Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 7 (4): 585-588, 2007
ISSN 1812-5654

Multiple Correspondence Analysis Technique Used in Analyzing the Categorical Data in Social Sciences

[1]Duygu Aktürk, [2]Sema Gün and [1]Taner Kumuk
[1]Department of Agricultural Economy, Faculty of Agriculture,
Çanakkale Onsekiz Mart University, Çanakkale, Turkey
[2]Department of Agricultural Economy, Faculty of Agriculture, Ankara University, Ankara, Turkey

Abstract: It is observed that Chi-square, Fisher's Exact Probability Test, G-statistics and Z-test are frequently used in social sciences to interpret the data statistically. However, exploitation of these tests depends on some conditions. Even though these conditions are met there are still problems in interpretation of the results because the obtained data are general and limited. In this study, practical limitations of the above-mentioned tests are discussed. Then, an in-depth analysis follows on advantages and exploitation of Multiple Correspondence Analysis, which is suggested as the alternative technique that resolves the limitations of other techniques mentioned above.

Key words: Correspondence analysis, social sciences, categorical data, variable

INTRODUCTION

Research data can be reached in various modes depending on the main purpose of the research. For instance, while in certain cases it is necessary to measure, elaborate or analyze a feature, in other circumstances it is necessary to classify the data into categorical groups (or gather data directly in categories). On other occasions data is gathered in ordinal mode. Statistical interpretation of these data requires different statistics methods (Aktürk, 2004; Başpinar and Mendeş, 2000; Sokal and Rohlf, 1995).

In social sciences, categorical or categorized data is used frequently. Techniques widely used to analyze categorical data are Chi-Square analysis, Fischer's exact test, G-Statistics and Ratio Test (Z-test). However, the use of these statistical techniques depend on several assumption and sometimes these assumption can not be satisfied to use these techniques or even if the assumption are met results of the analysis are too general for interpretation (Aktürk, 2004; Başpinar and Mendeş, 2000).

In social sciences, after being coded, data are grouped in a two way cross table in order to do the χ^2 analysis. In order to have reliable results from χ^2 analysis and to have minimum missing data, expected frequencies in each cell of cross table should be at least 5. Besides, even in cases where frequencies in all cells of the cross table are 5 and more, results will indicate broad information such as whether or not the row and column variables are independent from each other. Whereas, researchers are interested also in issues other than whether or not row and column variables are independent from each other. For instance, they can be interested in both the relations between a variable's own sub-categories and the mutual relations between different levels of variables. In that case, χ^2 analyses can be insufficient for the researcher to reach his/her aim. If expected frequency in the cells of cross tables is less than 5, analysis of the table can be done by Fischer's exact test by transforming the cross table into various 2×2 tables. However, in case one or several of expected frequencies in the cells are zero, employing this test will not be valuable for the researcher. In such a case we can apply G-statistics to our table. Yet, in case there are zero frequency cells, reliability of results will diminish because degree of freedom will be reduced. There can be even negative degree of freedom in certain cases. Such circumstances restrict the employability of G-statistics. Ratio test, which is another technique that can be used in the analysis of coded (categorical) data, can be computed for situations with probabilities between (0,1) open gap. Ratio test can not be used when probability is exactly 1 or 0. Also in cases where there is zero frequency, results of ratio test can be misleading (Başpinar and Mendeş, 2000; Düzgüneş *et al*., 1983).

One of the analysis techniques developed for that aim is correspondence analysis (Greenacre, 1998). The object of correspondence analysis technique is to analyze

Corresponding Author: Sema Gün, Department of Agricultural Economy, Faculty of Agriculture, Ankara University, Ankara, Turkey Tel : +90-312-5961612 Fax : +90-312-3185360

categorical or categorized data which are transformed into cross tables in the form of r×c or r×c×k (r≥2, c≥2 and k≥2) and to demonstrate the results in graphs. With this technique, in comparison to other techniques, more detailed results can be obtained (both relations between row and column variables and relations between different levels of each variable) in a smaller dimensional space (Aktürk, 2004; Mendeş, 2002; Chou, 1994; Gifi, 1990).

The objective of this study is to discuss multiple correspondence analysis, which does not need any preconditions and which can be used in the analysis of categorical or categorized data when statistical techniques above are insufficient.

