Association between Two Categorical Variables and Information Channels Effectiveness Assessment

MILAN TEREK
Vysoká Vysoká škola manažmentu v Trenčíne / City University of Seattle programs
Trenčín, Slovakia

Abstract: The paper deals with the possibilities of association between two categorical variables analyses and the application of such analysis in the i3nformation channels effectiveness assessment in the case when the same information is provided by more different information channels to different groups of people in the framework of one time period. When the association is confirmed by a test, the use of adjusted standardized residuals enabling the pattern of the association among the contingency table cells recognition is appropriate and will be used in the analysis. The association strength measurement possibilities will be described and the odds ratios will be applied in the information channels effectiveness assessment.

Keywords: adjusted standardized residuals; odds ratio; information channels

1 Introduction

It happens frequently that the same information in the framework of one time period is provided by more different information channels. For example, a firm informs people about offered services by paper publicity materials, the internet and publicity in media. It is interested in the effectiveness of the information channels for potential clients from different regions. The council of a town informs the citizens about its activities by a network of billboards, its web page and communal media. It would like to know the effectiveness of the information channels for different age categories of citizens.

A natural first step of such an analysis is asking people through which information channel they usually obtain information, by conducting a statistical survey. The "information channels" and "groups of people" are generally categorical variables. The association between two categorical variables will be analyzed – categories of the first variable represent information channels, and categories of the second one represent different groups of people obtaining information through these information channels. The procedure of information channels effectiveness assessment for different groups of people will be suggested. The use of the proposed procedure of information channels effectiveness assessment will be illustrated by the problem of customer satisfaction and s study of their profiles in a firm.

2 Analyzing Association between Categorical Variables

Categorical variables take values (categories) enabling to identify an attribute of each element. When only identifying of an attribute is possible, the measurement scale of the variable is nominal. When the values of the variable exhibit properties of nominal data and the order or rank of the values is meaningful, the variable is called ordinal – it is measured by the ordinal scale.

When the results of one variable tend to change as the results of the other variable take different values, we conclude there exists an association between those variables. Data for categorical variables association analysis are summarized in contingency tables. The association between two nominal variables will be analyzed.

2.1 The Procedures of Association Analysis

Three common procedures of association analysis between two nominal variables are recommended in the literature. The first one consists of two steps – conducting of a statistical test revealing if an association between variables exists, and if that is the case, measuring of how strong the existing association is, with the aid of some summary measures of association, such as Cramer`s *V*, contingency coefficient or Goodman and Kruskal's lambda (for example, see [4, 7]). The second procedure includes testing of the association and using of adjusted standardized residuals enabling the study of the structure of association when the association has been confirmed (for example, see [8]). The adjusted standardized residuals serve for identification cells of a contingency table, which are "responsible" for the revealed association. The last procedure consists of association testing, using of adjusted standardized residuals in the case of confirmed association and measurement of the strength of association by odds ratios (for example, see [3]). For information on channels effectiveness assessment in the stated context, the last mentioned procedure will be useful.

2.1.1 Chi-squared Test of Homogeneity and Independence

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. We will look at the chi-squared test of homogeneity. Let *Y* is a response variable and *X* is explanatory variable. The categories of *X* define *r* different populations, for example different groups of people.

The random samples from *r* multinomial populations (Multinomial distribution, see for example in [5, 10]) with *c* different outcomes are sampled in that test.

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, and $n_{.j}$ be the sum of n_{ij} values in the *j*-th column of a contingency table. The sum of all n_{ij} values is the sample size *n*. In a contingency table, the row totals $n_{i.}$ in the last column are fixed, the column totals $n_{.j}$ are influenced by randomness of sampling.

For a fixed category of *X*, variable *Y* has a probability distribution. Let $\pi_{j|i}$ denote the probability of classification of the element in column *j* of *Y*, given that the element is classified in row *i* of *X* (the probabilities $\{\pi_{1|i},\ \pi_{2|i},\ \dots,\ \pi_{c|i}\}$ define the conditional probability distribution[1] of *Y* at category *i* of *X*). When a response variable is identified and the population conditional distributions are identical for all populations, they are said to be homogeneous ([3], p. 229).

