Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie

5 (941)

ISSN 1898-6447
Zesz. Nauk. UEK, 2015; 5 (941): 49–61
DOI: 10.15678/ZNUEK.2015.0941.0504

Justyna Brzezińska
Department of Economic and Financial Analysis
University of Economics in Katowice

The Problem of Zero Cells in the Analysis of Contingency Tables*

Abstract

Log-linear analysis is a statistical tool used to analyse the independence of categorical data in contingency tables. With this method, any number of nominal or ordinal variables can be analysed: interactions can be included in the model, various types of association can be analysed, and the analysis provides a formal model equation. Although log-linear analysis is a versatile statistical method, there are some limitations in using it due to zero cells. Zero cells in contingency table are of two types: fixed (structural) and sampling zeros. Fixed zeros occur when it is impossible to observe values for certain combinations of the variable. Sampling zeros are due to sampling variations and the relatively small size of the sample when compared with a large number of cells.

In the paper several options will be presented for how to deal with zero cells in a table. All calculations will be conducted in **R**.

Keywords: zero cell, categorical data analysis, contingency tables, log-linear analysis.

1. Introduction

The analysis of discrete multivariate data, especially in the form of cross--classification, has occupied a prominent place in multivariate statistical analysis.

* The article was written for a project financed by a National Science Centre grant based on decision DEC-2012/05/N/HS4/00174.

A variety of data from the social, medical, psychological and biological sciences come in the form of cross-classified table counts, commonly referred to as contingency tables. A two- or multi-way table gives the observed counts simultaneously for the categories of two- or more categorical variables.

One of the most useful and powerful methods for analysing qualitative data is log-linear analysis. This method enables examination of the relationship between categorical data. It includes the analysis of multi-way tables where the dimensionality of the table refers to the number of variables. And it is an appropriate modelling method concerning multi-way tables, including interactions which are useful in identifying. In log-linear analysis, the expected value of observation is given by a linear combination of a number of parameters. Maximum likelihood method is used to estimate the parameters, and the estimated parameter values may then be used in identifying which variables are of greatest importance in predicting the observed values (Everitt 1977). This method has a variety of advantages and can give more complex and detailed information about data structure and association type. Log-linear analysis can be used for nominal as well as ordinal variables, and it provides a variety of models describing the association path (Brzezińska 2015).

Categorical data are usually described in contingency tables (cross-table), and can be split into two parts in the table:

1) the fully classified cases where information on all the categories is available (complete tables),

2) the partially classified cases where information on some of the categories is zero (zero cell tables).

Zero cells may cause some problems with further categorical data analysis (Fienberg 1980, Andersen 1997, Smirnoff 2003).

In this paper I present an analysis of contingency tables with zero cells. All calculations are done in **R**.

2. Contingency Tables Containing Zero Cells

A contingency table is incomplete if one or more cells have a zero count. We distinguish between sampling (random) and structural (fixed) zeros.

Sampling zeros are due to sampling variation and the relatively small size of the sample when compared to the large number of cells. These zeros disappear when the sample size is increased sufficiently. Sampling zeros occur when there is no observation in the cell, i.e. $n_{hj} = 0$, but probabilistically there is a chance of observing this value and the probability of observation in a cell is $\pi_{hj} > 0$. Increasing the sample size might yield the result $n_{hj} > 0$. Sampling zeros typically correspond to small expected counts, however, so they can indicate that the usual

asymptotic approximations for goodness-of-fit tests, tests of significance, etc., might not be valid.

Structural zeros occur when it is impossible to observe values for a certain combination of the variable, i.e. $n_{hj} = 0$ and $\pi_{hj} = 0$. Tables with structural zeros are structurally incomplete and they are known as incomplete tables. Such cases are different from those in which it is not possible to completely cross-classify all individuals or units. When we deal with tables containing structural zeros, the cells cannot be filled in with zeros, nor can the table be collapsed until there are no zeros in the table, nor can the analysis simply be abandoned.

