II. ARAŞTIRMALAR

DİSKRİMİNANT ANALİZİ VE BUNUNLA İLGİLİ BAZI PROBLEMLER ÜZERİNDE BİR ARAŞTIRMA (1)

Aydın ÖZTÜRK (2)

ÖZET :

Hangi populasyona ait oldukları kesinlikle bilinmeyen fertlerin ayırım ve sınıflandırılmalarını konu edinen diskriminant analizinin genel teorisi incelenmiştir. Optimum sınıflandırma kaidesi çok değişkenli populasyonlar için izah edilerek teori kesikli dağılış gösteren populasyonlar için de genelleştirilmiştir.

Araştırmada diskriminant analizi ile ilgili yanlış sınıflandırma ihtimalleri, kalitatif verilerle yapılan sınıflandırma, değişkenlerin seçimi ve kayıp müşahadeler gibi konular ele alınarak bunlara ait çeşitli metodların mukayesesi yapılmıştır.

Giriş :

Populasyonun parametreleri hakkında bilgi edinmek amacıyla alınan örneklere giren fertler tek bir vasıf bakımından ölçüye vurulabilecekleri gibi, birçok hallerde aynı şey iki ve daha fazla haller için de yapılabilir. Meselâ, aralarındaki farkın önemini tesbit etmek için iki ayrı muamelenin tatbik edildiği bir bitki çeşidinde yaprak genişliği, gövde uzunluğu ve kalınlığı gibi ölçülerin alınması istenebilir. Böyle ölçülere konu olan değişkenler arasında genellikle bir ilgi mevcut olup, bunların ayrı ayrı analiz edilmeleri yanlış bir sonuca yol açar.

Çok değişkenli istatistikte, birden fazla ölçüye vurulan fertlerin ayrımı ve sınıflandırılmaları bahis konusu olduğu zaman bunun muayyen bir kriterc göre yapılması gerekir. Bu tip problemler çok-değişkenli istatistikte diskriminant analizi adı altında topanmıştır. Teorisi kısa zamanda geliştirilen diskriminant analizi tıp, ziraat, ve taksonomi gibi konularda geniş çapta uygulama alanı bulmuştur.

(1) Prof. Dr. Şaban Karataş, Prof. Dr. Fahrettin Tosun, ve Prof. Dr. İbrahim Aksöz'den müteşemkil bir jüri tarafından kabul edilen doktora tezinin özetidir.

(2) A)atürk Üniversitesi Ziraat Fakültesi, Zootekni Bölümü Doktor Asistanı.

Dergi Komisyonuna geliş tarihi: 28.7.1973.

Literatür Özeti:

Anderson (1951), diskriminiant analizi konusunda geniş açıklamalarda bulunmuştur. Fertlerin mensup oldukları populasyonlara yanlışlıkla sınıflandırılması ihtimallerinin minimum yapılmasını kriter alarak aşağıdaki istatistiği bulmuştur.

$$W(\underline{X}) = [\underline{u}(1) - \underline{u}(2)]' \underline{V}^{-1} \underline{X} - \frac{1}{2}[\underline{u}(1) + \underline{u}(2)]' \underline{V}^{-1} [\underline{u}(1) \underline{u}(2)]$$

$$> \ln \frac{q_2 c_{12}}{q_1 c_{21}}$$

Formüldeki $\underline{u}(1)$, $\underline{u}(2)$ birinci ve ikinci populasyonlara ait ortalama vektörlerini $\underline{V}$ varyans kovarıyans matrisini q_1, q_2 populasyonlara ait a priori ihtimalleri, c_{12} ve c_{21} ise birinci ve ikinci populasyonlara yanlışlakla yapılan sınıflandırmanın maliyetini göstermektedir. Bu istatistiğe göre $\underline{X}$ vektörünün müşahade edildiği ferde ait $W(\underline{X})$ değeri eşitsizliğin sağındaki değerden büyükse ilgili fert birinci, aksi halde ikinci populasyona dahil edilir.

$\underline{U}(1)$, $u(2)$ ve $\underline{V}$ parametleri bilinmedikleri takdirde bunların yerine örnek tahminleri olan $\bar{\underline{x}}(1)$, $\bar{\underline{x}}(2)$ ve S değerleri kullanılır. Böylece bulunan yeni istatistik $w(\underline{X})$, sapmasız değildir. Wald (1944) ve Sitgreaves (1952) $w(\underline{X})$ in örnek dağılışını bulmuşlardır. Anderson (1958) ikiden fazla olan çokdeğişkenli normal populasyonlarda ayırım ve sınıflandırmanın nasıl yapılalacağını göstermiştir.

Sürekli değişkenler için geliştirilen teori kesikli değişkenler içinde genelleştirilebilir. Eldeki verilerin multinomiyal dağılış göterdiği farzedilirse sürekli değişkenlerdeki gibi bulunan optimum kriterle sınflandırma yapılabilir (Cochran ve Hopkins, 1962). İki şıklı değişkenlerin (dichotomus) tarif ettikleri populasyonlar için yapılan sınıflandırmada Martin ve Bradley (1972) değişik bir ihtimal modeli kullanmıştı.