We are testing the following:

[1] For more information about joint, conditional and marginal probability distribution, see [10], pp. 89 – 92.

H_0: The population conditional distributions are identical for all *r* populations, formally:

H_0: $\pi_{j|1} = \pi_{j|2} = \ldots = \pi_{j|r}$ for $j = 1, 2, \ldots, c$

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

H_1: The population conditional distributions are not identical for all *r* populations, formally:

H_1: $\pi_{j|1}$, $\pi_{j|2}, \ldots,$ $\pi_{j|r}$ are not all equal for at least one value of *j*.

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

$$o_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

The value of test statistics is calculated according to the 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 the level of significance α is $\chi^2 \geq \chi^2_{1-\alpha}((r-1)(c-1))$, where $\chi^2_{1-\alpha}((r-1)(c-1))$ is (1 – α) – quantile of chi-squared distribution with $(r-1)(c-1)$ degrees of freedom. Due to the considered test statistic following the chi-squared distribution only approximately, it is obviously recommended to use this test only in the cases when no expected frequencies values are lower than 5. If not, joining of corresponding columns is necessary.

Let both *X* and *Y* be response variables. Then, one sample from multinomial population with *rc* different outcomes is sampled. The statistical independence is tested and the test is called the test of independence.

We are testing the following:

H_0: The variables are statistically independent, formally:

H_0: $\pi_{ij} = \pi_{i.} \cdot \pi_{.j}$ for $i = 1, 2, \ldots r$ and $j = 1, 2, \ldots c$

versus an alternative:

H_1: The variables are statistically dependent, formally:

H_1: $\pi_{ij} \neq \pi_{i.} \cdot \pi_{.j}$ for at least one pair of values of *i* and *j*,

where $\pi_{i.}$ is the marginal probability in row *i*, $\pi_{.j}$ is the marginal probability in column *j*.

Two categorical variables are statistically independent if the population conditional distributions on one of them are identical at each category of the other. The variables are statistically dependent if the population conditional distributions are not identical (in [3], p. 223). Statistical independence is a symmetric property between two variables: If the conditional distributions within rows are identical, so are the conditional distributions within columns ([3], p. 224).

In the chi-squared test, the value of χ^2 test statistic does not depend on which one is the response variable and which one is the explanatory variable (if either). The steps of the test procedure and the results are identical either way ([3], p. 229). The testing procedure is the same as was mentioned above in the case of homogeneity, differences among more proportions (when the required value of proportion is not specified) or statistical independence testing **(**For more information on Pearson chi-square test of homogeneity and independence, see [2, 7]).

It is clear that the homogeneity of conditional distributions in a test of homogeneity implies statistical independence of corresponding variables. So, equivalent interpretations of a test of homogeneity results are possible. When we reject the null hypothesis in a test of homogeneity, we can conclude that we have obtained the evidence that the conditional distributions of response variable *Y* on *X* are not identical or that variables *X* and *Y* are statistically dependent9 or simply that *X* and *Y* are associated.

2.1.2 Residual Analysis

Cell-by-cell comparison of observed frequencies n_{ij} and expected frequencies o_{ij} reveals the nature of the evidence about association between variables. The difference $\left(n_{ij}-o_{ij}\right)$ is called a residual. The adjusted standardized residuals for two nominal variables can be defined as (in [3], p. 230):

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

where $\frac{n_{i.}}{n}$ − an estimated marginal probability in row *i*,

$\frac{n_{.j}}{n}$ − an estimated marginal probability in column *j*.

The denominator in formula (1) is a standard error of random variable $\left(n_{ij}-o_{ij}\right)$, when null hypothesis H_0 about statistical independence of variables[2] is true. Adjusted standardized

[2] Or identicality of conditional distributions.

residuals r_{ij} asymptotically follow the standard normal distribution. They can be used to describe the pattern of the association among the table cells. A too large value of an adjusted standardized residual (greater than 2 in the absolute value) indicates a deviation from homogeneity in the cell.