3. Multi-way Frequency Analysis for Contingency Tables with Zero Cells

Correspondence analysis is a method applicable for analyses of contingency tables to analyse the relations between two or more categorical variables (Greenacre 1984). The method is performed in three steps. The first is to calculate the categorical profiles (i.e., the relative frequencies) and masses (marginal proportions). The next is to compute the chi-square distances between the points and find the *n*-dimensional space that best fits the points (Clausen 1998). The graphical representation of correspondence analysis is usually presented in a perception map. Because correspondence analysis is so widely known, details of its use are not presented here – only an application for zero-cell tables is shown in comparison to log-linear analysis.

Log-linear analysis is a standard tool for analysing the path of association between nominal or ordinal variables in a multi-way contingency table. The criteria to be analysed are the expected cell frequencies m_{hjk} represented as a function of all variables in the survey. There are several types of log-linear models related to several types of association, depending on the number of variables and interactions included. Models are built with the hierarchy principle saying that a parameter of lower order cannot be removed when there is still a parameter of higher order that concerns at least one of the same variable.

The goodness of fit of a log-linear model for a two-way table is tested using the Pearson's chi-square statistic or the likelihood ratio statistic:

$$G^2 = 2\sum_{h=1}^{H}\sum_{j=1}^{J} n_{hj} \ln\left(\frac{n_{hj}}{m_{hj}}\right). \tag{1}$$

Therefore, larger G^2 values indicate that the model does not fit the data well and thus it should be rejected. In order to find the best model from a set of possible models, additional measures should be considered (determination coef-

ficients, information criteria). It is also advisable to compute G^2/df where a value close to 1 indicates a model that fits well.

The Akaike Information Criterion *AIC* is based on information theory, but a heuristic way to think about it is a criterion that seeks a model that has a good fit to the truth but few parameters. The chosen model is the one that minimises the Kullback-Leibler distance between the model and the truth. Akaike information criterion refers to the information contained in a statistical model according to the equation (Akaike 1973):

$$AIC = G^2 - 2df, \tag{2}$$

where df is the residual degrees of freedom.

Another information measurement is Bayesian information criterion (Raftery 1986):

$$BIC = G^2 - df \cdot \ln n, \tag{3}$$

where n is the total sample size.

The model that minimises *AIC* and *BIC* will be chosen. A rule of thumb to determine the degrees of freedom for the table without zeros is df = number of cells – number of free parameters. In order to test the goodness of fit of a model that uses an observed set of marginal totals with at least one zero entry, we must reduce the degrees of freedom associated with the statistic. This is because if an observed margin entry is zero, both expected and the observed entries for all cells is known to be perfect once it is observed that the marginal entry is zero. As a result, we must delete those degrees of freedom associated with the fit of the zero cell values. A general formula for computing degrees of freedom in cases where some of the margins fitted contain sampling zeros is the following (Fienberg 1980):

$$df = (T_e - Z_e) - (T_p - Z_p), \tag{4}$$

where:

T_e – the number of cells in the table that are being fitted,

T_p – the number of parameters fitted by model,

Z_e – the number of cells containing zero estimated expected values,

Z_p – the number of parameters that cannot be estimated because of zero marginal totals.

Log-linear analysis is a widely known statistical method used to analyse categorical data in contingency tables. Although log-linear models are versatile statistical models, there are some limitations in using them, largely due to zero cells that may arise in the contingency table. There are two consequences of the zero cells problem. First, we cannot include many variables in the analysis. This is related to the second consequence: eventual collapsing of the variables in

order to avoid zero cells in the table, which may distort the process being modeled and may result in a loss of some valuable data. Also, odds and odds ratios are undefined with zeros in the denominator (Ishii-Kuntz 1994).

