DOKTORA TEZİ

Tez Yöneticisi
Prof. Dr. Şaban KARATAŞ

DİSKRİMİNANT ANALİZİ VE BUNUNLA İLGİLİ BAZI PROBLEMLER ÜZERİNDE BİR ARAŞTIRMA

Aydın ÖZTÜRK

Atatürk Üniversitesi Ziraat Fakültesi
Zootekni bölümü Asistanı

Erzurum - 1973

Ö N S Ö Z

Müsbet ilimlerin konu edindiği olaylar, tabiatları icabı birden fazla ölçü ile ifade edilmek zorundadır. Bu, olayların ve cisimlerin çok boyutlu bir uzuyda yer alan noktalar halinde bulunmalarının tabii bir sonuçudur.

İlimdeki gelişmeleri yakından takip etmek, veya bu gelişim sürecine bizzat katılmak isteyenlerin, ilim metodolojisinin gelişen dallarına âşina olmaları artık bir zaruret halini almaktadır.

Yurdumuzda çok değişkenli analiz henüz geniş bir uygulama alanı bulmamıştır. Fakat araştırıcı ve istatistikcilerimizin bu konuya ileride gereken ağırlığı vermeleri kaçınılmazdır. Elektronik hesap makinalarının yaygın olarak kullanılmasıyla, vektör ve matrislerle ilgili işlem zorlukları da ortadan kalkacaktır. Çalışmamızda çok -değişkenli istatistiğin diskriminant analizi konusunda, metodun teorik dayanakları yanında uygulanışında ortaya çıkan problemlerin yer almasına da bilhassa önem verilmiş böylece araştırıcıların baş vurabilecekleri bir kaynak olmasına çalışılmıştır.

Araştırmamızda değerli teşvik ve yardımlarını esirgemeyen hocam Prof. Dr. Şaban Karataş'a ve mesai arkadaşım Dr. Fatin Sezgin'e teşekkürü bir borç bilirim.

İ Ç İ N D E K İ L E R

Sayfa

1. G İ R İ Ş

Olayların tâbi olduğu kanunları aramayı konu edinen müsbet ilimlerde, metodun önemli bir yeri vardır. Mevcut bilgiler ve bunlara dayanan hipotezlerin kontrolü metodun esas konusunu teşkil eder. Müşahadelerden hareketle uygun bir teori geliştirerek bunun ışığı altında yeni olayların tahmini yapılır. Daha sonra bu tahminlerin ne derece doğru olduğunu anlamak için yeniden müşahadelerde bulunulur.

Güvenilir ve uygun metodların geliştirilmesini konu edinen istatistiğin başlıca üç görevi vardır. Bunlar verilerin özetlenmesi, analiz ve tahmindir. Bütün müşahadeleri tek tek incelemek yerine, bunların tümü hakkında bir fikir verebilecek ortalama, variyans ve çarpıklık gibi ölçüler kullanılarak veriler özetlenir. İstatistik analizlerin gayesi, müşahadelerin yapıldığı değişkenlerin meydana getirdiği örneğin, hangi populasyona ait olduğunu tahmin etmektir. Populasyonlar, kendilerine ait yoğunluk fonksiyonlarıyla tarif edilirler. Bu fonksiyonda bulunan parametreler genellikle bilinmediklerinden tahmin edilmeleri gerekir.

Yakın zamana kadar istatistik analizler, tahminler ve bunlarla ilgili hipotez kontrollerinin sadece bilinmeyen parametrelerin dağılışı hakkındaki faraziyelere dayanarak yapılabileceği sanılıyordu. Bu düşünce muhakkakki akla uygundur. Fakat her zaman için faraziyelerin kesinlikle doğru olduğu söylenemez. Bu sebeple bazı ön bilgilerinde kullanılması gerekmektedir. Tamamlanmış olan denemeler, daha sonraki yollar için bir basamak teşkil eder. Aynı amaca yönelen çeşitli yolları seçmekle değişik bir takım risklere girilmiş olunur. Bu sebeple riski en az olan yolun seçilmesi gerekir. İstatistikte bu riskleri minimum yapan fonksiyonlara karar fonksiyonu adı verilir.

Araştırıcılar populasyonun parametreleri hakkında bilgi edinmek amacıyla seçecekleri örneklere giren fertleri tek bir vasıf bakımından ölçüye vurabilecekleri gibi birçok hallerde aynı şeyi iki ve daha fazla vasıf için de yapabilirler. Meselâ öğrencilerin

üniversiteye girişte aldıkları puanlarla mezun oldukları lisedeki durumları arasındaki ilişkiyi incelemek istiyen bir araştırıcı bunların mezuniyet derecelerine, bölümlerine ve lisenin bulunduğu yere bakabilir. Aynı şekilde, aralarındaki farkın önemini tesbit etmek için iki ayrı muamelenin tatbik edildiği bir bitki çeşidinde yaprak genişliği, gövde uzunluğu ve kalınlığı, kök ağırlığı gibi ölçülerin alınması istenmiş olabilir. Böyle ölçülere konu olan değişkenler arasında genel olarak bir ilgi vardır. Yani değişkenler bağımsız değildir. Bu durumda bir fert üzerinden alınmış çok sayıdaki ölçüleri ayrı ayrı tek değişkenli istatistikte olduğu gibi incelemek doğru değildir. Değişkenler arasında korelasyonların mevcut olması fertlere ait ölçülerin bir bütün olarak ele alınmasını gerektirir.

İstatistiğin, birden fazla değişkenleri konu edinerek bunların analizleriyle uğraşan koluna çok-değişkenli analiz (multivariate analysis) denmektedir.

Memleketimizde henüz geniş çapta bir tatbikat alanı bulamıyan çok-değişkenli istatistikle ilgili bellibaşlı çalışmalar, ilk defa ondokuzuncu yüz yılın ikinci yarısında Francis Galton ile başlamıştır. Bir genetikçi olan Francis Galton, aralarında ilgi bulunan değişken çiftleri üzerinde çalışmıştır. Bağımsız olmayan iki normal değişkene ait birlikte yoğunluk fonksiyonu üzerinde daha önceleri Laplace ve Gauss tarafından çalışılmıştır. Fakat bu araştırıcılar, iki değişken arasındaki ilginin bir ölçüsü olan korelasyon katsayısını ve şartlı dağılışın bir ölçüsü olan regresyon katsayısını keşfedememişlerdi.

Daha sonra Karl Pearson ve diğer araştırıcılar, çeşitli korelasyon katsayılarının teorisini geliştirerek bunların biyoloji, genetik ve diğer bazı alanlarda kullanılmalarını mümkün kıldılar.

Variyans-kovariyans matrisi diye bilinen ve çok-değişkenli istatistikte önemli bir yeri olan şans matrisinin yoğunluk fonksiyonu 1928'de Wishart tarafından elde edilmiştir. Bu önemli aşamadan sonra tek değişkenli istatistiğin, çok-değişkenli analizin sadece özel bir hali olduğu anlaşılmıştır.

Botanik ve antropolojideki sınıflandırma problemleri çok-değişkenli istatistikte, diskriminant analizi diye yeni bir bahsin ortaya çıkmasına yol açmıştır. Diskriminant analizi, çok ölçülü fertlerle bunların ait oldukları populasyonlar arasındaki bağıntıları ayırım ve sınıflandırma yönünden inceler. Tek bir karaktere bakarak ayırım yapmak birçok hallerde imkânsızdır. Bu sebeple araştırıcı mümkün olduğu kadar çok sayıdaki ölçülere dayanan bir kriter geliştirmek zorundadır.

Fertler ait oldukları populasyonlara dahil edilirken, örnek değerleri ile çalışıldığından birtakım hatalar yapılır. Bu hataların sebep olduğu zarara yanlış sınıflandırma maliyeti (cost of miscilasssification) denmektedir. Sınıflandırma işleminde seçilen kriter, bu maliyetlerin minimum yapılmasıyla elde edilir. Buradan da anlaşılacağı üzere fertlerin sınıflandırılmasında takip edilen yol karar fonksiyonlarının bir çeşit uygulamasıdır.

Geniş bir tatbikat sahası olan diskriminant analizinin pratikteki önemi hakkında bir fikir verebilmek için, kullanıldığı yerlerden bazılarılarını misalleriyle sıralıyalım:

a) Ziraat : Biralık arpaların kalitesi danelerdeki bazı maddelerin nisbi miktarlarına bağlıdır. Fakat bu maddelerin miktarlarını fizik veya kimya metodlarıyla tesbiti hem masrafa hemde zaman kaybına sebep olmaktadır. Arpalarda kolaylıkla elde edilebilen sap uzunluğu, gövde kalınlığı ve dane ağırlığı gibi çok sayıdaki ölçülerle kalitelerin tayini mümkündür. Fakat bu ne şekilde olur? Hangi ölçüler kalite tesbitinde en etkili rolü oynamaktadır?

b) Tıp : Arazları birbirine benziyen iki çeşit hastalığın ayırt edilmesinde hastalar üzerinde çok sayıda klinik bulgular kayıt edilir. Bunların hiç biri tek başına hastalıkların tanınmasına kâfi gelmemektedir. Hastalar hakkındaki tıbbi teşhisin daha doğru olabilmesi için bulgular nasıl değerlendirilebilir?

c) Eğitim : Öğrencilerin lisedeki bölümleri, mezuniyet dereceleri ve bunun gibi bilgilerle, üniversiteye giriş imtihanlarında

çeşitli testler sonunda aldıkları puanlar verilmiş olsun. Buna göre fakültelere öğrenci alınırken gözönünde bulundurulacak esas kriter nasıl olmalıdır? Öğrencilerin sadece lisedeki durumlarınamı yoksa üniversiteye giriş imtahanlarında aldıkları puanlaramı bakmalı? Acaba bunların ikisinin kombinasyonu olan yeni bir kriter bulunabilir mi?

d) Taksonomi : Birbirine çok benzeyen fakat başka populasyonlara ait oldukları çok uzun ve masraflı çalışmalardan sonra anlaşılabilen iki böcek tipi düşünelim. Bu böcekler arasından dış görünüş bakımından bir farklılığın olduğu bilinmekte fakat bu, gözle açıkca tesbit edilememektedir. Böcekler üzerinde mümkün olan bütün ölçüler alındıktan sonra bunların ayırımı nasıl yapılır?

e) Teknoloji : Piyasada satılan balları, arıların şekerdenmi yoksa nektardanmı yaptığı çoğu zaman merak edilir. Bilindiği gibi bu iki çeşidi birbirinden ayıran en önemli özellikler; ihtiva ettikleri şeker miktarı, kül nisbeti ve pH dır. Fakat bunların hiç biri ile, tekbaşına kullanıldığında ballar arasında kesin bir ayırım yapılamaz. Ancak hepsinin muayyen bir kombinasyonu belirli bir kriterle mukayese edilirse ayırım mümkün olabilmektedir. Bu kriter nasıl olmalıdır?

f) Sosyoloji : Çeşitli sosyal gruplardaki fertlerle yapılan bir ankette, fertlerin verebilecekleri cevapların mensup oldukları gruplara göre değişmekte olması beklenmektedir. Bilgiler kira, tahsil derecesi ve evde radyonun bulunup bulunmadığı gibi muayyen konulardaki sorulardan elde edilmiştir. Fertlerin mensup oldukları sosyal gruplar bu materyale dayanarak nasıl tesbit edilebilir?

g) Antropoloji : Afrika'da yaşıyan kabilelere ait bazı karakteristik ölçüler verilmiş olsun. Meselâ dil yapısı, kabilenin yağmur tanrısının olup olmadığı, ziraatla olan ilişkileri gibi. Bu bilgilere dayanarak, ele alınan başka bir kabilenin önceki etnik gruplardan birine ait olup olmadığının bilinmesi istenebilir. Bunun için nasıl bir kriter kullanmak gerekir?

Diskriminant analizi tekniklerini kullanarak yukarıdaki soruların cevapları verilebilir.

Bu araştırmada, diskriminant analizin bazılarına yukarda kısaca işaret edilen uygulama alanlarında kullanılışına yardım edecek teorik esasları ortaya koymak ve bu esasların uygulanışını göstermek amacıyla plânlanmıştır. Araştırmamızda konu teorik dayanaklarıyla oldukça geniş bir literatür özeti içinde ele alınmış, evvelce yapılan çalışmalar son gelişmelerin ışığı altında takdim edilmiştir. Diskriminant analizi uygulıyacak olanların anlamakta güçlük çekecek esasları bu bölümde rahatça bulacaklarını ümit etmekteyiz.

Çalışmamızda diskriminant analizinin genel teorisi incelenerek değişkenlerin seçimi, hata nisbetleri ve kayıp müşahadelerin tahmini gibi bazı problemler üzerinde durulmuştur.

Literatür özetine ayrılan ikinci bölümde diskriminant analizi hakkında genel bilgi verildikten sonra sınıflandırmada kullanılan kriterin seçimine ait esaslar izah edilmiştir. Çok-değişkenli normal populasyonlarda a priori ihtimallerin bilindiği ve bilinmediği haller için diskriminant fonksiyonu bulunarak sınıflandırmanın nasıl yapılacağı gösterilmiş, ayrıca kalitatif verilerle yapılan sınıflandırma hakkında bazı açıklamalarda bulunulmuştur.

Araştırmada kullanılan veriler iki kısımda toplanabilir. Bunlardan birincisi, Atatürk Üniversitesindeki öğrenci kayıtlarından, diğeri ise normal dağılış gösteren şans sayılarından elde edilmiştir.

Araştırmalar ve sonuçları diye adlandırdığımız dördüncü bölüm, diskriminant analizi ile ilgili bazı problemlere ayrılmıştır. İlk olarak yanlış sınıflandırma ihtimallerini hesaplamada kullanılan beş ayrı metodun mukayesesi yapılarak, en iyi olanını tesbitine çalışılmıştır. İkinci çalışma, kalitatif verilerle sınıflandırmanın nasıl yapılacağı hakkındadır. Bu çalışmada ayrıca normal değişkenlerin kalitatif hale getirilmeleri sonunda ayırım gücünde meydana gelen değişiklik incelenmiştir. Bu bölümün bahislerinden biri de

kayıp müşahadeler üzerinedir. Kaybolmuş müşahadelerin tahmini için ileri sürülen variyans-kovariyans matrisinin izini ve determinantını kriter olarak alan iki metodun mukayesesi yapılmıştır. Son olarak, diskriminant fonksiyonu için değişkenlerin nasıl seçileceği izah edilmiştir.

2. LİTERATÜR ÖZETİ

2.1. Diskriminant Analizinin Tarihçesi

Diskriminant analizi fikri çok eskidir. Ondokuzuncu yüzyılda bu konuda başlıyan çalışmalar Karl Pearson'un yaptığı araştırmalarla hız kazanmıştır. Takriben 1920 sıralarında antropometrik verileri inceliyerek iki populasyon arasındaki mesafe için bir katsayı bulmuştur (Kendall, 1957).

Diğer taraftan Hintli bir istatistikçi olan Mahalanobis (1948) iki populasyon arasındaki mesafe için daha değişik bir ölçü ileri sürmüştür. D^2 diye adlandırdığı bu ölçüyü ilk defa 1925 te Bengal eyaletindeki çeşitli ırkların ayırımı için tatbik etmişti. Son zamanlarda birhayli geliştirilen diskriminant analizinde, çok önemli bir yeri olan D^2 istatistiği, iki populasyon arasındaki standardize edilmiş mesafenin karesine tekabül eder. Variyans-kovariyans matrisleri ($\underline{V}$) eşit ve ortalama vektörleri $\underline{u}^{(1)}$ ve $\underline{u}^{(2)}$ olan çok-değişkenli iki populasyon için "Mahalanobis'in genelleştirilmiş mesafesi"

$$\Delta^2 = (\underline{u}^{(1)} - \underline{u}^{(2)})' \, \underline{V}^{-1} \, (\underline{u}^{(1)} - \underline{u}^{(2)}) \quad * \qquad (2.1.1)$$

şeklinde yazılabilir. Bu bağıntı örnek tahminleri ile

$$D^2 = (\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)})' \, \underline{S}^{-1} \, (\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)}) \qquad (2.1.2)$$

dir.

Aynı tarihlerde Hotelling (1931), Student'in t- dağılışını çok-değişkenli istatistik için genelleştirerek

$$T^2 = \frac{D^2 \, n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 2)} \qquad (2.1.3)$$

bağıntısını bulmuştur. Eşitlikteki n_1 ve n_2 her iki örnekteki fert sayılarını göstermektedir.

---

* Altı çizilen harfler vektör veya matrisi, çizilmeyenler skalar'ı göstermektedir.

Fisher çoklu korelasyonun dağılışını bulduktan sonra bunun T^2 ve D^2 ninkilere benzediğini göstermiştir.

1936'da Fisher diskriminant fonksiyonu hakkındaki ilk makalesini neşretti. İleri sürülen fonksiyon, örneklerin teşkil ettiği yüzeyi ikiye bölerek, ele alınan fertlerin ilgili populasyonlara dahil edilmesinde kullanılıyordu. Fisher'in bu yaklaşımı bugünkü karar fonksiyonlarının modern teorisine çok yakındır. Fisher'in fonksiyonu halen diskriminant analizinde geniş çapta kullanılmaktadır.

2.2. Genel Bilgiler

Çok-değişkenli ölçülere vurulan fertleri sınıflandırma bahis konusu olduğu zaman, yapılan müşahadelerin tümü esas alınarak bir kriterin seçilmesi gerekir. Sınıflandırmada en önemli husus fertleri hata yapmadan mensup oldukları populasyonlara dahil etmek olduğundan, yanlış sınıflandırma ihtimallerinin en uygun kriter olduğu düşünülebilir. Buna göre herhangibir fert için ölçüleri esas alınarak yapılan sınıflandırmada yanlış bir populasyona dahil edilme ihtimalinin minimum olması gerekir.

