Proceedings of the 12th International Conference "Reliability and Statistics in Transportation and Communication" (RelStat'12), 17–20 October 2012, Riga, Latvia, p. 36–44. ISBN 978-9984-818-49-8
Transport and Telecommunication Institute, Lomonosova 1, LV-1019, Riga, Latvia

IDENTIFICATION OF REGIONAL DISPARITIES IN THE FIELD OF INTERNET SERVICES BY THE CHI-SQUARE TEST FOR INDEPENDENCE

Alena Košťálová

University of Zilina
Faculty of Operation and Economics of Transport and Communications
Department of Communications
Zilina, Slovakia, Univerzitna 8215/1, Postcode: 010 26
Ph.: +421 415133143. E-mail: Alena.Kostalova@fpedas.uniza.sk

In the field of internet services we can see more or less significant regional disparities. The goal of this paper is to show them in the base of research, which was realised in five districts in the Zilina region (the Slovak Republic). One of the many outputs of the realised survey was identification of regional disparities in the utilisation of internet services. One of the methods, which can be used for identification of regional disparities, is the methodology of the chi-square test for independence between values of two categorical variables. There is characterized the methodology of the chi-square test for independence between values of two categorical variables and the mechanism of the determination the measures of association in the first part of this paper. The null hypotheses, which were tested by the chi-square test for independence and regional disparities resulting from testing, are described in the next part of paper. The bases for determination null hypotheses were one-way and two-way tables.

Keywords: chi-square test for independence, null hypothesis, the degree of association, regional disparities, internet services

1. Introduction

The chi-square test for independence is a nonparametric statistical analysis method often used in experimental work where the data consist in frequencies or 'counts'. The most common use of the test is to assess the probability of association or independence of facts [1]. This paper deals with the application of a chi-square test to the results of a marketing survey.

The survey was focused on small and medium enterprises in five districts in the Zilina region. The questionnaire comprised 25 questions. It was completed by a total of 250 small and medium enterprises across the five districts. The questions were divided into five sets, each referring to each other:

Set 1 – Postal services
- ✓ postal services providers
- ✓ frequency of selected postal services utilization
- ✓ knowledge of selected postal services
- ✓ method of postal fee payment
- ✓ method of mail delivery
- ✓ average monthly expenditure on postal services
- ✓ satisfaction with postal services
- ✓ strengths and weaknesses of the postal services
- ✓ trends in postal services development

Set 2 – Telecommunication services
- ✓ telecommunication services providers
- ✓ frequency of utilization of selected telecommunication services
- ✓ average monthly expenditure on telecommunication services
- ✓ satisfaction with telecommunication services
- ✓ strengths and weaknesses of the telecommunication services
- ✓ trends in telecommunication services development

Set 3 – Internet services
- ✓ implementation of data transmission
- ✓ type of internet connection
- ✓ provider of connection to the internet

Set 4 – Complementary questions

- ✓ marketing communication
- ✓ method of money t ransfer in business
- ✓ factors determining the method of communication

Set 5 – Respondent's details

- ✓ legal status of business
- ✓ business activity
- ✓ number of employees
- ✓ area covered by business
- ✓ annual turnover of business
- ✓ district where the firm is located.

Members of the sample are double classified (i.e. classified in two separate ways) and results are arranged in contingency tables. Contingency tables are the basis for the application of the chi-square test for independence.

Table 1. Contingency table [11]

A	B						(a_i)
	b_1	b_2	...	b_j	...	b_c	
a_1	$(a_1\ b_1)$	$(a_1\ b_2)$	...	$(a_1\ b_j)$	...	$(a_1\ b_c)$	(a_1)
a_2	$(a_2\ b_1)$	$(a_2\ b_2)$	...	$(a_2\ b_j)$	...	$(a_2\ b_c)$	(a_2)
...	...	...	...	...	...	...	...
a_i	$(a_i\ b_1)$	$(a_i\ b_2)$	...	$(a_i\ b_j)$	...	$(a_i\ b_c)$	(a_i)
...	...	...	...	...	...	...	...
a_r	$(a_r\ b_1)$	$(a_r\ b_2)$	...	$(a_r\ b_j)$	...	$(a_r\ b_c)$	(a_r)
(b_j)	(b_1)	(b_2)	...	(b_j)	...	(b_c)	n

Legend: $(a_i) = \sum_j (a_i\ b_j)$, $(b_j) = \sum_i (a_i\ b_j)$, $n = \sum_i (a_i) = \sum_j (b_j) = \sum_i \sum_j (a_i\ b_j)$

2. Chi-Square Test for Independence

The chi-square test for independence is a nonparametric statistical test to determine if two or more classifications of the samples are independent or not. A common question with regards to a contingency table is whether it has independence. By independence, we mean that the row and column variables are unassociated (i.e. knowing the value of a row variable will not help us predict the value of a column variable, and likewise, knowing the value of a column variable will not help us predict the value of a row variable) [3].

