Analysis of the Association between Nominal Variables in Statistical Study of Information Flows about Academic Ethics[1]

Milan Terek
UNIVERSITY OF ECONOMICS IN BRATISLAVA, SLOVAKIA, e-mail: milan.terek@euba.sk
Peter Kročitý
SCHOOL OF MANAGEMENT IN TRENCIN, SLOVAKIA, e-mail: pkrocity@vsm.sk

Abstract

The paper deals with the possibilities of analysis the association between two nominal variables in statistical surveys. Analyzed data derived from random sampling are concentrated in contingency tables. The procedure of realization chi-square test of homogeneity is described. Then, residual analysis, concretely the possibilities of using adjusted standardized residuals in analysis of association between variables, is presented.   Finally, the measure of strength of association − odds ratio is described. All described methods and tools are applied in the study of information flows about academic ethics in one Slovak university. The study has been realized as a part of information flows and academic ethics knowledge management learning process in a wider context of knowledge management at this university.

Key words: *chi-square test of homogeneity, adjusted standardized residuals, odds ratio*

Introduction

Over the last few years the use of statistical survey-based sampling has spread enormously. A statistical survey can be characterized as a process of collecting data through the means of questioning subjects. It is a common practice to create a set of questions which are then presented in a questionnaire. Questionnaires often include questions, answers to which, from the formal-mathematical point of view, can be classified as nominal variables. We will concentrate on association description methods between two qualitative, nominal variables.

In general, when values of one variable tend to appear together with some values of the other variable, we conclude there is an association between those variables.

When the data for a variable consist of values used to identify an attribute of the element, the scale of measurement is considered a nominal scale. The scale of measurement for a variable is called an ordinal scale if the data exhibit the properties of nominal data and the order or rank of the data is meaningful.

Data for qualitative variable association analysis is summarized in contingency tables. The basic questions asked by an analyst while analyzing a contingency table are:

[1] *This paper was prepared with the support of the project VEGA, no. 1/0761/12: Alternative approaches to measurement of socioeconomic development and VEGA no. 1/0092/15: Modern Approaches to Complex Statistical Surveys Design.*

ISBN 978-602-19725-8-8

- Is there an association between the variables? The answer is revealed with the help of chi-square test. The lower the *p*-value, the more likely the association between the variables.
- How do the data differ from a situation where the variables are not associated? Adjusted standardized residuals identify cells which are more or less similar to a non-existing association situation.
- How strong is the existing association? To measure the association strength obviously odds ratio is used.

The use of the just-mentioned methods will be illustrated on a study of nominal variables associations that have been analyzed in a statistical study aimed to obtaining information about academic ethics at School of Management/City University of Seattle in Trencin and Bratislava sites. The study has been realized as a part of information flows and academic ethics knowledge management learning process in a wider context of knowledge management at this school.

Randomly selected students were sent a questionnaire with eight questions. Question number two read: „When did you first learn about the academic ethics rules at School of Management?“ The answers to this question in regards to the form of the study will be analyzed more closely now.

Chi-square test of homogeneity

In general, we speak of homogeneity in statistics when statistical characteristics of one part of a data set are the same as characteristics of another part of that data set.

A decision if often needed as to whether the observed differences among values of sampling proportions are statistically significant or whether they have been obtained due to randomness of sampling. We will look at tests regarding differences among proportions. We will analyze *r* x *c* (*c* > 2) contingency table. We will consider random samples from *r* populations with multinomial distribution[2]. In every experiment, *c* possible outcomes can be reached. In a contingency table, the sample sizes in the last column are fixed, the column totals are influenced by randomness of sampling.

Let π_{ij} be the probability of *j*-th outcome for *i*-th population. We are testing:

$$H_0 : \pi_{1j} = \pi_{2j} = \ldots = \pi_{rj} \qquad \text{for } j = 1, 2, \ldots c$$

meaning that random samples are from *r* populations with the same multinomial distribution versus an alternative

$$H_1 : \pi_{1j}, \pi_{2j}, \ldots \pi_{rj} \text{ are not all equal for at least one value of } j$$

Let n_{ij} be the observed frequency in the *i*-th row and *j*-th column, n_i be the sum of n_{ij} values in the *i*-th row, n_j be the sum of n_{ij} values in the *j*-th column. The sum of all n_{ij} values is *n*.

Assuming H_0 is true, then expected frequencies are calculated as follows:

[2] Multinomial distribution, see in Freund (1992), p. 216 – 218.

ISBN 978-602-17725-8-8

$$o_{ij} = \frac{n_i \cdot n_j}{n} \qquad (1)$$

The value of test statistic is calculated according to following relationship:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-o_{ij}\right)^2}{o_{ij}}$$

The critical region at level of significance α is $\chi^2 \geq \chi^2_{1-\alpha}\left((r-1)(c-1)\right)$.