Other than collapsing variable categories, several options are available for analysing a table with zero cells:

1) add a small value (0.5 is frequently suggested) to every cell in the table when fitting the saturated model (Goodman 1970),

2) add a small quantity (such as 0.2) only to zero cells (Evers & Namboodiri 1977),

3) add the value $\frac{1}{r}$ to zero cells, where r equals the number of response categories (Grizzle, Starmer & Koch 1969),

4) arbitrarily define zero divided by zero to be zero (Fienberg 1980),

5) increase the sample size sufficiently to remove all zero cells (Knoke & Burke 1980),

6) replace sampling zeros by 0.1×10^{-8}, or a smaller number and then check results against those obtained without such an adjustment (Clogg & Eliason 1988).

Technically, sampling and structural zeros are treated in the same way. The reason is that in any test statistic, a term corresponding to a cell with a zero count will cancel out. Only for the saturated model is it necessary that the table be complete with no zeros (Smirnoff 2003). Researchers need to consider carefully the limitations of log-linear models in order to analyse the categorical data in the table effectively.

A comparison of these approaches in log-linear analysis for a contingency table containing zeros will be presented.

4. Application in R

Data come from the Central Statistical Office of Poland, from the Local Data Bank, and show the number of individuals fatally injured in accidents at work in the first three quarters of 2013. A two-dimensional contingency table for 2 variables was worked out and analysed:

1) *Voivodeship* (1. Dolnośląskie, 2. Kujawsko-pomorskie, 3. Lubelskie, 4. Lubuskie, 5. Łódzkie, 6. Małopolskie, 7. Mazowieckie, 8. Opolskie, 9. Podkarpackie, 10. Podlaskie, 11. Pomorskie, 12. Śląskie, 13. Świętokrzyskie, 14. Warmińsko-mazurskie, 15. Wielkopolskie, 16. Zachodniopomorskie),

2) *Cause of the accident* (1. Electricity, 2. Explosion, fire, 3. Ignition, 4. Material slipping and falling or collapsing on the person, 5. The person slipping and falling).

Out of 80 cells (16×5), 48 cells contain zeros. The sample size is 55.

Now we compare some options of transformation for dealing with zero cells in log-linear analysis. As the saturated model should be built only for non-zero cells, this model will not be analysed. The effects of zero cell action for the independence model [*Voivodeship*][*Cause*] are presented in table 1.

Table 1. Adjustment and Goodness of Fit Criteria for a Two-way Zero-cells Table

Adjustment	G^2	*df*	G^2/df	*AIC*	*BIC*
No adjustment for zeros	47.380	60	0.790	–72.620	–193.060
$n \rightarrow n + 0.5$	21.331	60	0.356	–98.669	–219.109
$n = 0 \rightarrow n + 0.2$	21.331	60	0.356	–98.669	–219.109
$n = 0 \rightarrow n = 0.1 \times 10^{-8}$	21.331	60	0.356	–98.669	–219.109

Source: the author's own calculations in **R** based on data from the Central Statistical Office (www.stat.gov.pl).

The goodness of fit statistics for no adjustment seem to be the best for the independence model. The comparison for other transformations shows that the result for the adjustments conducted is the same and no significant differences are seen.

The second example is based on data `Titanic{datasets}` summarised in a four-way table. Data provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic summarised according to the variables *Class*, *Sex*, *Age* and *Survival*. Out of 32 cells ($4 \times 2 \times 2 \times 2$), 8 cells are zeros. The sample size is 2201.

The effects of zero cell action for the independence model [*Class*][*Sex*][*Age*] [*Survived*] are summarised in Table 2.

Table 2. Adjustment and Goodness of Fit Criteria for the [*Class*][*Sex*][*Age*][*Survived*] Model

Adjustment	G^2	*df*	G^2/df	*AIC*	*BIC*
No adjustment for zeros	1243.663	25	49.747	1193.663	1051.246
$n \rightarrow n + 0.5$	1216.387	25	48.655	1166.387	1023.970
$n = 0 \rightarrow n + 0.2$	1229.594	25	49.184	1179.594	1037.177
$n = 0 \rightarrow n = 0.1 \times 10^{-8}$	1243.663	25	49.747	1193.663	1051.246

Source: the author's own calculations in **R**.

