Analysis of Multiple-Response Data

Petr Vlach, Miroslav Plašil

Abstract

This paper presents new methods for analyzing categorical multiple-response data. We discuss the advantages of using multiple-response questions in surveys, testing for associations in contingency tables with multiple-response data and graphical analysis.

Key words

Categorical data, Multiple-response, Chi-square test Association, Rao-Scott Correction, Odds ratio

1. Introduction

It is often of interest to test for independence between two categorical variables. This article presents methods for testing of marginal association in 2-way tables in which one or both response categorical variables allow for multiple responses. Although the multiple response variables are far from being rare and appear in all possible fields of research, it is only recently that necessary tests have been developed. Traditional Pearson chi-squared tests for independence could not be used in presence of multiple response questions because of within-subject dependence among responses. We use an *easy-to-understand* example (genuine, however) to illustrate the existing tests and methodology.

When a respondent is provided with a list of possible items in a survey question, answers do not necessarily fall into one of several mutually exclusive categories. On the contrary, we are often faced with the situation when the individual is fully described only through a combination of the items in question. Sometimes we could limit ourselves to the *best-to-fit* response. However, we loose some part of the information available.

Data admitting more than one response from the list of items are referred to as *multiple-response data* or less frequently as *pick any out of c data*. Multiple-response questions are quite common in all fields of research, including marketing, education and social sciences as prominent examples.

The purpose of this article is to provide a reader with the elementary knowledge about methods developed in this area recently, namely new approaches to the testing of marginal independence between two multiple-response categorical variables. More detailed treatment of multiple-response tests for independence can be found in the influential articles Agresti and Liu (1999) and Tomas and Decady (2000, 2004).

If the in/dependency between two multiple-response variables is of interest, data can be presented quite naturally in a traditional 2-way contingency table. This time, however, we should not use the classical Pearson chi-squared test for independence because of the within-subject dependence among responses. Methods proposed to deal with the issue could be more or less precisely divided into two main classes: a) computer-intensive methods based on bootstrapping a suitable test statistic whose distribution is not known exactly, b) chi-squared approximation to the sampling distribution.

Following Agresti and Liu (1999) and Tomas and Decady (2000, 2004) we focus on the latter. Approximate chi-squared tests bear the advantage over the bootstrapping procedures chiefly because they are somewhat similar to the traditional chi-square test, thus they provide intuitive relation to the familiar technique of testing for independence. Easy applicability and low computational intensiveness play also an important role.

As in the traditional case, rejection of the hypothesis of marginal independence in favour of marginal association is an important first stage in the analysis of the data. However, the follow-up analysis determining the sources of association is not less important. We will show that this could be done via slightly different presentation of the original data - the *odds ratios*.

The article is organized as follows: section 2 describes basic concepts and tests, section 3 describes the application of the testing methods and provides the results. Final section concludes.

2. Basic concepts and tests

There are two basic concepts and hypothesis tests for multiple response variables: 1) multiple by multiple marginal independence test (MMI) and 2) single by multiple marginal independence test (SPMI). They only differ in whether we have one multiple-response variable (SPMI) or two (MMI). We briefly begin with the first situation.

The MMI reflects the situation that both variables in contingency table are of multiple-response nature. Following the notation of Bilder (2000) we briefly present the concept of MMI.

Suppose we have two multiple-response questions Q_i and Q_j. Let m_{ij} denote the number of observed positive responses to Q_i and Q_j. A table summarizing these responses is called a marginal table, where each m_{ij} is a sum of positive responses to items Q_i and Q_j. The marginal probability of a positive response to Q_i and Q_j is denoted by π_{ij} and its maximum likelihood estimate is $\hat{\pi}_{ij} = m_{ij}/n$. The hypotheses for the test of MMI are

