DEÜ MÜHENDİSLİK FAKÜLTESİ
FEN VE MÜHENDİSLİK DERGİSİ
Cilt: 9 Sayı: 3 sh. 1-12 Ekim 2007

APPLICATION OF A POPULATION BASED STUDY OF CORRESPONDENCE ANALYSIS IN CHOOSING A HEALTH INSTITUTION

(SAĞLIK KURUMU SEÇİMİNDE UYGUNLUK ANALİZİNİN TOPLUM TABANLI BİR ÇALIŞMAYA UYGULANMASI)

Aslı SÜNER*, Can Cengiz ÇELİKOĞLU*

ABSTRACT/*ÖZET*

Correspondence analysis is a method that makes it easy to interpret the categorical variables given in contingency tables, showing the similarities between row and column variables and associations as well as divergences among these variables via graphics on a lower dimensional space. It factors the data matrix into the rows and columns, shows each of them with a separate graphics; it releases important information about the structure of the data. Having lots of categorical data in applied science, such as in medicine, biology and economy, makes correspondence analysis a very popular method. In correspondence analysis, the aim is to explain the associations among the variables on a lower dimensional space. A multidimensional graphic consisting the points regarding the levels of categorical variables is drawn using mathematical methods to determine the structure of the contingency table. There are similarities between this analysis and other multivariate methods; mainly Principal Component Analysis, Log-Linear Models and Multi-dimensional scaling. In this study, correspondence analysis will be applied to the epidemiological research data obtained from the DEU School of Medicine Department of Public Health.

Uygunluk Analizi, kategorik değişkenlerin yorumlanmasını kolaylaştıran, çapraz tablolarda satır ve sütun değişkenleri arasındaki benzerlikleri ve bu değişkenlerin birlikte değişimlerini, daha az boyutlu bir uzayda grafiksel olarak gösteren bir yöntemdir. Bu teknik, veri matrisinin satır ve sütun bölgelerine ayrıştırılması üzerine yoğunlaşıp, elde edilen bileşenleri ayrı ayrı grafiklerle göstererek, veri setinin yapısına ilişkin önemli bilgiler vermektedir. Özellikle tıp, sağlık bilimleri, biyometri, ekonomi, pazarlama ve sosyal bilimler gibi kategorik verilerin analizine ihtiyaç duyulan alanlarda oldukça popüler bir yöntemdir. Uygunluk analizinde, değişkenler arasındaki ilişkilerin indirgenmiş boyutlu bir uzayda sunulması amaçlanmaktadır. Bu analizde, çapraz tabloların yapılarını belirlemek amacıyla matematiksel teknikleri kullanarak çok boyutlu uzayda değişkenlerin kategorilerini temsil eden noktaları içeren bir grafik oluşturulur. Uygunluk analizinin, çok değişkenli analiz tekniklerinden Temel Bileşenler Analizi, Log-Doğrusal ve Çok Boyutlu Ölçekleme yöntemleriyle de benzerliği bulunmaktadır. Bu çalışmada, DEÜ Tıp Fakültesi Halk Sağlığı Anabilim Dalında yürütülen araştırma verileri kullanılarak, uygunluk analizi ile ilgili bir uygulama yapılmıştır.

KEYWORDS/*ANAHTAR KELİMELER*

Correspondence analysis, Multivariate analysis methods, Analysis of categorical data
Uygunluk analizi, Uyum analizi, Çok değişkenli analiz yöntemleri, Kategorik veri analizi

* Dokuz Eylul University, Faculty of Art and Science, Department of Statistics, İZMİR

1. INTRODUCTION

In scientific studies depending on real life applications; the features of the objects under observation are related to each other since the events that are being discussed are usually influenced by many factors. It is an obligation to evaluate the objects under observation, from all possible aspects, to get valid and reliable results from the studies (Tatlıdil, 1996). So as to study on multivariate data and analysis, it is referred to the multivariate statistical methods.

Reduction of difficulties confronted in the commendation and summarization of the results related to the multiple variables is aimed with the help of multivariate statistical methods used in the analysis of large data matrix (Burcu and Çetin, 2004). Methods with one variable have a restricting assumption; they require experimental control of many factors in the event (and necessity) of observing the effects of each factor at one time (Çetin, 2003). On the other hand, in multivariate statistical methods, get the "event" analysis integrated and to explain the collinearity of the variables supplying this integrity is aimed.

The multivariate statistical methods are divided into two according to their interrelation as; methods used in the analysis of dependency structures and methods used in the analysis of interdependency structures. In the methods used in the analysis of dependency structures, one or more variable must be dependent on the other variables and their values must be predictable and explanatory (Bayram, 2000). In the methods used in the analysis of interdependency structures, there is the condition that the variables would not be explained with the help of other variables, the values would not be predicted or the variables would not be defined as dependent or independent, and the whole relation between the variables are considered (Bayram, 1996). The general tendency of where and when the multivariate statistical methods are used can be classified as in the Table 1 (Çetin, 2003). Basic multivariate statistical methods are explained and their classification is defined as is shown in Figure 1 (Hair et al., 1992).