Table 2 shows that the best transformation for zero cells is adding 0.5 to all cells. The likelihood criteria, as well as information criteria, are minimal for

this adjustment, and G^2/df is closest to 1. These results show that the best fit is obtained for the adjustment where we add 0.5 to each cell in the table.

As the interaction between *Class* and *Survived* seems to be interesting, the second model tested is the conditional association model [*ClassSurvived*][*Sex*][*Age*]. The goodness of fit statistics are summarised in Table 3.

Table 3. Adjustment and Goodness of Fit Criteria for the [*ClassSurvived*][*Sex*][*Age*] Model

Adjustment	G^2	*df*	G^2/df	*AIC*	*BIC*
No adjustment for zeros	1062.762	22	48.307	1018.762	893.435
$n \rightarrow n + 0.5$	1036.113	22	47.096	992.113	866.786
$n = 0 \rightarrow n + 0.2$	1049.425	22	47.701	1005.425	880.098
$n = 0 \rightarrow n = 0.1 \times 10^{-8}$	1062.762	22	48.307	1018.762	893.435

Source: the author's own calculations in **R**.

The result is the same as for the previous model. The best fit is obtained for the adjustment where 0.5 is added to each cell.

As the interaction between *Class*, *Sex* and *Survived* is interesting, the next model tested is the conditional association model [*ClassSurvivedSex*][*Age*]. The goodness of fit statistics are summarised in Table 4.

Table 4. Adjustment and Goodness-of-fit Criteria for the [*ClassSurvivedSex*][*Age*] Model

Adjustment	G^2	*df*	G^2/df	*AIC*	*BIC*
No adjustment for zeros	225.338	15	15.023	195.338	109.888
$n \rightarrow n + 0.5$	210.616	15	14.041	180.616	95.166
$n = 0 \rightarrow n + 0.2$	214.707	15	14.314	184.707	99.257
$n = 0 \rightarrow n = 0.1 \times 10^{-8}$	225.338	15	15.023	195.338	109.888

Source: the author's own calculations in **R**.

Table 4 shows a similar result: the best fit is obtained for the adjustment where we add 0.5 to each cell.

The results presented in Tables 2, 3 and 4 prove that the best transformation of the zero cell in a multi-way table is obtained when we add 0.5 to each cell. The other adjustments are not significantly worse and the difference in the fit is not large.

The differences between the presented adjustments for a two- and four-way table are not significantly different and the results obtained may not be general-

ised, as some of them show only very little improvement. However, for the four--way table, adding 0.5 to each cell in the table leads to the best fitting model.

5. Visualising Categorical Data

There are many techniques and methods for visualising categorical data in contingency tables. They may be defined as a recursive generalisation of bar charts. This visualisation in log-linear analysis is possible with the use of mosaic plot and association plot. Both are available in **R** in the `vcd` package (*visualising categorical data*), which was developed by *Visualising Categorical Data* (Friendly 1991, 2000; Cohen 1980). There are not many publications on visualising categorical data in the Polish literature (Brzezińska 2013).

Visualisation graphically brings out the departure of an observed table from the expected table. Mosaic plot is one of the most popular and useful methods for log-linear modelling, which generalises readily to multi-way tables. Friendly (2000) extended the use of mosaic plots for fitting log-linear models. A mosaic represents each cell of a table with a rectangle (or tile) whose area is proportional to the cell count. The mosaic is constructed by dividing a unit square vertically by one variable, then horizontally by the other. Further variables are introduced by recursively subdividing each tile by the conditional proportions of the categories of the next variable in each cell, alternating between the vertical and horizontal dimensions of the display. The mosaic plot shows that the difference between observed and expected cell counts are small and all Pearson's residuals ($d_{hj} = \frac{n_{hj} - m_{hj}}{\sqrt{m_{hj}}}$) are $|d_{hj}| < 2$ and the fit of the model is good. Visualising log-linear models in mosaic plot in **R** is possible with the use of the `mosaic{vcd}` function (Brzezińska 2014).