H_0: $\pi_{ij} = \pi_{i.}\pi_{.j}$ for $i = 1,\ldots r$ and $j = 1,\ldots c$

H_1: At least one equality does not hold

where $\pi_{ij} = P(Q_i = 1, Q_j = 1)$, $\pi_{i.} = P(Q_i = 1)$ and $\pi_{.j} = P(Q_j = 1)$. MMI can be rewritten in another way. Suppose we have *rc* 2 x 2 tables with cells in the following form:

π_{ij}	$\pi_{i.} - \pi_{ij}$
$\pi_{.j} - \pi_{ij}$	$1 - \pi_{.j} - \pi_{i.} + \pi_{ij}$

Provided none of these cells have probability zero, MMI can be written as follows:

H_0: $\Phi_{QiQj} = 1$ for $i = 1,\ldots r$ and $j = 1,\ldots c$

H_1: At least one equality does not hold

where Φ stands for odds ratio and $\Phi_{QiQj} = \pi_{ij}(1-\pi_{.j} - \pi_{i.} + \pi_{ij})/[(\pi_{i.} - \pi_{ij})(\pi_{.j} - \pi_{ij})]$. To develop a suitable test statistic we begin with the table above. It is obvious, that the association between row and column variables in the above table can be tested using a standard Pearson chi-square statistic, that will have the appropriate asymptotic chi-squared distribution on 1 degree of freedom. Therefore we can calculate chi-square test for each table as

$$\chi^2_{rc} = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\hat{\pi}_{ij} - \hat{\pi}_{i.}\hat{\pi}_{.j}\right)^2}{\hat{\pi}_{i.}\hat{\pi}_{.j}}, \tag{1}$$

and odds ratio as

$$\Phi = \frac{\pi_{11} - \pi_{12}}{\pi_{21} - \pi_{22}}. \tag{2}$$

Following Agresti and Liu (1999) we can calculate a test statistic for MMI by summing up the individual chi-square statistics corresponding to each cell of the *r* x *c* table, that is

$$\chi^2_{MMI} = \sum_{i=1}^{r}\sum_{j=1}^{c}\chi^2_{ij}. \tag{3}$$

Formula (3) can be also rewritten as

$$\chi^2_{MMI} = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\hat{\pi}_{ij} - \hat{\pi}_{i.}\hat{\pi}_{.j}\right)^{2\cdot}}{\hat{\pi}_{i.}\hat{\pi}_{.j}(1-\hat{\pi}_{i.})(1-\hat{\pi}_{.j})}. \tag{4}$$

Individual components of (3) have asymptotically chi-square distribution on 1 degree of freedom. However, the sum of the components is not asymptotically distributed as chi-square on *rc* degrees of freedom. The reason is mutual dependence of the individual chi-square tests.

To solve this problem, approximate procedures based on the Rao-Scott approach were developed and we will shortly describe them here. Approximate chi-squared tests have close relation to the classical Pearson chi-square test. There are two subtypes of Rao-Scott approach: first-order and second-order Rao-Scott tests.

The Rao-Scott approach is based on the fact that a test statistic of chi-squared form is, under certain mild condition asymptotically distributed as a weighted sum of random variables distributed chi-square on 1 degree of freedom. As demonstrated in Thomas and Decady (2000), formula (3) can be regarded as a member of Rao-Scott corrected chi-squared family of tests. It is *self-correcting* in the sense that its first-order Rao-Scott correction factor is equal to one. Briefly say, first-order test procedure consists in referring χ^2_{MMI} to $\chi^2_{\alpha}(\text{rc})$.

The second-order Rao-Scott test that is based on the statistic (4) consists in referring $\chi^2_{MMI}/(1+\hat{a}^2)$ to a chi-squared random variable on $rc/(1+\hat{a}^2)$ degrees of freedom, where the second-order correction factor can be expressed in the special form fully described in Rao-Scott (1981, 1984).

The SPMI applies when we have single by multiple response variable and we test for their independence. This hypothesis was first introduced by Agresti and Liu (1999) and was referred to as a *multiple marginal independence*. As stated earlier in our text, the SPMI marginal independence hypothesis cannot be tested using the Pearson chi-square test for the *r* x *c* table because subjects within the same row/column can be represented by more than one category of the multiple-response question.