Apart from many other multivariate statistical methods; in correspondence analysis, not only the inner relations among the levels of the rows and the columns in data matrix are given, but also their similarities and the differences are shown. Moreover, the data structure is interpreted by observing the levels of the variables, which state the features expressed in rows and columns (Behdioğlu, 2000).

Correspondence analysis is similar to principal component analysis. In both methods the dimension of the data matrix is aimed to be reduced and presented in a simple way in multidimensional space; the data is analyzed in reduced dimensions with the help of singular value decomposition (Etikan et al., 2000). Similar to the principal component analysis' decomposing of general change into units, data can be divided according to inertia in correspondence analysis. There is a difference between these two techniques on type of the data matrix. In principal component analysis, the data consist of continuous or discrete measured variables, which supply a multivariate normal distribution. In the other hand in correspondence analysis the data are categorical and there is no need for them to be distributed, according to a distribution assumption (Etikan et al., 2000). The non-linear method is neither wanted nor needed in correspondence analysis (Seyfullahoğulları, 2002). Correspondence analysis has some similarities to the Log-linear analysis and multidimensional scaling. But correspondence analysis differs from log-linear models since it does not need an assumption related to the distribution and differs from multidimensional scaling in the base of showing the relations between the categories and variables in the same space (Tuna and Kıroğlu, 1996).

Table 1. The general tendency of multivariate statistical methods

Multivariate Method	When It Is Used?	Function
Factor Analysis	It is used to find and discover a less number of meaningful variables (factors) by gathering a large number of variables.	It explains the inner relations between the multivariate variables and finds the common dimensions (factors) standing in the centre of the variables.
Cluster Analysis	It is used to classify the grouped data and to get proper, useful, summarized data.	It classifies the entities (individuals or objects) to smaller sub-groups by their similarities.
Discriminant Analysis	It is used to determine the discriminant factors effecting the discriminant of groups and to assign an object, whose source is not known, to a group.	It predicts the group differences and an entity's (an individual or an object) belonging to a group or a cluster defined by metric independent variables.
Multivariate Regression Analysis	It is used to observe and analyze the relationship between a dependent variable and one or more independent variables.	It predicts the changes of dependent variables as a response to independent variables.
Multidimensional Scaling Method	It is used in the analysis of behavioural data such as personal choices, manners, tendencies, beliefs and preferences.	It gives the structure of the objects as similar as possible to the original vision with as few dimensions as possible.
Correspondence Analysis	It is used in the analysis of the categorical data. The graphical outcome of the analysis has rich data that can be used for decision making.	It visually presents the closeness and similarities between the objects and gives the relations, which were not given in the tables.
Principal Component Analysis	It is used to reduce the data, to re-form the data set in order to be analyzed by some methods, to calculate the principal component scores and to order the units according to these scores.	It gives a chance to find out the relations, which were not reached before, and finding unusual results.
Logistic Regression Analysis	It is a method used when the dependent variable is a discrete variable composed of binary (0,1) or multi levels	It is a method to group the observations.
Canonical Correlation Analysis	It is an expansion and extension of the multiple regression analysis.	It relates various independent and dependent metric variables.

Correspondence analysis can be applied in two ways according to the number of variables and dimensions in the cross tables (Cangür et al., 2005). In its very simple form, it is called "Simple Correspondence Analysis" and is used in examination of two sided cross tables. When the number of variables is not limited, the variables are coded as a matrix and applied to the multi-way cross table, the case is called as a "Multiple Correspondence Analysis" (Greenacre and Hastie, 1987). Multiple correspondence analysis is also known as "Homogeneity Analysis", HOMALS (Homogeneity analysis by Alternating Least Squares) and it is named in America as "optimal scaling", "optimal scoring", "reciprocal averaging" and "appropriate scoring"; in Japan as "quantification methods"; in Holland as "homogeneity analysis"; in Canada as "dual scaling" and in Israel as "scalogram" (Gifi, 1990; Etikan et al., 2000; Bayram, 2000). In Turkish literature, these are used as "Uyum Analizi", "Karşılık Getirme Analizi" and "Görsel İlişki Analizi" (Akıncı and Atılgan, 2005).

