ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Iwona Kasprzyk**

GRAPHICAL PRESENTATION OF A MULTI-WAY CONTINGENCY TABLE IN THE R SOFTWARE

Abstract. The contingency table is a popular way of presenting categorical data. This paper presents the various types of the log-linear models, which describe the relationship between variables in the contingency table.

We can make a visualisation of data contained in the multi-way contingency table using the *vcd* and *graphics* packages in the R software. The main aim of this paper is to show the mosaic plots, which are the most popular ways of visualization of this kind of models. The mosaic display was proposed by Hartigan and Kleiner (1981).

This paper is a continuation of the paper titled "Visualizing of a two-way contingency table in the R software" delivered at the conference on Multivariate Statistical Analysis in 2006.

Key words: contingency table, log-linear models, association plot, mosaic display, sieve diagram.

I. INTRODUCTION

This paper provides various types of plots of visualization in a contingency table, especially in the multi-way table.

As in example, we present the analysis of a job satisfaction. The analysis are based on data of Polish General Social Survey (1992–2005). In this paper we take this data into consideration since 2005.

The job satisfaction analysis is shown on the strength of variables: age, sex and job satisfaction. We use the following variables:

* Ph.D. student, Department of Statistics, The Karol Adamiecki University of Economics, Katowice.

Table 1. The list of variable used for analysis

Name of variable	Categories of variables
Job satisfaction	a) very satisfied
	b) rather satisfied
	c) rather not satisfied
	d) very dissatisfied
Age	a) 18–24
	b) 25–29
	c) 30–39
	d) 50–59
	e) ≥ 60
Sex	a) woman
	b) man

Source: Own research.

II. LOG-LINEAR MODELS

Suppose we have a three – dimensional contingency table. Let e_{ijk} denote the theoretical cell frequencies. The saturated model contains all main effects and interaction effect for variables X, Y, Z is written:

$$\log(e_{ijk}) = u + u_i^X + u_j^Y + u_k^Z + u_{ij}^{XY} + u_{ik}^{XZ} + u_{jk}^{YZ} + u_{ijk}^{XYZ}, \tag{1}$$

where:

$e_{ijk} = n_{ijk}$,

u_i^X – the main effect for variable X,

u_{ij}^{XY} – the interaction effect for variables X and Y,

$$u = \frac{1}{rcl}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\log(e_{ijk}),$$

$$u_i^X = \frac{1}{cl}\sum_{j=1}^{c}\sum_{k=1}^{l}\log(e_{ijk}) - u, \quad u_j^Y = \frac{1}{rl}\sum_{i=1}^{r}\sum_{k=1}^{l}\log(e_{ijk}) - u$$

$$u_k^Z = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\log(e_{ijk}) - u,$$

$$u_{ij}^{XY} = \frac{1}{l}\sum_{k=1}^{l}\log(e_{ijk}) - u_i^X - u_j^Y - u, \quad u_{ik}^{XZ} = \frac{1}{c}\sum_{j=1}^{c}\log(e_{ijk}) - u_i^X - u_k^Z - u,$$

$$u_{jk}^{YZ} = \frac{1}{r}\sum_{i=1}^{r}\log(e_{ijk}) - u_j^Y - u_k^Z - u,$$

$$u_{ijk}^{XYZ} = \log(e_{ijk}) - u_i^X - u_j^Y - u_k^Z - u_{ij}^{XY} - u_{ik}^{XZ} - u_{jk}^{YZ} - u. \qquad (2)$$

Model (1) fulfils the following conditions:

$$\sum_{i=1}^{r} u_i^X = \sum_{j=1}^{c} u_j^Y = \sum_{k=1}^{l} u_k^Z,$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c} u_{ij}^{XY} = \sum_{i=1}^{r}\sum_{k=1}^{l} u_{ik}^{XZ} = \sum_{j=1}^{c}\sum_{k=1}^{l} u_{jk}^{YZ} = 0,$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} u_{ijk}^{XYZ}.$$

The independent model we can follow as:

$$\log(e_{ijk}) = u + u_i^X + u_j^Y + u_k^Z, \qquad (3)$$

Model (3) is called the mutually independent model.

