Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

05-12

Continuous Analogues of Cochran-Mantel-Haenszel Statistics

J.C.W. Rayner and D.J. Best

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Continuous Analogues of Cochran-Mantel-Haenszel Statistics

J. C. W. Rayner
Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia and
School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia (John.Rayner@newcastle.edu.au)

and

D. J. Best
School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia (John.Best@newcastle.edu.au)

Abstract

Cochran-Mantel-Haenszel (CMH) statistics are reviewed and continuous analogues of the CMH general association and mean score statistics are given. CMH generalised correlation statistics are defined and analogues given.

Key Words: CMH general association statistic; CMH mean score statistic; generalised correlation.

1. Introduction

Subsequently Cochran-Mantel-Haenszel (CMH) statistics are reviewed. They are defined for tables of counts, but it is helpful to define continuous analogues. These analogues will be indicative of the performance of CMH tests without requiring the assessment of a broad range of categorisations. CMH generalised correlation statistics are also defined, as are their continuous analogues. Davis (2002, section 8.7) recommends the use of continuous analogues of the CMH mean score and correlation statistics and gives examples of their use.

2. Review of the Cochran-Mantel-Haenszel Statistics

The Cochran-Mantel-Haenszel statistics apply to counts N_{ihj} in which $i = 1, \ldots, t$, $h = 1, \ldots, c$ and $j = 1, \ldots, b$. The layer index j reflects the *block* in which the treatment occurs; the row index i reflects the *treatment* of interest, and the column index h reflects the value of the *response* variable.

The marginal totals $\{n_{.hj}\}$ and $\{n_{i.j}\}$ for each of the b blocks are taken to be known. For each block, the vector of counts $N_j = (n_{11j}, \ldots, n_{1cj}, \ldots, n_{t1j}, \ldots, n_{tcj})^T$ has probability function

$$\left(\prod_{i=1}^{t} n_{i.j}!\right)\left(\prod_{h=1}^{c} n_{.hj}!\right) \Big/ n_{..j}! \prod_{i=1}^{t}\prod_{h=1}^{c} n_{ihj}!.$$

Initially no assumption is made about 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 *b* tables, is tested against its negation. In the following expectations and variances are taken under the null hypothesis.

2.1 The CMH General Association Statistic

Consider

- counts in *b* independent $t \times c$ nominal-nominal tables and
- the vectors of counts $N_j = (N_{ihj})$ for each of the *j* blocks, $j = 1, \ldots, b$, each being of dimension $(t-1)(c-1) \times 1$ after removal of the redundant counts.

Take

- $E[N_j]$ to be the expected value under the null hypothesis of no association,
- $G_j = N_j - E[N_j]$, observed minus expected for the counts for the *j*th block,
- $G = G_1 + \ldots + G_b$, to be the aggregation over all blocks of differences between observation and expectation, and
- the covariance matrix of *G* under the null hypothesis to be V_G say.

Then as the total sample size $n_{\ldots} = n_{..1} + \ldots + n_{..b}$ approaches infinity, the *CMH general association statistic* $Q_G = G^T V_G^{-1} G$ has asymptotic distribution $\chi^2_{(t-1)(c-1)}$.

2.2 CMH Mean Score Statistic

Assume now that the row variable is nominal, while the column variable is ordinal or interval, and that every observation in the *h*th column of the *j*th block is scored as b_{hj}, $j = 1, \ldots, c$. The null hypothesis of no association between row and column variables in any of the *b* tables, is now tested against the alternative that the *t* row mean scores differ, on average, across strata.

Take

- M_j to be the vector containing the first $t-1$ row means for the *j*th block
- $M = \sum_j (M_j - E[M_j])$, where the $E[M_j]$ are expectations under the null hypothesis,
- V_M to be the covariance matrix of *M* under the null hypothesis.

Then as the total sample size $n_{\ldots} = n_{..1} + \ldots + n_{..b}$ approaches infinity $Q_M = M^T V_M^{-1} M$, the *CMH mean score statistic*, has asymptotic distribution χ^2_{t-1}.

2.3 CMH Correlation Statistic

Assume now that the row and column variables are both ordinal or interval, and that the *i*th treatment row scores are a_{hi}, $i = 1, \ldots, t$, and the *j*th block column scores are b_{hj}, $j = 1, \ldots, b$.