MATERIALS AND METHODS

Material of this study is a hypothetical sample which is composed of three variables, namely variable of enterprise size (small, medium and large enterprises), variable related to information sources (formal and informal sources) to be informed about the novelties and variable of education level of the entrepreneur (0-5, 6-11 and 11-> years). Data of 70 companies related to these variables are produced in the computer with the simulation method. Multiple Correspondence analysis technique was used in the analysis of relationship between three variables as this technique provides more detailed results in comparison to other techniques that can be used for the same purpose. Variables of this study, there are levels of; enterprise size i = 3, education level j = 3, information source k = 2. Obtained data is demonstrated in a table that has i×j×k dimensions. In order to apply multiple correspondence analysis, first of all indicator matrix is formed (Mendeş, 2002; Gifi, 1990). This matrix which is shown as L is expressed as in the following:

$$L = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}_{70\times 8} \quad (1)$$

In the columns of L matrix, total level number of the variables analyzed (3+3+2 = 8), while in the rows interview number (number of tested units) (70) are placed. Consequently, in this case, L matrix becomes a matrix of 70×8 dimensions. L matrix is formed by coding categories of three variables in the questionnaires as 1 and others as 0. In this case, in the L matrix total sum of rows is equal to 1 in a variable's subcategories, while within all categories it is equal to variable number (p). Analysis of L matrix is based on Burt table or Burt matrix, which is formed of inner multiplication of L matrix (Mendeş 2002; Gifi, 1990). This matrix is formed by the following equation:

$$B = L'L \quad (2)$$

Analysis of Burt table is done through Singular Value Decomposition (SVD)method:

$$C_B^{-1} B C_B^{-1} = U\Lambda U' \quad (3)$$

This equation gives the cumulative cluster answer for all categories (levels) of variables (Gifi, 1990).

In the Eq. 3, C_B matrix can be written in the following way while p represents the varying number:

$$C_B = P\begin{bmatrix} C_i & 0 & 0 \\ 0 & C_j & 0 \\ 0 & 0 & C_k \end{bmatrix} \quad (4)$$

Since, in this study, there are 3 coded variables diagonal elements of Burt Matrix are consecutively: $3L_i$, $3L_j$ and $3L_k$.

In this study, MINITAB for windows (version 14.0) statistical package program was used to data analysis.

RESULTS AND DISCUSSION

Table 1 shows the values related to enterprise size, information sources and education level of entrepreneur for total seventy hypothetical enterprises.

In order to apply multiple correspondence analysis, first of all initial matrix is made and then, based on inner multiplication of this matrix, Burt Table is made (Table 2). Diagonal elements of this matrix show sums of three variable levels. As can be seen in Table 2, 30 of the enterprises are small, 25 medium and 15 of them are large sized companies. Elements other than diagonal, for instance, among enterprises at all educational levels that benefited from formal information sources, 10 are small, 11 are medium and 10 are large size companies.

Thirty one of the enterprises have benefited from formal information sources while 39 have benefited from informal information sources. Thirty one of the enterprise owners have education periods of 0-5 years, 26 of them have 6-11 years, while 13 of them have more than 11 years.

Table 3 shows value changes of each dimension according to L Matrix analysis results, within the inertia, which is regarded as average value of variations among the levels of three variables of this data. Total variation of each dimension is calculated by dividing inertia of each dimension by variation. As can be seen from Table 3, first

Table 1: Enterprise size, information sources to follow novelties and education levels

Education level	Small		Medium		Large		
	Formal	Informal	Formal	Informal	Formal	Informal	Total
0-5 year	4	10	0	8	7	2	31
6-11 year	6	7	5	4	1	3	26
11->	0	3	6	2	2	0	13
Total	10	20	11	14	10	5	70

Table 2: Forming of burt table by level of variables

	Small	Medium	Large	Formal	Informal	0-5	6-11	11->
Small	30	0	0	10	20	14	13	3
Medium	0	25	0	11	14	8	9	8
Large	0	0	15	10	5	9	4	2
Formal	10	11	10	31	0	11	12	8
Informal	20	14	5	0	39	20	14	5
0-5	14	8	9	11	20	31	0	0
6-11	13	9	4	12	14	0	26	0
11->	3	8	2	8	5	0	0	13

Table 3: Results of L matrix

Dimensions	Inertia	Explanatory percentages of total variance of dimensions	
		Each dimension (%)	Summative share (%)
1	0.4562	0.2737	0.2737
2	0.4034	0.2420	0.5158
3	0.3438	0.2063	0.7221
4	0.2612	0.1567	0.8788
5	0.2020	0.1212	1.0000
Total	1.6667		

Table 4: Contribution of variable categories in each dimension

Varying levels	1st dimension	2nd dimension
Small	0.191	-0.010
Medium	-0.115	-0.172
Large	-0.033	0.454
Formal	-0.174	0.060
Informal	0.138	-0.047
0-5	0.065	0.140
6-11	0.009	-0.052
11->	-0.275	-0.065

dimension has the highest explanatory ratio (45.62%). Explanatory ratios decrease less and less in each of the other dimensions. When shares of total variation in explanation, total variation in the first two dimensions is 51.58%. In other words, it is possible to explain only 51.58% of total variation, in case distances between levels of variables are demonstrated by reducing from 6 dimensional space, to 2 dimensional space. Showing the relationship between different levels of these variables in two dimensional space is not enough to explain total variation. However, only two dimensions are taken into account in order to show interpretation of the results of this study. On the other hand, share of total variation of three dimensions in the explanation is 72.21%. This percentage also shows that only 72.21% of total variation can be explained by reducing distances between levels of variables from six dimensional space, to three dimensional space. This explanatory ratio can be accepted