Although one can intuitively use the Pearson chi-square test, this will not be correct. First difference is in marginal totals. They do not sum up to expected number of subject of the particular row/column. Simply say, they usually exceed this value.

The first-order Rao-Scott correction is obtained by dividing the original statistic by the mean of the weights in the linear combination, resulting in a corrected statistic that has the same mean as the corresponding chi-squared random variable, at least asymptotically. Again it was shown (see Thomas and Decady (2000) for the proof), that statistic (4) is self correcting, i.e. its first-order Rao-Scott correction factor is equal to one.

The second order Rao-Scott correction was designed for situations of large inter-item correlations and is fully described in Thomas and Decady (2004).

3. Application

Let us assume that we have two multiple response questions Q1 and Q2 consisting of *r* and *c* items. As an example of such multiple response questions suppose we have a survey questionnaire where a total of 100 individuals responded to the questions presented below:

Q1: *Which of the following media do you prefer?*

☐ TV ☐ Newspapers ☐ Radio ☐ Internet

Q2: *What is your favourite leisure time activity?*

☐ Cinema ☐ Reading ☐ Music ☐ Computers ☐ Sport

Aggregated data are shown in Tab. 1. The multiple-response nature of the data is evident. The totals show the number of cases with *Yes* response for the corresponding item in the row/column. The sum of totals (in this example 81 + 47 + 42 + 27 = 197 and 34 + 54 + 41 + 24 + 40 = 193) show the total count of responses and it is almost double as high as the number of respondents in both cases. Ratio of this value to the total number of respondents can be interpreted as an average number of preferred items per respondent, which is then 1,97 and 1,93 respectively.

Tab. 1: Contingency table with multiple response data

Which of the following media do you prefer?	What is your favourite leisure activity?					
	Cinema	Reading	Music	Computers	Sport	Total ($\pi_{i.}$)
TV	32	44	32	20	35	81
Newspapers	15	39	17	7	18	47
Radio	9	19	29	12	18	42
Internet	8	17	15	16	7	27
Total ($\pi_{.j}$)	34	54	41	24	40	100

To compute the test statistic (4), one can create 4 x 5 = 20 2x2 tables and sum up the traditional Pearson chi-square tests defined by (3). An example of such 2 x 2 table computed for the frequency n_{22} = 34 (shaded grey in Tab. 1) is shown in Tab. 2.

Tab. 2: Marginal table

39	8
15	38

The chi-square statistic is then the sum of 20 individual components. For the data in Tab. 2 $\chi^2 = 29{,}98$. We have computed chi-square test (4) with SPSS 14.2. The result is shown in Tab. 3. Evaluation of the formula (3) yields $\chi^2_{MMI} = 109{,}78,$ which exceeds the $\chi^2_{0,95}(20) = 31{,}4.$ Under the simple random sampling assumption on which the above test is developed, the observed value for the statistic χ^2_{MMI} leads to a strong rejection of the null hypothesis of marginal independence between Q1 and Q2.

Tab. 3: Pearson Chi-Square Tests

		Q2
Q1	Chi-square	109,782
	df	20
	Sig.	,000(*)