The methodology of the chi-square test for independence between values of two categorical variables is divided into four steps. The first step is the expression of the null and alternative hypothesis. The second step is to determine the significance level (α). The third step is to calculate the chi-square test statistic (χ^2). The fourth step is to compare the computed (χ^2) with the critical value in the table for the significance level (α) and then to make a statistical decision in regard to the null hypothesis.

2.1. Expression of Hypotheses

The null hypothesis H_0 expresses the independence of variables. In contrast, the alternative hypothesis H_a, which we want to prove to be true in the majority of cases, mostly expresses a statistical association of the variables. The truth of the alternative hypothesis is always shown only indirectly, in a way that will show that the null hypothesis is unlikely, and that the alternative hypothesis (the only remaining hypothesis) is therefore likely. Independence is tested by a chi-square test, which based on the chi-square distribution. The chi-square distribution has paramount importance in a dependency analysis in the association and contingency tables [3]. The chi-square distribution is mainly used in comparison of observed frequencies $O_{ij} = (a_i\ b_j)$ from the contingency table with expected (theoretical) frequencies $E_{ij} = (a_i)*(b_j) / n$.

In a chi-square test for independence the null and alternative hypotheses are expressed:

H_0: The values of two categorical variables A and B are independent;

H_a: The values of two categorical variables A and B are related.

2.2. Determination of the Significance Level (α)

If we reject the null hypothesis when it is in fact valid, we make a 'type 1' error (i.e. we come to the conclusion that there is a relationship between the variables when in fact there is none). The significance level (α) is the probability of committing a 'type 1' error. We can reduce the chances of making a 'type 1' error by selecting a smaller value for (α). This makes it more likely that the null hypothesis will be accepted, but it also increases the risk of making a 'type 2' error (i.e. incorrectly concluding that there is no relationship between the variables). Determining the significance level is thus a sort of compromise between these two types of errors, and its choice depends on the type of tested facts, on the experience of the researcher, etc. [4]. Alpha (α) is traditionally set at 5% (= 0.05), or 1% (= 0.01). Variations that occur with a probability less than the chosen significance level are called statistically significant at the selected significance level. In this paper the significance level $\alpha = 0.05$ is used.

2.3. Calculation of the Chi-Square Test Statistic

The chi-square test statistic that asymptotically approaches a chi-square distribution was first introduced by the British statistician Karl Pearson in 1900. A chi-square distribution is mostly used in the testing of a compliance table with some theoretical model. This involves comparing observed and expected frequencies. Expected frequencies are those which should be observed if the values of two categorical variables A and B are independent [5]. In order to compare the observed and expected frequencies we produce the chi-square value (χ^2) using the formula in equation 1:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1}$$

where

χ^2 … the test statistic that asymptotically approaches a chi-square distribution,
O_{ij} … the observed frequency of the i^{th} row and j^{th} column,
E_{ij} … the expected (theoretical) frequency of the i^{th} row and j^{th} column,
r … the number of rows in the contingency table,
c … the number of columns in the contingency table.

A second important part of determining the chi-square test statistic is to define the degrees of freedom (df) of the test. The degree of freedom of a contingency table with *r* rows and *c* columns is computed using the following formula given in equation 2:

$$df = (r-1)*(c-1). \tag{2}$$

When using the chi-square test in tables larger than 2 by 2 Cochran suggests that no more than 20% of the expected frequencies should be less than 5 and that all individual expected frequencies should be 1 or greater [6, 7]. This suggestion, which can be written by using probability P (formula 3), is an assumption/restriction on the use of the chi-square test in contingency tables:

$$P\left[E_{ij} < 5\right] \le 0{,}2 \wedge E_{ij} > 1. \tag{3}$$

If any expected frequencies in 2 by 2 tables are less than 10, but greater than or equal to 5, some authors suggest that Yates´ Correction of Continuity should be applied [8]. This is done by subtracting 0.5 from the absolute value of (O_{ij} - E_{ij}) before squaring (equation 4). However, the use of Yates´ Correction of Continuity is controversial, and is not recommended by all authors.

$$\chi^2_{(Yates)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\left|O_{ij} - E_{ij}\right| - 0{,}5)^2}{E_{ij}}. \tag{4}$$

