Liet. matem. rink., T. 43, spec. nr., 486–491
© 2003 Matematikos ir informatikos institutas

Housing and health in Vilnius 2002

Silvija JUNEVIČIŪTĖ, Danutė KRAPAVICKAITĖ (VGTU)
e-mail: s.juneviciute@vb.lt, krapav@ktl.mii.lt

1. Introduction

With a fast development of new technologies, great attention has also to be paid to the health of society. The society stimulates the development of technologies, but the aim of this development is a comfortable life of society. It cannot be achieved without a healthy society.

The World Health Organization carried out the second international survey on housing and health in Europe in 2002. 8 countries took part in this survey: France, Germany, Hungary, Italy, Lithuania, Portugal, Slovakia, and Switzerland. The aim of the survey was to investigate the relationship between housing, socioeconomic status and health of the residents.

Samples of households were surveyed in some town of each country. Vilnius was chosen as a representative of Lithuania.

The paper presents statistical methods as a tool for the analysis of survey data which can be used for any other social survey for investigating categorical data, and presents some results.

2. Statistical methods

2.1. *Sampling design*

The population register served as a sampling frame for sample selection of the Vilnius population. A simple random sample of 1 100 individuals was drawn from the register and people living at the same address with the selected individuals were included into the sample. The cluster sample of individuals was obtained. A probability for the household (as well as for an individual) to belong to the sample depends on the size of the household, and can be expressed as

$$\pi_i = P(i \in s) = \frac{nm_i}{L}, \quad i = 1, \ldots, n,$$

with L – a size of the sampling frame of individuals, s – the sample of the households of size n, m_i, $i = 1, \ldots, n$ – the household size.

The sample design weight for the household is defined by $d_i = 1/\pi_i$, $i \in s$. For any study variable y with the values $y_1, y_2, \ldots, y_N$ in the population of households, the population total $t_y = y_1 + y_2 + \cdots + y_N$ can be estimated by the Horvitz–Thompson estimator

$$\widehat{t}_y = \sum_{i \in s} d_i y_i.$$

Denoting the study variable in the population of individuals by z with the value z_{ij} for the jth person in the ith household, the same is valid for the total

$$t_z = \sum_{i=1}^{N} \sum_{j=1}^{m_i} z_{ij} \text{ of individuals: } \widehat{t}_z = \sum_{i \in s} \sum_{j=1}^{m_i} d_{ij} z_{ij}, \quad d_{ij} = d_i.$$

This is a standard weighting system.

2.2. *Calibration of design weights*

The sampling frame suffers from the undercoverage and overcoverage, the survey has a high nonresponse rate, about 40%. As a result the distribution of the sample, weighted by the design weights, differs greatly from the population distribution by age and sex (see Table 1). Health characteristics depend on age, a psycho-emotional feeling depends on sex, and in order to get correct results, the sample has to be reweighted. In order to adjust the sample to the demographical data and to take nonresponse into account, the calibration of sampling weights on age and sex is used ([2]).

Let us define J age and sex groups and construct a vector of auxiliary information $\mathbf{x} = (x_1, \ldots, x_J)^T$ with the values $\mathbf{x}_k = (x_{1k}, \ldots, x_{Jk})^T$, $k = 1, \ldots, n$. The components x_{jk} for each household $k = 1, \ldots, n$ equal the number of individuals in the household belonging to the corresponding age and sex group $j = 1, \ldots, J$. Suppose that the vector of totals $t_{\mathbf{x}} = (t_{x1}, \ldots, t_{xJ})^T$, $t_{xj} = \sum_{k=1}^{n} x_{jk}, \; j = 1, \ldots, J$ is known from the demographical data.

Table 1

Distribution of design-based estimates and demographical data by age and sex

Age	Men		Women	
	Demography	Estimate	Demography	Estimate
< 15	49 034	22 377	43 012	23 581
16–25	38 748	43 533	45 178	48 339
26–35	41 829	32 068	44 198	34 833
36–45	40 632	25 383	47 337	35 174
46–60	43 072	49 911	57 306	73 036
> 60	33 448	56 706	57 991	97 758