Materyal ve Metod :

Araştırmada iki gurup veri kullanılmıştır. Bunlardan birincisi şans sayıları kullanılarak elde edilmiştir. Ortalaması ve variyans bilinen bağımsız normal değişkenler elde edildikten sonra çok - değişkenli dağılışa geçilmiştir. Diğer guruptaki veriler ise Atatürk Üniversitesi, Öğrenci İşlerindeki kayıtlardan alınmıştır.

Bu verileri kullanmak suretiyle disriminant analizi ile ilgili bazı problemler için şimdiye kadar ileri sürülmüş olan çeşitli metodların mukayesesi yapılmıştır.

Araştırmalar ve Sonuçları :

Literatür özetinde verilen teorinin ışığı altında yapılan çalışmalarda sırasiyle, yanlış sınıflandırma ihtimalleri, kalitatif verilerle yapılan sınıflandırma,

değişkenleri seçimi ve kayıp müşahadeler gibi konular ele alınmıştır.

Hata nisbetlerini tahmin için geliştirilen metodlardan beş tanesi alınarak mukayese edilmiştir. Bunlardan $D^2 = [\underline{\bar{x}}(1) - \underline{\bar{x}}(2)]' \underline{S}^{-1} [\underline{\bar{x}}(1) - \underline{\bar{x}}(2)]$ si alınarak istatistiğine dayanan iki tanesinin en iyi olduğu sonucuna varılmıştır.

Öğrenci kayıtlarından alınan kalitatif verilerle yapılan sınıflandırmada Cochran ve Hopkins metoduyla Martin ve Bradley metodunun uygulaması gösterilmiştir. Ayrıca bir normal değişkenin kalitatif şekle getirilmesinden dolayı ayırım gücünde meydana gelen değiddeğişiklik incelenmiştir. İki populasyon arasındaki mesafe çok küçük olduğu zaman bu değişikliğin takriben % 63,6 olduğu gösterilmiştir.

Kayıp müşahadelerin tahmini için varyans - kovarıyans matrisinin izinin ve determinatının minimum yapılmasını kriter olarak alan iki metodun mukayesesi yapılarak bunların birbirlerine olan üstünlükleri izah edilmiştir.

Son olarak diskriminant fonksiyonu için kullanılan değişkenlirin seçiminde bazı metodlar incelenmiştir. Diskriminant fonksiyonu ile çoklu reğresyon arasındaki ilişkide gösterilerek değişkenlerin seçimi için regresyon analizindeki tekniklerden birinin kullanılmasının mümkün olabileceği belirtilmiştir.

A RESEARCH ON DISCRIMINANT ANALYSIS AND SOME RELATED TOPICS

In this reseacrh the general theory of discriminant analysis and some related topics such as probalitiy of misclassification, choosing the best variables and the estimation of the missing observations were considered.

The poblem of classification aries when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several pupulations on the basis of these measurements . In costructing a precedure of calassification it is desired to miniinise the probalitiy of misclassification. For this purpose a calissification procedure was obtained by minimising the expected cost of misclassification.

Anderson propesed the following statistics for the classifictaion of the individuals into one of two multivariate normal populations.

$$W(\underline{X}) = [\underline{u}(1) - \underline{u}(2)]' \underline{V}^{-1} \underline{X} - \frac{1}{2} [\underline{u}(1) + \underline{u}(2)]' \underline{V}^{-1} [\underline{u}(1) - \underline{u}(2)]$$

where $\underline{u}(1)$, $\underline{u}(2)$ are the mean vectors and $\underline{V}$ is the variance - covariance matrix. When the parameters $\underline{u}(1)$, $\underline{u}(2)$ and $\underline{V}$ are unknown, the sample estimates of these are used. The above mentioned statistical formula is the same as Fisher's linear discriminant function, except the constant term included in $W(\underline{X})$. In the case where there are more than two populations, the allocation rule was generalised. The same rule was used to show how to classify individuals with the multivariate qualitative data.

The data used in the research was obtained by inkoving the simula-

tion techniques and gathering observations on the student records.

Comparing five of the methods developed to estimate the probalitiy of misclassification, two of them were found to be the best.

Application of two different was illustrated for the classification with the multivariate qualitative data. The effect of replacing a normal variate by a qualitative variate was also mentioned for a special case.

The advantages and disadvantages of the two methods proposed to estimate the missing observations were discussed.

In the last part of the resarch some which are used to sellect the best variables for discriminant function were considered. The relation between the multiple regression equation and the discriminant function was illustrated, and the possibility of the use of one of the regression techniques was pointed out for selecting the best variables of the discriminant function.

LİTERATÜR

Anderson, T. W. (1951), Classification By Multivariate Analysis. Psychometrika, 16, 31.

Anderson, T. W. (1958), An introduction to Multivariate Statistical Analysis. john Wiley and Sons Inc. USA.

Cochran, W. G. ve Hopkins, C. E. (1961). Some Classification Problems With multivariate Qualitative Data. Biometrics. 17, 10.

Martin, D. C. ve Bradley, R. A. (1972), Probability Models. Estimation, and Classification For Multivariate Dichotomus Populations. Biometrics, 28, 203.

Sitgreaces, R. (1952), On the Distribution of Two Random Matrices Used in Classification Proceoures. Annals of Math. Statist. 23,263.

Wald, A. (1944) On a Statistical Problem Arising in the Classification of an Individual Into One of Two Groups. Annals of Math. Statist. 15,145.