Konunun daha iyi izah edilebilmesi bakımından, önce sadece iki populasyonun mevcut olduğu farzedilecek, daha sonra bazı haller için teori genelleştirilecektir.

Birinci populasyonu Π_1, ikinci populasyonu da Π_2 ile gösterelim. $\underline{X}' = (X_1 \ldots X_p)$ müşahadelerinin yapıldığı bir fert Π_1 veya Π_2 populasyonlarından birine ait olsun. Bu ferdi mensup olduğu populasyona mümkün olduğu kadar az bir hata ile sınıflandırabilmek için $X_1 \ldots X_p$ ölçülerinin muayyen bir kombinasyonunun değeri, seçilen optimum kriterle mukayese edilir. Meselâ ilgili değer kriterin üzerinde ise fert birinci, altında ise ikinci populasyona sınıflandırılır.

Kriterin tabiatı ne olursa olsun, örnekleme yapıldığından bazı fertler sınıflandırma esnasında mensup olmadıkları populasyonlara dahil edilebilirler. Bu gibi hatalar iki grupta incelenebilir:

a) π_1 populasyonuna ait bir ferdin yanlışlıkla π_2'ye sınıflandırılması.

b) π_2 populasyonuna ait bir ferdin yanlışlıkla π_1'e sınıflandırılması.

Yanlış sınıflandırmadan dolayı meydana gelen zarar, yukarıda izah edilen iki tip hata için de her zaman aynı değildir. "Yanlış sınıflandırma maliyeti"olarak tabir edeceğimiz bu zararları aşağıdaki sembollerle gösterelim :

c_{21} : Birinci tip yanlış sınıflandırmanın maliyeti.

c_{12} : İkinci tip yanlış sınıflandırmanın maliyeti.

İleride de bahsedileceği gibi c_{21} ve c_{12} nin seçilen birimleri ne olursa olsun, nisbetleri alındığından optimum kriterin değerini etkilemez.

Sınıflandırmadaki hata ve bunların maliyetlerinden başka kriterin seçiminde gözönünde bulundurulması gereken en önemli husus populasyonu meydana getiren fertlerin nisbî frekansları veya a priori ihtimalleridir. Bu ihtimallerin bilinip bilinmemelerine göre sınıflandırma kriteri farklı olur.

2.3. Sınıflandırmada Kullanılan Kriterin Seçimi

Anderson (1958), bundan önce belirtilen sebeplerden dolayı, kriterin seçiminde esas teşkil edecek kaidelerin nasıl olabileceğini a priori ihtimallerin bilindiği ve bilinmediği haller için ayrı ayrı incelemiştir. Bunlar aşağıda sırasıyla açıklanacaktır.

2.3.1. A priori ihtimallerin bilindiği haller

Fertlerin π_1 populasyonundan gelmiş olma ihtimali q_1 ve π_2 den gelmiş olma ihtimali de q_2 olsun. Dağılışın bir yoğunluk fonksiyonunun olduğunu farzederek π_1 populasyonununkini $f_1(\underline{X})$ ve π_2'ninkini de $f_2(\underline{X})$, $\underline{X}' = (X_1 \ldots X_p)$ ile gösterelim. Ayrıca bu populasyonlara ait sınıflandırma bölgeleri sırasıyla R_1 ve R_2 olsun. Buna göre aslında π_1 populasyonuna mensup olan bir ferdin gene π_1'e doğru olarak sınıflandırma ihtimali

$$P(\pi_1|\pi_1,R) = \int_{R_1} f_1(\underline{X})\, d\underline{X}, \quad d\underline{X} = dX_1 \ldots dX_p \qquad (2.3.1)$$

şeklindedir. Aynı ferdin π_1'den, yanlış sınıflandırılarak, π_2'ye dahil edilmesi ihtimali

$$P(\pi_2|\pi_1, R) = \int_{R_2} f_1(\underline{X})\, d\underline{X} = 1 - \int_{R_1} f_1(\underline{X})\, d\underline{X} \qquad (2.3.2)$$

olarak bulunur.

Benzer yoldan π_2'ye ait olan bir ferdin doğru olarak sınıflandırılma ihtimali

$$P(\pi_2|\pi_2,R) = \int_{R_2} f_2(\underline{X})\, d\underline{X} \qquad (2.3.3)$$

ve yanlış sınıflandırma ihtimali de

$$P(\pi_1|\pi_2,R) = \int_{R_1} f_2(\underline{X})\, d\underline{X} = 1 - \int_{R_2} f_2(\underline{X})\, d\underline{X} \qquad (2.3.4)$$

olarak bulunur.

π_1 populasyonundan bir fert çekme ihtimali q_1 idi. Onun için π_1'den bir fert alıp bunu doğru bir şekilde sınıflandırma ihtimali $q_1P(\pi_1|\pi_1,R)$ olur. Yanlış sınıflandırma ihtimali ise $q_1P(\pi_2|\pi_1,R)$ dir. Aynı şekilde π_2'den alınan bir ferdin doğru olarak sınıflandırılma ihtimali $q_2P(\pi_2|\pi_2,R)$ ve yanlış sınıflandırma ihtimali de $q_2P(\pi_1|\pi_2,R)$'dir.

Şimdi yanlış sınıflandırma maliyetinin beklenen değerini bulalım. Bu maliyetin beklenen değeri her iki populasyon için yanlış sınıflandırma ihtimalleri, malıyetleri ve a priori ihtimalleri

çarpılarak toplanır. Bunu bir formülle ifade edersek

$$\varphi(\underline{X}) = c_{21}P(\pi_2|\pi_1,R)q_1 + c_{12}P(\pi_1|\pi_2,R)q_2 \qquad (2.3.5)$$

bulunur.

R_1 ve R_2 bölgelerini kesin olarak belirtebilmek için yukarıdaki fonksiyonu minimum yaparak optimum kriter bulunur. Verilen q_1 ve q_2 değerleri için $\varphi(\underline{X})$'i minimum yapmada takip edilecek yola Bayes usulü denir.

Müşahade edilmiş değişkenler verildiğine göre, bir ferdin π_1 populasyonundan gelmiş olmasının şartlı ihtimali

$$\frac{q_1 f_1(\underline{X})}{q_1 f_1(\underline{X}) + q_2 f_2(\underline{X})} \qquad (2.3.6)$$

ve π_2 den gelmiş olmasının şartlı ihtimali

$$\frac{q_2 f_2(\underline{X})}{q_1 f_1(\underline{X}) + q_2 f_2(\underline{X})} \qquad (2.3.7)$$

şeklindedir.

Yanlış sınıflandırma maliyetlerinin eşit olması halinde R_1 ve R_2 bölgelerini tayin etmek için

$$\varphi(\underline{X}) = q_1 \int_{R_2} f_1(\underline{X})\, d\underline{X} + q_2 \int_{R_1} f_2(\underline{X})\, d\underline{X} \qquad (2.3.8)$$

ifadesi minimum yapılır. Kolayca anlaşılacağı gibi fertler şartlı ihtimallerinin daha yüksek olduğu populasyonlara sınıflandırıldığında yukarıdaki ifade minimum yapılmış olur. Yani

$$\frac{q_1 f_1(\underline{X})}{q_1 f_1(\underline{X}) + q_2 f_2(\underline{X})} \geq \frac{q_2 f_2(\underline{X})}{q_1 f_1(\underline{X}) + q_2 f_2(\underline{X})} \qquad (2.3.9)$$

ise π_1 populasyonu,

$$\frac{q_1 f_1(\underline{X})}{q_1 f_1(\underline{X}) + q_2 f_2(\underline{X})} < \frac{q_2 f_2(\underline{X})}{q_1 f_1(\underline{X}) + q_2 f_2(\underline{X})} \qquad (2.3.10)$$

ise π_2 populasyonu seçilir. Bu eşitsizliklere göre R_1 ve R_2 bölgeleri aşağıdaki gibi tarif edilir.

$$R_1 : q_1 f_1(\underline{X}) \geq q_2 f_2(\underline{X})$$
$$R_2 : q_1 f_1(\underline{X}) < q_2 f_2(\underline{X}) \qquad (2.3.11)$$

$q_1 f_1(\underline{X}) = q_2 f_2(\underline{X})$ olduğunda fertler π_1 veya π_2 ye rastgele sınıflandırılırlar.

Yanlış sınıflandırma maliyetleri dikkate alındığı taktirde $\Omega(\underline{X})$ gene aynı şekilde bulunur :

$$\Omega(\underline{X}) = c_{21} q_1 \int_{R_2} f_1(\underline{X}) d\underline{X} + c_{12} q_2 \int_{R_1} f_2(\underline{X}) d\underline{X} \qquad (2.3.12)$$

şeklindedir. Buna göre $\Omega(\underline{X})$'i minimum yapan bölgeler aşağıdaki gibidir.

$$R_1 : c_{21} q_1 f_1(\underline{X}) \geq c_{12} q_2 f_2(\underline{X})$$
$$R_2 : c_{21} q_1 f_1(\underline{X}) < c_{12} q_2 f_2(\underline{X}) \qquad (2.3.13)$$

2.3.2. A priori ihtimallerin bilinmediği haller

Birçok durumlarda a priori bilgiye sahip olunmadığından bunlarla ilgili ihtimalleri kullanmak mümkün değildir. Böyle hallerde fertleri yanlış sınıflandırma ihtimallerinin iki populasyonda da eşit olduğu farzedilir (Kendall ve Stuart, 1969). Yani

$$\int_{R_1} f_2(\underline{X}) d\underline{X} = \int_{R_2} f_1(\underline{X}) d\underline{X} \qquad (2.3.14)$$
$$= 1 - \int_{R_1} f_1(\underline{X}) d\underline{X}$$

olarak bulunur. Buradan

$$\int_{R_1} \left[f_2(\underline{X}) + f_1(\underline{X}) \right] d\underline{X} = 1 \qquad (2.3.15)$$

yazılabilir. Böylece bu şartın kullanılarak $\int_{R_1} f_2(\underline{X}) d\underline{X}$ in minimum

yapılması gerekir. Yanlış sınıflandırma maliyeti dikkate alındığında, R_1 ve R_2 bölgeleri

$$\int_{R_1} \left[f_2(\underline{X}) - \alpha (f_1(\underline{X}) + f_2(\underline{X})) \right] d\underline{X} \qquad (2.3.16)$$

ifadesi minimum yapılmak suretiyle tarif edilir. Yukarıdaki α bir sabiteyi göstermektedir. Bu ifade biraz daha sadeleştirilirse

$$\int_{R_1} \left[\beta f_2(\underline{X}) - f_1(\underline{X}) \right] d\underline{X} \qquad (2.3.17)$$

elde edilir. β bir sabitedir. Yukarıdaki ifade yanlış sınıflandırma ihtimalleri arasındaki farkı göstermektedir. Buna göre

$$\beta f_2(\underline{X}) - f_1(\underline{X}) \leq 0 \qquad (2.3.18)$$

şartını yerine getiren noktalar R_1 bölgesine dahil edilirse bu ifade minimum yapılmış olur. O halde R_1 ve R_2 bölgeleri aşağıdaki gibi seçilir :

$$R_1 : \frac{f_1(\underline{X})}{f_2(\underline{X})} \geq \beta$$

$$R_2 : \frac{f_1(\underline{X})}{f_2(\underline{X})} < \beta$$

$\frac{f_1(\underline{X})}{f_2(\underline{X})} = \beta$ olduğu zaman fertlerin sınıflandırılmaları, bundan önceki halde olduğu gibi rastgele yapılır.

Eğer örnek yeteri kadar büyük ise sınıflandırmanın belirli bir safhasından sonra fertlerin π_1 veya π_2 populasyonlarından gelme ihtimalleri tahmin edilebilir. Bu ihtimaller daha sonraki denemeler için a priori ihtimaller olarak alınırlarsa, sınıflandırmada kullanılacak kriterin doğruluk derecesi daha da artırılmış olur.

Genel bir sınıflandırma işleminde seçilecek kriterin elde edilişinin ana hatlarının böylece izah ettikten sonra bu, çok yaygın olan iki hal için kullanılacaktır. Bunlardan biri kantitatif ölçülerin bahiskonusu olduğu çok-değişkenli normal dağılış gösteren populasyonlar, diğeri ise kalitatif ölçülerin bahiskonusu olduğu çok-değişkenli kesikli dağılış gösteren populasyonlardır.

2.4. Çok-Değişkenli Normal Populasyonlarda Sınıflandırma

Çok-değişkenli normal populasyonlardaki sınıflandırma için yukarıda belirtilen usuller kullanılacaktır. Teori, önce iki populasyonun mevcut olduğu haller için izah edilecek, daha sonra da ikiden fazla haller için genelleştirilecektir.

2.4.1. Çok değişkenli iki normal populasyonda sınıflandırma

Anderson (1958), fertlerin sınıflandırılmalarında kullanılan diskriminant fonksiyonunun çıkarılışını göstermiştir. Bu konuda ayrıca Kendall (1957) ve Rao (1965) bazı açıklamalarda bulunmuştur.

Her iki populasyonda da çok-değişkenli normal dağılış gösteren p-elemanlı bir $\underline{X}$ vektörü olsun. Bunun birinci ve ikinci populasyondaki yoğunluk fonksiyonu

$$f_1(\underline{X}) = \frac{1}{(2\pi)^{p/2}|V_1|^{1/2}} \exp\left[-\frac{1}{2}(\underline{X}-\underline{u}^{(1)})'\underline{V}_1^{-1}(\underline{X}-\underline{u}^{(1)})\right]$$

$$f_2(\underline{X}) = \frac{1}{(2\pi)^{p/2}|V_2|^{1/2}} \exp\left[-\frac{1}{2}(\underline{X}-\underline{u}^{(2)})'\underline{V}_2^{-1}(\underline{X}-\underline{u}^{(2)})\right] \qquad (2.4.1)$$

şeklinde yazılabilir. $\underline{V}_1$ ve $\underline{V}_2$ variyans-kovariyans matrisleridir.

(2.3.13) deki eşitsizlikler çözülüp, her iki tarafın logaritması alınırsa R_1 ve R_2 bölgeleri aşağıdaki gibi bulunur:

$$R_1 : \ln \frac{f_1(\underline{X})}{f_2(\underline{X})} \geq \ln \frac{c_{12}q_2}{c_{21}q_1}$$

$$R_2 : \ln \frac{f_1(\underline{X})}{f_2(\underline{X})} < \ln \frac{c_{12}q_2}{c_{21}q_1} \qquad (2.4.2)$$

(2.4.1) deki değerler kullanılırsa

$$\ln \frac{f_1(\underline{X})}{f_2(\underline{X})} = \frac{1}{2}\ln \frac{|\underline{V}_2|}{|\underline{V}_1|} - \frac{1}{2}\left[(\underline{X}-\underline{u}^{(1)})'\underline{V}_1^{-1}(\underline{X}-\underline{u}^{(1)})\right]$$

$$+ \frac{1}{2}\left[(\underline{X}-\underline{u}^{(2)})'\underline{V}_2^{-1}(\underline{X}-\underline{u}^{(2)})\right] \qquad (2.4.3)$$

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

eşitsizliği elde edilir. Dikkat edilirse eşitsizliğin sol tarafı ikinci dereceden p-değişkenli bir polinomdur. Bu polinom bir nevi diskriminant fonksiyonudur. Pratikte ayırım veya sınıflandırmada kullanılan bu gibi fonksiyonların mümkün olduğu kadar basit olmaları arzu edilir. İki populasyonunda variyans-kovariyans matrislerinin eşit yani $\underline{V}_1 = \underline{V}_2 = \underline{V}$ olduğu kabul edilirse (2.4.3) deki eşitsizlik aşağıdaki şekli alır.

$$-\frac{1}{2}\left[(\underline{X}-\underline{u}^{(1)})'\underline{V}^{-1}(\underline{X}-\underline{u}^{(1)}) - (\underline{X}-\underline{u}^{(2)})'\underline{V}^{-1}(\underline{X}-\underline{u}^{(2)})\right] \geq \ln k \qquad (2.4.4)$$

Eşitsizlikteki $k = c_{12}q_2/c_{21}q_1$ dir. Bunun sol tarafı açılıp sadeleştirilirse ikinci dereceden olan terimler gider ve (2.4.4) aşağıdaki son şeklini alır.

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

$$\geq \ln k \qquad (2.4.5)$$

Eşitsizliğin sol tarafındaki fonksiyona Anderson'un sınıflandırma istatistiği denir. Bu Fisher'in diskriminant fonksiyonu olan

$$D(\underline{X}) = \underline{X}'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) \qquad (2.4.6)$$

den sadece ihtiva ettiği bir sabit terim bakımından farklıdır.

Böylece yoğunluk fonksiyonları $f_1(\underline{X})$ ve $f_2(\underline{X})$ olan Π_1 ve Π_2 populasyonları için yapılan sınıflandırmada seçilecek en iyi R_1 ve R_2 bölgeleri

$$R_1: \underline{X}'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) - \frac{1}{2}(\underline{u}^{(1)} + \underline{u}^{(2)})'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) \geq \ln k$$

$$R_2: \underline{X}'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) - \frac{1}{2}(\underline{u}^{(1)} + \underline{u}^{(2)})'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) < \ln k \qquad (2.4.7)$$

şeklinde olmalıdır.

$q_1=q_2$ ve $c_{12}=c_{21}$ olduğu zaman $k= 1$, dolayısiyle $\ln 1 = 0$ olur. O zaman (2.4.7) deki R_1 ve R_2 bölgeleri

$$R_1: \underline{X}'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) - \frac{1}{2}(\underline{u}^{(1)} + \underline{u}^{(2)})'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) \geq 0$$

$$R_2: \underline{X}'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) - \frac{1}{2}(\underline{u}^{(1)} + \underline{u}^{(2)})'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) < 0 \qquad (2.4.8)$$

şeklini alır. Pratikte ençok son olarak bulunan bu diskriminant fonksiyonu kullanılır.