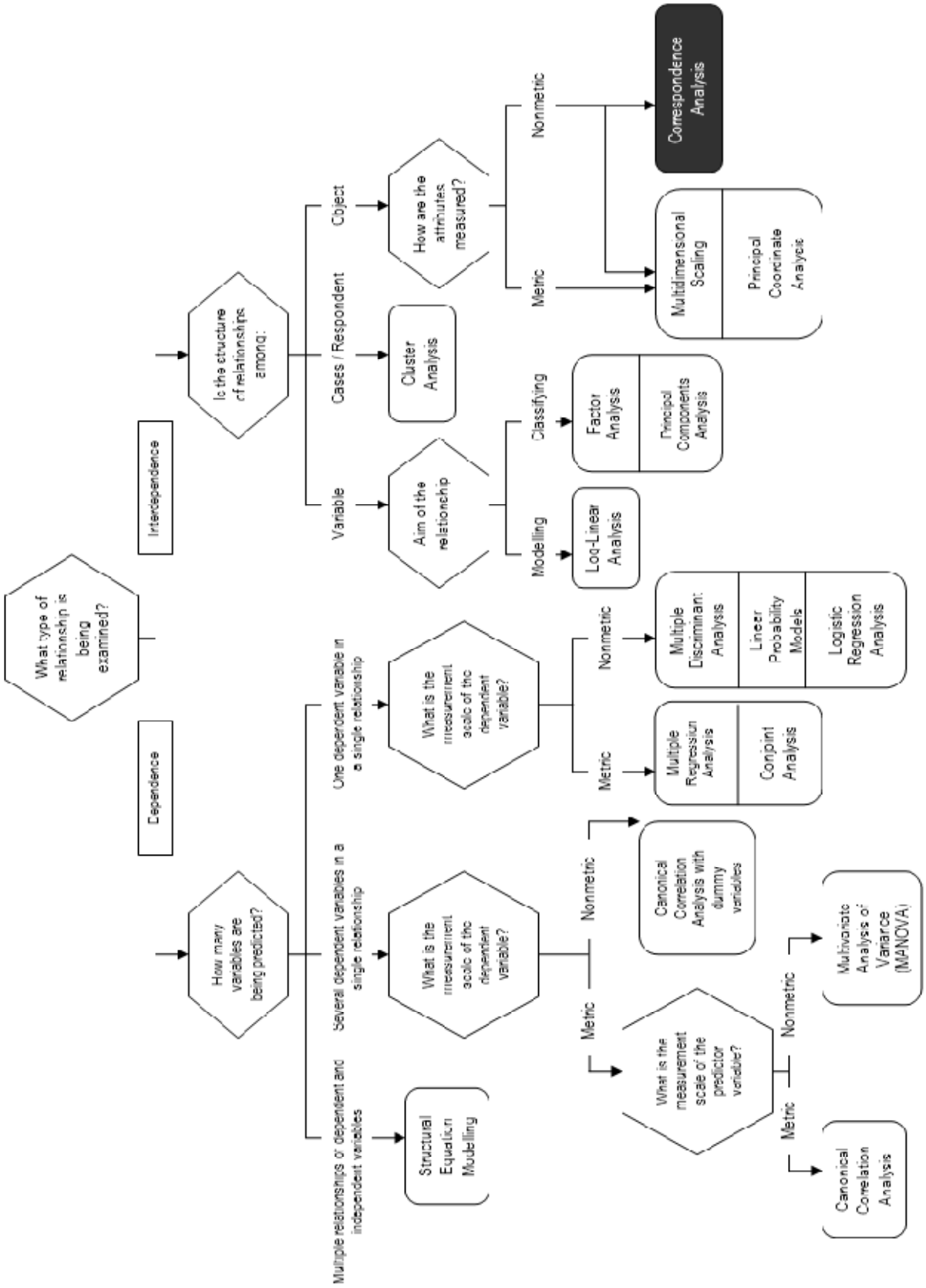

Figure 1. Selecting a multivariate technique

The relationships among categorical variables in large tables can be summarily described with correspondence analysis. The associations between rows and columns in a frequency table are mapped graphically, as points in a space of few dimensions. Conceptually, the analysis steps are simple. Category profiles and marginal proportions are computed, the distances between these points are calculated, and the best-fitting space of n dimensions is located. Correspondence analysis is preferred since it is a data reduction and residual analysis and it gives information about two or more dimensional graphs and categorical variables, since it has the ability to explain the lack of homogeneity in row profiles or the (dependency or interaction) between the rows and columns of the cross tables in a fewer dimensions (Cangür et al.., 2005).

Since the studies on categorical data analysis high in number, the use of correspondence analysis is popular. It is a widespread method in the fields where the categorical data analysis is needed, such as medicine, medical sciences, biometry, economy, marketing and social sciences. Because it is useful in the analysis of cross tables and it presents easy, understandable and easily interpretable graphical visualizations, this method has gained a place in marketing research lessons' course plans at universities. Especially, related to the increasing rate of computer use, it has become an optimal scaling method which can be found in statistical packages such as SPSS, MINITAB and SAS (Clausen, 1998).