2.1.3 Measures of Association Strength for Nominal Variables

A measure of the association strength is a statistic or parameter that indicates the strength of an association between two variables (in [3], p. 233). There are more summary measures of association strength between two nominal variables. Two approaches to summarize the strength of the association between nominal variables are known:

- Coefficients based on the χ^2
- Coefficients based on proportional reduction of prediction error (PRE)

Association Coefficients based on the χ^2

The χ^2 statistic as such is not used to measure association between two variables, but it serves as the element in association coefficients construction. The following association coefficients based on χ^2 are suggested in the literature: Phi–square (φ^2), Cramer’s *V* and contingency coefficient.

The Cramer`s *V* is most frequently used. It is defined as

$$V = \sqrt{\frac{\chi^2}{n \cdot h}},$$

where *h* is the minimum from (*r* – 1) and (*c* – 1). Cramér's *V* varies from 0 (corresponding to no association between the variables) to 1 (complete association) and can reach 1 only when the two variables are equal to each other.

Association Coefficients based on PRE

Goodman and Kruskal introduced the idea of proportional reduction in error of prediction. Two association coefficients based on this idea are known − the Goodman and Kruskal's lambda and Goodman and Kruskal's tau.

The Goodman and Kruskal's lambda (λ) measures the percentage improvement in predictability of the response variable (row variable or column variable), given the value of the other variable (column variable or row variable).

The value of λ_r for row response variable is

$$\lambda_r = \frac{\sum_i \max n_{ij} - \max\left(n_{.j}\right)}{n - \max\left(n_{.j}\right)}$$

The value of λ_c for column response variable is

$$\lambda_c = \frac{\sum_j \max n_{ij} - \max(n_{i.})}{n - \max(n_{i.})}$$

The symmetric (non-directional) lambda (λ) can be also calculated. It lies between the values of λ_r and λ_c. The symmetric lambda is defined as

$$\lambda = \frac{\sum_i \max n_{ij} + \sum_j \max n_{ij} - \max(n_{i.}) - \max(n_{.j})}{2n - \max(n_{i.}) - \max(n_{.j})}$$

Asymmetric and symmetric lambda take values from the interval [0, 1]. The information obtained by summary measures of association is interesting but not so useful for managing the information flows (for more details see [6]).

Other Measures of Association Strength

When 2 *x* 2 contingency table is analyzed, the difference of proportions can be used as measure of association (see in [3], p. 234 or in [10], p. 298 − 301).

Odds Ratio

The odds ratio is the measure of association that can be used in all contingency tables. We will use success to denote the outcome of interest and failure to denote the other outcome. For a response variable with two values, the odds for success is defined as:

$$\text{Odds} = \frac{\text{Probability of success}}{\text{Probability of failure}}$$

The estimated odds for a response variable with two values equals the number of successes divided by the number of failures. The odds ratio θ in 2 *x* 2 contingency table equals 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 2 *x* 2 sub-table. In the context of the stated problem, we will use the odds ratios as the measure of relative effectiveness of information channels.

2.2 Procedure of Information Channels Effectiveness Assessment

We will suggest a procedure of information channels effectiveness assessment. In general, the different information channels are expressed by the *c* values of the first categorical variable, while the *r* populations − groups of people obtaining information through these information channels - are represented by values of the second categorical variable. Then the answers of respondents sampled by simple random sampling from each of *r* populations in the case of homogeneity testing are obtained. Alternatively, one sample from multinomial population with *rc* different outcomes can be realized. Then the independence is tested. The sampled respondents indicate the information channel they obtained information through in their answers.

Then, the Pearson chi-squared test of homogeneity or independence is applied depending on how the random sampling was conducted. When the null hypothesis is not rejected, we did not obtain the evidence that the effectiveness of information channels is different for different groups of people. When the null hypothesis is rejected we conclude that the effectiveness of information channels for different groups of people is not identic. Then the residual analysis is effectuated. This analysis determines the cells of contingency table "causing" association. In the context of information channels effectiveness assessment, the relative effectiveness of information channels for different groups of people can be determined by that analysis.

In the third step, information channels effectiveness assessment based on odds ratios is realized (more in details about third step see in [9]).

3 Using of the Procedure of Information Channels Effectiveness Assessment

The use of the just-mentioned procedure will be illustrated on a study of the problem of the information channels effectiveness assessment based on hypothetical data from a statistical survey of a firm.