Mosaic plots are presented for different adjustments of zero cells for the independence model [*Voivodeship*][*Cause*] (Figures 1–4). Voivodeship names are coded from 1 to 16, according to the notation in the beginning of Section 4.

The result of the mosaic plots is very similar for every adjustment except the first one. The largest departure of an observed cell from a theoretical one exists in Fig. 1, in the cell for the Zachodniopomorskie region for the cause of the accident: explosion or fire. The other plots (Figures 2, 3 and 4) do not show significant departures between observed and expected cell counts and they indicated that the fit of those models is very good. Similar results occur for the tables of higher dimensions for the second dataset.

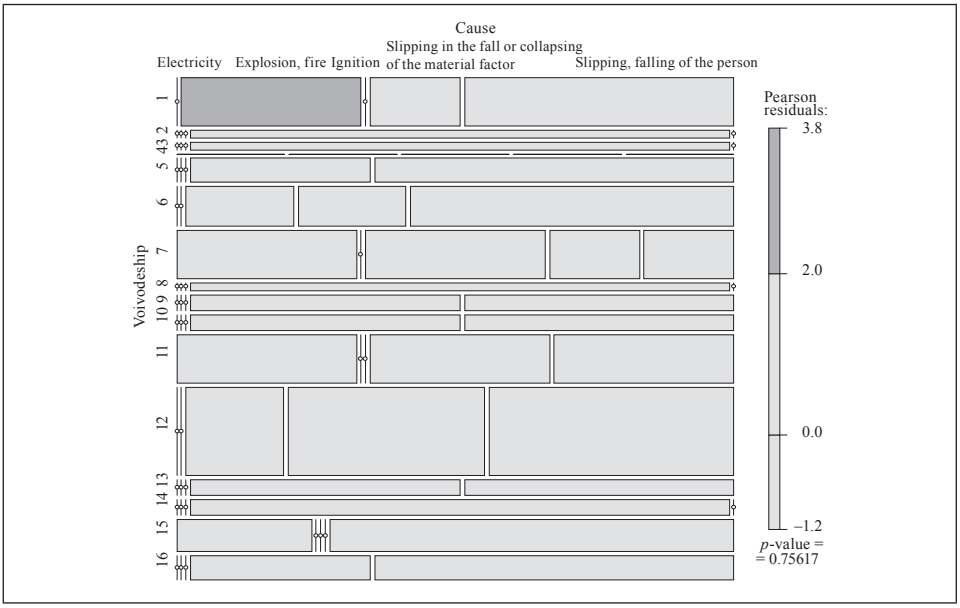

Fig. 1. Mosaic Plot for the Independence Model for No Adjustment of the Zero Cell
Source: the author's own calculations in **R**.