The null hypothesis of no association between row and column variables in any of the *b* tables, is tested against the alternative that across blocks there is a consistent association, positive or negative, between the row scores and column scores.

Take

- $C_j = \sum_i \sum_h a_{hi} b_{hj} \{N_{ihj} - E[N_{ihj}]\}$
- $C = \sum_j C_j$

- V_C to be the covariance matrix of C under the null hypothesis

Then as the total sample size $n_{\cdot\cdot\cdot} = n_{\cdot\cdot 1} + \ldots + n_{\cdot\cdot s}$ approaches infinity $Q_C = C^T V_C^{-1} C$, the *CMH correlation statistic*, has asymptotic distribution χ^2_1.

3. Continuous Analogues of the CMH Statistics

In constructing continuous analogues the CMH general association statistic is of no interest.

3.1 Continuous Analogue of the CMH Mean Score Statistic

The CMH mean score statistic Q_M is based on the treatment means $\bar{Y}_{i\cdot}$, obtained by averaging over blocks. In the following the marginal totals are assumed to be unknown and hence there are no redundant responses. If the marginal totals were known then the means would no longer be mutually independent.

As in Best et al. (2012, section 2) we assume the model

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

where τ_i are treatment effects with $\sum_i \tau_i = 0$, β_j are block effects with $\sum_j \beta_j = 0$, μ is an overall mean and the ε_{ij} are independent random variables with mean 0 and variance σ^2. For this model the treatment means over blocks, $\bar{Y}_{i\cdot}$, are, by the Central Limit Theorem, approximately N($\mu + \tau_i$, σ^2/b) for the number of blocks, b, sufficiently large. Now we may apply the result that if $X_1, \ldots, X_n$ are IN(μ, σ^2) then $\sum_i (X_i - \bar{X})^2 / s^2$ is χ^2_{n-1} distributed. It follows that under the null hypothesis that $\tau = (\tau_i) = 0$, $b\sum_{i=1}^{t} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 / s^2 = M'$ say, is χ^2_{n-1} distributed. Further routine analysis reveals that when testing $\tau = 0$ against $\tau \neq 0$ the null hypothesis is rejected for large values of M'.

In practice is σ^2 unknown. On the jth block, because the sample variance is an unbiased estimator of the population variance, $\mathrm{E}[\sum_{i=1}^{t} (\bar{Y}_{ij} - \bar{Y}_{\cdot j})^2 / (t-1)] = \sigma^2$. Writing $V = \{\sum_{i=1}^{t}\sum_{j=1}^{b} Y_{ij}^2 - t\sum_{j=1}^{b} \bar{Y}_{\cdot j}^2\} / (t-1)$ as in Best et al. (2012, section 1) and summing over blocks gives $\mathrm{E}[V] = b\sigma^2$. In M' replacing σ^2 by its unbiased estimate V/b gives $M = b^2 \sum_{i=1}^{t} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 / V$ as a test statistic for testing $\tau = 0$ against $\tau \neq 0$. Its approximate distribution is χ^2_{n-1}.

3.2 Continuous Analogue of the CMH Correlation Statistic

First a contrast that can be the basis of an analogue of Q_C is constructed. If X_i has mean μ_i and standard deviation σ_i then a *contrast* in these random variables is a function $\sum_i a_i X_i$ such that $\mathrm{E}[\sum_i a_i X_i] = 0$ and $\sum_i a_i^2 = 1$. Thus if $\lambda_1, \ldots, \lambda_t$ are such that $\lambda_1 + \ldots + \lambda_t = 0$ then because $\mathrm{E}[\sum_i a_i Y_{i\cdot}] = \mu \sum_i a_i = 0$, $C' = \sum_i l_i Y_{i\cdot} / \sqrt{\sum_i l_i^2}$ is a contrast.

From 3.1 immediately above we know that the $\bar{Y}_{i\cdot}$ are, under the null hypothesis $\tau = 0$, approximately distributed as N(μ, σ^2/b). Hence C' is approximately distributed as N(0, σ^2/b) and $b(C')^2/\sigma^2$ is approximately distributed as χ^2_1. A further approximation is introduced if, again as in 3.1, σ^2 by its unbiased estimate V/b. Then $b^2(C')^2/V = b^2(\sum_i l_i Y_{i\cdot})^2/\{V\sum_i l_i^2\} = C$ say is approximately distributed as χ^2_1. So C is approximately the square of a contrast in the response means with approximate distribution χ^2_1.