Çok-değişkenli bir populasyondaki herhangibir fertte ölçülen $\underline{x}$ şans vektörünü ele alarak a priori ihtimallerin bilinmediği haller için R_1 ve R_2 bölgelerini bulalım. $\underline{x}$, Fisher'in diskriminant fonksiyonunda yerine konduğu zaman

$$D(\underline{x}) = \underline{x}'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) \qquad (2.4.9)$$

bulunur. $D(\underline{x})$ in beklenen değeri $\underline{x}$ in mensub olduğu populasyona göre değişir. Bu ortalamaları Π_1 ve Π_2 de sırasiyle A_1 ve A_2 ile

gösterelim.

$$A_1 = E\left[D(\underline{x}|\pi_1)\right] = \underline{u}^{(1)'}\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)})$$

$$A_2 = E\left[D(\underline{x}|\pi_2)\right] = \underline{u}^{(2)'}\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) \qquad (2.4.10)$$

bulunur. Variyansı ise

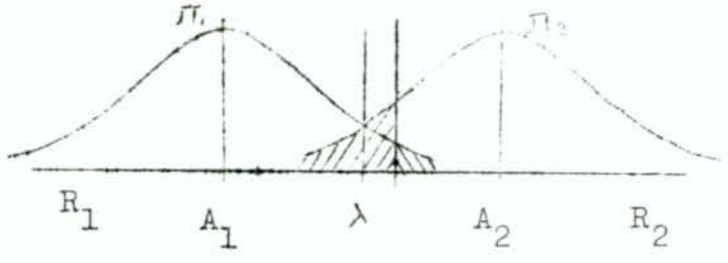
$$\begin{aligned}\text{Var}\left[D(\underline{x}|\pi_1)\right] &= E\Big\{\left[\underline{x}'\ \underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) - \underline{u}^{(1)'}\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)})\right]' \\ &\quad \left[\underline{x}'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) - \underline{u}^{(1)'}\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)})\right]\Big\} \\ &= (\underline{u}^{(1)} - \underline{u}^{(2)})'\underline{V}^{-1'}E\left[(\underline{x} - \underline{u}^{(1)})(\underline{x} - \underline{u}^{(1)})'\right]\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) \\ &= (\underline{u}^{(1)} - \underline{u}^{(2)})'\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) \qquad (2.4.11)\end{aligned}$$

şeklinde elde edilir. $\underline{x}$ in variyans-kovariyans matrisi iki populasyonda da aynı olduğundan

$$\text{Var}\left[D(\underline{x}|\pi_2)\right] = \text{Var}\left[D(\underline{x}|\pi_1)\right] = \text{Var}\left[D(\underline{x})\right] \qquad (2.4.12)$$

dir.

Özetliyecek olursak örnekteki fert sayısı çok büyükse $D(\underline{x})$, π_1 populasyonunda ortalaması A_1, variyansı $\text{Var}\left[D(\underline{x}|\pi_1)\right]$ ve π_2 de ortalaması A_2, variyansı $\text{Var}\left[D(\underline{x}|\pi_2)\right]$ olan normal dağılış gösterir. Durum aşağıdaki şekilde olduğu gibidir.



Şekil.2.1. π_1 ve π_2 Populasyonlarına Ait R_1 ve R_2 bölgeleri

R_1 ve R_2 bölgelerini birbirinden ayıran sınır noktası λ olsun. Yanlış sınıflandırma ihtimalleri aşağıdaki gibi bulunur :

$$P(\pi_1 | \pi_2, R) = \int_{-\infty}^{\lambda} \frac{1}{\sqrt{2\pi \, Var[D(\underline{x})]}} \exp\left\{ - \frac{[D(\underline{x}) - A_2]^2}{2Var[D(\underline{x})]} \right\} dD(\underline{x})$$

$$= \Phi\left(\frac{\lambda - A_2}{\sqrt{Var[D(\underline{x})]}} \right)$$

$$P(\pi_2 | \pi_1, R) = \int_{\lambda}^{\infty} \frac{1}{\sqrt{2\pi \, Var[D(\underline{x})]}} \exp\left\{ - \frac{[D(\underline{x}) - A_1]^2}{2 \, Var[D(\underline{x})]} \right\} dD(\underline{x})$$

$$= 1 - \Phi\left(\frac{\lambda - A_1}{\sqrt{Var[D(\underline{x})]}} \right) \qquad (2.4.13)$$

λ yı hesaplamak için

$$c_{12} \, P(\pi_1 | \pi_2, R) = c_{21} \, P(\pi_2 | \pi_1, R)$$

eşitliği çözülür. Buna göre bilinen c_{12} ve c_{21} değerleri için

$$c_{12} \Phi\left(\frac{\lambda - A_2}{\sqrt{Var[D(\underline{x})]}} \right) = c_{21} \left[1 - \Phi\left(\frac{\lambda - A_1}{\sqrt{Var[D(\underline{x})]}} \right) \right] \qquad (2.4.14)$$

eşitliğini sağlıyan λ değeri standart normal dağılış cetvellerinden hesaplanabilir. Dikkat edilirse $c_{12}=c_{21}$ olduğu zaman yukarıdaki eşitlik

$$\Phi\left(\frac{\lambda - A_2}{\sqrt{Var[D(\underline{x})]}} \right) = 1 - \Phi\left(\frac{\lambda - A_1}{\sqrt{Var[D(\underline{x})]}} \right) \qquad (2.4.15)$$

şeklini alır. Standart normal dağılış eğrisinin simetri özelliğinden faydalanarak

$$\frac{\lambda - A_2}{\sqrt{\text{Var}\,[\,D(\underline{x})\,]}} = -\frac{\lambda - A_1}{\sqrt{\text{Var}\,[\,D(\underline{x})\,]}}$$

$$\lambda = \frac{1}{2}\,(A_1 + A_2) \qquad (2.4.16)$$

bulunur. Böylece yanlış sınıflandırma maliyetleri eşit olduğu zaman R_1 ve R_2 yi birbirinden ayıran sınır noktası Π_1 ve Π_2 populasyonlarına ait $D(\underline{x})$ in ortalamalarının orta yerinde olur.

2.4.2. Diskriminant fonksiyonunun başka bir yoldan çıkarılışı

Lineer diskriminant fonksiyonunun

$$D(\underline{X}) = \underline{X}'\,\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)})$$

olduğu yukarıda belirtilmişti. Bu fonksiyon aşağıdaki ifadenin maksimum yapılması ile de elde edilebilir(Fisher, 1936).

$$\frac{\left[E(\underline{X}'\,\underline{d}\,|\Pi_1) - E(\underline{X}'\,\underline{d}\,|\Pi_2)\right]^2}{\text{Var}\,(\underline{X}'\,\underline{d})} \qquad (2.4.17)$$

İfadedeki $\underline{d}$, herhangibir vektördür. Görüldüğü gibi (2.4.17), lineer kombinasyonu alınan $\underline{X}$ in birinci ve ikinci populasyonlardaki ortalamaları arasındaki farkın karesinin variyansına oranıdır. Şimdi bu ifadeyi maksimum yapan $\underline{d}$ yi bulalım.

$$\left[E(\underline{X}'\,\underline{d}\,|\Pi_1) - E(\underline{X}'\,\underline{d}\,|\Pi_2)\right]^2 = (\underline{u}^{(1)'}\,\underline{d} - \underline{u}^{(2)'}\,\underline{d})^2$$

$$= \underline{d}'\,(\underline{u}^{(1)} - \underline{u}^{(2)})'\,(\underline{u}^{(1)} - \underline{u}^{(2)})'\,\underline{d}$$

$$\text{Var}(\underline{X}'\underline{d}) = E\left[(\underline{X}'\underline{d}) - \underline{u}^{(1)'}\,\underline{d}\right]\left[(\underline{X}'\underline{d}) - \underline{u}^{(1)'}\,\underline{d}\right]$$

$$= \underline{d}'\,\underline{V}\,\underline{d} \qquad (2.4.18)$$

eşitlikleri yazılabilir. (2.4.17) yi maksimum yapmak için payda sabit tutulur. $\underline{d}'\underline{V}\,\underline{d} = 1$ şartı ile Lagrange çarpanları metodu kullanılırsa

$$\psi = \underline{d}' \left[(\underline{u}^{(1)} - \underline{u}^{(2)})(\underline{u}^{(1)} - \underline{u}^{(2)})' \right] \underline{d} + \lambda(\underline{d}'\,\underline{V}\,\underline{d} - 1) \tag{2.4.19}$$

ifadesinin $\underline{d}$ ye göre türevini alıp sıfıra eşitlemek gerekir. Böylece

$$\frac{\partial \psi}{\partial \underline{d}} = 2 \left[(\underline{u}^{(1)} - \underline{u}^{(2)})(\underline{u}^{(1)} - \underline{u}^{(2)})' \right] \underline{d} + 2\lambda\,\underline{V}\,\underline{d} = 0 \tag{2.4.20}$$

bulunur. $(\underline{u}^{(1)} - \underline{u}^{(2)})'\underline{d}$ ifadesi v gibi bir sabite olduğundan

$$(\underline{u}^{(1)} - \underline{u}^{(2)}) = -\frac{\lambda}{v}\,\underline{V}\,\underline{d} \tag{2.4.21}$$

elde edilir. O halde $\underline{d}$ vektörü $\underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)})$ ye orantılıdır. Dolayısıyle $\underline{X}'\underline{d}$ de $D(\underline{X})$ e orantılı olur.

2.4.3. Parametrelerin tahmini

Şimdiye kadar populasyonun parametreleri olan $\underline{u}^{(1)}$, $\underline{u}^{(2)}$ ve $\underline{V}$ nin bilindiği farzedilmişti. Pratikte çoğu zaman bu değerler bilinmediklerinden örnek tahminlerinin kullanılması gerekir. Her populasyondan alınan belirli genişlikteki örnek değerlerine göre yukarıda bahsedilen fonksiyon hesaplanarak daha sonra elde edilecek fertlerin sınıflandırılmalarında kullanılır. Bunun için kabul edilen en önemli faraziye fonksiyonun hesaplanmasında kullanılan fertlerin hangi populasyondan geldiklerinin hatasız olarak bilinmesidir.

π_1 populasyonundan $\underline{X}_1, \ldots, \underline{X}_{n_1}$ ve π_2 den de $\underline{X}_{n_1+1}, \ldots, \underline{X}_{n_1+n_2}$ müşahadeleri yapılmış olsun. Doğru olduğu kabul edilen bu bilgilere dayanarak, hangi populasyona ait olduğu bilinmeyen $\underline{X}_{n_1+n_2+1}$ ferdi sınıflandırılır. Bunun için (2.4.5) deki sınıflandırma istatistiği veya (2.4.6) daki Fisher'in diskriminant fonksiyonu kullanılır.

$$\hat{\underline{u}}^{(1)} = \bar{\underline{X}}^{(1)} = \frac{1}{n_1} \sum_{\alpha=1}^{n_1} \underline{X}_\alpha$$

$$\hat{\underline{u}}^{(2)} = \bar{\underline{X}}^{(2)} = \frac{1}{n_2} \sum_{\alpha=n_1+1}^{n_1+n_2} \underline{X}_\alpha .$$

$$\hat{\underline{V}} = \underline{S} = \frac{1}{n_1+n_2-2} \left[\sum_{\alpha=1}^{n_1} (\underline{X}_\alpha - \bar{\underline{X}}^{(1)})(\underline{X}_\alpha - \bar{\underline{X}}^{(1)})' + \sum_{\alpha=n_1+1}^{n_1+n_2} (\underline{X}_\alpha - \bar{\underline{X}}^{(2)})(\underline{X}_\alpha - \bar{\underline{X}}^{(2)})' \right] \qquad (2.4.22)$$

yazılırsa $W(\underline{X})$ in örnek tahmini olan $w(\underline{X})$,

$$w(\underline{X}) = \underline{X}'_{n_1+n_2+1} \underline{S}^{-1}(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) - \frac{1}{2}(\bar{\underline{X}}^{(1)} + \bar{\underline{X}}^{(2)})' \underline{S}^{-1}(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) \qquad (2.4.23)$$

şeklinde hesaplanır (Sitgreaves, 1952).

Bilindiği üzere, yukarıdaki $\bar{\underline{X}}^{(1)}$, $\bar{\underline{X}}^{(2)}$ ve $\underline{S}$ nin beklenen değerleri sırasıyla $\underline{u}^{(1)}$, $\underline{u}^{(2)}$ ve $\underline{V}$ yi verirler. Bu tahminler kullanılarak elde edilen $w(\underline{X})$ şüphesiz $W(\underline{X})$ in sapmasız (unbiased) tahmini değildir. y ve z gibi iki değişken bağımsız olmadıkça genellikle

$$E(y.z) \neq [E(y)][E(z)] \qquad (2.4.24)$$

bağıntısı vardır. Aynı şekilde iki değişkenin çarpımı gibi terimleri ihtiva eden $w(\underline{X})$ in beklenen değeri $W(\underline{X})$ e eşit olmaz. Ancak, parametrelerin bilindiği hal için en iyi kriter olarak bulunan $W(\underline{X})$ istatistiğinin bu şartlar altında yaklaşık tahmini $\bar{\underline{X}}^{(1)}$, $\bar{\underline{X}}^{(2)}$

ve $\underline{S}$ yi kullanmak suretiyle bulunur.

2.4.4. $D(\underline{X})$ in örnek dağılışı

$D(\underline{X})$ in $W(\underline{X})$ den sadece sabit bir terim bakımından farklı olduğu belirtilmişti. $W(\underline{X})$ de olduğu gibi $D(\underline{X})$ in örnek tahmini

$$d(\underline{X}) = \underline{X}'_{n_1+n_2+1}\ \underline{S}^{-1}(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) \qquad (2.4.25)$$

olarak bulunur.

$$\underline{A} = (n_1+n_2-2)\underline{S}$$

$$\underline{Z} = \left(\sqrt{\frac{n_1 n_2}{n_1+n_2}}\right)(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) \qquad (2.4.26)$$

yazılırsa

$$d(\underline{X}) = \frac{(n_1+n_2-2)\sqrt{n_1+n_2}}{\sqrt{n_1 n_2}}\ \underline{X}'_{n_1+n_2+1}\ \underline{A}^{-1}\ \underline{Z} \qquad (2.4.27)$$

elde edilir. $d(\underline{X})$ bu haliyle

$$G = k\ \underline{Y}_1'\ \underline{A}^{-1}\ \underline{Y}_2 \qquad (2.4.28)$$

nin özel bir halidir. Yukarıda k bir sabiteyi, $\underline{Y}_1$ ve $\underline{Y}_2$ ise beklenen değerleri) ve & olan şans vektörlerini göstermektedir.

Wald (1944), G nin örnek dağılışını incelemiştir.

$$G = k\ \frac{m_{12}}{(1-m_{11})(1-m_{22}) - m_{12}^2} \qquad (2.4.28)$$

şeklinde bir eşitlik bularak G nin m_{11}, m_{12} ve m_{22} gibi üç değişkenin bir fonksiyonu olduğunu göstermiştir.

Sitgreaves (1952), bu değişkenlerin

$$\underline{M} = \underline{Y}'\ \underline{B}^{-1}\ \underline{Y} \qquad (2.4.29)$$

şeklindeki 2x2 boyutlu simetrik matrisin elemanları olduğunu göstererek bunun dağılışını bulmuştur. Yukarıda $\underline{Y} = (\underline{Y}_1, \underline{Y}_2)$ ve $\underline{B} = \underline{A} + \underline{Y}\,\underline{Y}'$ dür. Çok karışık bir formda olan $\underline{M}$ nin dağılışının çıkarılışını ana hatlariyle kısaca izah edelim.

$\underline{\delta} = k_1\underline{\delta}$ ve $\underline{\delta} = k_2\underline{\delta}$ olduğunu kabul ederek önce $\underline{Y}_1$ ve $\underline{Y}_2$ nin birlikte yoğunluk fonksiyonunu bulalım. $\underline{Y}_1$ ve $\underline{Y}_2$ çok-değişkenli normal dağılış gösterdiklerinden

$$f(\underline{Y}_1,\underline{Y}_2) = f(\underline{Y}_1)\, f(\underline{Y}_2)$$

$$= \frac{1}{(2\pi)^{p}|\underline{V}|^{+1}} \exp\left\{ -\frac{1}{2}\left[(\underline{Y}_1-k_1\underline{\delta})'\, \underline{V}^{-1}(\underline{Y}_1-k_1\underline{\delta}) \right.\right.$$

$$\left.\left. +(\underline{Y}_2-k_2\underline{\delta})'\, \underline{V}^{-1}(\underline{Y}_2-k_2\underline{\delta}) \right]\right\}$$

$$= \frac{1}{(2\pi)^{p}|\underline{V}|^{+1}} \exp\left\{ -\frac{1}{2}\left[\underline{\delta}'\,\underline{V}^{-1}\underline{\delta}\,(k_1^2 + k_2^2) \right.\right.$$

$$\left.\left. -2\underline{\delta}'\,\underline{V}^{-1}(k_1\underline{Y}_1 + k_2\,\underline{Y}_2)+\mathrm{Tr}(\underline{Y}'\underline{V}^{-1}\underline{Y}) \right]\right\} \qquad (2.4.30)$$

bulunur. Yukarıdaki $\mathrm{Tr}(\underline{Y}'\underline{V}^{-1}\underline{Y})$ terimi $\underline{Y}'\,\underline{V}^{-1}\underline{Y}$ matrisinin köşegen elemanlarının toplamını göstermektedir.