2.4. Conclusion of the Null Hypothesis

The last step is comparing the calculated (χ^2) with the critical value in the table at the significance level (α). The statistical decision in regard to the null hypothesis depends on validity of the inequality that is shown in formula 5:

$$\chi^2 \ge \chi^2_{1-\alpha}\left[(r-1)*(c-1)\right]. \tag{5}$$

There may be two variants [4]:

- ✓ formula (5) is valid – the computed (χ^2) is equal to or greater than the critical value. We reject the null hypothesis, and confirm the alternative hypothesis. The different between the observed and expected frequencies is statistically significant. Therefore we conclude that there is a relationship between the variables.
- ✓ formula (5) is invalid – the computed (χ^2) is less than the critical value. The null hypothesis cannot be rejected. The difference between the observed and expected frequencies is not statistically significant. However, this does not mean that the null hypothesis is true. It indicates that there is insufficient evidence of a relationship between the variables. So, if there is a small probability of a 'type 2' error, hypothesis H_0 could be accepted.

3. Measures of Association

Use of the chi-square test for independence can provide information on whether the association between values of two categorical variables A and B can be regarded as statistically significant or not. In the case of statistical significance of this association, it is possible to evaluate the degree of association only indirectly. Direct evaluation of the degree of association can be done using measures of association [3, 9], which are based on the computed chi-square value (χ^2). The measures of association are: the Phi coefficient, the Contingency coefficient C, Cramer's V, and Tschuprow's coefficient τ. The nearer the value of the measure of association is to 0, the greater the degree of independence between the values of two categorical variables A and B is confirmed (interpretation of the measure of association values is shown in table 2). There are some restrictions on the use of measures of association [9]:

- ✓ Phi coefficient is used in 2 by 2 tables,
- ✓ Contingency coefficient C (Pearson's C) is only used for 5 by 5 tables or larger,
- ✓ Cramer's V is the most popular measure of association regardless of table size (in this paper it was used in tables larger than 2 by 2 but smaller than 5 by 5),
- ✓ Tschuprow's coefficient τ is used only in square tables where row marginals are identical to column marginals (this coefficient is little used and is not supported by major statistical packages).

Table 2. The measures of association value and its interpretation [10]

Measure of association value	Degree of association
0,0	perfect independence
0,0 – 0,1	trivial association
0,1 – 0,3	small association
0,3 – 0,5	moderate association
0,5 – 0,7	large association
0,7 – 0,9	very large association
0,9 – 1,0	nearly perfect association
1,0	perfect association

4. Application of the Chi-Square Test for Independence to the Results of the Marketing Survey

One of the many outputs of the realized survey was identification of regional disparities in the utilization of internet services by application of chi-square test for independence between values of two categorical variables. Several pairs of categorical variables were identified within the framework of the marketing survey. For each of the identified pairs of categorical variables (A and B) the chi-square test for independence was applied. It defined the null hypothesis concerning their independence. Then the criteria for rejecting or not rejecting the null hypothesis were defined. In the case of rejecting the null hypothesis, the degree of association was determined by one of the measures of association (according to the restrictions in chapter 3).

The next three chapters include results, which are presented in three groups: internal test for independence, cross test for independence and test for independence in the regional context.

5. Internal Test for Independence inside Set 3 – Internet Services

In the internal test for independence inside set 3, both categorical variables relate to set 3 – internet services. Three pairs of categorical variables were identified. The following hypotheses were devised to assess their independence:

H_0: Implementation of data transmission is independent on the provider of connection to the internet

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A moderate association between categorical variables was confirmed by the Cramer's V (0,466).

It means, that implementation of data transmission depended on the provider of connection to the internet. Dependence, which has been established, results from perfect association between two variables:

- ✓ data transmission is not realised,
- ✓ respondent does not use any of the providers of connection to the internet.

H_0: Implementation of data transmission is independent on the type of internet connection

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A moderate association between categorical variables was confirmed by the Cramer's V (0,477).

It means, that implementation of data transmission depended on the type of internet connection. As in the previous case, perfect association between two variables, mentioned above, influenced the test result towards to moderate association.

H_0: The type of internet connection is independent on the provider of connection to the internet

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A moderate association between categorical variables was confirmed by the Cramer's V (0,393).

It means that the type of internet connection depended on the provider of connection to the internet. This fact relates to the different technical possibilities of connection to the internet offered by providers.

6. Cross Test for Independence

In the cross test for independence, the first categorical variable relates to set 3 – internet services, and the second categorical variable relates to the one of the rest sets (1, 2, 4, or 5).