2. APPLICATION

In this research, data concerning adults who reside in Narlıdere, İzmir were examined in detail. The utilized data belongs to 348 people, who have illnesses, and was gathered for adults' mental health research at Narlıdere education and research area by the DEU School of Medicine Department of Public Health. Application of a population based study of correspondence analysis in choosing a health institution with these data was aimed. Correspondence analysis was carried out so as to analyze the medical establishment, consulted by people who reside in Narlıdere according to their disease group, their age classification and the health insurance they have.

This data set contains information on the characteristics of ages, health insurance, health institute and disease. The following tables show the variables, along with their variable labels, and the value labels assigned to the categories of each variable in the data set. Distribution of categories of age is given in Table 2. Looking at this table, 39.7% of patients (138 people) are in 50-64 age group, 27.0% (94 people) are in 35-49 age group, 21.6% (75 people) are at 65 or above and 11.8% (41 people) are in 18-34 age group. Distribution of the categories of age group variable can be seen in Figure 2.

Table 2. The frequency distribution of age variable

Age	Frequency	Percent	Cumulative Percent
18 - 34	41	11,7	11,7
35 - 49	94	27,0	38,7
50 - 64	138	39,7	78,4
65+	75	21,6	100,0
TOTAL	348	100,0	

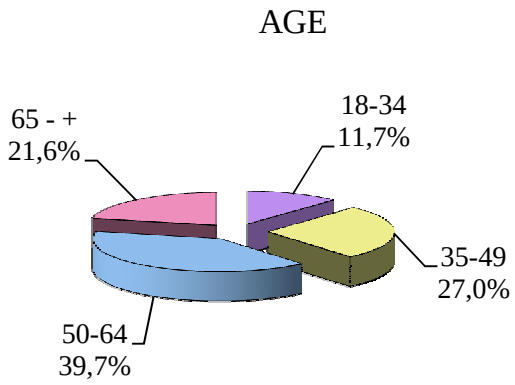

Figure 2. Pie chart of age groups

When looking at Table 3, distribution of categories of patients' health insurance variable can be seen. In this table, 38.8% of patients (135 people) have health insurance from social security organization, 36.5% (127 people) have health insurance from retirement fund, 10.1% (35 people) have health insurance from Bağkur, 3.7% (13 people) have green card coverage and 2.3% have commercial health insurance while 8.6% (30 people) do not have any health insurance. Distribution of the categories of patients' health insurance variable is illustrated in Figure 3.

Table 3. The frequency distribution of health insurance variable

Health Insurance	Frequency	Percent	Cumulative Percent
None	30	8,6	8,6
Social security organization	135	38,8	47,4
Retirement fund	127	36,5	83,9
Bağkur	35	10,1	94,0
Commercial	8	2,3	96,3
Green card coverage	13	3,7	100,0
Total	348	100,0	

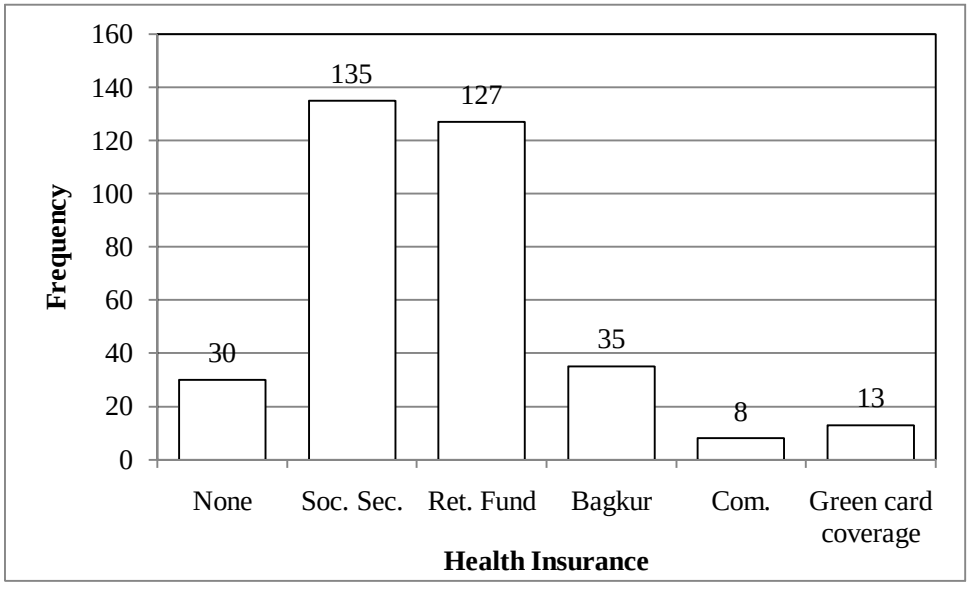

Figure 3. Bar chart of health insurances

Distribution of categories of patients' referred health institute variable is given in Table 4. In this table, 25.3% of patients (88 people) enter university hospitals, 20.1% (70 people) enter social insurance institute hospitals, 9.2% (32 people) enter public hospitals, 4.6% (16 people) enter health unit and 4.0% (14 people) enter private hospitals. It is unknown which medical establishment 36.8% of patients (128 people) enter. Frequency distribution of health institute variable is examined in detail in Figure 4.