$\underline{A}$ matrisi bilindiği gibi Wishart dağılışı gösterir. Yani,

$$f(\underline{A}) = \frac{|\underline{A}|^{\frac{1}{2}(n-p-1)} \exp\left[-\frac{1}{2}\mathrm{Tr}(\underline{V}^{-1}\underline{A}) \right]}{2^{\frac{1}{2}np}\,\pi^{\frac{1}{4}p(p-1)}\,|\underline{V}|^{\frac{1}{2}n}\prod_{i=1}^{p}\left[\Gamma\left[\frac{1}{2}(n+1-i)\right]\right]}$$

(2.4.31)

dır. Böylece $\underline{Y}$ ve $\underline{A}$ nın birlikte dağılışları (2.4.30) ve (2.4.31) den

$$f(\underline{Y},\,\underline{A}) = \frac{|\underline{V}|^{-\frac{1}{2}(n+2)}\,|\underline{A}|^{\frac{1}{2}(n-p-1)}}{2^{\frac{1}{2}p(n+2)}\,\pi^{p+p(p-1)/4}\prod_{i=1}^{p}\left[\Gamma\left[\frac{1}{2}(n+1-i)\right]\right]}$$

$$.\exp\left\{-\frac{1}{2}\lambda^2(k_1^2+k_2^2)-\frac{1}{2}\mathrm{Tr}\left[\underline{V}^{-1}(\underline{A}+\underline{Y}\,\underline{Y}')\right]\right.$$

$$\left.+\,\underline{\delta}'\,\underline{V}^{-1}(k_1\underline{Y}_1+k_2\underline{Y}_2)\right\} \qquad (2.4.31)$$

olarak bulunur. Yukarıdaki $\lambda^2=\underline{\delta}'\,\underline{V}^{-1}\,\underline{\delta}$ dır. $\underline{M}$ matrisi $\underline{Y}$ ve $\underline{A}$ cinsinden ifade edildiğinden, $f(\underline{Y},\underline{A})$ dan gidilerek oldukça uzun cebir işlemleri ve transformasyonlardan sonra

$$f(\underline{M})=\frac{\Gamma\left[\frac{1}{2}(n+1)\right]\exp\left[-\frac{1}{2}\lambda^2(k_1^2+k_2^2)\right]|\underline{M}|^{(p-3)/2}|\underline{I}-\underline{M}|^{(n-p-1)/2}}{\Gamma\left[\frac{1}{2}(n-p+2)\right]\Gamma\left[\frac{1}{2}(n-p+1)\right]\Gamma\left[\frac{1}{2}(p-1)\right]\Gamma\left(\frac{1}{2}\right)}$$

$$\sum_{j=0}\frac{\Gamma\left[\frac{1}{2}(n+2)+j\right]}{\Gamma\left(\frac{1}{2}p+j\right)j!}$$

$$\left[\left(\frac{1}{2}\lambda^2\right)^j(k_1^2\,m_{11}+2k_1k_2m_{12}+k_2^2\,m_{22})^j\right] \qquad (2.4.32)$$

elde edilir. Bu fonksiyon sadece $\underline{M}$ nin elemanlarının dağılışını ifade etmektedir. $\underline{V}$ nin dağılışını bulmak için (2.4.28) deki gibi bir transformasyon yapmak gerekir. O zaman fonksiyonun çok daha karışık bir durum göstereceği aşikardır. Wald, büyük n değerleri için $G\cong -km_{12}$ olduğunu belirtmiştir. G nin bu şekli kullanıldığı takdirde basit bir transformasyonla yoğunluk fonksiyonu (2.4.32) den bulunabilir.

Anderson (1951), kendisinin ileri sürdüğü $w(\underline{X})$ in dağılış fonksiyonunu bulmuştur. Burada sadece asimtotik dağılışı üzerinde durulacaktır.

2.4.5. $w(\underline{X})$ in asimtotik dağılışı

Örnek büyüklüğü arttıkça, tahmin edilen parametreler populasyondaki gerçek değerlerine yaklaşırlar. n, n' birer tamsayı ve

δ, ϵ sıfır ile bir arasında küçük sayılar olsun. $n' \geq n$ olmak üzere

$$Pr(|\bar{y}_{n'} - \mu| > \epsilon) < 1-\delta$$

bağıntısını sağlıyacak büyüklükte n gibi bir tam sayı vardır. Diğer bir ifade ile n veya daha büyük sayıda müşahadeler yapıldığı takdirde, örnek ortalamasının populasyon ortalamasından farklı olmama ihtimali bire yaklaşır. Aynı şey vektör ve matrisler için de düşünülebilir. O halde, ihtimal limitleri

$$\begin{aligned}
&Pr \lim_{n_1 \to \infty} \underline{\bar{X}}^{(1)} = \underline{u}^{(1)} \\
&Pr \lim_{n_2 \to \infty} \underline{\bar{X}}^{(2)} = \underline{u}^{(2)} \\
&Pr \lim_{n_1, n_2 \to \infty} \underline{S} = \underline{V} \\
&Pr \lim_{n_1, n_2 \to \infty} \underline{S}^{-1} = \underline{V}^{-1}
\end{aligned} \qquad (2.4.34)$$

olur. İhtimal limitlerinin çarpım ve toplamları, çarpım ve toplamların ihtimal limitlerine eşit olduklarından (Cramér 1946).

$$\begin{aligned}
&Pr \lim_{n_1, n_2 \to \infty} \underline{S}^{-1}(\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)}) = \underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)}) \\
&Pr \lim_{n_1, n_2 \to \infty} (\underline{\bar{X}}^{(1)} + \underline{\bar{X}}^{(2)})', \underline{S}^{-1}(\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)}) \\
&\qquad = (\underline{u}^{(1)} + \underline{u}^{(2)})', \underline{V}^{-1}(\underline{u}^{(1)} - \underline{u}^{(2)})
\end{aligned} \qquad (2.4.35)$$

bulunur. Görüldüğü gibi yeteri kadar büyük örnekler için $w(\underline{X})$ in dağılışı $W(\underline{X})$ in dağılışına yaklaşmaktadır. O halde Π_1 ve Π_2 populasyonlarından alınmış büyük örnekler için, parametrelere ait tahminler, gerçek parametre değerleri imiş gibi alınarak $w(\underline{X})$ in

normal dağılış gösterdiği kabul edilebilir. Böyle bir kabüllenme ile muhakkak ki bir hataya yol açılmış olur fakat bu ihmal edilecek kadar küçüktür (Anderson 1958).

2.4.6. İkiden fazla çok-değişkenli populasyonlarda sınıflandırma

Çok-değişkenli populasyon sayısı ikiden fazla olduğu takdirde daha önce izah edilen teori bu gibi haller için de genelleştirilebilir. $\pi_1,\ldots,\pi_m$ populasyonlarına ait yoğunluk fonksiyonları $f_1(\underline{X}),\ldots, f_m(\underline{X})$ olsun. Fertleri bu populasyonlara sınıflandırmak için p-boyutlu uzay $R_1,\ldots,R_m$ bölgelerine ayrılır. Böylece R_k bölgesine düşen bir fert π_k populasyonuna sınıflandırılır. π_i deki ferdi yanlışlıkla π_j ye sınıflandırmanın maliyeti c_{ji} ve π_i nin a priori ihtimali q_i olsun. O zaman yanlış sınıflandırma ihtimali

$$P(\pi_j \mid \pi_i, R) = \int_{R_j} f_i(\underline{X})\, d\underline{X} \qquad (2.4.36)$$

ve yanlış sınıflandırmanın maliyetinin beklenen değeri

$$\sum_{i=1}^{n} q_i \left\{ \sum_{\substack{j=1 \\ j\neq i}} c_{ji} P(\pi_j \mid \pi_i, R) \right\} \qquad (2.4.37)$$

olur. $R_1,\ldots R_m$ bölgelerini seçmek için bu ifade minimum yapılır.

A priori ihtimalleri bilindiğine ve $\underline{X}$ vektörünün elemanları verildiğine göre, herhangibir müşahadenin i'inci populasyondan gelmiş olma ihtimali aşağıdaki gibi hesaplanabilir.

$$\frac{q_i f_i(\underline{X})}{\sum_{k=1}^{m} q_k f_k(\underline{X})} \qquad (2.4.38)$$

yanlışlıkla π_j ye sınıflandırmanın maliyetinin beklenen değeri ise

$$\sum_{\substack{i=1\\ i\neq j}}^{m} \frac{q_i f_i(\underline{X})}{\sum_{k=1}^{m} q_k f_k(\underline{X})} c_{ji} \qquad (2.4.39)$$

olur. O halde

$$\sum_{\substack{i=1\\ i\neq j}}^{m} q_i f_i(\underline{X}) c_{ji} \qquad (2.4.40)$$

ifadesinin minimum yapılması gerekir. Yani

$$\sum_{\substack{i=1\\ i\neq k}}^{m} q_i f_i(\underline{X}) c_{ki} < \sum_{\substack{i=1\\ i\neq j}}^{m} q_i f_i(\underline{X}) c_{ji} \qquad j\neq k \qquad (2.4.41)$$

kaidesine göre $\underline{X}$ ferdi R_k ya dahil edilir.

Yanlış sınıflandırma maliyetlerinin bütün populasyonlarda eşit olduğunu ve a priori ihtimallerin bilindiğini farzederek bu son bahiste anlatılanların çok-değişkenli normal populasonlar için tatbik edelim. Π_i ye ait ortalama $\underline{u}^{(i)}$ ve variyans-kovariyans matrisi $\underline{V}$ olsun. Π_j ve Π_k için kullanılacak fonksiyon aşağıdaki gibidir.

$$W_{jk}(\underline{X}) = \left[\underline{X} - \frac{1}{2}(\underline{u}^{(j)} + \underline{u}^{(k)})\right]' \underline{V}^{-1}(\underline{u}^{(j)} - \underline{u}^{(k)}) \qquad (2.4.42)$$

A priori ihtimalleri bilindiğinden R_j bölgesi

$$R_j : W_{jk}(\underline{X}) \geq \ln \frac{q_k}{q_j} \qquad j \neq k \qquad (2.4.43)$$

şeklinde tarif edilir.

A priori ihtimaller bilinmediği zaman R_j bölgesi daha değişik bir şekilde tarif edilir:

$$R_j : W_{jk}(\underline{X}) \geq c_j - c_k \qquad (2.4.44)$$

c_j ve c_k her populasyon için yapılacak doğru sınıflandırma ihtimallerini eşit kılma şartını yerine getiren sabitelerdir. Doğru sınıflandırma ihtimalini bulmak için önce $W_{jk}(\underline{X})$ lerin birlikte yoğunluk fonksiyonu bulunur. Kolayca görüleceği üzere

$$E\left[W_{jk}(\underline{X})\right] = \frac{1}{2}(\underline{u}^{(j)} - \underline{u}^{(k)})' \underline{V}^{-1} (\underline{u}^{(j)} - \underline{u}^{(k)}) \quad (2.4.45)$$

$$Kov\left[W_{jk}(\underline{X}), W_{jl}(\underline{X})\right] = (\underline{u}^{(j)} - \underline{u}^{(k)})' \underline{V}^{-1}(\underline{u}^{(j)} - \underline{u}^{(l)}) \quad (2.4.46)$$

eşitlikleri vardır. $W_{jk}(\underline{X})$ $(i = 1, \ldots, m;\ k \neq j)$ nın birlikte yoğunluk fonksiyonu L_j ile gösterilirse, c_j ve c_k sabiteleri aşağıdaki integralleri $j = 1, \ldots, m$ için eşit yapan değerler olarak hesaplanır.

$$f(\pi_j \mid \pi_j, R) = \int_{c_j - c_m}^{\infty} \cdots \int_{c_j - c_1}^{\infty} L_j dW_{jl} \cdots dW_{j(j-1)}\, dW_{j(j+1)} \cdots dW_{jm} \quad (2.4.47)$$

Bu ifadeden de anlaşıldığı gibi populasyon sayısı arttıkca c_j değerlerini hesaplamak daha da güçleşir.

(2.4.42) de belirtilen diskriminant fonksiyonunun parametrelerini, bundan önceki hallerde olduğu gibi örnek değerlerini kullanarak tahmin etmek gerekir. Onun için

$$\bar{\underline{X}}^{(i)} = \frac{1}{n_i} \sum_{\alpha=1}^{n_i} \underline{X}_{\alpha}^{(i)} \quad (2.4.48)$$

istatistiği $\underline{u}^{(i)}$ nin tahmini,

$$\underline{S} = \left[\left(\sum_{i=1}^{m} n_i - m\right)\right]^{-1} \left[\sum_{i=1}^{m} \sum_{\alpha=1}^{n_i} (\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}}^{(i)})(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}}^{(i)})'\right] \quad (2.4.49)$$

ise $\underline{V}$ nin tahmini için kullanılır. O zaman $w_{ij}(\underline{X})$ aşağıdaki şekli

alır.

$$w_{ij}(\underline{X}) = \left[\underline{X} - \frac{1}{2}(\bar{\underline{X}}^{(i)} + \bar{\underline{X}}^{(j)}) \right]' \underline{S}^{-1}(\bar{\underline{X}}^{(i)} - \bar{\underline{X}}^{(j)}) \quad (2.4.50)$$

2.5. Çok-Değişkenli Kalitatif Verilerle Diskriminant Analizi

Değişkenler sürekli olmadığı zaman diskriminant analizinin nasıl yapılacağını Cochran ve Hopkins (1961) geniş bir şekilde incelemiştir.

Bundan önceki bölümlerde Fisher ve Anderson'un diskriminant fonksiyonları ve bunlarla ilgili diğer bahisler ele alınmıştı. Bu fonksiyonlar, sürekli değişkenlerden elde edilen ölçüler kullanılmak suretiyle bulunuyordu. Değişkenler sürekli olmadığı zaman ayırım veya sınıflandırma problemi daha değişik bir durum gösterir.

Bir fert üzerinde yapılan müşahadelerden elde edilen veriler bazan renk, kıvam, şekil... gibi kalitatif ölçüler olabilirler. Her ölçü belirli sayıda değişik değerler alır. Meselâ renk ile ilgili ölçüler kırmızı, beyaz, yeşil ve sarı; kıvam ile de akışkan, katı, sert gibi olabilirler.

Bu tip kalitatif verilerle yapılacak sınıflandırma için her ölçüye bir değer (meselâ renk için 0,1,2,3) verilmek suretiyle sürekli değişkenlerin kullanılması mümkün olmakla beraber, bu durum sadece iki değer alan ölçüler için incelenecektir.

Sınıflandırmadaki optimum kaideyi bulmak için takip edilecek yol esas olarak çok-değişkenli kantitatif verilerde olduğu gibidir. Sürekli halde en önemli faraziyelerden biri değişkenlerin her iki populasyonda da çok-değişkenli dağılış gösterdikleri idi. Ayrıca varıyans-kovariyans matrislerinin iki populasyonda da aynı olduğu kabul edilmişti. Kalitatif değişkenlerde ise bu biraz daha değişiktir. i'inci kalitatif değişkenin d_i tane ayrı "durumu" olsun. Meselâ renk değişkeninde kırmızı, beyaz, yeşil, sarı olmak üzere dört ayrı durum vardır. Herhangibir fertte bu değişken ölçüldüğünde o ferdin hangi "duruma" gireceği belli olur. Aynı şey diğer

değişkenler içinde yapılır. Misal olarak 2,4, ve 5 durumlu değişkenleri ele alalım. Bunlar 2x4x5 = 40 tane "çok-değişkenli durum" meydana getirirler. Herbir fertte bu üç değişken ölçüldüğünde, o ferdin hangi çok-değişkenli duruma dahil edileceği belli olur. Böylece bunu genelleştirirsek, k- tane kalıtatif değişkenli ölçüler bir ferdi $d_1 d_2 \ldots d_k$ tane çok-değişkenli durumlardan birine sınıflandırır. Sürekli değişkenlerdeki birinci faraziyeye karşılık kalitatif değişkenlerde de fertlerin j'inci populasyondaki i'inci duruma dahil olma ihtimali p_{ji} nin bilindiği farzedilir ve $\sum_j p_{ji}=1$ dir.

İkinci husus olarak; sınıflandırma esnasında fertlerin birkısmı yanlışlıkla mensup olmadıkları populasyonlara dahil edilirler. Sürekli değişkenlerde olduğu gibi yanlış sınıflandırmadan dolayı bir "maliyet" bahis konusudur. Eğer yanlış sınıflandırmanın bu maliyeti her populasyonda bilinir veya tahmin edilebilirse, optimum kaidenin bunu da dikkate alması gerekir.

Son olarak, optimum kaidenin bulunabilmesi için fertlerin mensup oldukları populasyonların nisbî frekanslarının bilinmesi gerekmektedir. Sınıflandırma yapılırken ele alınan fert, mevcut bütün populasyonlardan eşit şansla gelmeyebilir.

Önemli olan hususları böylece açıkladıktan sonra optimum kaideyi bulabiliriz. Bunun için yanlış sınıflandırma maliyetinin beklenen değeri minimum yapılır. j'inci populasyona ait a priori intimal q_j olsun. O zaman i'inci duruma sınıflandırılanların nisbeti $q_j p_{ji}$ olarak bulunur. c_{kj}, j'inci populasyondaki ferdin yanlışlıkla k'ncı populasyona sınıflandırılmış olması maliyetini göstersin. Buradan, maliyetin beklenen değeri $q_j p_{ji} c_{kj}$ bulunur. Bu, bütün $j \neq k$ için toplanırsa $\sum_{j \neq k} q_i p_{ji} c_{kj}$ elde edilir. Bu ifade herbir populasyon için ayrı ayrı hesaplanarak fertler en küçük değere sahip olan populasyona dahil edilirler.