6.1. Cross Test for Independence between Set 3 – Internet Services, and Set 1 – Postal Services

Three pairs of categorical variables were identified between set 3 – internet services, and set 1 – postal services. The following hypotheses were devised to assess their independence:

H_0: Frequency of selected postal service's utilization (2^{nd} class letter) is independent on the implementation of data transmission

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,256).

Respondents were asked what the frequency in their utilization of selected postal services is. The selected postal services were: 1^{st} class letter, 2^{nd} class letter, 1^{st} class parcel, 2^{nd} class parcel, express mail service, direct mail advertising, business parcel and postal money order. Statistically significant dependence was confirmed only between variables expressed in this H_0.

H_0: Satisfaction with postal services is independent on the implementation of data transmission

(assumption: H_0 cannot be rejected)

The assumption was confirmed. The null hypothesis could be accepted.

H_0: Average monthly expenditure on postal services is independent on type of internet connection

(assumption: H_0 cannot be rejected)

The assumption was not confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,218).

It means that respondents' average monthly expenditure on postal services depended on type of internet connection. It results from the fact: missing internet connection influenced respondents' average monthly expenditure on postal services.

6.2. Cross Test for Independence between Set 3 – Internet Services, and Set 2 – Telecommunication Services

Three pairs of categorical variables were identified between set 3 – internet services, and set 2 – telecommunication services. The following hypotheses were devised to assess their independence:

H_0: The trend in telecommunication services development, which respondent assumed, is independent on the implementation of data transmission

(assumption: H_0 will be rejected)

It was not possible to confirm or refute this assumption. The pair of categorical variables was not suitable for the application of a chi-square test for independence. The clause (3) was not fulfilled.

H_0: Average monthly expenditure on telecommunication services is independent on the provider of connection to the internet

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,256).

It means that respondents' average monthly expenditure on telecommunication services depended on the provider of connection to the internet, because providers offer different prices for connection and using of the internet.

H_0: Frequency of selected telecommunication service's utilization (fax service) is independent on the type of internet connection

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,296).

It means that the frequency in respondent's utilization of fax service depended on the type of internet connection. It results from the fact: if respondent haven't internet connection, he more frequently uses fax service.

6.3. Cross Test for Independence between Set 3 – Internet Services and Set 4 – Complementary Questions

Three pairs of categorical variables were identified between set 3 – internet services, and set 4 – complementary questions. The following hypotheses were devised to assess their independence:

H_0: The type of internet connection is independent on method of money transfer in business

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,222).

One of the methods of money transfer in business is internet banking, which directly depends on the use of internet connection. This fact influenced the test result towards to (only) small association.

H_0: Respondent's selection of the provider of connection to the internet is independent on factors determining the method of his communication

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,188).

It means that respondent's selection of the provider of connection to the internet depended on factors determining the method of his communication.

H_0: The type of internet connection is independent on form of marketing communication

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,176).

In marketing communication is better, though not necessary, to have internet connection. This fact influenced the test result towards to small association.

6.4. Cross Test for Independence between Set 3 – Internet Services and Set 5 – Respondent's Details

Three pairs of categorical variables were identified between set 3 – internet services, and set 5 – respondent's details. The following hypotheses were devised to assess their independence:

H_0: Implementation of data transmission is independent on number of employees

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,206).

It means that implementation of data transmission depended on number of employees. The test result was influenced by these facts:

- ✓ respondents with a smaller number of employees (0 – 9) don't realize the data transmission,
- ✓ respondents with a higher number of employees this option less marked.

H_0: The type of internet connection is independent on number of employees

(assumption: H_0 cannot be rejected)

It was not possible to confirm or refute this assumption. The pair of categorical variables was not suitable for the application of a chi-square test for independence. The clause (3) was not fulfilled.

H_0: Respondent's selection of the provider of connection to the internet is independent on number of his employees

(assumption: H_0 cannot be rejected)

The assumption was not confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,224).

The test result was significantly influenced by this fact: for application of the chi-square test for independence was necessary to combine several classes with low frequency counts.

7. Test for Independence in the Regional Context

As mentioned above (chapter 1), the respondents in the marketing survey were small and medium enterprises located in the five districts of the Zilina region (Bytca, Cadca, Kysucke Nove Mesto, Namestovo and Zilina). There were identified regional disparities in the field of internet services by the test for independence in the regional context.

In the test for independence in the regional context, the first categorical variable relates to set 3 – internet services, and the second categorical variable relates to the selected district. Three pairs of categorical variables were identified. The following hypotheses were devised to asses their independence:

H_0: Implementation of data transmission is independent on the district where the respondent's firm is located

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,213).