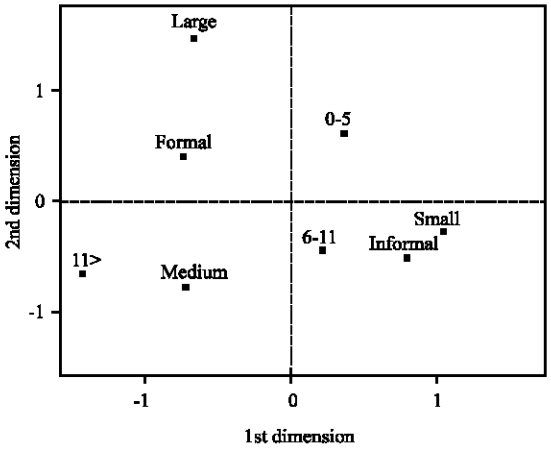

Fig. 1: Multiple Correspondence analysis diagram

as satisfactory (Greenacre, 1998; Gifi, 1990). On the other hand, particularly in cases when variable numbers are greater than small percentages of dimensions in explaining the variance is a drawback. This situation in turn leads to certain difficulties in interpretation chart (Mendeş, 2002; Kaciak and Louviere, 1990).

As shown in Table 4, the first level of enterprise size variable (small) can be placed in the first dimension while the second (medium) and the third (large) can be placed in the second dimension. Both levels of information source variable (formal and informal) can be placed in the first dimension. The first level of education variable (0-5) can be placed in the second dimension while the third level (11->) can be placed in the first dimension. It can be concluded that since the second level (6-11) has a value close to zero in both dimensions this level of education variable has the same level of effect with other variables.

According to the evaluation of three variables together, large firms use formal information source, small enterprises which had 6-11 years of education opt for mainly informal information sources. Medium sized enterprises which have 11 or more years of education can be regarded as having higher education. It can be argued that enterprise owners who had 0-5 years of education have been less affected by the enterprise size and information sources to be informed about novelties in comparison to other levels. Such conclusions can be reached also by applying Multiple Correspondence analysis diagram Fig. 1.

CONCLUSIONS

In the social science researches, assessment of the relationship among coded variables by multiple correspondence analysis technique allows the analysis of relations between the variables and between different

levels of one variable. At the same time, this technique in comparison to other methods statistical results can be seen both analytically and visually. In this way researchers can have more detailed information about the relationship between different variables and it will be easier to interpret the results. Researchers choose correspondence analysis technique also because there are not any preconditions that are necessary in other social sciences methods.

REFERENCES

Aktürk, D., 2004. Multiple correspondence analysis technique: Its application in social science researches. J. Agri. Sci., 10: 218-221. (Çoklu uyum analiz tekniğinin sosyal bilim araştirmalarinda kullanimi, tarim bilimleri dergisi, 10: 218-221).

Başpinar, E. and M. Mendeş, 2000. The usage of correspondence analysis technique at the contingency tables. J. Agric. Sci., 6: 98-106. (İki yönlü tablolarda uyum analizi tekniğinin kullanimi, tarim bilimleri dergisi, 6: 98-106).

Chou, R.J., 1994 Multivariate Analysis and Its Application, pp: 194-210, USA.

Düzgüneş, O., T. Kesici and F. Gürbüz, 1983. Statistical Methods I, Ankara Üniversity, Faculty of Agriculture Pablications 861 (İstatistik Metodlari I, Ankara Üniversitesi Ziraat Fakültesi Yayinlari: 861, Ankara.).

Gifi, A., 1990. Nonlinear Multivariate Analysis. John Willey and Sons Ltd. West Sussex, England, pp: 579.

Greenacre, M., 1998. Visualization of Categorical Data, pp: 107-112, San Diego, USA.

Kaciak, E. and J. Louviere, 1990 Multiple Correspondence Analysis of Multiple Choice Experiment Data JMR, pp: 455-466.

Mendeş, M., 2002. The usage of multiple correspondence analysis technique. J. Agrci. Eng. No. 337. (Çoklu uyum analizi tekniğinin kullanimi. Ziraat mühendisliği dergisi, sayi 337).

Sokal, R.R. and F.J. Rholf, 1995. Biometry. W.H. Freeman and Company New York, pp: 887.