Eğer fertlerin yanlış sınıflandırılmalarının maliyeti sadece mensup oldukları populasyona bağlı ise ($c_{kj}=c_j$), yukardaki ifade

$$\sum_{j\neq k} q_j p_{ji} c_{kj} = \sum_{j=1}^{m} q_j p_{ji} c_j - q_k p_{ki} c_k \qquad (2.5.51)$$

şekline dönüşür. Eşitliğin sağındaki birinci terim k ya bağlı olmadığından, i'nci durumdaki fertler $q_k p_{ki} c_k$ değeri en büyük olan populasyona sınıflandırılırlar.

Eğer c_k her populasyonda eşit olarak alınırsa, i'nci durumdaki fertler $q_j p_{ji}$ değeri en büyük olan populasyona sınıflandırılır. Aynı şekilde, q_j ihtimalleri de bütün populasyonlarda eşit olduğu zaman sadece p_{ji} değerine bakılarak sınıflandırma yapılır.

Kalitatif verilerle yapılan diskriminant analizinde a priori ihtimallerin bilinmediği haller için genel bir çözüm yolu yoktur. Cochran ve Hopkins (1961) sadece iki populasyonda maliyetlerin eşit olduğu haller için nisbî frekansları tahmin etmişlerdir.

i'nci durumdaki fertlerin birinci populasyondaki nispetleri p_i, ikincide ise p_i' olsun. Ayrıca i'nci durumdaki fert sayısı r_i olsun.

$$P_i = q_1 p_i + q_2 p_i' \qquad (2.5.52)$$

yazılırsa, bütün müşahadeler için en yüksek ihtimal (maximum likelihood) fonksiyonu aşağıdaki gibi bulunur.

$$L = \prod_i (q_1 p_i + q_2 p_i')^{r_i} \qquad (2.5.53)$$

her iki tarafın logaritması alanırsa

$$\ln L = \sum_i r_i \ln(q_1 p_i + q_2 p_i') \qquad (2.5.54)$$

eşitliği bulunur. Ortalama ve variansı hesaplamak için birinci ve ikinci kısmî türevler alınır.

$$\frac{\partial \ln L}{\partial q_1} = \sum_i \frac{r_i (p_i - p_i')}{q_1 p_i + q_2 p_i'} \qquad (2.5.55)$$

$$\frac{\partial^2 \ln L}{\partial q_1^2} = \sum_i \frac{r_i (p_i - p_i')^2}{(q_1 p_i + q_2 p_i')^2} \qquad (2.5.56)$$

Buradan q_1 in en yüksek ihtimal tahmini, $\partial \ln L/\partial q_1$ i sıfıra eşitleyip çözmekle bulunur. Fakat $\partial \ln L/\partial q_1$ bilinen basit şekilde olmayıp q_1 in değeri deneme yoluyla bulunur. q_1 in tahmini olan $\hat{q}_1$ in varyansı

$$\text{Var}(\hat{q}_1) = -\left[E\left(\frac{\partial^2 \ln L}{\partial q_1^2}\right) \right]^{-1}$$

$$= \left[n \sum_i \frac{(p_i - p_i')^2}{\hat{q}_1 p_i + \hat{q}_2 p_i'} \right]^{-1} \qquad (2.5.57)$$

şeklinde bulunur.

2.6. İki Şıklı (Dichotomus) Verilerle Diskriminant Analizi

Birçok hallerde kalitatif değişkenler 0 ve 1 gibi sadece iki değer alırlar. Bu şekildeki değişkenlerin ölçüldüğü fertleri ihtiva eden populasyonlara iki şıklı (dichotomus) populasyonlar denir. Bunlar bir önceki bölümde bahsedilenlerin özel bir halidir. Diskriminant analizinde izah edilen genel metodlar burada aynı şekilde kullanılabilir. Fakat değişkenlere 0 ve 1 değeri verilmek suretiyle kantitatif hale getirildiklerinden daha değişik metodların uygulanması mümkün olmaktadır(Martin ve Bradley, 1972). $\underline{z}'=(z_1, z_2, \ldots z_I)$, $z \in \mathcal{Z}$, vektörü bir fert üzerinden elde edilmiş ölçüleri göstersin. $\mathcal{Z}$, $2x2x\ldots x2 = 2^I$ tane çok değişkenli durumdan ibarettir.

Π_m populasyonuna ait ihtimal fonksiyonu $p_m(\underline{z})=P(\underline{Z}=\underline{z} \mid \Pi_m)$ olsun. $\mathcal{Z}$ yüzeyi üzerinde

$$f(\underline{z}) = \sum_{m=1}^{M} w_m p_m(\underline{z}), \quad w_m \geq 0, \quad \sum_{m=1}^{M} w_m = 1 \qquad (2.6.58)$$

gibi bir fonksiyon tarif edilebilir. w_m, m'inci populasyonun nisbî frekansına ait bir sabitedir. Buna göre ele alınan ihtimal modeli

$$p_m(\underline{z}) = f(\underline{z}) \left[1 + h(\underline{a}_m, \underline{z}) \right] \qquad m = 1, \ldots, M \qquad (2.6.59)$$

şeklindedir. $h(\underline{a}_m, \underline{z})$; $\underline{z}$ elemanları cinsinden bir polinomu göstermektedir.

$$\phi_o(\underline{z}) = 1, \quad \phi_i(\underline{z}) = 2z_i - 1 \qquad i = 1, \ldots, I \qquad (2.6.60)$$

transformasyonları ile

$$\phi_\delta(\underline{z}) = \prod_{s=1}^{S} \phi_{\delta_s}(\underline{z}), \qquad \delta = (\delta_1, \ldots, \delta_s) \qquad (2.6.61)$$

$$S = 2, \ldots, I, \quad \delta_s = (1, \ldots, I)$$

ortogonal polinomu tarif edilirse bütün değişkenler kullanılarak 2^I tane ortogonal polinom elde edilebilir.

Ortogonallik şartı, bilindiği gibi

$$\sum_{\underline{z} \in Z_\delta} \phi_\gamma(\underline{z}) \, \phi_\delta(\underline{z}) = 2^I \Delta(\gamma, \delta) \qquad (2.6.62)$$

dır. Eğer $\gamma = \delta$ ise $\Delta(\gamma, \delta)$ bire, aksi halde sıfıra eşittir. Böylece ihtimal fonksiyonu $p_m(\underline{z})$ için

$$h(\underline{a}_m, \underline{z}) = \sum_{\delta \in \Omega_1} a_{m\delta} \, \phi_\delta(\underline{z}) \qquad (2.6.63)$$

yazılabilir. (2.6.59) numaralı eşitlik $h(\underline{a}_m, \underline{z})$ için çözülürse

$$h(\underline{a}_m, \underline{z}) = \frac{p_m(\underline{z}) - f(\underline{z})}{f(\underline{z})} \qquad (2.6.64)$$

bulunur. Ayrıca (2.6.62) ve (2.6.63) ün yardımıyla

$$a_{m\delta} = 2^{-I} \sum_{\underline{z} \in Z_\delta} \phi_\delta(\underline{z}) \, \frac{p_m(\underline{z}) - f(\underline{z})}{f(\underline{z})} \qquad (2.6.65)$$

eşitliği elde edilir. Eğer her populasyonda $p_m(\underline{z}) > 0$ ise $f(\underline{z})$ de pozitiftir. $f(\underline{z})= 0$ olduğu haller için (2.6.64) den dolayı $h(\underline{a}_m,\underline{z})$ belirsiz olur. Bu gibi hallerde $h(\underline{a}_m,\underline{z})$'e sıfır değeri verilir. (2.6.59) da tanımlanan modelde $h(\underline{a}_m,\underline{z})$ nin yüksek dereceden bir polinom oluşu, bu modelin pratikte kullanılmasını güçleştirir. Çünkü polinomun derecesi yükseldikçe, modelde tahmin edilecek olan parametrelerin de sayısı artar. Bu bakımdan pratikte en kullanılışlı modeller $h(\underline{a}_m,\underline{z})$ nin açınımında belirli bir dereceden sonraki terimler atılmakla elde edilir. Meselâ birinci dereceye kadar olan terimler alınırsa

$$h(\underline{a}_m,\underline{z}) \simeq h_s(\underline{a}_m,\underline{z}) = \sum_{\delta\in\Omega_s} a_{m\delta_s}\,\phi_{\delta_s}(\underline{z}),\quad s=1 \qquad (2.6.66)$$

elde edilir. Sınıflandırmada kullanılacak kriter

$$d_k(\underline{z}) = \sum_{m=1}^{M} L_{mk} q_m \left[1+h(\underline{a}_m,\underline{z})\right] \qquad (2.6.67)$$

dir. L_{mk}, doğru sınıflandırma halinde sıfır, aksi halde 1 değerini alan bir değişken, q_m ise Γ_m ye ait a priori ihtimali göstermektedir. Buna göre $d_k(\underline{z})$ değeri $[\,d_1(\underline{z}),\ldots,d_M(\underline{z})\,]$ içerisinde en küçük ise ilgili fert Γ_k populasyonuna dahil edilir. Dikkat edilirse sınıflandırmada optimum kaide olarak kullanılan kriter $f(\underline{z})$ ye bağlı değildir. Pratikte q_m ve $h(\underline{a}_m,\underline{z})$ değerleri bilinmediklerinden tahmin edilmeleri gerekmektedir. q_m nin tahmini $\hat{q}_m$ ve $\underline{a}_m$ nin tahmini $\hat{\underline{a}}_m$, (2.6.67) de yerine koyulursa

$$\hat{d}_k(\underline{z}) = \sum_{m=1}^{M} L_{mk}\,\hat{q}_m\left[1+h(\hat{\underline{a}}_m,\underline{z})\right] \qquad 2.6.68)$$

bulunur.

(2.6.66) daki polinom kullanılarak elde edilen modelden aşağıdaki şartlar çıkartılabilir:

1. (2.6.58) deki eşitliğin her iki tarafı toplanırsa, $\sum_{\underline{z}\in\Omega} p_m(\underline{z})$ = 1 şartından dolayı

$$\sum_{z \in z_0} f(\underline{z}) = 1 \qquad (2.6.69)$$

bulunur.

2. Bütün populasyonlara ait a priori ihtimallerin toplamı bire eşittir. Yani,

$$\sum_{m=1}^{M} q_m = 1 \qquad (2.6.70)$$

dir.

3. (2.4.58) ve (2.4.59) dan

$$f(\underline{z}) = \sum_{m=1}^{M} w_m f(\underline{z}) \left[1 + h_s(\underline{a}_m, \underline{z}) \right]$$

yazıp (2.6.69) kullanılırsa, $\underline{z} \notin z_0$ için

$$\sum_{m=1}^{M} w_m h_s(\underline{a}_m, \underline{z}) = 0 \qquad (2.6.71)$$

bağıntısı bulunur.

4. (2.6.65) in her iki tarafı w_m ile çarpılıp toplanırsa

$$\sum_{m=1}^{M} w_m a_{m\delta} = 2^{-I} \sum_{z \in z_0} \frac{\phi_\delta(\underline{z})}{f(\underline{z})} \left[\sum_{m=1}^{M} w_m p_m(\underline{z}) - f(\underline{z}) \right] \qquad (2.6.72)$$

bulunur. (2.6.58) den dolayı

$$\sum_{m=1}^{M} w_m a_{m\xi} = 0 \qquad (2.6.73)$$

bağıntısı elde edilir.

5. (2.6.59) un her iki tarafı toplanırsa

$$\sum_{\underline{z} \in z_0} p_m(\underline{z}) = \sum_{\underline{z} \in z_0} f(\underline{z}) + \sum_{\underline{z} \in z_0} f(\underline{z}) h_s(\underline{a}_m, \underline{z}) \qquad (2.6.74)$$

bulunur. (2.6.69) kullanılırsa

$$\sum_{\underline{z} \in \mathcal{Z}} f(\underline{z})\, h_s\,(\underline{a}_m, \underline{z}) = 0 \qquad (2.6.75)$$

olduğu sonucuna varılır.

Bu şartlar tam manasiyle bağımsız değillerdir. Meselâ (2.6.75); m = 1, ...,M-1 için doğru ise (2.6.73)den dolayı M için de doğru olmak zorundadır.

Beş şart böylece izah edildikten sonra, parametreleri en yüksek ihtimal metodu ile tahmin etmek mümkündür. Burada w_m in bilindiği haller ele alınacaktır. Diğer haller için $w_m = q_m$ alınabilir.

m'inci populasyonda $\underline{z}$ durumundaki müşahade sayısı $c_m(\underline{z})$ ile gösterilirse en yüksek ihtimal fonksiyonu

$$L = \prod_{m=1}^{M} \left[q_m^{N_m} \right] \prod_{\underline{z} \in \mathcal{Z}} \left\{ f(\underline{z}) \left[1 + h_s\,(\underline{a}_m, \underline{z}) \right] \right\}^{c_m(\underline{z})} \qquad (2.6.76)$$

şeklinde olur. Logaritması alındığı takdirde

$$\ln L = \sum_{m=1}^{M} N_m \ln q_m + \sum_{\underline{z} \in \mathcal{Z}} c(\underline{z}) \ln f(\underline{z})$$

$$+ \sum_{m=1}^{M} \sum_{\underline{z} \in \mathcal{Z}} c_m(\underline{z}) \ln \left[1 + h_s(\underline{a}_m, \underline{z}) \right] \qquad (2.6.77)$$

bulunur. Yukarıda, $c(\underline{z}) = \sum_{m=1}^{M} c_m(\underline{z})$ ve $N_m = \sum_{\underline{z} \in \mathcal{Z}} c_m(\underline{z})$ dir. Modeldeki parametrelerin en yüksek ihtimal tahminlerini bulabilmek için (2.6.77) nin bu parametrelere göre kısmî türevlerinin alınarak sıfıra eşitlenip çözülmesi gerekir. Yani ln L yi maksimum yapan değerler hesaplanır. Bunun için daha önce izah edilen şartlar esas alınıp Lagrange çarpanları metodu kullanılırsa

$$\Psi = \ln L + \Delta\left(1 - \sum_{m=1}^{M} q_m\right) + \lambda \left[1 - \sum_{\underline{z} \in \mathcal{Z}} f(\underline{z}) \right]$$

$$+ \sum_{m=1}^{M} \lambda_m \left[\sum_{\underline{z} \in \mathcal{Z}} f(\underline{z})\, h_s(\underline{a}_m, \underline{z}) \right] + \sum_{s} \mu_s \sum_{m=1}^{M} w_m a_{ms} \qquad (2.6.78)$$

ifadesini maksimum yapmak gerekir. Yukarıdaki Δ, λ, λ_m ve μ_δ Lagrange çarpanlarıdır. ψ nin q_m ve $f(\underline{z})$ ye göre kısmî türevleri alınırsa

$$\frac{\partial \psi}{\partial q_m} = \frac{N_m}{q_m} - \Delta = 0$$

$$\frac{\partial \psi}{\partial f(\underline{z})} = \frac{c(\underline{z})}{f(\underline{z})} - \lambda + \sum_{m=1}^{M} \lambda_m h_s(\underline{a}_m, \underline{z}) \qquad (2.6.79)$$

bulunur. Birinci denklemin her iki tarafı m= 1,...,M için; ikinci denklemin her iki tarafı $f(\underline{z})$ ile çarpılıp bütün $\underline{z} \in \underline{z}_0$ için toplandığında $\Delta = N$ ve $\lambda = N$ bulunur. Böylece q_m ve $f(\underline{z})$ nin tahminleri

$$\hat{q}_m = \frac{N_m}{N} \qquad m = 1,\ldots,M$$

$$\hat{f}(\underline{z}) = \frac{c(\underline{z})}{\left[N - \sum_{m=1}^{M} \lambda_m h_s(\hat{\underline{a}}_m, \underline{z}) \right]} \qquad \underline{z} \in \underline{z}_0 \qquad (2.6.80)$$

şeklinde bulunurlar. ψ nın $a_{m\delta}$ ya göre kısmî türevi aşağıdaki gibidir.

$$\frac{\partial \psi}{\partial a_{m\delta}} = \sum_{\underline{z} \in \underline{z}_0} \frac{c_m(\underline{z})\, \phi_\delta(\underline{z})}{1+h_s(\hat{\underline{a}}_m, \underline{z})} + \lambda_m \sum_{\underline{z} \in \underline{z}_0} \hat{f}(\underline{z})\, \phi_\delta(\underline{z}) + \mu_\delta w_m$$

(2.6.81)

dır. m=M olduğu zaman $\lambda_M = 0$ alınırsa

$$\mu_\delta = - \frac{1}{w_M} \sum_{\underline{z} \in \underline{z}_0} \frac{c_M(\underline{z})\, \phi_\delta(\underline{z})}{1 + h_s(\hat{\underline{a}}_M, \underline{z})} \qquad \delta \in s \quad (2.6.82)$$

bulunur. Ayrıca (2.6.81) de, tarifte belirtildiği gibi $\delta = 0$ için $\phi_0(\underline{z}) = 1$ bağıntısını kullanarak

$$\lambda_m = \frac{w_m}{w_M} \sum_{\underline{z} \in \underline{z}_0} \frac{c_M(\underline{z})}{1+h_s(\hat{\underline{a}}_M, \underline{z})} - \sum_{\underline{z} \in \underline{z}_0} \frac{c_m(\underline{z})}{1+h_s(\hat{\underline{a}}_m, \underline{z})} \qquad (2.6.83)$$

$$m=1,\ldots,M-1$$

elde edilir. Böylece $\hat{f}(\underline{z})$, $\underline{\hat{a}}_m$ cinsinden ifade edilebilir. $\hat{a}_{M\delta}$ ise

$$\hat{a}_{M\delta} = - \sum_{m=1}^{M-1} \frac{w_m}{w_M} \hat{a}_{m\delta} \qquad (2.6.84)$$

olur. Böylece

$$\varphi_{m\delta} (\underline{\hat{a}}_1, \ldots, \underline{\hat{a}}_{M-1}) = 0 \qquad (2.6.85)$$

formundaki denklem sistemleri çözülerek $\hat{a}_{m\delta}$ değerleri hesaplanır. Bu denklemler (2.6.75)'i kullanarak

$$\varphi_{mo} = \sum_{\underline{z} \in Z_\delta} \hat{f}(\underline{z})\, h_s(\underline{\hat{a}}_m, \underline{z}) \qquad m = 1, \ldots, M-1 \qquad (2.6.86)$$

ve (2.6.81)'i kullanarak

$$\varphi_{m\delta} = \sum_{\underline{z} \in Z_\delta} \frac{c_m(\underline{z})\, \phi_\delta(\underline{z})}{1 + h_s(\underline{\hat{a}}_m, \underline{z})} + \lambda_m \sum_{\underline{z} \in Z_\delta} \hat{f}(\underline{z}) \phi_\delta(\underline{z}) + \mu_\delta w_m$$

$$m = 1, \ldots, M-1 \qquad (2.6.87)$$

şekline getirilir. λ_m ve μ_δ , (2.6.83) ve (2.6.82) den bulunur.

