2015

A note on the CMH general association statistic and square contingency tables

D. J Best
University of Newcastle

J.C.W Rayner
University of Wollongong

A note on the CMH general association statistic and square contingency tables

Abstract

In this expository note a simplified formula for the CMH general association statistic applicable to repeated categorical response data is given and applied to three-way square contingency tables.

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

06-15

A Note on the CMH General Association Statistic and Square Contingency Tables

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

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

A Note on the CMH General Association Statistic and Square Contingency Tables

D. J. Best[1] and J. C. W. Rayner[1,2]
University of Newcastle[1]*, Australia and University of Wollongong*[2]*, Australia*

Abstract. In this expository note a simplified formula for the CMH general association statistic applicable to repeated categorical response data is given and applied to three-way square contingency tables.

AMS Subject Classification: 62-07, 62G10

Keywords: Nonparametric statistics; Just about right sensory evaluation repeated trinary response data; Repeated binary response data; Repeated categorical response data.

1. Introduction

Square contingency tables can be formed from repeated categorical response data. McNemar's (1947) statistic can be used to analyse such 2×2 square tables and Stuart's (1955) statistic can be used to analyse such $k \times k$ square tables for $k \geq 2$. Cochran's (1950) Q statistic can be used to analyse binary data for t treatments where $t \geq 2$. There are many applications of these three nonparametric tests in elementary nonparametric or general texts. However there is often the need to compare more than two treatments ($t > 2$) or to use more than two categories ($k > 2$).

These three nonparametric statistics are all special cases of the CMH general association statistic, S say. We now define S for more general repeated categorical square contingency table response data, although we note that the formulae for the McNemar, Stuart and Cochran statistics are well known, simpler and can be used if applicable.

If there are t products to compare given k categories with data obtained from c subjects or judges then define $X_{uvw} = 1$ if product u is assessed by judge v as being in category w and $X_{uvw} = 0$ otherwise. Then

$$S = \frac{t-1}{t}\sum_{u=1}^{t} q_u^T V^{-1} q_u \ ,$$

[1] School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia

[2] Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia

in which, for u = 1, 2, ..., t, the vector q_u has wth element $X_{u\bullet w} - X_{\bullet\bullet w}/t$, for w = 1, …, k – 1 and the (w, z)th element of matrix V is, for w, z = 1, …, k – 1,

$$v_{wz} = -\sum_{v=1}^{c} X_{\bullet vw} X_{\bullet vz} / t^2 \text{ for } w \neq z \text{ and}$$

$$v_{ww} = X_{\bullet\bullet w} / t - \sum_{v=1}^{c} X_{\bullet vw}^2 / t^2 \text{ for } w = z.$$

A dot subscript indicates summation over that subscript. S can be shown to have an approximate $\chi^2_{(k-1)(t-1)}$ distribution. A derivation of S is given in Rayner and Best (2001).

We now give two examples. The first example involves likely to purchase market research data and the second GSS social survey data. Both data sets are presented as three-way square tables. For very small data sets Best et al. (2014) give an example of calculating S by hand. However for larger data sets, as here, a computer routine such as that described in the appendix is needed.

2. Likely to purchase and GSS trinary response data

The following table gives counts for three repeated responses to t = 3 hot ‘chips’ or ‘fries’. We note that the square contingency table presentation is convenient and that this convenience is more pronounced as c increases.

Each of c = 55 consumers rated their likelihood of purchase of the potato products A, B and C using k = 3 categories: would not buy (NB), undecided (U) or would buy (WB) with the repeated trinary responses shown in the three-way square contingency table below. This data is derived from Rayner *et al.* (2005, Table 6.14). Categories 1 and 2 as well as categories 4 and 5 were combined so as to avoid many zeroes and to make presentation easier. The products differed as they were prepared with different cooking oils. We find S = 1.888 with χ^2_4 p-value of 0.756 indicating similar marginal counts.

Likely to purchase counts

	A NB			A U			A WB		
B	NB	U	WB	NB	U	WB	NB	U	WB
C NB	6	5	2	3	1	6	3	1	2
C U	4	1	1	4	1	0	0	0	1
C WB	5	2	0	0	2	1	2	1	1
Sum B	15	8	3	7	4	7	5	2	4
Sum A	26			18			11		

Marginal likely to purchase counts

	NB	U	WB
A	26	18	11
B	27	14	14
C	29	12	14

As an example of larger c consider the 2006 American General Social Survey (GSS) data taken from Agresti (2013, Table 13.10). Here c = 1478. Subjects were asked their opinion on how much government funding should be spent on t = 3 areas: environment (EN), health (H) or education (ED). The k = 3 responses were 1 = more, 2 = same and 3 = less. See the following table.

Opinions on Environment, Education and Health Spending

	EN 1			EN 2			EN 3		
H	1	2	3	1	2	3	1	2	3
ED 1	651	45	15	304	59	10	92	24	17
ED 2	57	10	3	50	35	12	15	14	6
ED 3	7	1	5	7	10	4	6	3	16
Sum H	715	56	23	361	104	26	113	41	39
Sum EN	794			491			193		

We find S = 470.30 with X_4^2 p-value of less than 0.001. This very highly significant p-value is mostly due to EN getting less 1 and more category 2 and 3 responses. See the table below. Would this be the global response?

Marginal GSS counts

	1	2	3
ED	1217	202	59
H	1189	201	88
EN	794	491	193

3. Conclusion

We noted that the McNemar, Stuart and Cochran statistics are special cases of S. We gave a definition of S that is somewhat simpler than that usually given and applied the formula to three-way square contingency tables of repeated categorical data. We are not aware that the CMH analysis given here has been used before, although Darroch (1981) has suggested it could be. Model based analyses are also available; see, for example, the generalised linear models discussed in Agresti (2013). These require more assumptions than the essentially nonparametric approach given here.

References

Agresti, A. (2013). *Categorical Data Analysis,* 3rd ed., New York: Wiley.

Best, D.J., Rayner, J.C.W. and Allingham, David (2014). A note on formulae for CMH statistics for JAR data. *Food Quality & Preference,* 31, 19-21.

Cochran, W. G. (1950). The comparison of percentages in matched samples, *Biometrika,* 37, 256-266.

Darroch, J. N. (1981). The Mantel-Haenszel test and tests of marginal symmetry: fixed-effects and mixed models for a categorical response. *International Statistical Review,* 49, 285-307.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, 12, 153-157.

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

Rayner, J.C.W., Best, D.J., Brockhoff, P.B. and Rayner, G.D.(2005). *Nonparametrics for Sensory Science - A More Informative Approach.* Ames : Blackwell.

Stuart, A. (1955). A test of homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412-416.

Appendix

A computer routine, SCMH.EXE, for calculating S given square contingency table data is available from the first author. The routine needs the availability of the command line which is part of "Accessories" on Windows XP and Windows 7. On Windows 8 use the command line App. Suppose the user is in the directory where the files SCMH.EXE and STRI.DAT are located. For Example 1 STRI.DAT, using code 1 = NB, code 2 = U and code 3 = WB, is

```
3  55  3
26*1  18*2  11*3
15*1  8*2  3*3  7*1  4*2  7*3  5*1  2*2  4*3
6*1  4*2  5*3  5*1  2  2*3  2*1  2  3*1  4*2  1  2  2*3  6*1  3  3*1  2*3  1  3  2*1   2  3
```

These counts are as given in the table for the first example above. The first line of STRI.DAT is " t c k " using the notation above. On the command line enter "SCMH < STRI.DAT".