Example. A firm carried out a statistical survey focused on customer satisfaction and their profiles study. The 400 customers were randomly selected and asked to fill in a questionnaire. One of the closed questions in the questionnaire was: "From which information source did you obtain the first information about the services provided by our firm?" The answer distribution is shown in Table 1 (in parentheses there are the expected frequencies).

Tab. 1 Distribution of Information Channels according to Permanent Residence of the Customer

Information Channels **Permanent residence of customer**	**1st Information Channel (paper publicity materials)**	**2nd Information Channel (internet)**	**3rd Information Channel (publicity in media or other sources)**	$n_{i.}$
Bratislava city	150 (134.375)	74 (74.375)	26 (41.250)	250
Region Bratislava (except Bratislava city)	20 (31.713)	21 (17.553)	18 (9.735)	59
Elsewhere	45 (48.913)	24 (27.073)	22 (15.015)	91
$n_{.j}$	215	119	66	400

Source: own

The Pearson chi-squared test of independence offered *p*-value = 0.000106. That means that H_0 can be rejected in favor of the alternative hypothesis. We can conclude that variables „information channels“ and „permanent residence of customer“ are associed. The effectiveness of the information channels about services provided by the firm differs by the permanent residence of a customer.

3.1 Residual Analysis and Information Channels Relative Effectiveness

The adjusted standardized residuals were calculated according to (1), to find the cells causing the association. The results are in Table 2.

Tab. 2 Information Channels and Adjusted Standardized Residuals for the Different Permanent Residence of the Customer

Information Channels **Permanent residence of customer**	**1st Information Channel (paper publicity materials)**	**2nd Information Channel (internet)**	**3rd Information Channel (publicity in media or other sources)**
Bratislava city	3.24	− 0.08	− 4.24
Region Bratislava (except Bratislava city)	− 3.31	1.06	3.14
Elsewhere	− 0.94	− 0.80	2.24

Source: own

Based on results in Table 2, we can conclude that there are more customers from the city Bratislava and fewer customers from the region of Bratislava who obtained information through the first information channel than it is suggested by the independency hypothesis. There are more customers from the region of Bratislava and fewer customers from the city Bratislava who obtained information through the third information channel than it is suggested by the independency hypothesis. We can conclude that the first information channel is relatively more effective for customers from the city Bratislava than for customers from the region of Bratislava and the third information channel is relatively more effective for customers from the region of Bratislava than for customers from the city Bratislava.

3.2 Relative Effectiveness Assessment of Information Channels

The odds ratio analysis will be applied. The association strength measured by odds ratio will be understood and interpreted as a value of relative effectiveness. In general, an arbitrary 2 *x* 2 sub-table can be analyzed by the odds ratio. The only appropriate approach is to analyze the sub-tables for which the corresponding residuals are greater than 2 in the absolute value.

We will analyze the following sub-table (in Table 3).

Tab. 3 Sub-table of Tab. 1

Information Channels **Permanent residence of customer**	**1st Information Channel (paper publicity materials)**	**3rd Information Channel (publicity in media or other sources)**	**Total**
Bratislava city	150	26	176
Region Bratislava (except Bratislava city)	20	18	38

Source: own

The relative effectiveness assessment of the first information channel for the customers from the city Bratislava in comparison to the customers from the region of Bratislava will be realized. The first information channel (the second column) will represent a success and the third one (the third column) will represent a failure.

The estimated odds for customers from the city of Bratislava is

$$\frac{\frac{150}{176}}{\frac{26}{176}} = \frac{150}{26} \approx 5.7692$$

There are about 5.7692 of customers from the city of Bratislava who obtained information through the first information channel per 1 customer who obtained the information through the third information channel.

The estimated odds for customers from the region of Bratislava is

$$\frac{\frac{20}{38}}{\frac{18}{38}} = \frac{20}{18} \approx 1.1111$$

There are about 1.1111 of customers from the region of Bratislava who obtained information through the first information channel per 1 customer who obtained the information through the third one.