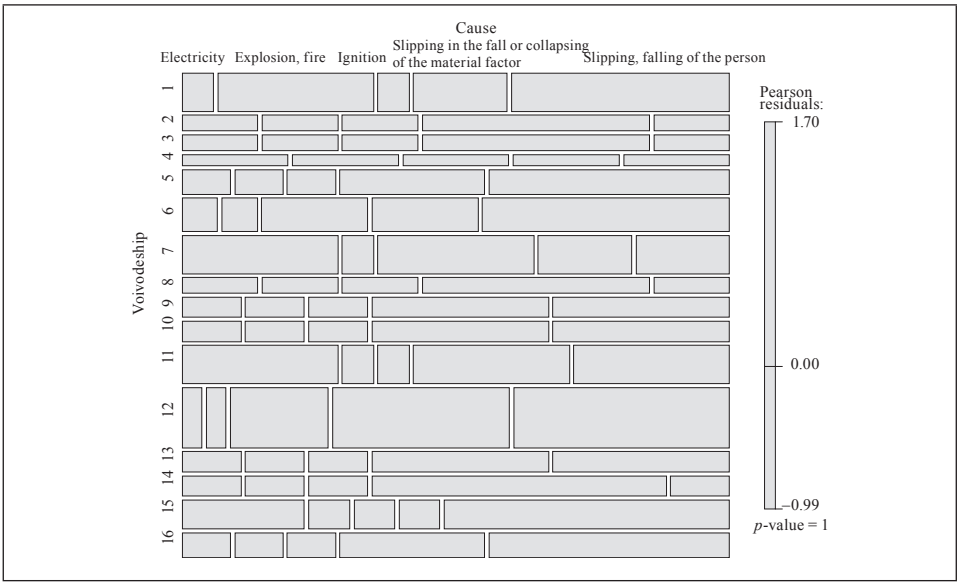

Fig. 2. Mosaic Plot for the Independence Model for Adjustment $n \rightarrow n + 0.5$
Source: the author's own calculations in **R**.

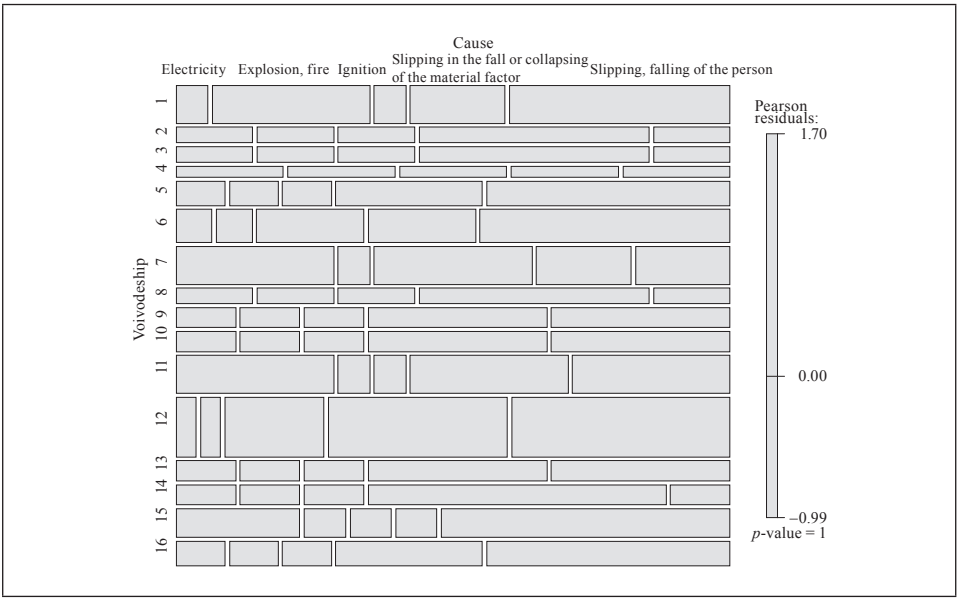

Fig. 3. Mosaic Plot for the Independence Model for Adjustment $n = 0 \rightarrow n + 0.2$
Source: the author's own calculations in **R**.