Table 4. The frequency distribution of health institute variable

Health Institute	Frequency	Percent	Cumulative Percent
Health unit	16	4,6	4,6
Social insurance inst.	70	20,1	24,7
Public	32	9,2	33,9
University	88	25,3	59,2
Private	14	4,0	63,2
Unknown	128	36,8	100,0
TOTAL	348	100,0	

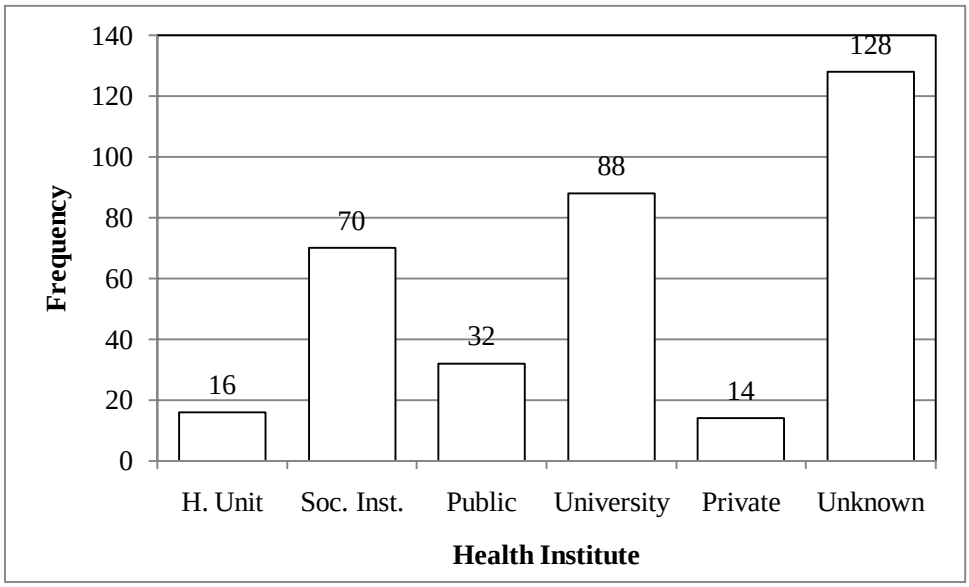

Figure 4. Bar chart of health institutes

When looking at Table 5, distribution of categories of disease variable is seen. In this table, 27.0% of patients (94 people) have cardiovascular diseases (hypertension, heart attack, heart failure and apoplexy), 15.8% (55 people) have both cardiovascular and other diseases except diabetes, 14.1% (49 people) have rheumatism, 12.6% (44 people) have other diseases (tuberculosis, epilepsy, having obstacles, cancer, etc.), 8.0% (28 people) have diabetes, 7.8% (27 people) have gut-rot, 7.5% o (26 people) cardiovascular disease, diabetes and any other diseases, 4.0% (14 people) have asthma, 3.2% (11 people) have depression. Distribution of the categories of the disease variable can be evidently seen in Figure 5.

Table 5. The frequency distribution of disease variable

Disease	Frequency	Percent	Cumulative Percent
Cardiovascular	94	27,0	27,0
Diabetes	28	8,0	35,1
Asthma	14	4,0	39,1
Gut-rot	27	7,8	46,8
Rheumatism	49	14,1	60,9
Depression	11	3,2	64,1
Other	44	12,6	76,7
Cardiovascular, Diabetes and Any Other	26	7,5	84,2
Cardiovascular and Any Other (except diabetes)	55	15,8	100,0
Total	348	100,0	