Böylece $\hat{a}_{m\delta}$ nın nümerik çözümünü aşağıdaki gibi özetliyebiliriz:

1. $\underline{\hat{a}}_1, \ldots \underline{\hat{a}}_{M-1}$ için uygun başlangıç değerleri seçilir.
2. $\underline{\hat{a}}_M$ (2.6.84) den çözülür.
3. λ_m, $m = 1, \ldots, M-1$; (2.6.83) den hesaplanır. λ_M için sıfır değeri alınır.
4. $\hat{f}(\underline{z})$, $\underline{z} \in Z_\delta$ (2.6.80) den hesaplanır.
5. $\underline{\hat{a}}_1, \ldots \underline{\hat{a}}_{M-1}$ in daha doğru tahmini için (2.6.85) çözülür.
6. Arzu edilen bir doğruluk derecesine varıncaya kadar işlemlere devam edilir.

2.7. Diskriminant Analiziyle İlgili Diğer Konular

Diskriminant analizi üzerine birçok çalışmalar yapılmıştır. İki populasyon için bulunan diskriminant fonksiyonunun doğru olup olmadığını anlamak maksadiyle Barlett (1938) tarafından bir test geliştirilmiştir. Dana sonra Rao (1946, 1948) bu testi genişleterek eldeki verilerin iki grubu ayırdedebilecek bir durumda olup olmadıkları hakkında bir kriter ileri sürmüştür.

Burnaby (1966) büyüme ve bunun gibi faktörlerin diskriminant fonksiyonuna olan tesirlerini nispeten azaltan bir metod geliştirmiştir.

Sınıflandırma esnasında yapılan hata nispetleri üzerine de geniş çapta çalışmalar yapılmıştır (Hills, 1966; Lachenbruch ve Mickey 1968). Okamoto (1963) ise bu hata nispetlerinin asimtotik değerlerini hesaplamıştır.

Ölçülerin dağılışları ve populasyonlara ait a priori ihtimallerinin bilinmemesi sebebiyle diskriminant fonksiyonlarının pratiğe aktarılması büyük çapta olmamıştır. Bununla beraber tıp (Warner, Torantı, Veesey ve Stephenson, 1961), antropoloji (Ashton, Healy ve Lipton, 1957), ziraat (Blackith, 1960) ve taksonomi gibi (Lubischew, 1962) alanlardaki uygulamalar geleceğe açık gelişme temayüllerine ciddi işaretler olarak görülebilir.

3. MATERYAL VE METOD

Araştırmamızda iki grup veri kullanılmıştır. Bunlardan birincisi Atatürk Üniversitesi Öğrenci İşlerindeki kayıtlardan alınmış, diğeri ise şans sayılarından elde edilmiştir. Bunların alınış şekilleri aşağıda kısaca izah edilmiştir.

3.1. Öğrenci Kayıtlarından Elde Edilen Veriler

Bu gruptaki veriler, öğrencilerin mezun oldukları lisedeki durumları ile üniversiteye girişte aldıkları fen puanları arasındaki ilişkiyi araştırmak maksadıyla alınmıştır. Fen puanına ağırlık veren fakültelere aynı yılda giren 290 öğrenci üzerinde yapılan incelemede; lisedeki bölümü (F: fen veya E: edebiyat), mezuniyet derecesi (O: orta veya P: iyi ve daha yukarı) ve mezun olduğu lisenin bulunduğu yer (K: ilçe veya I: il) ile üniversiteye giriş imtahanında alınan fen puanları arasındaki ilişki araştırılacaktır. Diğer bir ifade ile ortalamanın üzerinde ve altında puan alan öğrencilerin mensup oldukları gruplar tesbit edilecektir. Bu hususta düşük bir hata ile kesin bir hükme varabilmek için şüphesiz daha fazla sayıdaki öğrenci üzerinde çalışmak gerekirdi. Fakat tatbik edilen metodların sadece tanıtılması ve mukayesesi bahis konusu olduğundan 290 öğrencinin kâfi gelebileceği düşünülmüştür. Bu gruptaki veriler cetvel 3.1. deki gibidir.

3.2. Normal Dağılış Gösteren Şans Sayılarından Elde Edilen Veriler

Bu gruptaki verilerin elde edilişinde simülasyonda kullanılan tekniklerden faydalanılmıştır.

Bilindiği üzere belirli bir aralıktaki üniform şans değişkeninin şans örneklerini elde etmek çeşitli metodlarla mümkün olmaktadır. Bir defa değişkenler bulunduktan sonra muayyen transformasyonlarla diğer dağılışları elde etmek zor değildir. Meselâ üniform dağılıştan normal dağılışa geçmek için aşağıdaki gibi bir yol takip edilir (Sezgin, 1972).

X_i üniform dağılış gösteren bir şans değişkeni olsun.

$$w_n = \sum_{i=1}^{n} X_i$$

şans değişkeni büyük n değerleri için standardize edilirse, standart normal dağılış gösterir. Çünkü

$$\frac{w_n - E(w_n)}{\sqrt{Var\ (w_n)}} \qquad (3.2.1)$$

değişkeni merkezi limit teoremine göre standart normaldır. Standart normalden diğer herhangi bir normal dağılışa geçiş ise ortalama ve variyanslarda yapılacak uygun değişimlerle mümkündür. Ortalama ve variyans-kovariyans matrisi verilen çok-değişkenli normal dağılışa da aynı şekilde geçilebilir.

Cetvel 3.1. Öğrenci Kayıtlarından Alınan Veriler

Durumlar	Vasatın altında	Vasatın üstünde
O K F	20	19
O K E	11	5
O I F	67	38
O I E	13	9
P K F	14	34
P K E	2	6
P I F	16	33
P I E	2	1
Toplam	145	145

Çalışmamızda standart normal dağılış gösteren rakkamlar hazır cetvellerden alınmıştır (Dikson ve Massey, 1969). Bunlar kullanılarak çok-değişkenli normal dağılışa geçiş aşağıdaki gibi olmuştur.

X'_{1i} ve X'_{ki},$(k = 2,3,\ldots,p)$ ortalama ve variyansları bilinen bağımsız normal değişkenler olsun. a_i ve b_j birer sabiteyi göstermek

üzere

$$X_{1i} = a_1 X'_{1i}$$
$$X_{ki} = b_k X'_{1i} + c_k X'_{ki} \quad , (k=2,3,\ldots,p) \tag{3.2.2}$$

şeklinde tarif edilen X_{ki} $(k = 1,2,\ldots,p)$ değişkenleri için

$$\begin{aligned} Var(X_{1i}) &= a_1^2 \, Var\,(X'_{1i}) \\ Var(X_{ki}) &= b_k^2 \, Var\,(X'_{ki}) + c_k^2 \, Var\,(X'_{ki}) \\ Kov(X_{1i},X_{ki}) &= a_1 b_k \, Var(X'_{1i}) \\ Kov(X'_{1i},X'_{ki}) &= 0 \end{aligned} \tag{3.2.3}$$

bağıntıları yazılabilir. Böylece X_{ki} $(k= 1,\ldots,p)$, çok-değişkenli normal dağılış gösterir.

Bu esaslara göre cetvel 3.2. ve 3.3. deki veriler aşağıdaki gibi elde edilmiştir.

a) Cetvel 3.2. deki veriler : Bu gruptaki veriler için ortalaması 5 ve variyansı 1 olan X'_{1i} ve X'_{2i} değişkenleri alınarak birinci populasyondaki veriler

$$X_{1i} = 0.3 X'_{1i}$$
$$X_{2i} = 0.1 X'_{1i} + 0.2 X'_{2i} \tag{3.2.4}$$

modellerine göre elde edilmiştir. Bu değişkenlere ait ortalama ve variyans-kovariyans matrisi aşağıdaki gibidir.

$$\underline{u}^{(1)'} = (1.5 \quad 1.5)$$
$$\underline{V} = \begin{bmatrix} 0.09 & 0.03 \\ 0.03 & 0.05 \end{bmatrix} \tag{3.2.5}$$

İkinci populasyondaki verileri elde etmek için de

Cetvel 3.2. Normal Dağılış Gösteren Şans Sayılarından Elde Edilen Veriler

Populasyon	I		II	
Değişken	X_1	X_2	X_1	X_2
1	1.639	1.573	2.355	1.487
2	1.518	1.001	2.087	1.032
3	1.946 (y_1)	1.578	2.197	1.219
4	1.807	1.508	2.499	1.020
5	1.918	1.528	2.918	1.106
6	1.772	1.488	1.933	0.959
7	1.854	1.407	2.469	1.119
8	1.050	1.252	2.098	1.481
9	1.293	1.582	2.812	1.411(y_2)
10	1.912	1.682	2.584	0.817

$$X_{1i} = 1.00 + 0.3\, X'_{1i}$$
$$X_{2i} = -0.30 + 0.1\, X'_{1i} + 0.2\, X'_{2i} \qquad (3.2.6)$$

modelleri kullanılmıştır. Bu gruptaki değişkenlerin ortalaması

$$\underline{u}^{(2)'} = (2.5 \quad 1.2) \qquad (3.2.7)$$

olup variyans-kovariyans matrisi birinci populasyondaki gibidir.

b) Cetvel 3.3. deki veriler : Önceki halde olduğu gibi birinci populasyon için ortalaması 5 ve variyansı 1 olan X'_{1i}, X'_{2i}, X'_{3i} ve X'_{4i} değişkenleri alınarak

$$X_{1i} = 0.3\, X'_{1i}$$
$$X_{2i} = 0.1\, X'_{1i} + 0.3\, X'_{2i}$$
$$X_{3i} = 0.2\, X'_{1i} + 0.3\, X'_{3i} \qquad (3.2.8)$$
$$X_{4i} = 0.2\, X'_{1i} + 0.3\, X'_{4i}$$

Cetvel 3.3. Normal Dağılış Gösteren Şans Sayılarından Elde Edilen Veriler

Pop.	I				II			
Değ.	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	1.403	1.948	2.524	2.349	2.134	2.391	3.122	2.892
2	1.442	2.144	1.994	2.517	2.225	1.660	2.685	2.379
3	1.709	2.248	3.052	2.875	1.976	2.499	3.038	2.470
4	1.507	2.173	1.949	2.562	2.437	2.404	3.120	2.903
5	1.596	2.915	3.156	2.487	2.244	2.649	2.331	2.943
6	1.679	2.324	2.339	3.093	2.210	1.865	3.007	2.632
7	1.731	2.368	2.868	2.981	2.270	2.326	2.949	2.621
8	1.459	2.296	2.534	2.607	1.979	2.738	2.655	2.581
9	1.397	1.811	1.816	2.294	1.838	2.109	2.593	2.201
10	1.728	2.130	2.431	2.940	2.053	2.117	3.091	2.072
11	1.131	1.731	2.511	2.107	2.167	2.335	2.608	3.455
12	1.332	1.867	2.324	2.454	2.372	2.422	2.561	2.358
13	1.979	2.180	2.945	2.769	2.047	2.872	2.543	2.556
14	1.283	2.272	2.319	2.684	2.318	2.226	3.132	2.962
15	1.869	2.063	2.672	3.118	2.743	3.094	3.337	3.327
16	1.814	1.953	2.217	2.665	2.218	2.656	2.683	3.124
17	1.608	1.738	2.537	2.063	2.053	2.651	2.662	2.459
18	1.627	2.333	2.243	2.273	1.844	2.608	2.343	2.343
19	1.913	2.433	2.376	3.261	1.971	2.513	2.963	2.991
20	1.238	1.584	1.906	2.639	1.738	2.783	3.104	2.255
21	1.663	2.129	2.558	2.618	2.498	2.567	3.034	2.944
22	1.765	2.468	2.615	2.721	1.778	1.885	2.034	2.756
23	1.137	1.601	2.386	2.345	2.049	2.609	3.113	2.506
24	1.767	2.021	2.859	2.941	2.061	2.184	2.589	2.346
25	1.305	1.762	2.441	2.283	2.450	2.706	2.862	2.439

modellerine göre X_{1i}, X_{2i}, X_{3i} ve X_{4i} değişkenleri elde edilmiştir. Bunların ortalama ve variyans-kovariyans matrisleri şöyledir :

$$\underline{u}^{(1)'} = (1.5 \quad 2.0 \quad 2.5 \quad 2.5)$$

$$\underline{V} = \begin{bmatrix} 0.09 & 0.03 & 0.06 & 0.06 \\ 0.03 & 0.10 & 0.02 & 0.02 \\ 0.06 & 0.02 & 0.13 & 0.04 \\ 0.06 & 0.02 & 0.04 & 0.13 \end{bmatrix} \qquad (3.2.9)$$

ikinci populasyondaki verileri elde etmek için de

$$\begin{aligned} X_{1i} &= 0.7 + 0.3\, X'_{1i} \\ X_{2i} &= 0.5 + 0.1\, X'_{1i} + 0.3\, X'_{2i} \\ X_{3i} &= 0.3 + 0.2\, X'_{1i} + 0.3\, X'_{3i} \\ X_{4i} &= 0.1 + 0.2\, X'_{1i} + 0.3\, X'_{4i} \end{aligned} \qquad (3.2.10)$$

modelleri kullanılmıştır. Bu populasyondaki değişkenlerin ortalamaları

$$\underline{u}^{(2)'} = (2.2 \quad 2.5 \quad 2.8 \quad 2.6) \qquad (3.2.11)$$

olup, variyans-kovariyans matrisleri birincideki gibidir.

3.3. M e t o d

Diskriminant analiziyle ilgili bazı problemler için şimdiye kadar ileri sürülmüş olan metodların bazılarının mukayesesi, araştırmamızın ağırlık noktasını teşkil etmektedir. Literatür özetinde verilen teorinin ışığı altında yapılan çalışmalarda sırasiyle, yanlış sınıflandırma ihtimalleri, kalitatif verilerle yapılan sınıflandırma, değişkenlerin seçimi ve kayıp müşahadeler gibi konular ele alınmıştır.

Yanlış sınıflandırma ihtimallerinin tahmini için ileri sürülen metodların en çok tatbik edilenlerinden beş tanesi alınarak bunların mukayesesi yapılmıştır. İkisi ampirik ve üçü teorik esaslara

dayanan bu metodların mukayesesi için cetvel 3.3. deki verilerden faydalanılmıştır. Önceden de belirtildiği gibi verilere ait parametrelerin gerçek değerleri bilinmektedir. Bu sebeple, çeşitli metodlara göre bulunan yanlış sınıflandırma ihtimalleri gerçek değerleriyle mukayese edilerek, metodların etkinlik durumları hakkında bir bilgi edinmek mümkündür. Yanlış sınıflandırma ihtimallerine ait rakkamların dağılışı hakkında kesin bir hüküm verilemediğinden parametrik olmayan (non-parametric) metodların bu amaç için kullanılmasının uygun olabileceği düşünülmüştür.

Çalışmamızın ikinci kısmı, kalitatif verilerle yapılan sınıflandırmaya ayrılmıştır. Sınıflandırmanın esas teorisi, kesikli değişkenlerin mevcut olduğu haller için de genelleştirilebildiği, ikinci bölümde belirtilmişti. Kalitatif verilerle sınıflandırmanın iki ayrı metoda göre nasıl yapılacağı cetvel 3.1. deki veriler kullanılmak suretiyle gösterilmiştir. Ayrıca, bir normal değişken kalitatif şekle çevrildiğinde ayırım gücünde meydana gelecek değişiklik incelenmiştir.

Kayıp müşahadelerin tahmini için iki metod üzerinde durulmuştur. Bunlardan birincisi, çok-değişkenli lineer modeller için geliştirilen bir kriterin diskriminant analizine tatbikiyle ilgilidir. Diğeri ise ileri sürdüğümüz genelleştirilmiş variyans kriterine dayanmaktadır. Bu iki metodun esasları anahatlariyle izah edilerek, birbirlerine göre avantaj ve dezavantajları belirtilmiştir. Mevcut imkânlarla elde edilen neticelerin genelleştirilmesi mümkün olamadığından, sadece cetvel 3.2. deki verilere dayanan çok özel bir misal üzerinde çalışılmıştır.

Son olarak, değişkenlerin seçimi için ileri sürülen bazı metodların mukayesesi yapılmıştır. Diskriminant fonksiyonundaki bağımlı değişkene, tarif edilen populasyonlara bağlı olmak üzere $-\frac{1}{2}$ ve $\frac{1}{2}$ değerleri verildiği taktirde, aynı verilerle hesaplanan diskriminant fonksiyonunu katsayılariyle, çoklu regresyon denkleminin katsayıları birbirleriyle orantılı olur. Bu noktadan hare-

ketle, şimdiye kadarki metodlara ilâveten çoklu regresyon teorisindeki değişkenlerin seçimine ait tekniklerden birinin, diskriminant analizi için de tatbik edilebileceği görüşü ileri sürülmüştür. Cetvel 3.3. deki veriler kullanılarak çeşitli metodların mukayesesi yapılmıştır.