For instance, if variables X and Y are mutually independent of Z, that this model we can show in following form:

$$\log(e_{ijk}) = u + u_i^X + u_j^Y + u_k^Z + u_{ij}^{XY}, \qquad (4)$$

Model (4) is called the partial independent model.

Another example of the independent model is a conditional independence model, where we can find, for instance, the interaction effect for variables X and Z and the interaction effect for variables Y and Z. As an example of such a model is:

$$\log(e_{ijk}) = u + u_i^X + u_j^Y + u_k^Z + u_{ik}^{XZ} + u_{jk}^{YZ}, \tag{5}$$

In **R** software, function `loglin` in MASS package realizes the log-linear models.

III. THE ASSOCIATION PLOT

The association plot has been proposed by Cohen (1980). The height of each rectangle is proportional to the Pearson residual e.t.:

$$r_{ijk} = \frac{n_{ijk} - e_{ijk}}{\sqrt{e_{ijk}}}, \tag{6}$$

where:

$$e_{ijk} = \frac{n_{i..} n_{.j.} n_{..k}}{n}.$$

The width of each rectangle is proportional to $\sqrt{e_{ijk}}$, and the area of the rectangle is proportional to $n_{ijk} - e_{ijk}$. If the difference is positive, the rectangle is filled with black colour, if negative – the colour is red.

Figure 1 presents the job satisfaction. In the R software, the commands can be saved as follow:

```
> library(vcd)
> dat<-read.table("data-satisfaction.R", header=TRUE)
> tab <-xtabs (~age + satisfaction + sex, data=dat)
> assoc(aperm(tab), expected = ~ (zadowolenie + wiek) *
płeć, labeling_args = list(just_labels = c(age = "left"),
offset_labels = c(right = -0.5), offset_varnames = c(right
= 1.2), rot_labels = c(right = 0), tl_varnames = c(age =
TRUE)))
```

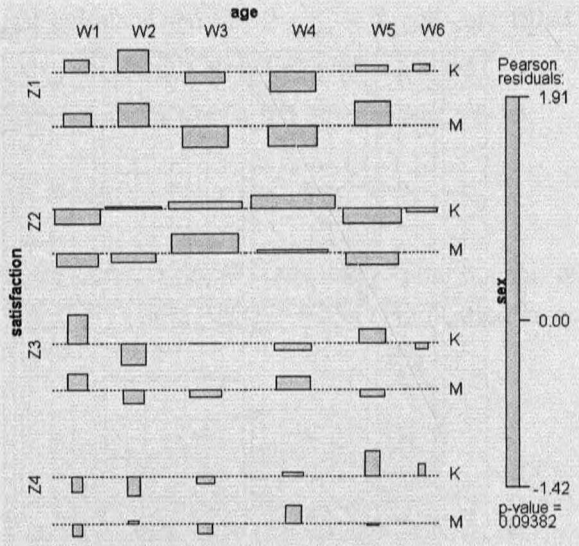

Figure 1. The association plot for job satisfaction, age and sex
Source: Own research.

III. THE SIEVE DIAGRAM

The sieve diagram has been proposed by Riedwyl and Schüpbach (1983) and in 1994 it was called a parquet diagram. This kind of plot divides a square unit into rectangles. The area of each rectangle is proportional to the expected frequency (e_{ijk}).

If the difference between the observed and expected frequency is positive, the rectangle is filled with blue colour, but if it is negative, the rectangle is red. Using these colours in one time can indicate whether the deviation from independence is positive or negative. The inside of each of the rectangles are drawn in squares, which reflect to the observed frequency contained in the contingency table.