Due to the considered test statistic following the chi-square distribution only approximately, it is recommended to use this test only in the cases when no o_{ij} values are lower than 5. These tests are often referred to as tests of homogeneity [3].

Residual analysis

A comparison of the observed and expected frequencies allows to analyze the direction of the association between variables. The difference $\left(n_{ij}-o_{ij}\right)$ is called a residuum. The adjusted standardized residuals can be defined as:

$$r_{ij} = \frac{n_{ij}-o_{ij}}{\sqrt{o_{ij}\left(1-\frac{n_i}{n}\right)\left(1-\frac{n_j}{n}\right)}} \qquad \text{for } i = 1,2...r; \ \ j = 1, 2, ...c \qquad (2)$$

where $\frac{n_i}{n}$ is an estimated marginal probability in row *i,*

$\frac{n_j}{n}$ is an estimated marginal probability in column *j.*

The denominator in formula (2) is a standard error of random variable $\left(n_{ij}-o_{ij}\right)$, when H_0 is true (Agresti & Finlay, 2014, p.230). Adjusted standardized residuals r_{ij} follow asymptotically the standard normal distribution. They can be used informally to describe the relationship among the table cells. A too large value of an adjusted standardized residuum indicates a deviation from homogeneity in the cell. If H_0 is true, there is approximately just a 0,05 probability that the adjusted standardized residuum exceeds 2 in its absolute value. Absolute values over 3 clearly indicate an association in a cell.

Measures of association

[3] More on chi-square test of homogeneity and independence see in Freund (1992) p. 477 – 487 and all internet sources in references.

A measure of the association strength is a statistic or parameter that indicates the strength of an association between two variables. (Agresti & Finlay, 2014, p. 233).

When 2 x 2 contingency table is analyzed, the difference of proportions can be used as measure of association. The odds ratio is the measure of association that can be used in all contingency tables.

Odds ratio

For a response variable with two values, the odds for success is defined as:

$$odds = \frac{probability\ of\ success}{probability\ of\ failure}$$

In general, the estimated odds[4] for a response variable with two values equals the number of successes divided by the numbers of failures. The odds ratio θ in 2x2 contingency tables equals to the ratio of the 1st row odds to the 2nd row odds.

In *r x c* contingency tables, odds ratio can be calculated in any 2x2 sub-table.

Analysis of association in the study of information flows about academic ethics

Randomly selected students of the School of Management in Trencin and Bratislava were sent a questionnaire with eight questions. One of the questions was question number 2: „When did you first learn about the academic ethics rules at our school?“ The answers to this question in regards to the form of the study will be analyzed more closely now.

We are interested in whether the method of obtaining information about academic ethics significantly differs between online and in-class students. Using random sampling with replacement we obtained answers to question number 2: „When did you first learn about the academic ethics rules at our school?“ from 67 online students and 56 in-class students. The final results (columns with numbers less than 5 were combined) are presented in Table 1 with expected frequencies in parentheses.

Table 1 Distribution of answers to question no.2 for online and in-class students

Answers to question no. 2 **Study form**	Before application submission or during studies in the first year (later than in the 1st trim.), 2nd or 3rd years of studies, other method, no info received 1	At the beginning of studies – during new student orientation 2	During studies in the first trimester 3	Total n_i
Online	10 (10,89)	22 (27,78)	35 (28,33)	67
In-class	10 (9,11)	29 (23,22)	17 (23,67)	56
Total n_j	20	51	52	123

The chi-square test was performed using the CHISQ.TEST statistical function in Excel. The *p*-value obtained in the test was 0.043915. That means that at the 0.05 level of

[4] Calculated value is only estimate of the real unknown value of odds in statistical population, that is why it is called estimated

significance, we reject the hypothesis that random samples from online and in-class student populations come from the same probability distribution of a random variable – The method of academic ethics information obtaining, and we accept the alternative hypothesis that they are not from the same distribution. In other words, the method of academic ethics information obtaining is significantly different for online and for in-class students at the 0.05 level of significance. Thus, we accept an assumption that there is an association between the method of academic ethics information obtaining and the form of study.

The test itself, however, doesn't say anything about the direction or the strength of the association. The test doesn't indicate whether all cells significantly differ from the homogeneity or whether it is just one or two cells. To find out, we perform the residual analysis. We will calculate adjusted standardized residuals r_{ij} following the (2) formula.

