variables are ordinal or interval, $a_{hi} = a_u(i)$ for all h with $\sum_i a_r(i) a_s(i) N_{i..}/n_{...} = \delta_{rs}$, and $b_{hj} = b_v(j)$ for all h with $\sum_j b_r(j) b_s(j) N_{.j.}/n_{...} = \delta_{rs}$. Then $C = \sum_i \sum_j a_u(i) b_v(j)\{N_{ij.} - E[N_{ij.}]\}$. It follows from Rayner and Best [16, Section 8.2] that C standardised is the uvth component of Pearson's X^2, detecting departures of the data from the model of independence in the uvth bivariate moment. As previously noted, if natural linear scores are used for both row and column variables, then C is Spearman's ρ. However, if one set of scores is linear while the other is quadratic, this leads to interesting tests of bivariate skewness.

3. Randomized blocks

Possibly, the most commonly used experimental design is the randomized block design. We begin this section by illustrating how the three CMH statistics Q_G, Q_M, and Q_C introduced in the previous section are equivalent to three nonparametric rank statistics for randomized blocks.

Suppose, as in Bradley [4, page 127], that we consider measures of visual acuity for five subjects which have been given drugs designated as A, B, C, and D. The data are presented in Table 3.1. Suppose further that we wish to carry out nonparametric tests based on the within blocks (subjects) rankings for this data set. These rankings are given in parentheses in Table 3.1. We wish to use these ranks to test for equality of median drug effects, that is, to test $H_0 : \tau_A = \tau_B = \tau_C = \tau_D$ against K: not H_0, that at least two medians differ. Friedman's [8] test statistic T takes the value 8.28 with corresponding p-value 0.04 based on an χ_3^2 approximation.

Table 3.1. Visual acuity data from Bradley [4].

Drug\subject	1	2	3	4	5
A	0.39 (3)	0.21 (2)	0.73 (1)	0.41 (2)	0.65 (1)
B	0.55 (1)	0.28 (1)	0.69 (2)	0.57 (1)	0.57 (3)
C	0.33 (4)	0.19 (3)	0.64 (3)	0.28 (4)	0.53 (4)
D	0.41 (2)	0.16 (4)	0.62 (4)	0.35 (3)	0.60 (2)

Table 3.2. Stratum 1 contingency table for visual acuity data.

Drug\rank	1	2	3	4
A	0	0	1	0
B	1	0	0	0
C	0	0	0	1
D	0	1	0	0

Table 3.3. Partition of A for visual acuity data.

Source	df	SS	p-value
Friedman	3	8.28	0.04
Dispersion	3	0.60	0.90
Residual	3	1.32	0.72
Anderson	9	10.20	0.33

If we wish to test $H_0 : \tau_A = \tau_B = \tau_C = \tau_D$ against $K : \tau_A > \tau_B > \tau_C > \tau_D$, then Page's [14] test is appropriate. We find the Page test statistic L takes the value 4.7 with corresponding p-value 0.03 based on an χ_1^2 approximation. To test for the equality of the distributions of the ranks for the four drugs, we use Anderson's [3] test based on A, which here takes the value 10.20 with p-value 0.33 based on an χ_9^2 approximation.

The T statistic is simply Q_M, the L statistic is Q_C, and the A statistic is Q_G. To calculate the three CMH statistics, we need to form five 4×4 tables of counts. For subject or block 1, the 4×4 stratum table is shown as Table 3.2. Notice that each row and each column sum is one.

Software for calculating Q_G, Q_M, and Q_C is available in the IMSL, SAS, and StatXact (version 6) computer packages. To calculate Q_M and Q_C, scores are needed. To obtain T and L, the scores 1, 2, 3, and 4 are required. The usual parametric F test for mean drug differences gives $F_{3,12}$ with p-value 0.014.