By using the following commands in the R software, one receives the sieve diagram for two variables: the age, sex and the job satisfaction are shown in Figure 2.

```
> library(vcd)
> dat<-read.table("data-satisfaction.R", header=TRUE)
> tab <-xtabs (~age + satisfaction + sex, data=dat)
> sieve(tab, pop = FALSE, shade = TRUE)
> labeling_cells(text = tab, gp = gpar(fontface = 2))(tab))
```

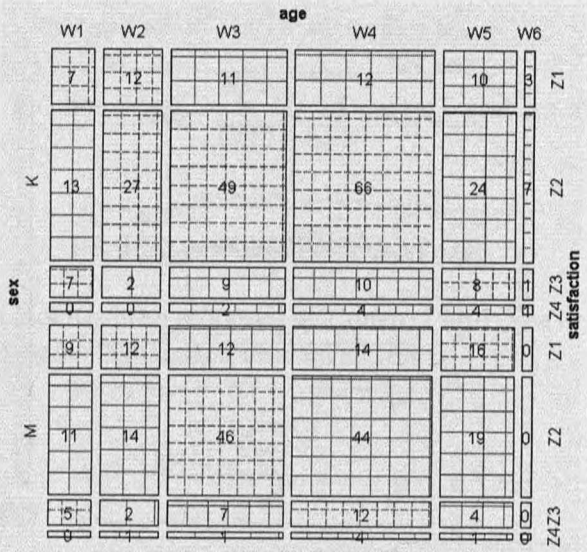

Figure 2. The sieve diagram for age, sex and job satisfaction
Source: Own research.

IV. THE MOSAIC DISPLAY

The mosaic display was proposed by Hartigan and Kleiner (1981) and later considered by Friendly (1994). This plot is a graphical method for visualizing n-way contingency table.

For the three-way table, the width of each rectangle is proportional to the marginal probabilities ($p_{ij.} = n_{ij.}/n$) and the height of the rectangle is proportional to the conditional probabilities for the columns given in rows i ($p_{k/ij} = n_{ijk}/n_{ij.}$).

The area of the rectangle depends on a kind of the log-linear model, for instance, for the conditional independent model this area is proportional to the observed frequency and the given probabilities are:

$$p_{ijk} = p_{ij.} \cdot p_{k/ij} = \frac{n_{ij.}}{n} \cdot \frac{n_{ijk}}{n_{ij.}} \tag{7}$$

In the mosaic display colour is of great significance. The $|r_{ijk}| \leq 2$ cells are filled with grey colour and the $r_{ijk} \leq -4$ cells are filled with navy red, the $r_{ijk} \geq 4$ cells are filled with navy blue. It is very specific for this kind of plot. First of all,

we use blue and red colour. Then the $2 < r_{ijk} < 4$ cells are filled with light red colour and the $-2 < r_{ijk} < -4$ cells are filled with navy red.

The best log-linear model is the partial independent model:

$$\log(e_{ijk}) = u + u_i^W + u_j^Z + u_k^P + u_{ij}^{WZ},$$

where Z denote the job satisfaction, W – age and P- sex. For this model, the likelihood ratio L^2 is 20,777 on 23 df ($p = 0.5947$) indicating an acceptable overall fit.

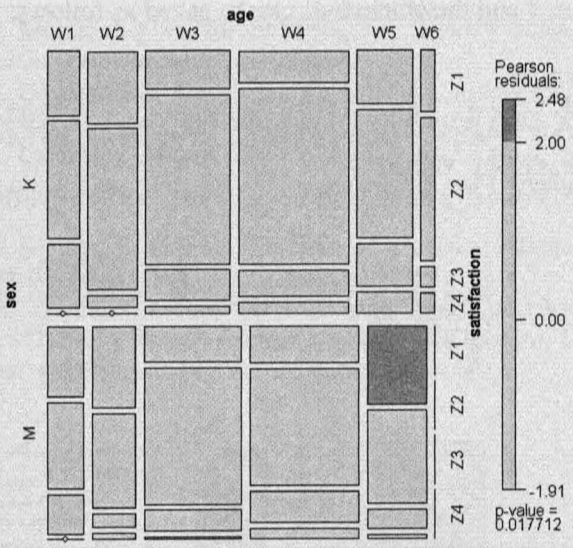

Figure 3. The mosaic display for age, sex and job satisfaction.
Source: Own research.

The mosaic display in Figure 3, can be obtained using the following commands in the R software:

```
> library(vcd)
> dat<-read.table("data-satisfaction.R", header=TRUE)
> tab <-xtabs (~age + satisfaction + sex, data=dat)
> mosaic(~ age + satisfaction | sex, data = dat, shade =
TRUE)
```

Analyzing the example of the job satisfaction, we can observe that the women, aged between 40 to 49 is rather satisfied with their work and men between the age of 30 to 39 is also rather satisfied with their work.