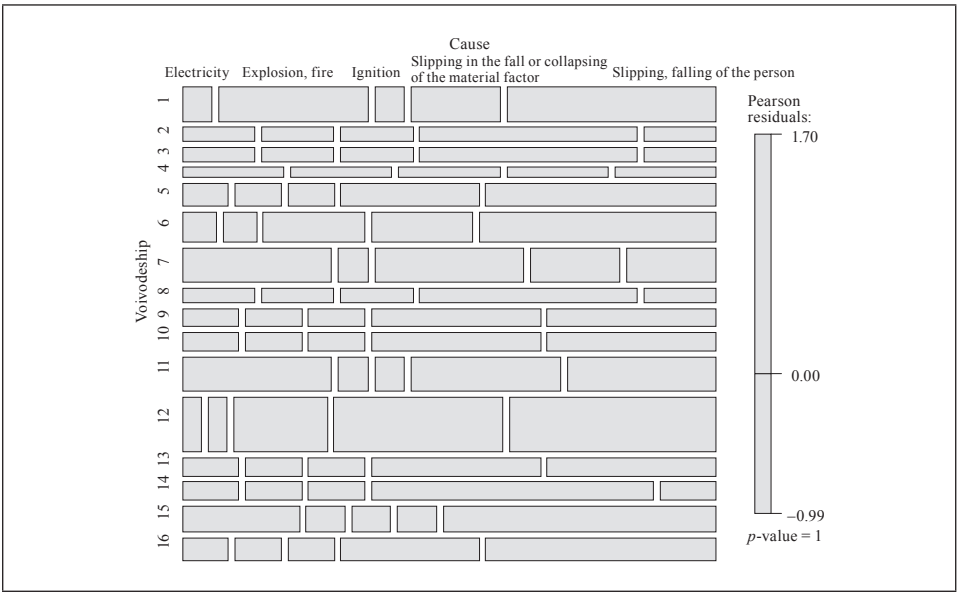

Fig. 4. Mosaic Plot for the Independence Model for Adjustment $n = 0 \rightarrow n = 0.1 \times 10^{-8}$
Source: the author's own calculations in **R**.

6. Conclusions

Log-linear models are a standard tool used to analyse structures of dependency in multi-way contingency tables. The criteria to be analysed are the expected cell frequencies in the table as a function of all the variables in the survey. The analysis of such a table may be troublesome when some cells are zeros. For log-linear models, most of the derivations of expected frequencies and other quantities assume $n_{hj} > 0$; however, in research we may have tables containing zeros. Zero frequencies may occur in a contingency table for two reasons: sampling and structural zeros. To avoid this problem, steps must be taken.

A mosaic plot graph represents a contingency table, with each cell corresponding to a piece of the plot, the size of which is proportional to the cell entry. Extended mosaic displays show the standardised residuals of a log-linear model of the counts by the colour and outline of the mosaic's tiles (standardised residuals are often referred to as standard normal distribution). Negative residuals are drawn in shades of red and with broken outlines; positive ones are rendered in blue with solid outlines. Thus, mosaic plots are perfect to visualise associations within a table and to detect cells which create dependencies. The association plot puts deviations from independence in the foreground: the area of each box is made proportional to the observed – expected frequency. Both log-linear analysis and visualising categorical data tools are a useful and practical tool that help to analyse the independence between the categorical data in a contingency table.

In this paper some solutions for zero-cell frequencies are presented. For two- and four-way tables, different solutions are compared with the use of likelihood statistics and information criteria (*AIC*, *BIC*). Visualising tools with the use of mosaic plots for contingency tables with zero cells have also been presented. This analysis may be helpful in looking for the best solution in the analysis of high dimensional tables.

Bibliography

Akaike H. (1973), *Information Theory and an Extension of the Maximum Likelihood Principle*, Proceedings of the 2nd International Symposium on Information, Budapest.

Andersen E. B. (1997), *Introduction to the Statistical Analysis of Categorical Data*, New York, Springer.

Brzezińska J. (2013), *Metody wizualizacji danych jakościowych w programie R*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 279, Klasyfikacja i analiza danych – teoria i zastosowania, 21: 182–190.