Araştırmamızda, incelenmesi istenen konuların teorisi, oldukça geniş literatür çalışması ile ortaya konulmuştur. Diğer taraftan üzerinde durulan problemler değişik olup, bunların herbiri için ayrı bir yol takip edilmiştir. Bu sebeple her konu incelenirken ilgili teori de birlikte verilecektir. Böylece meselenin daha kolay anlaşılabileceği düşünülmüştür.

4. ARAŞTIRMALAR VE SONUÇLARI

4.1. Hata Nispetleri

Sınıflandırmada kriter olarak kullanılan aşağıdaki istatistik

$$w(\underline{X}) = \left[\underline{X}^{(1)} - \frac{1}{2} (\bar{\underline{X}}^{(1)} + \bar{\underline{X}}^{(2)}) \right]' \underline{S}^{-1} (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})$$

örnekten elde edilen bilgilere göre bulunmuştu. Bunun π_1 ve π_2 populasyonlarındaki beklenen değerleri ve variyanslarının, Δ^2 Mahalonobis mesafesi olmak üzere,

$$E\left[w(\underline{X}) / \pi_1 \right] = \frac{1}{2} (\underline{u}^{(1)} - \underline{u}^{(2)})' \underline{V}^{-1} (\underline{u}^{(1)} - \underline{u}^{(2)}) = \frac{1}{2} \Delta^2$$

$$E\left[w(\underline{X}) / \pi_2 \right] = - \frac{1}{2} (\underline{u}^{(1)} - \underline{u}^{(2)})' \underline{V}^{-1} (\underline{u}^{(1)} - \underline{u}^{(2)}) = - \frac{1}{2} \Delta^2$$

$$\mathrm{Var}\left[w(\underline{X}) / \pi_1 \right] = \mathrm{Var}\left[w(\underline{X}) / \pi_2 \right]$$

$$= (\underline{u}^{(1)} - \underline{u}^{(2)})' \underline{V}^{-1} (\underline{u}^{(1)} - \underline{u}^{(2)}) = \Delta^2 \qquad (4.1.1)$$

olduğu görülür.

$w(\underline{X})$, örnek büyüklüğüne bağlı bir tahmin olduğundan sınıflandırma yapılırken, fertlerin yanlışlıkla mensup olmadıkları populasyonlara dahil edilmesi sık sık karşılaşılan bir problemdir. Populasyonların parametrelerinin bilindiği durumlarda bile, değişkenlerin değişim sınırlarının birbirlerinin içine girmiş olmaları sebebiyle, yanlış sınıflandırma yapmak daima mümkündür.

Önce parametreleri bilinen çok-değişkenli iki normal populasyonu alarak yanlış sınıflandırma ihtimalini bulalım. Bu gibi durumlarda $W(\underline{X})$, π_1 populasyonunda ortalaması $\frac{1}{2} \Delta^2$ ve variyansı Δ^2; π_2 de de ortalaması $-\frac{1}{2} \Delta^2$ ve variyansı Δ^2 olan normal dağılış gösterir. Maliyet ve a priori ihtimallerin eşit olduğu kabul edilirse yanlış sınıflandırma ihtimalleri aşağıdaki gibi

bulunur.

$$P(\pi_2/\pi_1) = P\left[W(\underline{X}) / \pi_1 < 0 \right]$$

$$= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\,\Delta} \exp\left\{ - \frac{1}{2\Delta^2} \left[W(\underline{X}) - \frac{1}{2}\Delta^2 \right]^2 \right\} dW(\underline{X})$$

$$= \Phi\left(-\frac{1}{2}\Delta\right)$$

$$P(\pi_1/\pi_2) = P\left[W(\underline{X}) / \pi_2 > 0 \right]$$

$$= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}\,\Delta} \exp\left\{ - \frac{1}{2\Delta^2} \left[W(\underline{X}) + \frac{1}{2}\Delta^2 \right]^2 \right\} dW(\underline{X})$$

$$= \Phi\left(-\frac{1}{2}\Delta\right) \qquad (4.1.2)$$

Böylece standart normal dağılış eğrilerinin hazır cetvelleri kullanılarak yanlış sınıflandırma ihtimalleri hesaplanabilir.

Parametreler bilinmediği zaman, yanlış sınıflandırma ihtimallerinin hesaplanması daha değişik olur. $W(\underline{X})$ fonksiyonu $w(\underline{X})$ ile tahmin edildiğinden bu ihtimallerin hesaplanmasında $w(\underline{X})$ in yoğunluk fonksiyonunun kullanılması gerekir. Fakat önceki bölümlerde gösterildiği gibi $w(\underline{X})$ in yoğunluk fonksiyonu aşırı derecede karışık bir formdadır. Bu bakımdan ihtimal fonksiyonunu kullanarak yanlış sınıflandırma ihtimallerini bulmak pratikte kolay değildir. Bununla beraber yaklaşık değerler veren çeşitli metodlar mevcuttur:

1. Metod : Diskriminant fonksiyonunun hesaplanmasında kullanılan örnek yeteri kadar büyükse bu iki kısma ayrılır. Birinci örnekten diskriminant fonksiyonu hesaplanır. Bu fonksiyon kullanılarak ikinci örnekteki fertler sınıflandırılırsa, bu işlem sırasında hasıl olacak hata nispetinden yanlış sınıflandırma ihtimali bulunur. Bu metod küçük örneklerde pek kullanılmaz. Tıbbî araştırmalarda olduğu gibi bazan çok sayıdaki fertlerin örneğe alınması mümkün değildir. Bu gibi hallerde daha değişik metodları kullanma yollarına gidilmesi lâzımdır.

2. Metod : Bu metodda örneğin tamamı diskriminant fonksiyonunu hesaplamada kullanılır. Daha sonra örnek tekrar ele alınarak fertler yeniden sınıflandırılmaya tabi tutulurlar. Birinci metodda olduğu gibi yapılan toplam hata nispetlerinden yanlış sınıflandırma ihtimalleri hesaplanabilir. Diskriminant fonksiyonu bulunduktan sonra aynı örnekteki fertlerin yeniden sınıflandırılmaları, bilhassa küçük örnekler için, daha düşük yanlış sınıflandırma ihtimali verir.

3. Metod : Mahalonobis mesafesi olan Δ^2, örnek değerlerini kullanmak suretiyle tahmin edilir. Yani Δ^2 için

$$D^2 = (\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)})' \, \underline{S}^{-1} (\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)}) \qquad (4.1.3)$$

ifadesi kullanılır. Bu değer (4.1.2) de yerine konulursa

$$P(\pi_2/\pi_1) = P(\pi_1/\pi_2) = \Phi(-\frac{1}{2} D) \qquad (4.1.4)$$

bulunur. Bunda da, örnek büyük olmazsa hesaplanan yanlış sınıflandırma ihtimali hatalı olur. Çünkü D^2; Δ^2 nin sapmasız bir tahmini değildir.

4. Metod : Lachenbruch ve Mickey (1968) tarafından ileri sürülen bu metod üçüncü metoda benzer. Sadece Δ^2 nin sapmasız tahmini hesaplanarak aşağıdaki istatistik kullanılmıştır.

$$D^{*2} = \left[D^2 - \frac{(n_1+n_2-2)(n_1+n_2)\,p}{n_1 n_2 (n_1+n_2-p-3)} \right] \frac{n_1+n_2-p-3}{n_1+n_2-2} \qquad (4.1.5)$$

Fakat p ye nispetle n_1 ve n_2 küçük olduğu zaman D^{*2} değerleri çoğu zaman negatif çıkar. Bu durumu önlemek için D^{*2} yerine

$$D_s = \frac{n_1+n_2-p-3}{n_1+n_2-2} D^2 \qquad (4.1.6)$$

tahmini kullanılır. Dikkat edilirse D_s, (4.1.5) de ikinci terimin

çıkarılmasiyle elde edilmiştir.

5. Metod : Son olarak üzerinde durulacak bu metod Lachenbruch (1968) tarafından ileri sürülmiştir. $\bar{\underline{X}}^{(1)}$, $\bar{\underline{X}}^{(2)}$ ve $\underline{S}$ verildiğine göre $w(\underline{X})$ şartlı olarak normal dağılış gösterir. Ortalama ve variyansları aşağıdaki gibidir.

$$E\left[w(\underline{X})/\pi_1\right] = \frac{n_1+n_2-2}{2(n_1+n_2-p-3)}\left[\Delta^2 - \frac{p(n_1-n_2)}{n_1n_2}\right]$$

$$E\left[w(\underline{X})/\pi_2\right] = \frac{n_1+n_2-2}{2(n_1+n_2-p-3)}\left[-\Delta^2 - \frac{p(n_1-n_2)}{n_1n_2}\right]$$

$$\mathrm{Var}\left[w(\underline{X})\right] = \left[\Delta^2 + \frac{p(n_1+n_2)}{n_1n_2} \cdot \frac{(n_1+n_2-3)(n_1+n_2-2)^2}{(n_1+n_2-p-2)(n_1+n_2-p-3)(n_1+n_2-p-5)}\right] \qquad (4.1.7)$$

Bu değerler alındığında yanlış sınıflandırma ihtimalleri

$$P(\pi_2/\pi_1) = \Phi\left\{\frac{E\left[-w(\underline{X})/\pi_1\right]}{\sqrt{E\left[\mathrm{Var}\,(w(\underline{X}))\right]}}\right\}$$

$$P(\pi_1/\pi_2) = \Phi\left\{\frac{E\left[w(\underline{X})/\pi_2\right]}{\sqrt{E\left[\mathrm{Var}\,(w(\underline{X}))\right]}}\right\} \qquad (4.1.8)$$

şeklinde bulunur. $w(\underline{X})$ in beklenen değer ve variyanslarında yer alan Δ^2 nin tahmini olarak (4.1.5) de belirtilen D^2 kullanılırsa (4.1.7) deki beklenen değerler aşağıdaki şekli alırlar.

$$\widehat{E\left[\, w(\underline{X}) \mid \Pi_1 \right]} = \frac{1}{2} D^{*2} - \frac{(n_1+n_2-2)p}{(n_1+n_2-p-3)n_1}$$

$$\widehat{E\left[\, w(\underline{X}) \mid \Pi_2 \right]} = -\frac{1}{2} D^{*2} - \frac{(n_1+n_2-2)\,p}{(n_1+n_2-p-3)\,n_1} \qquad (4.1.9)$$

$$\widehat{Var\left[\, w(\underline{X}) \right]} = D^{*2} \frac{(n_1+n_2-3)(n_1+n_2-2)}{(n_1+n_2-p-2)(n_1+n_2-p-5)} \,.$$

Böylece $P(\Pi_2 \mid \Pi_1)$ ın tahmini

$$\hat{P}(\Pi_2 \mid \Pi_1) = \Phi\left\{ \frac{\widehat{E\left[\, -w(\underline{X}) \mid \Pi_1 \right]}}{\sqrt{\widehat{Var\left[\, w(\underline{X}) \right]}}} \right\} \qquad (4.1.10)$$

şeklindedir. $P(\Pi_1 \mid \Pi_2)$ de aynı yoldan bulunur.

Şimdi cetvel 3.3 deki verileri alarak yukarıda açıklanan metodları birbirleriyle karşılaştıralım. Cetveldeki çok-değişkenli verilere pratikte birçok misaller bulunabilir. Meselâ iki bitki türünü ayırdetmek için X_1 = yaprak genişliği, X_2= yaprak sapı uzunluğu, X_3=gövdenin çevresi, X_4= kökün uzunluğu gibi karakteristik vasıflar ölçülür.

Bu gruptaki değişkenlere ait 4x4 boyutlarındaki variyans-kovariyans matrisi Doolitlle metoduyla veya doğrudan doğruya ters çevrilebilir (Karataş, 1967). Böylece

$$\underline{V}^{-1} = \frac{100}{27} \begin{vmatrix} 6 & -1 & -2 & -2 \\ -1 & 3 & 0 & 0 \\ -2 & 0 & 3 & 0 \\ -2 & 0 & 0 & 3 \end{vmatrix} \qquad (4.1.11)$$

simetrik matrisini kullanarak, maliyet ve nisbî frekansların eşit olduğu kabul edilirse diskriminant fonksiyonu

$$W(\underline{X}) = 10.7407\ X_1 + 2.9630\ X_2 - 1.8519\ X_3 - 4.0740\ X_4 - 11.2408 \qquad (4.1.12)$$

şeklinde çıkar. Mahalonobis mesafesi ise $\Delta^2 = 8.0370$ bulunur. Buna göre hesaplanan fonksiyon değeri sıfırdan büyükse fertler ikinci populasyona aksi halde birinci populasyona sınıflandırılırlar. Birinci ve ikinci populasyona ait

$$\underline{u}^{(1)'} = (1.5 \quad 2.0 \quad 2.5 \quad 2.5)$$
$$\underline{u}^{(2)'} = (2.2 \quad 2.5 \quad 2.8 \quad 2.6) \qquad (4.1.13)$$

değerleri kullanılırsa

$$W(\underline{u}^{(1)}) = 7.2223 - 11.2408 = -4.0185$$
$$W(\underline{u}^{(2)}) = 15.2593 - 11.2408 = 4.0185 \qquad (4.1.14)$$

bulunur. Diğer taraftan

$$W\left[\frac{1}{2}(\underline{u}^{(1)} + \underline{u}^{(2)})\right] = 11.2408 - 11.2408 = 0.0000 \qquad (4.1.15)$$

olduğundan, populasyonun ortalamalarından hesaplanan fonksiyon değerleri, genel ortalamadan hesaplanana göre simetrik bir durum gösterirler. Diğer bir ifade ile sınıflandırma kriteri iki değerin ortasında yer alır. Yanlış sınıflandırma ihtimali

$$\Phi\left(-\frac{1}{2}\Delta\right) = \int_{-\infty}^{-1.417} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt$$
$$= 0.0778 \qquad (4.1.16)$$

olarak bulunur. Bu değer gerçek değeri göstermekte olup, bundan sonra çeşitli metodlara göre örnekten hesaplanacak yanlış sınıflandırma ihtimallerinin mukayesesinde kriter olarak kullanılacaktır.

1. Metod : Bu metodda her gruptaki fertler iki kısma ayrılarak ilk 12 müşahade diskriminant fonksiyonunu, geriye kalan 13 müşahade de hata nispetlerini hesaplamak için kullanılmıştır. Durum cetvel 4.1. de gösterilmiştir.

Cetvel 4.1. Birinci Metoda Göre Her Gruba Düşen Fert Sayıları

Populasyonlar	I	II
Değişkenler	X_1, X_2, X_3, X_4	X_1, X_2, X_3, X_4
Diskriminant fonksiyonu için kullanılan fert sayısı	12	12
Hata nispetleri için kullanılan fert sayısı	13	13
Toplam	25	25

Populasyona ait variyans-kovariyans matrisinin örnek tahmini (2.4.22) kullanılarak hesaplanmıştır. Bunun tersi aşağıdaki gibidir.

$$\underline{S}^{-1} = \begin{bmatrix} 53.2283 & -1.4502 & -5.2073 & -15.8946 \\ -1.4502 & 13.3138 & -3.1175 & -4.2325 \\ -5.2073 & -3.1175 & 9.4068 & 1.8859 \\ -15.8946 & -4.2325 & 1.8859 & 14.9209 \end{bmatrix} \qquad (4.1.17)$$

Ayrıca her gruptaki ortalamalar

$$\underline{\bar{X}}^{(1)'} = (1.510 \quad 2.171 \quad 2.458 \quad 2.606)$$
$$\underline{\bar{X}}^{(2)'} = (2.159 \quad 2.293 \quad 2.813 \quad 2.626) \qquad (4.1.18)$$

olarak bulunmuştur. Bu değerler kullanılarak hesaplanan diskriminant fonksiyonu

$$w(\underline{X}) = 32.2018X_1 - 0.5082X_2 - 0.3828X_3 - 9.8641X_4 - 31.1266 \qquad (4.1.19)$$

şeklindedir.

Buna göre, eğer $\underline{X}$ müşahadesinin fonksiyon değeri sıfırdan büyükse ilgili fert ikinci, aksi halde birinci populasyona sınıflandırılır. Her iki gruptaki toplam 26 fert bu şekildeki bir işleme tabi tutularak birinci grupta bir ve ikinci grupta iki ferdin yanlış olarak sınıflandırıldığı görülmüştür. Bunların kendi gruplarındaki toplam fert sayılarına göre nispetleri sırasiyle 0.0769 ve 0.1538 dir. Ortalama yanlış sınıflandırma ihtimali böylece 0.1154 olarak bulunur.