Before proceeding to use the CMH approach to obtain analogues of T, L, and A for more complicated designs, we note as an aside that using the orthogonal polynomial methods of Rayner and Best [16] and Rayner et al. [17], we can partition the statistic A for randomized block designs. Results for Table 3.1 data are given in Table 3.3. The CMH approach can be used to obtain the dispersion statistic in Table 3.3 by using Q_M with the quadratic scores 9, 1, 1, 9.

4. Balanced incomplete blocks

We now illustrate the CMH approach for data from a balanced incomplete block design.

Off-flavour in six ice cream samples was rated by 15 subjects tasting four samples each. A seven-point scale was used, with "1" meaning little off-flavour and "7" meaning considerable off-flavour. The data were given in Meilgaard et al. [13, Table 7.11] and are shown here in Table 4.1. Notice that the original data in Meilgaard et al. [13, Table 7.11] is in error for subject 14, in that, a rating of "1" should be given to ice cream F, not to ice cream E. All six ice cream samples were not given to each subject as it was thought six samples were too many to evaluate at once. Sensory fatigue is well documented and often only three or four samples are judged at one sitting.

To apply CMH statistics, we form an $r \times c$ contingency table for each of the s subjects. Here r is the number of ice creams and c is the number of categories, so that $s = 15$, $r = 6$, and $c = 7$. The rows relate to ice creams and the columns to categories. Thus for each subject, a 6×7 contingency table of 0 s and 1s is formed. For example, for subject 1 the contingency table is given by Table 4.2. Of course, for a complete block design, rows E and F would have a "1" in one of the columns. Summing the 6×7 contingency tables for all 15 subjects, we obtain Table 4.3. Notice that not all rows and columns sum to one as they did for randomized blocks. Tied data would give us some column sums greater than one.

Are the six histograms whose counts are given in Table 4.3 significantly different? To answer this, we calculate Q_G, the generalized association CMH statistic, or the mean scores CMH statistic, Q_M. This can easily be done by using as data the 15 (0, 1) subject tables for IMSL [9] routine CTRAN. The SAS and StatXact routines for generalized CMH statistics will not now do all the analysis needed.

For the Table 4.3 data, we find $Q_G = 32.86$ with an approximate p-value, based on the χ^2_{29} distribution, of 0.28. Note that because here the covariance matrix is a generalized inverse of rank 29, the degrees of freedom are 29, not 30. It appears that Q_G is not too sensitive for these data. Perhaps this is because Q_G does not take into account that the data are ordered. If we use the category identifiers as scores, then we find $Q_M = 19.8$ with a p-value of 0.001 based on the χ^2_5 approximation. An F test using the same scores gives a p-value less than 0.001 according to Meilgaard et al. [13]. The F test relies on more assumptions than the test based on Q_M.

Rayner et al. [17] give an alternative analysis of the Table 3.1 data using ranks. Also note that for $r = 2$, Q_G is the Stuart [18] test of marginal homogeneity.

5. Missing values

Alvo and Cabilio [2] derive a nonparametric ranks-based test for an ordered alternative $\tau_1 \geq \tau_2 \geq \tau_3 \geq \cdots$ when the data are from a randomized block design with missing values. We now illustrate how to apply Q_C to obtain an alternative test statistic. We consider the same lymph heart pressure (in mm of Hg) as did Alvo and Cabilio [2]. These data are reproduced in Table 5.1 and concern measurements on eight toads which were dehydrated for 6-, 12-, 18-, and 24-hour periods. Biologists expect that on average, a toad's lymph heart pressure will decrease with increasing dehydration.

Table 4.1. Off-flavour ratings for six ice creams.

Subject\ice cream	A	B	C	D	E	F
1	6	1	1	2	—	—
2	6	—	—	1	3	3
3	—	4	2	—	5	2
4	7	2	3	—	2	—
5	3	5	—	1	—	1
6	—	—	1	1	3	2
7	7	4	4	—	—	3
8	2	—	1	1	1	—
9	—	2	—	2	2	3
10	4	2	—	2	5	—
11	5	—	3	—	1	1
12	—	3	2	1	—	2
13	4	2	—	—	1	1
14	5	—	2	2	—	1
15	—	2	4	5	3	—

Table 4.2. Off-flavour ratings of six ice creams for subject 1.