Brzezińska J. (2014), *Visual Models for Categorical Data in Economic Research*, M. Spiliopoulou, L. Schmids-Thieme, R. Janning (eds), Studies in Classification,

Data Analysis, and Knowledge Organization: Data Analysis, Machine Learning and Knowledge Discovery, Springer: 33–40.
Brzezińska J. (2015), *Analiza logarytmiczno-liniowa. Teoria i zastosowania z wykorzystaniem programu R*, C.H. Beck, Warszawa.
Clausen S. E. (1998), *Applied Correspondence Analysis. An Introduction*, Sage Publications, Thousand Oaks.
Clogg C. C., Eliason S. R. (1988), *Some Common Problems in Log-linear Analysis* (in:) J. S. Long (ed.), *Common Problems/Propers Solutions*, Sage, Newbury Park, CA.
Cohen A. (1980), *On the Graphical Display of the Significant Components in a Two-way Contingency Table*, "Communications in Statistics – Theory and Methods", 9(10): 1025–1041, http://dx.doi.org/10.1080/03610928008827940.
Everitt B. (1977), *The Analysis of Contingency Tables*, Chapman and Hall, London.
Evers M., Namboodiri N. K. (1977), *A Monte Carlo Assessment of the Stability of Log-linear Estimates in Small Samples*, Proceedings of the American Statistical Association, Social Statistics Section, American Statistical Association, Washington, DC.
Fienberg S. E. (1980), *The Analysis of Cross-classified Categorical Data*, MIT Press, Cambridge.
Friendly M. (1991), *The SAS System for Statistical Graphics*, SAS Institute Inc., Carry, NC.
Friendly M. (2000), *Visualizing Categorical Data*, SAS Institute Inc., Carry, NC.
Goodman L. A. (1970), *The Multivariate Analysis of Qualitative Data: Interaction among Multiple Classifications*, "Journal of the American Statistical Association", 65(329): 226–256, http://dx.doi.org/10.1080/01621459.1970.10481076.
Greenacre M. J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.
Grizzle J. E., Starmer C. F., Koch G. C. (1969), *Analysis of Categorical Data by Linear Models*, "Biometrics", 25(3): 489–504, http://dx.doi.org/10.2307/2528901.
Ishii-Kunts M. (1994), *Ordinal Log-linear Models*, Sage University Paper Series on Quantitative Applications in the Social Science, series no. 07-097, Sage, Beverly Hills–London.
Knoke D., Burke P. J. (1980), *Log-linear Models*, Sage University Paper Series on Quantitative Applications in the Social Science, series no. 07-020, Sage, Beverly Hills–London.
Raftery A. E. (1986), *Choosing Models for Bross-classification*, "American Sociological Review", 51: 145–146.
Smirnoff J. S. (2003), *Analyzing Categorical Data*, Springer Texts in Statistics, Springer, New York.
Snee R. D. (1974), *Graphical Display of Two-way Contingency Tables*, "The American Statistician", 28(1): 9–12, http://dx.doi.org/10.2307/2683520.

Problem zerowych liczebności w analizie tablic kontyngencji

(Streszczenie)

Analiza logarytmiczno-liniowa jest metodą badania zależności pomiędzy zmiennymi niemetrycznymi w tablicy kontyngencji, która pozwala analizować dowolną liczbę zmiennych nominalnych i porządkowych. Pomimo że jest ona uniwersalną metodą

analizy zmiennych niemetrycznych, występują jednak pewne ograniczenia w jej stosowaniu ze względu na zerowe liczebności. Zera występujące w tablicy mogą być dwojakiego rodzaju: strukturalne lub związane ze schematem losowania. Zera strukturalne pojawiają się wtedy, gdy nie jest możliwa obserwacja kategorii zmiennej, a zera związane ze schematem losowania występują w małych próbach i znikają, gdy próba zostanie odpowiednio zwiększona.

W artykule zaprezentowano sposoby radzenia sobie z zerowymi liczebnościami w tablicy kontyngencji. Wszystkie obliczenia przeprowadzono w programie **R**.

Słowa kluczowe: zerowe liczebności, analiza danych jakościowych, tablice kontyngencji, analiza logarytmiczno-liniowa.