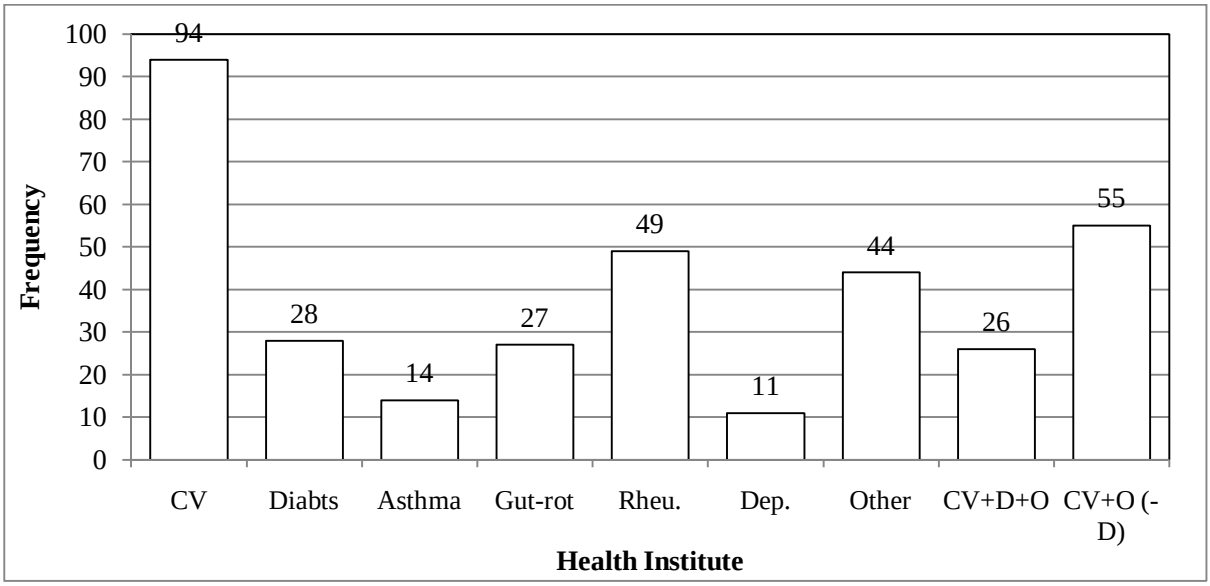

Figure 5. Bar chart of diseases

3. RESULTS OF ANALYSIS

In the application, medical establishment consulted according to disease type was examined thoroughly by simple correspondence analysis as a sample in the first place. Distribution of age group, health insurance and disease type according to medical establishment were studied by multiple correspondence analysis.

The symmetrical normalization makes it easy to examine the relationship between referred health institute of patient and diseases of patients categories. For example, diabetes are near the social insurance institute hospital category, while other diseases are closest to health unit. Cardiovascular diseases and other diseases seem to be associated with public hospitals; gut-rot and depression are not strongly associated with any particular referred health institute of patient category. This relationship is shown in Figure 6.

At the graphic obtained as a result of multiple correspondence analysis in Figure 7, it was seen that medical establishments consulted by adults, who are at 34 and under, having gut-rot disease and not having any health insurance was unknown. It is seen that individuals, who are at 35-49 age group and having rheumatism and asthma, prefer going private hospitals and individuals, who have diabetes and social insurance institute health insurance, prefer going to social insurance institute hospitals. It was seen that adults, who are at least 50 and above and have cardiovascular disease (hypertension, heart attack, heart failure and apoplexy) and other diseases and have retirement fund health insurance or private health insurance, prefer going to university hospitals. Besides, individuals who have other diseases (tuberculosis, epilepsy, having obstacles, cancer, etc.) and have Bağkur health insurance prefer going to public hospitals and health unit. When analyzing the graphic above, it is seen that categories of individuals, who have green card coverage, depression and cardiovascular, diabetes and other diseases, is distant from origin. This denotes that marginal frequencies of these categories are less then the others.