The CMH correlation statistic, Q_C, is based on the correlation between the treatment scores and the responses aggregated over blocks. First we construct the sample correlation between the treatment scores $\{\lambda_i\}$ and the response means $\{\bar{y}_{i\cdot}\}$. The treatments scores are assumed to sum to zero; their sample variance is $\sum_i l_i^2/t$. The ith treatment has population variance σ^2/b, which is estimated by V^2/b^2. It is this quantity that is used instead of the sample variance of the $\{\bar{y}_{i\cdot}\}$. The sample correlation between the $\{a_i\}$ and the $\{\bar{y}_{i\cdot}\}$ is, subject to this adjustment, $\sum_i l_i \bar{y}_{i\cdot} / \sqrt{\{(\sum_i l_i^2)V^2/b^2\}}$. The square of the random variable corresponding to this approximate correlation is C, a continuous analogue of Q_C.

4. CMH Generalised Correlation Statistics and their Analogues

The CMH correlation statistic C is defined in section 2.3 as $\sum_{i,h,j} a_{hi}b_{hj}\{N_{ihj} - E[N_{ihj}]\}$. Suppose that $a_{hi} = a_u(i)$ for all h with $\sum_i a_r(i)a_s(i)N_{i\cdot\cdot}/N_{\cdots} = \delta_{rs}$, and that $b_{hj} = b_v(j)$ for all h with $\sum_j b_r(j)b_s(j)N_{\cdot j\cdot}/N_{\cdots} = \delta_{rs}$. See the discussion in Rayner and Beh (2009) about generalised correlations. Call the modified statistic C_{uv} say, given by

$$C_{uv} = \sum_{i,h,j} a_u(i)b_v(j)\{N_{ij\bullet} - E[N_{ij\bullet}]\}.$$

Suppose V_{Cuv} is the covariance matrix of C_{uv} under the null hypothesis Then as the total sample size $n_{\cdots} = n_{\cdot\cdot 1} + \ldots + n_{\cdot\cdot s}$ approaches infinity $Q_{uvC} = C_{uv}^T V_{uvC}^{-1} C_{uv}$, the *CMH generalised correlation statistic*, has asymptotic distribution χ^2_1.

From Rayner and Best (2001, section 8.2) C_{uv} standardised, that is, Q_{uvC}, will detect departures of the data from the model of independence in the uvth bivariate moment.

From a data analytic perspective analogues of the generalised correlations that are of most interest are of order (u, 1) between the treatment scores and the responses. It may be of interest to know how quadratic and cubic treatment scores correlate with responses, but the correlation between quadratic and cubic treatment scores and quadratic responses is of far less interest. Hence an analogue of interest is $\sum_i a_u(i)\bar{y}_{i\bullet}/(bV)$ that reflects polynomial effects of degree u in treatments. The square of this random variable is a continuous analogue of Q_{uvC}, and is approximately distributed as χ^2_1.

5. Conclusion

Suppose now that Y_{ij} is a (continuous) measurement on the ith of t treatments in the jth of b blocks. Suppose that the ordered values of the observed Y_{ij} are $y_1 < y_2 < \ldots < y_{bt}$. If $a_0, a_1, \ldots, a_{bt}$ are such that $a_0 < y_1 < a_1 < y_2 < \ldots < y_{bt} < a_{bt}$ then if a response falls in (a_{h-1}, a_h) it is given the score y_h, $h = 1, \ldots, c$. Thus the observations $y_1, \ldots, y_{bt}$ generate counts N_{ihj} that are 1or 0 as the response for the ith treatment on the jth block falls in the hth category (a_{h-1}, a_h) or not.

If the data provide the categories in the sense just described, the for all data sets we have tested the analogue M gives the same value as Q_M and the analogue C gives the same value as Q_C.

References

Best, D.J., Rayner, J.C.W. and Thas, O. (2012). Comparing nonparametric tests of trend and equality of means for randomized block designs. Submitted.

Davis, C.S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

Rayner, J.C.W. and Beh, Eric J. (2009). Towards a better understanding of correlation. *Statistica Neerlandica*. 63(3), 324-333.

Rayner, J.C.W. and Best, D.J. (2001). *A Contingency Table Approach to Nonparametric Testing*. Boca Raton: Chapman & Hall/CRC.