The odds ratio for customers from the city Bratislava and for customers from the region of Bratislava can be calculated as follows:

$$\theta = \frac{5.7692}{1.1111} \approx 5.1923$$

A customer from the city Bratislava has a 5.1923 times greater chance to obtain information through the first information channel than a customer from the region of Bratislava. The first information channel is relatively 5.1923 times more effective for customers from the city Bratislava than for customers from the region of Bratislava.

It can be proven that when the third information channel represents success and the first one failure, the same odds ratio for the customers from the region of Bratislava will be obtained. This relation concerning odds ratios in contingency sub-tables is generally valid.

4 Conclusions

It was shown how to use data from a statistical survey and some methods of analysis of association between two nominal variables in the information channels effectiveness assessment. The testing of association enables us to make a decision about whether there exists an association between variables, where one variable represents information channels and the other one represents different groups of people obtaining information through these information channels.

When the null hypothesis is not rejected, the evidence about an association between variables is not obtained and we cannot conclude that there is a difference in relative effectiveness of information channels for different groups of people. When the null hypothesis is rejected, we can conclude that there is an association between variables. Once an association between variables is established, the question which combinations of variable values cause the identified association is interesting. Using of residual analysis is recommended for that purpose. Identification of cells "responsible" for an association in a contingency table enables us to recognize the relative effectiveness of information channels for different groups of people. When there is a great positive value of the adjusted standardized residual in the cell, the corresponding information channel is more effective for the corresponding group of people. When there is a great negative value of the adjusted standardized residual in the cell, the corresponding information channel is less effective for the corresponding group of people.

The use of odds ratios is recommended for information channels relative effectiveness assessment. In the application of the proposed procedure in the above mentioned example, the results showing that the first information channel is relatively 5.19 times more effective for customers from the city of Bratislava than for customers from the region of Bratislava and that the third information channel is relatively 5.19 times more effective for customers from the region of Bratislava than for customers from the city of Bratislava were obtained. In general, such an analysis based on odds ratios can be carried out for all 2 *x* 2 sub-tables of the contingency table of two nominal variables.

The described procedure can be also used in a lot of other contexts (see in [9]). The procedure does not require any professional software. MS Office with Excel is sufficient.

Acknowledgments

The paper was supported by grants from Grant Agency of VEGA no. 1/0393/16 entitled „European Union in the Post Crisis Period – Macro and Microeconomic Aspects“.

References

1. AGRESTI, A., 2010. *Analysis of Ordinal Categorical Data. Second Edition*. Hoboken: Wiley and Sons.
2. AGRESTI, A., 2013. *Categorical Data Analysis. Third Edition*. Hoboken: Wiley and Sons.
3. AGRESTI, A, FINLAY, B., 2014. *Statistical Methods for the Social Sciences. Fourth Edition*. Essex: Pearson.
4. DAGNELIE, P., 1998. *Statistique théorique et appliquée. Tome 2.* Paris: de Boeck and Larcier s.a.
5. FREUND, J. E., 1992. *Mathematical Statistics. Fifth Edition*. Englewood Cliffs: Prentice – Hall.
6. GOODMAN, L. A., KRUSKAL, W. H., 1954. Measures of association for cross classifications. Part I. *Journal of the American Statistical Association*, 1954, number 49, pp. 732 – 764.
7. MILLER, I., MILLER, M., JOHN, E., 2004. *Freund`s Mathematical Statistics with Applications. Seventh Edition.* Upper Saddle River: Pearson Prentice Hall.
8. SHARPE, N., DE VEAUX, R. D., VELLEMAN, P., 2010. *Business Statistics. Second Edition*. Boston: Pearson.
9. TEREK, M., 2016. Information Channels Effectiveness Assessment on the Basis of Data from Statistical Survey. *Scientific Annals of Economics and Business*, 63 (2), pp. 225 – 235.
10. TEREK, M., 2017. *Interpretácia štatistiky a dát. Piate doplnené vydanie*. Košice: Equilibria.

Contact data:

Milan Terek, prof., Ing., PhD.
Vysoká škola manažmentu v Trenčíne / City University of Seattle programs
Panónska cesta 17, 851 04 Bratislava, Slovakia
mterek@vsm.sk