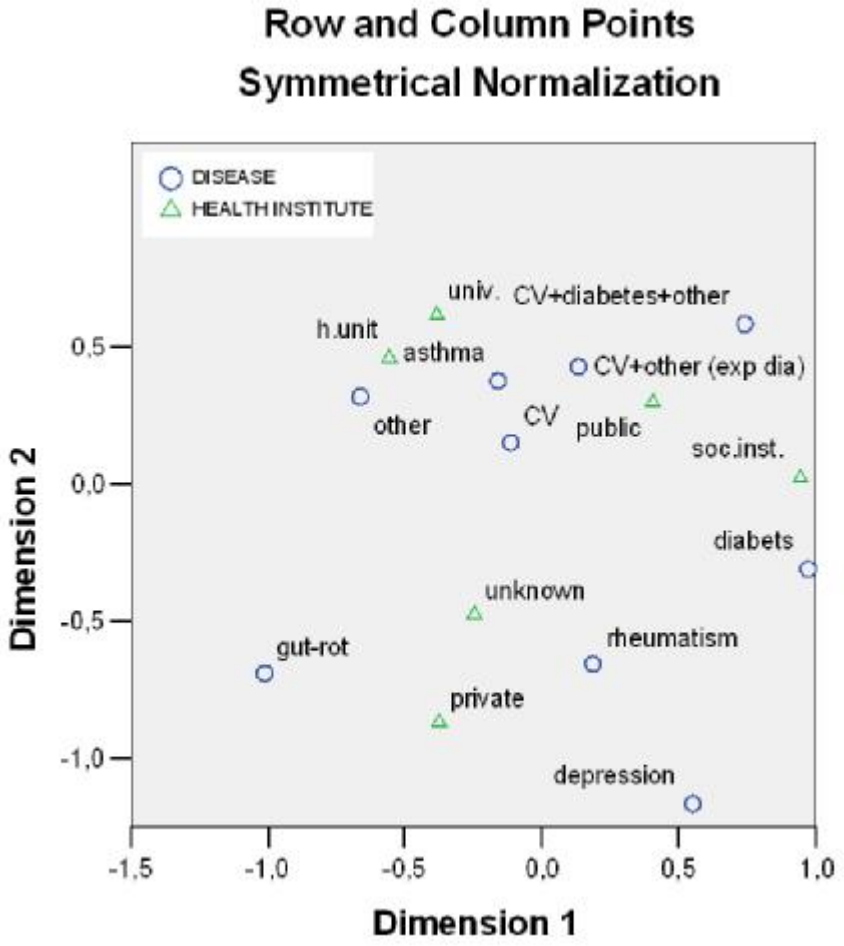

Figure 6. The symmetrical normalization graph of disease and health institute

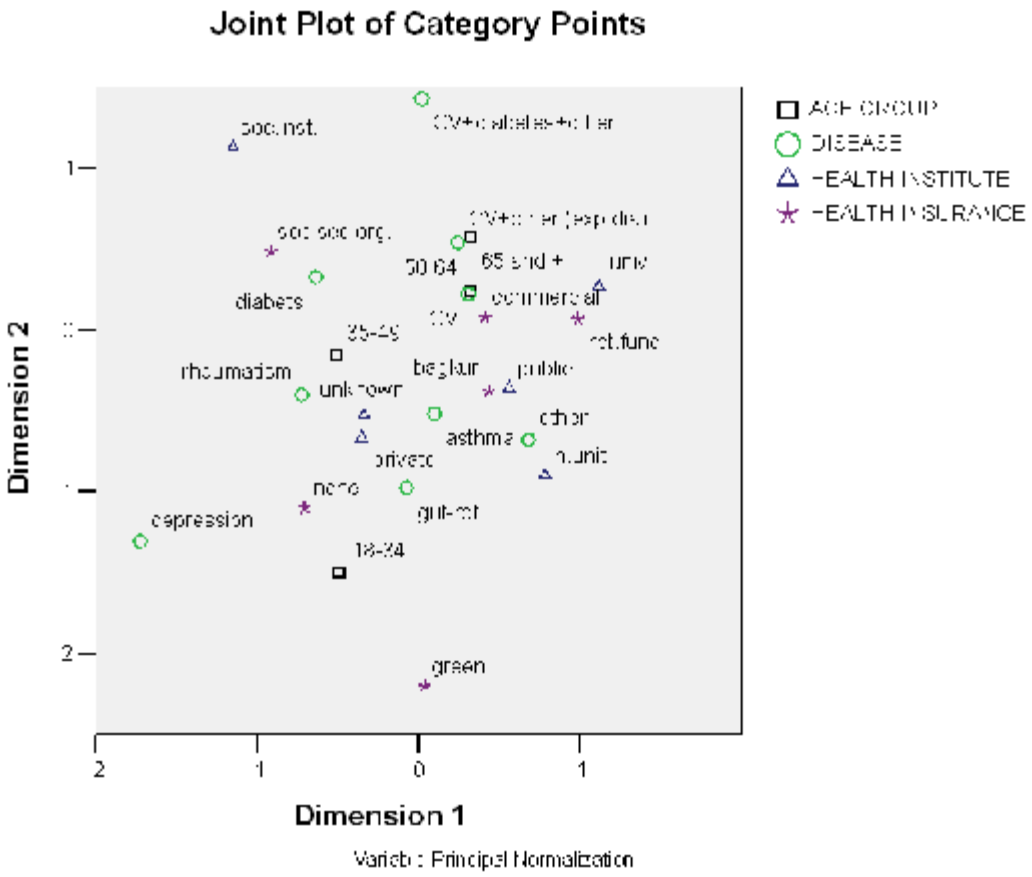

Figure 7. Discriminant measures graph of variables

4. CONCLUSIONS

Correspondence analysis is one of the multivariate statistical methods, in which the corresponding relationships between categorical variables. This analysis, which started to be used after 1930's, differs from other types of multivariate analyses, such that, it does not only help investigating the corresponding relationships defined in the columns or rows of the data matrices, but also investigates the similarities and differences between them. In addition, data structure is interpreted by investigating all of the relationships, defined as the columns or rows in the data matrix. It is a widely preferred method, especially because, it has no pre-assumptions about the distribution.

Correspondence analysis is a very flexible method. It imposes few requirements in regards to the data and it allows the use of supplementary information. It is an exploratory technique and can thus serve as a useful supplement to other methods based on hypothesis testing. The graphical display of the analysis is very beneficial for communicating complex relations between variables/categories (Clausen, 1998).

Most of the multivariate statistical methods have many similarities and differences. After the general evaluation of these methods it has been observed that component analysis and principal component analysis (PCA) are types of dimension decrease and dependence structure elimination methods and also, that the factor analysis has a property of defining common components by variable grouping. The analyses used in grouping the observations are cluster analysis, discriminant analysis and logistic regression analysis. In multivariate regression analysis, cluster analysis, discriminant analysis and canonical regression analysis,

the criteria of "suitability to the assumptions" is wanted. There are no limitations of assumptions in multivariate regression analysis, principal components analysis, multifunctional scaling method and logistic regression analysis.