V. ANOTHER PLOTS IN R SOFTWARE

In R software for visualizing conditional independence models we can use the `pairs` and the `cotabplot` function. The `pairs` function creates plots for all pair wise variable contented in the contingency table and bar plots in the diagonal to visualize the absolute frequencies of the variables. The pairs plot is shown in Figure 4 and the commands can be saved as follows:

```
> library(vcd)
> dat<-read.table("data-satisfaction.R", header=TRUE)
> tab <-xtabs (~age + satisfaction + sex, data=dat)
> pairs(tab,upper_panel =
pairs_assoc,lower_panel=pairs_sieve, shade=TRUE)
```

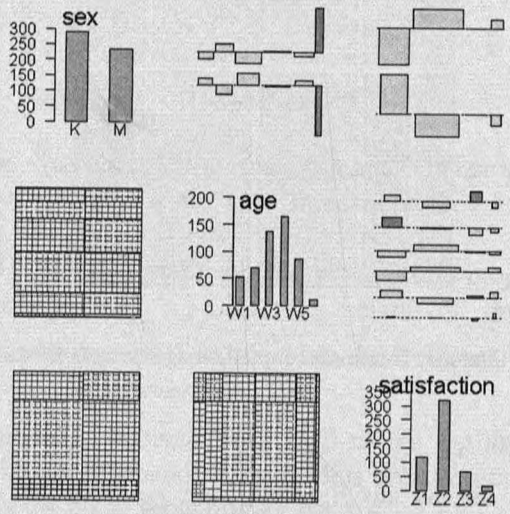

Figure 4. The pairs plot for age, sex and job satisfaction.
Source: Own research.

In the figure 4 we can see, that more women take part in this research than men.

The majority respondents were in the age of 40 to 49 and they were rather satisfied with their work.

VI. CONCLUSION

All types of plots that have been shown in this article present the degree of which the variables in the three-way contingency table are independent or not. The main aim of this method is not only to research association between the two or more variables, but also to show the relationship between categories of variables. In this paper was showed another plots based on the mosaic display, the association plot and sieve diagram.

REFERENCES

Friendly M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association, 89*, p.190–200.

Friendly M. (1998), *Conceptional Models for Visualizing Contingency Table Data*, in: Blasius J., Greenacre M. (eds.), *Visualization of Categorical Data*, Academic Press.

Friendly M.(1999), Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data, *Journal of Computational and Graphical Statistics*, 8, p. 373—395.

Hartigan, J. A., and Kleiner, B. (1984). A mosaic of television ratings. *The American Statistician, 38*, p.32–35.

Mayer D., Zeileis A., Hornik K. (2006). The Structplot Framework: Visualizing Multi-way Contingency Tables with *vcd. Journal of Statistical Software*, 10, vol. 17, Issue 3, p. 1–48, http://www.jstatsoft.org/v17/i03/paper.

Iwona Kasprzyk

GRAFICZNA PREZENTACJA WIELOWYMIAROWYCH TABLIC KONTYNGENCJI W PAKIECIE STATYSTYCZNYM R

Tablica kontyngencji jest częstym sposobem przedstawiania danych mierzonych zarówno na skali nominalnej jak i porządkowej. W referacie zostaną przedstawione różne typy modeli log-liniowych, które pozwalają na badanie zależności między zmiennymi zawartymi w tablicy kontyngencji.

Za pomocą pakietu *vcd* oraz *graphics* w programie *R* zostanie dokonana wizualizacja danych zawartych w wielowymiarowej tablicy kontyngencji Zostaną przedstawione przede wszystkim wykresy mozaikowe, które to są najczęstszym sposobem wizualizacji modeli log-liniowych. Tego typu wykresy mozaikowe zostały zaproponowane przez Hartigan i Kleiner [1981].

Referat jest kontynuacją referatu „Wizualizacja dwuwymiarowych tablic kontyngencji w pakiecie statystycznym R" wygłoszonego na XXV Konferencji MSA 2006.