Ice cream\category	1	2	3	4	5	6	7
A	0	0	0	0	0	1	0
B	1	0	0	0	0	0	0
C	1	0	0	0	0	0	0
D	0	1	0	0	0	0	0
E	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0

Table 4.3. Off-flavour ratings combined for subjects.

Ice cream\category	1	2	3	4	5	6	7
A	0	1	1	2	2	2	2
B	1	5	1	2	1	0	0
C	3	3	2	2	0	0	0
D	5	4	0	0	1	0	0
E	3	2	3	0	2	0	0
F	4	3	3	0	0	0	0

To find Q_C, we need to rank the data within toads and we use eight indicator matrices or contingency tables which are similar in form to Tables 3.2 and 4.2. For toad 24, this indicator table is given as Table 5.2.

Table 5.1. Lymph heart pressure (in mm of Hg) data of Alvo and Cabilio [2].

Toad\dehydration time	6 hours	12 hours	18 hours	24 hours
21	11.9	9.8	7.6	10.2
22	5.6	4.9	4.0	3.1
23	—	14.4	14.2	7.8
24	13.3	—	—	10.0
25	8.0	7.9	—	7.6
27	17.7	16.6	15.3	11.6
28	9.0	8.0	11.9	6.8
29	9.8	8.0	7.7	7.8

Table 5.2. Rankings for toad 24 in a 4×4 table.

Hours\rank	1	2	3	4
6	1	0	0	0
12	0	0	0	0
18	0	0	0	0
24	0	1	0	0

Table 6.1. Growth of strawberry plants after applying pesticides.

Block I	Block II	Block III	Block IV
C, 107	A, 136	B, 118	O, 173
A, 166	O, 146	A, 117	C, 95
D, 133	C, 104	O, 176	C, 109
B, 166	B, 152	D, 132	A, 130
O, 177	D, 119	B, 139	D, 103
A, 163	O, 164	O, 186	O, 185
O, 190	D, 132	C, 103	B, 147

Using routine CTRAN from IMSL [9], we find $Q_C = 11.9$ with p-value 0.0006 based on an χ_1^2 approximation. Alvo and Cabilio [2] found that for these data, their recommended test statistic took the value 226.75 and quoted exactly the same p-value as do we, namely 0.0006.

6. Supplemented balance designs

Pearce [15] suggested the use of supplemented balanced designs and used these designs to analyse data when pesticides designated as A, B, C, D, and O are applied to strawberry plants. The pesticides were intended to control weeds and allow the strawberry plants to grow bigger and presumably produce more strawberries. However, while eradicating the weeds, do the pesticides inhibit strawberry growth? Pearce [15] gave the results and the design that we reproduce in Table 6.1. The figures quoted represent the spread of the strawberry plants. Pesticide "O" is a control.

Table 6.2. Rankings for block I in a 5×8 table.

Drug\rank	1	2	3	3.5	4	5	6	7
A	0	0	0	1	0	1	0	0
B	0	0	0	1	0	0	0	0
C	0	0	0	0	0	0	0	1
D	0	0	0	0	0	0	1	0
O	1	1	0	0	0	0	0	0

To use CMH to obtain an analogue of Friedman's T for this more complex design, we proceed as before, ranking within blocks and forming four 5×8 indicator matrices. Notice in block I, there are two tied observations. Table 6.2 shows the indicator matrix for this block.