Table 2 Adjusted standardized residuals

Answers to question no. 2 **Study form**	Before application submission or during studies in the first year (later than in the 1st trim.), 2nd or 3rd years of studies, other method, no info received 1	At the beginning of studies – during new student orientation 2	During studies in the first trimester 3
Online	−0,44	−2,12	2,44
In-class	0,44	2,12	−2,44

In the table 2, there are considerably large positive values of residuals concerning in-class students who obtained the information at the beginning of their studies – during an orientation for new students as well as concerning online students who obtained the information during the first trimester of their studies. That means there are more in-class students who obtained the information at the beginning of their studies - during an orientation for new students as well as more online students who obtained the information during the first trimester of their studies than it is suggested by the hypothesis of homogeneity.

Similarly, there are considerably large negative values of residuals concerning online students who obtained the information at the beginning of their studies – during an orientation for new students as well as concerning in-class students who obtained the information during the first trimester of their studies. That means there are less online students who obtained the information at the beginning of their studies - during an orientation for new students as well as less in-class students who obtained the information during the first trimester of their studies than it is suggested by the hypothesis of homogeneity. In-class students were more likely to obtain the information at the beginning of their studies – during an orientation for new students and online students were more likely to obtain the information during the first trimester of their studies.

Now the strength of association will be measured by odds ratios. In table 1 we will concentrate on columns 2 and 3 only in which residuals indicate a strong association. Possibility 2 will represent a success and possibility 3 will represent a failure. The data are presented in table 3. We will calculate the estimated odds and the odds ratio.

Table 3 Possibilities 2 and 3 from Table 1

Answers to question no. 2 **Study form**	At the beginning of studies – during new student orientation 2	During studies in the first trimester 3	Total n_i
Online	22	35	57
In-class	29	17	46
Total n_j	51	52	103

$$\text{The estimated odds for online students} = \frac{\frac{22}{57}}{\frac{35}{57}} = \frac{22}{35} \approx 0{,}629$$

As far as online students are concerned, there is about 0.629 of a student who learned about academic ethics rules using possibility 2 per 1 student who learned about the rules using possibility 3.

$$\text{The estimated odds for in-class students} = \frac{\frac{29}{46}}{\frac{17}{46}} = \frac{29}{17} \approx 1{,}706$$

Concerned in-class students, there is about 1.706 of a student who learned about academic ethics rules using possibility 2 per 1 student who learned about the rules using possibility 3.

Let's find out the odds ratio for in-class students:

$$\theta = \frac{1.706}{0.629} \approx 2.712$$

An in-class student has a 2.712 times bigger chance of learning about the academic ethics rules via possibility 2 than an online student.

In *r x c* contingency tables, odds ratio can be calculated in any 2x2 sub-table.

Conclusion

We showed in the paper how useful could be analysis of association between two nominal variables in the study of information flows in the context of knowledge management in an organization. The chi-square test of homogeneity was described. The test enables to make decision about whether there exists an association between variables. It is not required to use specialized software in order to perform the test, Excel is sufficient. Specifically, the CHISQ.TEST function was used.

Once an association between variables is established, it is worth to analyze which combinations of variable values cause it. The article presented possibilities of the adjusted standardized residuals usage which allows for identification of cells which more or less resemble a non-association state in a contingency table.

At last, it is useful to seek answers to a question: "How strong is the association?" To measure the strength, the odds ratio was used. This measure is easily interpreted and providing a clear apprehension of the association strength.

We only discussed nominal variables in the paper. As far as ordinal variables are concerned, it is possible to implement "stronger statistical methods" intended for this kind of higher-level measurement.

This paper was prepared with the support of the project VEGA, no. 1/0761/12: Alternative approaches to measurement of socioeconomic development and VEGA no. 1/0092/15: Modern Approaches to Complex Statistical Surveys Design.

References

AGRESTI, A., FINLAY, B. (2014), *Statistical Methods for the Social Sciences,* Pearson, Essex.

ANDERSON, C. J. (n. d.), *Two-Way Tables: Chi-Square Tests Edpsy/Psych/Soc 589,* available at http://courses.education.illinois.edu/EdPsy589/lectures/2way_chi-ha-online.pdf

DIPANKAR BANDYOPADHYAY, *Lecture 10: Partitioning Chi Squares and Residual Analysis,* available at http://www.biostat.umn.edu/~dipankar/bmtry711.11/lecture_10.pdf

FREUND, J. E. (1992), *Mathematical Statistics.* Prentice-Hall International, Englewood Cliffs.

Analysing Tables. Part V. Interpreting Chi-Square, available at

http://www.helsinki.fi/~komulain/ Tilastokirjat/09.%20Ristiintaulukko.pdf

Chi-square Test of Independence, available at

http://www.geneseo.edu/~bearden/socl211/chisquareweb/chisquare.html