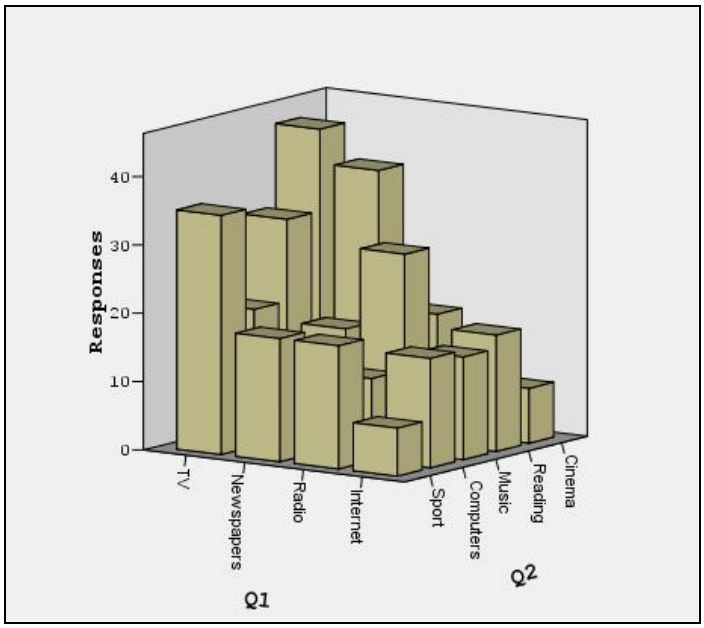
The follow-up analysis is similar to the traditional Pearson chi-square test. Analyst must determine the sources of the rejection of the omnibus hypothesis. First we compute the table of odds ratios using the formula (2).

Tab. 4: Odds ratios fot the survey data

Which of the following media do you prefer?	What is your favourite leisure activity?				
	Cinema	Reading	Music	Computers	Sport
TV	5,55	1,07	0,73	1,23	2,13
Newspapers	0,84	12,35	0,68	0,37	0,87
Radio	0,36	0,54	8,55	1,53	1,23
Internet	0,76	1,65	2,26	11,82	0,42

One can see immediately, that the strongest associations are between categories *Newspapers* and *Reading* ($\Phi = 12{,}35$), *Internet* and *Computers* ($\Phi = 11{,}82$) and *Radio* and *Music* ($\Phi = 8{,}55$). These results are not surprising and are strongly in agreement with the a priori hypothesis on the association between presented categories. The other odds ratios which exceed value of two are also notable. These conclusions can be also seen on the graphical representation of survey data (see figure 1).

Figure 1: Graphical representation of survey data



4. Conclusion

In this article we presented some of the recent developments in the area of association between two categorical variables, when one or both variables admit multiple responses. We illustrated the tests for marginal independence in presence of such variables on a real-life example and indicated a possible follow-up analysis determining sources of association. To achieve this goal we employed existing software, although it implicitly supports the first order Rao-Scott tests only. Under some conditions (possibly implausible in most cases), these tests provide very poor control over the test levels (i.e. Type I errors) and consequently second-order Rao-Scott tests should be employed. As far as the authors know no such tests are available in standard statistic packages, but they can be computed after some additional effort in any routine that supports essential matrix manipulation. For more comprehensive and detailed treatment of the issue we refer the reader to the existing literature cited below.

References

[1] Agresti, A, - Liu, L.M.: Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics* 55, 936- 943, 1999.

[2] Bilder, C. R. - Loughin, T.M.,: Testing for marginal independence between two categorical variables with multiple responses*, Biometrics*, 60(1), 241-8, 2004.

[3] Nettleton, D.: Testing for Association between Categorical Variables with multiple-Response Data.

[4] Rao, J. N. K. - Scott, A. J.: The analysis of categorical data from complex surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* 76, 221-230, 1981.

[5] Thomas, R. D. - Decady, Y. J.: Analyzing categorical data with multiple responses per subject. *Statistics Society of Canada Proceedings of the Survey Methods Section*, 121-130, 2000.

[6] Thomas, R. D. - Decady, Y. J.: Testing for Association Using Multiple Response Survey Data: Approximate Procedures Based on the Rao-Scott Approach. *International journal of Testing*, 4, 43-59, 2004.

Petr Vlach
University of Economics Prague
Department of Statistics and Probability
W. Churchill Sq. 4
130 67 Prague 3
Czech Republic
vlach@vse.cz

Miroslav Plašil
University of Economics Prague
Department of Statistics and Probability
W. Churchill Sq. 4
130 67 Prague 3
Czech Republic
plasil@vse.cz