There were confirmed regional disparities in the form of small association between implementation of data transmission and the district where the respondent's firm is located. Detailed examination of the results of double classification in the contingency table showed that:

- ✓ respondents from district Namestovo marked at least the option "we don't realize the data transmission",
- ✓ respondents from three districts – Bytca, Cadca and Kysucke Nove Mesto – marked in equal measure the option "we realize the data transmission by CD, USB,…".

H_0: The type of internet connection is independent on the district where the respondent's firm is located

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A moderate association between categorical variables was confirmed by the Contingency coefficient C (0,427).

Significant regional disparities in this area, which were confirmed, relate to the technical options of internet connectivity in mentioned districts.

H_0: Respondent's selection of the provider of connection to the internet is independent on the district where the respondent's firm is located

(assumption: H_0 will be rejected)

The assumption was confirmed. The null hypothesis was rejected; dependence between variables is statistically significant. A small association between categorical variables was confirmed by the Cramer's V (0,205).

As in the previous case, regional disparities in this case were confirmed too. It means that respondent's selection of the provider of connection to the internet depended on the district where the respondent's firm is located.

8. Conclusions

The marketing survey realized in a regional entrepreneurial environment, on which this paper is based, was conducted in order to obtain relevant information about existing postal, telecommunication and internet services and about customer satisfaction with these services. The results of the marketing survey were double classified and then the chi-square test for independence was applied to the results. This methodology can also be used for other questionnaire surveys focused on other areas. The results of regular surveys can be used as groundwork for suggestions and recommendations concerning improving the quality of the postal, telecommunication and internet services, to identify regional disparities in their utilization and also for market segmentation. The assumed regional disparities were confirmed, but also highlighted the need for further investigation of factors influencing customer behaviour.

References

1. Maxwell, A. E. (1971). *Analysing Qualitative Data*. 4th Edition. Chapman and Hall Ltd., Library of Congress Catalog, Card Number 75–10907.
2. Zibran, M. F. (2007). *CHI-Squared Test of Independence*. Department of Computer Science, University of Calgary, Alberta, Canada. [online]. Retrieving date of access August 12, 2010, from <http://pages.cpsc.ucalgary.ca/~saul/wiki/uploads/CPSC681/topic-fahim-CHI-Square.pdf>.
3. Chajdiak, J., Komorník, J., Komorníková, M. (1999). *Štatistické metódy*. *(Statistical methods).* Bratislava: STATIS.
4. Linczényi, A. (1974). *Inžinierska štatistika*. *(Engineering Statistics)*. Bratislava: ALFA.
5. Sojková, Z. *Materiál na prednášky z predmetu Ekonomická štatistika*. *(Material for lectures on the subject Economic Statistics)*. Department of Statistics and Operation Research, Faculty of Economics and Management, Slovak University of Agriculture in Nitra, Slovak Republic. [online]. Retrieving date of access August 12, 2010, from <http://www.fem.uniag.sk/ACADEMIC/depart/ksov/predmety/statistika/subory/prednaska6a.ppt#26>.
6. Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics,* 10, 417–451.
7. Yates, D., Moore, D., McCabe, G. (1999). *The Practice of Statistics*, 1st edition. New York: W. H. Freeman.
8. Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society,* 1(2), 217–235. JSTOR Archive for the journal
9. *Nominal Association: Phi, Contingency Coefficient, Tschuprow's T, Cramer's V, Lambda, Uncertainty Coefficient*. [online]. Retrieving date of access August 12, 2010, from <http://faculty.chass.ncsu.edu/garson/PA765/assocnominal.htm>.
10. *Cohen's scale for correlation coefficient.* [online]. Retrieving date of access August 16, 2010, from <http://www.sportsci.org/resource/stats/effectmag.html>.
11. Pacáková, V. et al. (2003). *Štatistika pre ekonómov*. *(Statistics for Economists).* Bratislava: Edícia EKONÓMIA.

Grant Support
KOŠŤÁLOVÁ, A. et al.: Aplikácia diagnostiky v riadení kvality v službách, so zameraním na poštové služby *(Application of Diagnostics in Quality Management in Services Focused on the Postal Services)*. Institutional project No. 4/KS/2012. University of Zilina, Faculty of Operation and Economics of Transport and Communications, 2012.

KOLAROVSZKI, P. et al.: Podpora vzdelávania pomocou výučbových multimediálnych aplikácií *(Support education by educational multimedia applications)*. Institutional project No. 3/KS/2012. University of Zilina, Faculty of Operation and Economics of Transport and Communications, 2012.

The paper was conducted within the project VEGA 1/0687/11 "Assessment of the Business Excellence status".