In order to get a calibrated estimator of the total $t_y = y_1 + \cdots + y_N$, we have to find the weights w_k, which differ as little as possible from the design weights d_k with arbitrary positive q_k in the sense

$$\sum_{k\in s} \frac{d_k}{q_k}\Big(\frac{w_k}{d_k} - 1\Big)^2 = \sum_{k\in s} \frac{(w_k - d_k)^2}{d_k q_k} \to \min,$$

and which satisfy the calibration equation

$$\sum_{k\in s} w_k \mathbf{x}_k^T = \Big(\sum_{k\in s} w_k x_{1k}, \ldots, \sum_{k\in s} w_k x_{Jk}\Big)^T = (t_{x1}, \ldots, t_{xJ})^T.$$

The weights w_k are said to be calibrated and the estimator

$$\widehat{t}_y^{(cal)} = \sum_{k\in s} w_k y_k$$

is called calibrated. The same weights can be used for the members of household in the sample of individuals. The design weights in the Vilnius housing and health survey in 2002 are calibrated by 12 age and sex groups, used in Table 1.

The calibration of weights is performed by the computer program SAS macro CLAN of Statistics Sweden.

2.3. *Logistic regression*

When investigating how some health characteristics of an individual depend on the housing characteristics, logistic regression is used.

Let us denote for each health characteristic the value of a study variable y for the individual i as

$$y_i = \begin{cases} 1, & \text{if the individual has a health characteristic,} \\ 0, & \text{if the individual does not have a health characteristic,} \end{cases}$$

$\mathbf{x} = (1, x_1, \ldots, x_p)^T$ is the vector of categories of explanatory variables (characterizing housing, different from that in Section 2.2), $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of unknown parameters. The health characteristic, for example, may be suffering from asthma.

The logistic regression model takes the form

$$P(y = 1|\mathbf{x}) = \frac{e^{\beta t}}{1 + e^{\beta t}}$$

and expresses the probability for the individual to have the health characteristic $y = 1$, if the value of the explanatory vector $\mathbf{x}$ is known. The finite population parameter $\mathbf{B} = (B_0, B_1, \ldots, B_p)$ can be found as a maximum likelihood estimate of the model parameter β, using population data, and a design based estimate $\widehat{\mathbf{B}} = (\widehat{B}_0, \widehat{B}_1, \ldots, \widehat{B}_p)$ of $\mathbf{B}$ can be obtained using the sample data.

The members of household form a cluster, and, sharing the same apartment, they are dependent via the housing characteristics. So the clustering has to be taken into account in the data analysis.

The hypotheses H_0: $\beta_i = 0$ against the alternative H_1: $\beta_i \neq 0$ for $i = 0, 1, \ldots, p$ are tested using the modified t-test, taking the sampling design into account ([3]).

2.4. *Correspondence analysis*

Correspondence analysis ([1], [4]) is an algebraic technique analogous to principal components analysis but appropriate to categorical rather to continuous variables.

Let several categorical variables have K categories totally. A Burt contingency table $\mathbf{A}$, based on the values of those variables, is built. It is a square $K \times K$ table, having rows and columns corresponding to each of the categories of the variables chosen. An element of this table equals to the weighted number of the sample elements with the categories in the corresponding row and column.

Suppose the matrix $\mathbf{A}$ is of the rank K and denote $\mathbf{Y} = \frac{1}{N}\mathbf{A}$, here N is the population size. Multiple correspondence analysis is based on the decomposition $\mathbf{Y}^T\mathbf{Y} = \mathbf{B}\mathbf{D}\mathbf{B}^T$, matrix $\mathbf{D}$ being a diagonal matrix of the eigenvalues of $\mathbf{Y}^T\mathbf{Y}$, and matrix $\mathbf{B}$ being a square matrix consisted of corresponding eigenvectors. Based on this decomposition, the system of the orthogonal factors is build, and associations in the Burt table are expressed through those factors.

3. Some results

Three extensive questionnaires are used in the survey: housing and health questionnaire, housing inspection questionnaire, and inhabitant questionnaire.

Among other things, the relationship between the health and housing characteristics is studied. Asthma, chronic bronchitis, chronic anxiety and depression, headache, nasal allergy are investigated, using logistic regression. The models developed are weak, they explain only about 10% of the total variance. Significant factors with the level of significance 0.05 are presented in Table 2.

Suffering from bronchitis at least once in relation with the flat features was studied using correspondence analysis. The notations of the categories of the variables studied are presented in the Table 3.

Multiple correspondence analysis locates all the categories in a Euclidean space. The first two dimensions of this space are plotted in Fig. 1 to examine the association among the categories. The interpretation is based on points found in approximately the same direction from the origin and in approximately the same region of the space. Distances between points do not have a straightforward interpretation in multiple correspondence analysis.