We find $Q_M = 20.1$ with a p-value of 0.0005 based on an χ^2_4 approximation. An F test based on a regression routine gives, for these data, $F_{4,20} = 24.6$ with p-value less than 0.0001. Desu and Raghavarao [7] give an analogue of Friedman's T for general block designs that have the same asymptotic chi-squared distribution as T. For Table 6.1 data, their statistic has the value 20.0, almost identical to Q_M. Perhaps the difference is in the treatment of the tied observations.
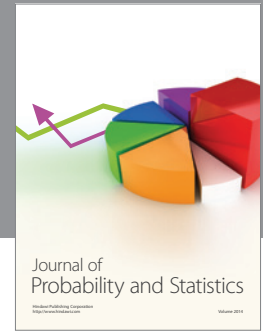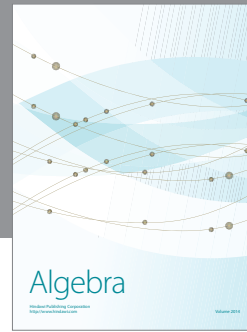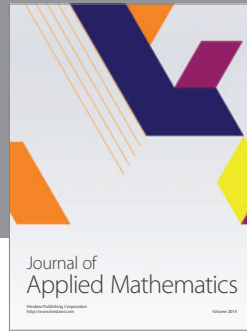
References

[1] A. Agresti, *Categorical Data Analysis*, 2nd ed., Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 2002.

[2] M. Alvo and P. Cabilio, *Testing ordered alternatives in the presence of incomplete data*, Journal of the American Statistical Association **90** (1995), no. 431, 1015–1024.

[3] R. L. Anderson, *Use of contingency tables in the analysis of consumer preference studies*, Biometrics **15** (1959), 582–590.

[4] J. V. Bradley, *Distribution-Free Statistical Tests*, Prentice-Hall, New Jersey, 1968.

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

[6] C. S. Davis, *Statistical Methods for the Analysis of Repeated Measurements*, Springer Texts in Statistics, Springer, New York, 2002.

[7] M. M. Desu and D. Raghavarao, *Nonparametric Statistical Methods for Complete and Censored Data*, Chapman & Hall/CRC, Florida, 2004.

[8] M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, Journal of the American Statistical Association **32** (1937), no. 200, 675–701.

[9] IMSL, *IMSL User's Guide-Mathematical & Statistical Functions*, Houston: Visual Numerics. 1995.

[10] W. H. Kruskal and W. A. Wallis, *Use of ranks in one-criterion analysis of variance*, Journal of the American Statistical Association **47** (1952), no. 260, 583–621.

[11] N. Mantel and W. Haenszel, *Statistical aspects of the analysis of data from retrospective studies of disease*, Journal of the National Cancer Institute **22** (1959), 719–748.

[12] Q. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages*, Psychometrika **12** (1947), no. 2, 153–157.

[13] M. Meilgaard, G. V. Civille, and B. T. Carr, *Sensory Evaluation Techniques*, 3rd ed., CRC Press, Florida, 1999.

[14] E. B. Page, *Ordered hypotheses for multiple treatments: a significance test for linear ranks*, Journal of the American Statistical Association **58** (1963), no. 301, 216–230.

[15] S. C. Pearce, *Supplemented balance*, Biometrika **47** (1960), no. 3-4, 263–271.

[16] J. C. W. Rayner and D. J. Best, *A Contingency Table Approach to Nonparametric Testing*, Chapman & Hall/CRC, Florida, 2001.

[17] J. C. W. Rayner, D. J. Best, P. B. Brockhoff, and G. D. Rayner, *Nonparametrics for Sensory Science: A More Informative Approach*, Blackwell, Iowa, 2005.

[18] A. Stuart, *A test for homogeneity of the marginal distributions in a two-way classification*, Biometrika **42** (1955), no. 3-4, 412–416.

[19] F. Yates, *The analysis of contingency tables with groupings based on quantitative characters*, Biometrika **35** (1948), no. 1-2, 176–181.

D. J. Best: School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia
E-mail address: donald.j.best@newcastle.edu.au

J. C. W. Rayner: School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia
E-mail address: john.rayner@newcastle.edu.au

O. Thas: Department of Applied Mathematics, Biometrics and Process Control, Ghent University, 9000 Gent, Belgium
E-mail address: olivier.thas@ugent.be