In order to apply correspondence analysis to a population based study, purpose of which is the selection of the health institution, individuals at the age of 18 or more who live in Narlıdere town of Izmir city was investigated as the sample population. The data used in this study was taken from the project "Research of Mental Health of Adults at Narlıdere Education and Research Area" designed by DEU School of Medicine Department of Public Health. In this study, different attributes of the patients, was applied to different health institutions. When the results of this study are investigated and documented, the conclusions below were reached:

- It was seen that medical establishments consulted by adults, who are at 34 and under, have gut-rot disease and do not have any health insurance, was unknown.
- It is seen that individuals who are at 35-49 age group and have rheumatism and asthma, prefer going to private medical establishments
- Individuals, who have diabetes and have social insurance institute health insurance, prefer going to social insurance institute hospitals.
- It was seen that adults, who are at 50 and above and have cardiovascular disease (hypertension, heart attack, heart failure and apoplexy) and other diseases except diabetes and have retirement fund health insurance or private health insurance, prefer going to university hospitals.
- Individuals, who have other diseases (tuberculosis, epilepsy, having obstacles, cancer, etc…) and have Bağkur health insurance, prefer going to public hospitals and health unit.

The main problem of correspondence analysis encountered in the application was the formation of correct clusters in the mapping procedure. In the future studies, application of different kinds of clustering methods on the resulting maps is planned. Also, evaluations of the results of this application with other multivariate analysis methods and comparison of their efficiencies on this application are highly recommended to the researchers who are interested in the subject investigated in this study.

REFERENCES

Akıncı S., Atılgan E. (2005): "Pazarlama Araştırmalarında Kategorik Verilerin Haritalanması: Görsel İlişki Analizi ve Uygulama Örneği", Akdeniz İİBF Dergisi, Cilt 9, s. 1-17.

Bayram N. (1996): "Minres Tekniği ile Faktör Analizi üzerine bir Uygulama Denemesi", Bursa: Basılmamış Y.L. Tezi.

Bayram N. (2000): "Karşılık Getirme Analizi ve Bankacılık Sektörüne Uygulanması", Uludağ Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı, İstatistik Bilim Dalı, Bursa: Doktora Tezi.

Behdioğlu S. (2000): "Çok Değişkenli Veri Yapısının Yorumlanmasında Olumsallık Tablolarının Uygunluk Çözümlemesi ve Bir Uygulama", Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Uygulamalı İstatistik Bilim Dalı, Bursa: Doktora Tezi.

Burcu E., Çetin M. C. (2004): "Özürlülüğe İlişkin Düşüncelerin Homojenleştirme Analizi ile İncelenmesi: Ankara Örneği", 4. İstatistik Günleri Sempozyumu.

Cangür Ş., Sığırlı D., Ediz B., Ercan İ., Kan İ. (2005): "Türkiye'de Özürlü Grupların Yapısının Çoklu Uyum Analizi ile İncelenmesi", Uludağ Üniversitesi, Tıp Fakültesi Dergisi, Cilt 31, No. 3, s. 153–157.

Clausen S. E. (1998): "Applied Correspondence Analysis: an Introduction", USA: Sage Publications.

Çetin E. İ. (2003): "Çok Değişkenli Analizlerin Pazarlama ile İlgili Araştırmalarda Kullanımı: 1995–2002 Arası Yazın Taraması", Akdeniz İ.İ.B.F. Dergisi, Cilt 5, s. 32-47.

Etikan İ., Uysal M., Sanisoğlu Y., Dirican B. (2000): "Uygunluk Analizi ile Kanser Vakalarının Çözümlenmesi", 5. Ulusal Biyoistatistik Kongresi.

Gifi A. (1990). "Non-linear Multivariate Analysis", John Wiley and Sons. Ltd., Chichester.

Greenacre M. J., Hastie T. (1987): "The Geometric Interpretation of Correspondence Analysis", Jasa, 82 (398), s. 437–447.

Hair F. J., Anderson E. R., Tahtam L. R., Black C. W. (1992): "Multivariate Data Analysis", New Jersey: Prentice Hail Inc.

Seyfullahoğulları A. (2002): "Uygunluk Analizi ve Tekstil Sektöründe Toplam Kalite Yönetimi Anlayışı Üzerine bir Uygulama", Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı, İstatistik Bilim Dalı, İstanbul, Doktora Tezi.

Tatlıdil H. (1996): "Uygulamalı Çok Değişkenli İstatistiksel Analiz", Ankara: Akademi Matbaası.

Tuna M., Kıroğlu, G. (1996): "Uygunluk Analizi Üzerine bir Uygulama", Araştırma Sempoyumu'96.