The quadrants of the plot show the following associations: suffering from bronchitis, being problems of temperature in the flat during summer, transition period and winter, visual mould growing and nevertheless satisfaction with the flat are associated categories;

Table 2

Relationship between health and housing characteristics

Health indicator	**Housing and psycho-emotional features**
Chronic bronchitis, emphysema	Age, satisfaction with the flat, perceiving the temperature in the flat during winter and transient seasons as a problem, mould growth in the flat, smoking in the flat
Asthma	Having problems with draught in the flat because doors/windows cannot be closed tightly, cat at home, satisfaction with the flat
Headache	Age, sleep disturbed by noise, satisfaction with the flat, perceiving the temperature in the flat during summer as a problem, problems with draught in the flat because doors/windows cannot be closed tightly, satisfaction with the flat and with the air quality in the flat, evaluation of this residential area by others
Nasal allergy	Thermo-isolation in the flat, problems of hot water-supply
Accidents at home	Considering some flat construction features, like balcony and elevator, dangerous for the residents at home
Chronic anxiety and depression	Age, sleep disturbed by noise at night, problems of temperature in the flat in winter, thermo-isolation of flat, humidity problems in the flat, satisfaction with the equipment and installation in the bathroom, satisfaction with the flat

Table 3

Categories being studied with suffering from bronchitis

Variable	No	Sometimes	Yes
Temperature problems in the flat in summer	$S1$	$S3$	$S5$
Temperature problems in the flat in transition period	$T1$	$T3$	$T5$
Temperature problems in the flat in winter	$W1$	$W3$	$W5$
Thermo-isolation	$I1$	$I3$	$I5$
Visual mould growth in the flat	$M1$	$M3$	$M5$
Satisfaction with the flat	$F1$	$F3$	$F5$
Age group	A1 (Young)	A3 (Medium)	A5 (Old)
Smoking	N (No)	Y (Yes)	
Suffering from bronchitis	0 (No)	1 (Yes)	

bad thermo-isolation and having no problems of temperature in summer, winter and transition period and nevertheless un-satisfaction with the flat are also associated categories; very good thermo-isolation is associated with sometimes occurring problems of temperature in winter, transition period and summer and visual mould growth.

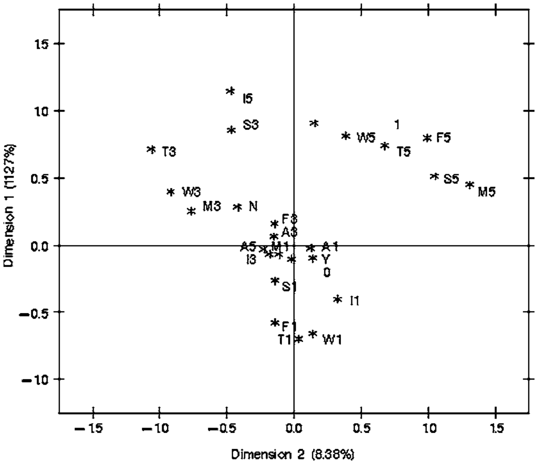

Fig. 1. Chronic bronchitis and flat features.

4. Conclusions

1. Despite that the models obtained are not very strong, the results of analysis show a significant relationship between the health of residents and the temperature problems, humidity in the flat, tightness of doors/windows. The results also support the relationship between the psycho-emotional feeling of housing and health.
2. The correspondence analysis can give a picturesque graphical representation of the survey data illustrating the relationship between the categorical variables.
3. The calibration of the weights is recommended for the social surveys having a complex sampling design and a high nonresponse level, as well as for public opinion surveys.

Acknowledgement. The authors are thankful to the State public health center for supplying with the survey data.

References

[1] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York (1990).

[2] J. Deville, C.-E. Särndal, and O. Sautory, Generalized raking procedures in survey sampling, *JASA*, **88**, 1013–1020 (1993).

[3] S.L. Lohr, *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove (1999).

[4] М. Жамбю, *Иерархический кластер-анализ и соответствия*, Финансы и статистика, Москва (1988).

Būstas ir sveikata Vilniuje 2002-aisiais

S. Junevičiūtė, D. Krapavickaitė

Straipsnyje pateikiami statistiniai metodai, naudojami Pasaulinės Sveikatos Organizacijos 2002-aisiais metais Vilniuje atlikto tyrimo duomenų analizei, ir kai kurie šios analizės rezultatai.