2. Metod : Bu sefer, diskriminant fonksiyonunu hesaplamak için bütün müşahadeler kullanılmıştır. Toplam olarak 50 müşahade alınarak hesaplanan variyans-kovariyans matrisinin tersi

$$\underline{S}^{-1} = \begin{vmatrix} 33.0846 & -4.1810 & -6.7942 & -10.9572 \\ -4.1810 & 14.2747 & -3.0819 & -2.1851 \\ -6.7942 & -3.0819 & 12.0051 & 1.3824 \\ -10.9572 & -2.1851 & 1.3824 & 13.7601 \end{vmatrix} \qquad (4.1.20)$$

ve gruplardaki ortalamalar

$$\underline{\bar{X}}^{(1)'} = (1.563 \quad 2.104 \quad 2.463 \quad 2.626)$$
$$\underline{\bar{X}}^{(2)'} = (2.147 \quad 2.435 \quad 2.806 \quad 2.661) \qquad (4.1.21)$$

şeklindedir. Böylece diskriminant fonksiyonu aşağıdaki gibi bulunur.

$$w(\underline{X}) = 15.2236X_1 + 1.4960X_2 - 0.8218X_3 - 6.1665X_4 - 12.3827 \qquad (4.1.22)$$

Bu fonksiyonun hesaplanmasında kullanılan 50 fert tekrar ele alınarak hata nispetleri hesaplanmıştır. İkinci gruptaki fertlerin hepsi doğru sınıflandırıldıkları halde birinci gruptaki fertlerden biri yanlış olarak sınıflandırılmıştır. Ortalama yanlış sınıflandırma ihtimali böylece 0.020 bulunur. Diğer bir ifade ile yukarıdaki diskriminant fonksiyonu kullanılarak yapılan sınıflandırmada ortalama olarak her 50 fertten biri mensup olmadıkları populasyonlara dahil edilirler.

3. Metod : İkinci metodda bulunan variyans-kovariyans matrisi ile ortalamalar kullanılarak Mahalonobis mesafesinin tahmini olan D^2,

$$D^2 = (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})' \, \underline{S}^{-1} \, (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) \qquad (4.1.23)$$
$$= 8.7734$$

şeklinde hesaplanır. Yanlış sınıflandırma ihtimali $\Phi(-\frac{1}{2} D) = 0.0681$ bulunur.

4. Metod: Mahalonobis mesafesini daha iyi tahmin edecek istatistik $\overset{*}{D}^2 = 7.5395$ olarak bulunur. Buradan, yanlış sınıflandırma ihtimali $\Phi(-\frac{1}{2} \overset{*}{D}) = 0.0853$ olur.

5. Metod : Ortalama ve variyansların sabit olduğu kabul edilerek $w(\underline{X})$ in beklenen değeri ve variyansının tahminleri aşağıdaki gibidir.

$$\widehat{E\left[w(\underline{X}) \mid \Pi_1 \right]} = 3.5912$$

$$\widehat{Var\left[w(\underline{X}) \right]} = 9.4281$$

Yanlış sınıflandırma ihtimali buradan

$$\Phi(-\frac{3.5912}{3.070}) = 0.1210$$

olarak hesaplanır.

5 ayrı metoda göre yanlış sınıflandırma ihtimalleri böylece hesaplandıktan sonra hangisinin daha etkili olduğu hakkında mukayese ile bir fikir edinilebilir. Yukarıda sadece 4 değişken kullanılmıştı. Yanlış sınıflandırma ihtimalleri değişken sayısına da bağlı olduğuna göre herhangi bir metodun üstünlüğü çeşitli değişken sayılarına göre aynı kalmayabilir. Bu bakımdan değişken sayısının da dikkate alınarak mukayesenin ona göre yapılması gerekir.

Eldeki misalde bütün metodlar dört değişkenli olduğu gibi, üç değişken (X_1, X_2, X_3), iki değişken (X_1, X_2) ve bir değişken (X_1) halleri için ayrı ayrı tatbik edilmiştir. Elde edilen sonuçlar

cetvel 4.2. deki gibidir.

Cetvel 4.2. Çeşitli Metodlara Göre Hesaplanan Yanlış Sınıflandırma İhtimalleri

Değiş.say. / Metodlar	1	2	3	4
1	0.1923	0.2692	0.2692	0.1154
2	0.1400	0.1400	0.1200	0.0200
3	0.1093	0.1093	0.1093	0.0681
4	0.1314	0.1210	0.1251	0.0853
5	0.1469	0.1446	0.1587	0.1210
Populasyonda	0.1210	0.1056	0.1003	0.0778

Cetvelin son sırasındaki değerler populasyon için hesaplanan yanlış sınıflandırma ihtimallerini göstermektedir. Bunlar gerçek ihtimaller olup, diğer metodlarla bulunan değerlerin mukayesesi için kullanılacaktır.

$$P_x = | P_\mu - P | \qquad (4.1.24)$$

formülüne göre P_x değerlerini hesaplıyalım. Yukarıdaki P_μ, populasyon değerlerine göre; P ise herhangi bir metoda göre hesaplanan yanlış sınıflandırma ihtimallerini göstermektedir. Bununla ilgili değerler cetvel 4.3. de gösterilmiştir.

Cetvelden de anlaşıldığı gibi metodların birbirlerine olan üstünlükleri, diskriminant fonksiyonunda kullanılan değişken sayısına göre değişmektedir. Onun için ilk bakışta hangi metodun en iyi olduğunu söylemek mümkün olamaz. Eğer rakkamlar variyans analizinin faraziyelerini yerine getirmiş olsalardı, çoklu mukayese yolu ile önemli fark gösteren metodlar tesbit edilebilirdi. Bu durumda, parametrik olmıyan metodlardan birini kullanmak en uygun bir yol olarak görünmektedir. Bu maksatla cetvel 4.3. deki veriler,

Cetvel 4.3. Çeşitli Metodlara Göre Tahmin Edilen Yanlış Sınıflandırma İhtimallerinin Gerçek Değerlerinden Ayrılışlarının Mutlak Değerleri

Değiş.say. / Metodlar	1	2	3	4
1	0.0713	0.1636	0.1689	0.0376
2	0.0190	0.0344	0.0197	0.0578
3	0.0117	0.0037	0.0090	0.0097
4	0.0104	0.0154	0.0248	0.0075
5	0.0259	0.0390	0.0584	0.0432
Populasyonda	0.0000	0.0000	0.0000	0.0000

her değişken için ayrı ayrı alınarak büyüklük sırasına göre sıra numaraları verilmiştir. Meselâ bir değişkenle ilgili olan sütunda en küçük değere (0.0104) sahip olan dördüncü metoda 1, bundan sonra gelen üçüncü metoda 2, ikinci metoda 3, beşinci metoda 4 ve birinci metoda 5 sıra numaralarının verildiği gibi. Neticeler cetvel 4.4. de gösterilmiştir.

Cetvel 4.4. Yanlış Sınıflandırma İhtimallerinin Sıra Numaraları

Değiş.say. / Metodlar	1	2	3	4	Toplam
1	5	5	5	3	18
2	3	3	2	5	13
3	2	1	1	2	6
4	1	2	3	1	7
5	4	4	4	4	16

Cetveldeki toplam değeri küçük olan metodların iyi, diğerlerinin daha kötü olması gerekir. Bu rakkamlar esas alınarak iki yönlü parametrik olmıyan variyans analizi yapmak için Friedman testi uygulanmıştır (Siegel, 1956). Bu testte kullanılan

$$\chi_r^2 = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 - 3n\,(k+1) \qquad (4.1.25)$$

değişkeni yaklaşık olarak k-1 serbestlik dereceli ki-kare dağılışı gösterir. Formüldeki n değişken sayısını; k metod sayısını, R_i ise her metoda ait toplamı göstermektedir. n= 4 ve k= 5 olduğundan χ_r^2 = 11.4 bulunur. P= 0.05 ihtimal sınırındaki 4 serbestlik dereceli ki-karenin cetvelden bulunan değeri 9.49 olduğundan, metodların aynı olduklarına dair sıfır hipotezi reddedilir. O halde metodlar arasında bir farklılık vardır.

Parametrik olmıyan istatistikte çoklu mukayese yapmak diğerlerinde olduğu gibi pek kolay değildir. Gayemiz en iyi metodu tespit etmek olduğundan, eldeki özel hal için aşağıdaki gibi bir yol takip edilmiştir.

Cetvel 4.4. den anlaşılacağı üzere birinci ve beşinci metodlar genel sıralamada daima en sonlarda yer almaktadır. O halde bunlar dikkate alınmadan geriye kalan üç metod içinden bir seçim yapılabilir. Bu üç metodun farklı olup olmadığı bir önceki problemde olduğu gibi Friedman testi ile anlaşılabilir. k= 3 ve n = 4 olduğundan χ_r^2 = 15.5 bulunur. Friedman testine göre serbestlik derecesi üçten küçük olduğu zaman χ_r^2 nin dağılışı daha başka bir durum göstermektedir. Bu bakımdan kesin ihtimal değerleri (exact probabilities) kullanılır. k=3 ve n=4 olduğu zaman $P(\chi_r^2 \geq 15.5) < 0.005$ bulunur (Siegel, 1956). Yani χ_r^2 değerinin 15.5 veya daha fazla olması ihtimali 0.005 den küçüktür. Diğer bir ifade ile metodlar arasındaki farklılık önemlidir. Hangisinin daha farklı olduğunu anlamak için ikişer kombinasyonlarını alarak parametrik olmıyan testlerden birini uygulamak mümkündür. Her toplamı meydana getiren

müşahadelerin sayısı küçük olduğundan, Mann-Whitney testi tatbik edilecektir.

Önce üçüncü ve dördüncü metodların alarak hipotezleri kuralım.

Ho : İki metodla hesaplanan yanlış sınıflandırma ihtimalleri aynıdır.

Hı : Üçüncü metodla hesaplanan yanlış sınıflandırma ihtimalleri diğer metodla hesaplanandan daha küçük olmaktadır.

Cetvel 4.3. deki değerler büyüklüklerine göre sıraya dizilirse

0.0037	0.0075	0.0090	0.0097	0.0104	0.0117	0.0154	0.0248
(3)	(4)	(3)	(3)	(4)	(3)	(4)	(4)

elde edilir. Her değer için (4) ün kaç defa (3) ü geçtiği sayılır. 0.0075'e tekabül eden (4), (3) ü bir defa; 0.0104'e tekabül eden (4), üç defa diğerleri ise dörder defa geçer. Böylece Mann-Whitney istatistiği

$$U = 1 + 3 + 4 + 4 = 12$$

olarak hesaplanır. Hazır cetvellerden U= 12 için P= 0.171 bulunur. Bu ihtimal 0.05 in çok üzerinde olduğu için sıfır hipotezi kabul edilir. Bu durumda farklılığın ikinci metoddan ileri gelmiş olması gerekir. Gerçekten, aynı şekilde yapılan testte ikinci metodun 0.05 ihtimal sınırında farklı olduğu bulunmuştur.

Yukarıdaki mukayeselerden, yanlış sınıflandırma ihtimallerini hesaplamada kullanılan üçüncü ve dördüncü metodların diğerlerine nazaran daha doğru sonuç verdiği anlaşılır. Birinci, beşinci ve bunları takip eden ikinci metod en kötü neticeleri vermektedir.

Ampirik hesaplara dayanan ilk iki metodun yanlış sonuç vermeleri iki sebepten ileri gelebilir. Birincisi, diskriminant fonksiyonu, populasyonun ortalaması ve variyans-kovariyans matrisinin örnek tahminleri kullanılarak elde edilmiştir. Böyle bir fonksiyon sapmasız değildir. İkinci olarak, fonksiyonun hesaplandığı müşahadeler tekrar yanlış sınıflandırma ihtimallerinin bulunması için

kullanılmıştır.

Lachenbruch tarafından ileri sürülen beşinci metodun da iyi bir sonuç vermemesinin sebebinin örnek ortalaması ve variyans-kovariyans matrisinin sabit olarak kabul edilmelerinden ileri geldiği düşünülmektedir.

Ele aldığımız misalde iyi sonuç veren dördüncü metodun herzaman için aynı durumunu muhafaza edebileceği şüphelidir. İki populasyon arasındaki standardize edilmiş mesafe küçük olduğu zaman, D^2 nin sapmasız tahmini olan $\hat{D}^2$, çoğu zaman negatif çıkar. Bunu önlemek için, daha önce bahsedilen başka bir tahminin kullanılması, şüphesiz bu metodla elde edilecek sonuca tesir eder.

4.2. Kalitatif Verilerle Yapılan Diskriminant Analizi Üzerine Bir Çalışma

Bu kısımda cetvel 3.1. deki rakkamlar alınarak, esaslarını (2.5) ve (2.6) da verdiğimiz kalitatif verilerle yapılan sınıflandırma izah edilecektir.

Üniversiteye girişte alınan fen puanlarının medyanı bulunarak öğrenciler "vasatın üstünde" ve "vasatın altında" olmak üzere iki gruba ayrılmıştır. Bu öğrencilerin lise mezuniyet dereceleri (O: orta veya P: iyi veya daha yüksek), mezun oldukları lisenin bulunduğu yer (I: il veya K: ilçe) ve lisedeki şubeleri (F: fen veya E: edebiyat) ile üniversiteye girişte aldıkları fen puanları arasındaki ilişkilerin incelenmesi istenmektedir.

4.2.1. Cochran ve Hopkins metoduna göre sınıflandırma

İki populasyonun nisbî frekansları medyanın özelliğinden dolayı aynıdır. Yanlış sınıflandırma maliyetlerinin eşit olduğu farzedilerek, daha önce bahsedilen kaideye göre öğrenciler, lisedeki durumlarına dayanarak sınıflandırılmaktadır.

Cetvel 4.5. de çeşitli durumlardaki öğrencilerin ihtimalleri verilmiştir. Buna göre $p_{1i} > p_{2i}$, i=1,2,3,4,8 olduğundan OKF, OKE, OIF, OIE ve PIE durumlarındaki öğrenciler birinci populasyona, diğerleri ise ikinci populasyona sınıflandırılmıştır.

Cetvel 4.5. Çeşitli Durumlardaki Öğrencilere Ait Nisbî Frekanslar ve Sınıflandırıldıkları Populasyonlar

($\underline{x}$) Durumlar	p_{1i} Vasatın altında	p_{2i} Vasatın üstünde	Sınıflandırıldığı Populasyon
OKF	0.138	0.131	I
OKE	0.076	0.035	I
OIF	0.462	0.262	I
OIE	0.089	0.062	I
PKF	0.097	0.234	II
PKE	0.014	0.041	II
PIF	0.110	0.228	II
PIE	0.014	0.007	I
Toplam	1.000	1.000	

Yanlış sınıflandırılma ihtimalleri cetvelden kolayca hesaplanabilir. Birinci ve ikinci populasyona yapılan yanlış sınıflandırma ihtimalleri aşağıdaki gibi bulunur.

$$P(\Pi_2 \mid \Pi_1) = 0.097 + 0.014 + 0.110 = 0.221$$

$$P(\Pi_1 \mid \Pi_2) = 0.131 + 0.035 + 0.262 + 0.062 + 0.007 = 0.497$$

Ortalama yanlış sınıflandırma ihtimali 0.359 dur.

Sadece öğrencilerin lise mezuniyet derecelerine bakarak sınıflandırma yapmanın mümkün olup olmadığını anlamak için cetvel 4.5a yapılır.

Böylece orta derece ile mezun olanlar birinci populasyona, diğerleri ise ikinci populasyona sınıflandırılırlar. Diğer bir ifade ile liseden orta derece ile mezun olanlar iyi ve daha yüksek

Cetvel 4.6. Mezuniyet Derecesine Göre Her Duruma Düşen Öğrenci Nispetleri

Durumlar	p_{1i} Vasatın altında	p_{2i} Vasatın üstünde	Sınıflandırılan Populasyon
0	0.765	0.490	I
P	0.235	0.510	II
Toplam	1.000	1.000	

derece ile mezun olanlara nazaran üniversite giriş imtahanlarında daha düşük fen puanı almaktadırlar. Bu şekilde yapılan sınıflandırmada

$$P(\pi_2|\pi_1) = 0.235$$

$$P(\pi_1|\pi_2) = 0.490$$

ihtimalleri bulunur. Ortalama hata nispeti 0.363 dür. Bu daha önce bulunan ortalama yanlış sınıflandırma ihtimalinden çok farklı değildir. Aynı şey il-ilçe ve fen-edebiyat kriterlerine göre yapılarak yanlış sınıflandırma ihtimallerinin çok daha büyük olduğu görülmüştür. Bu sebeple lisenin bulunduğu yer ve mezun olunan bölüm dikkate alınmadan da öğrencilerin sınıflandırılması mümkün olabilir. Diğer bir ifade ile sadece lise mezuniyet derecelerine bakmak suretiyle üniversite giriş imtahanlarında alınan fen puanlarının vasatın altında veya üstünde olacağı takriben 0.64 ihtimalle doğru olarak tahmin edilebilir.

4.2.2. Martin ve Bradley Metoduna göre sınıflandırma

Bir önceki misalde, eldeki değişkenlerin her biri iki ayrı değeri almaktadır. Meselâ, öğrencinin lisedeki bölümü fen ve edebiyat diye iki kısımda incelenmiştir. Diğer değişkenlerde aynı şekildedir. Bu bakımdan her değişkene 0 ve 1 değerleri verilirse, yukarıda tarif edilen populasyonlar iki şıklı (dichotomus) populasyonları olur. Onun için aynı misal ele alınarak, daha önce iki şıklı

populasyonlar için izah edilen teorinin tatbikatı gösterilecektir.

$O= 0,\ P= 1;\ K= 0,\ I= 1;\ F= 0,\ E= 1$

değerlerini almış olsunlar. Misalde $n = 290$, $m= 2$, $I= 4$, $n_1= n_2 = 145$, $w_1=w_2= \frac{1}{2}$ dir. Ayrıca $L_{11}= L_{22}= 0$ ve $L_{21}=L_{12}= 1$ olarak alınacaktır.

Kullanılan ihtimal modeli aşağıdaki gibidir.

$$p_m(\underline{z}) = f(\underline{z})\left[1+ h_1(\underline{a}_m, \underline{z})\right] \qquad m= 1,2;\ \underline{z} \in \underline{z} \qquad (4.2.26)$$

Bütün $\underline{z}$ değerleri için $p_m(\underline{z})$ ayrı ayrı yazılırsa aşağıdaki eşitlikler elde edilir.

$$\begin{aligned}
p_1(000) &= f(000)(1+ a_o- a_1- a_2 - a_3)\\
p_1(001) &= f(001)(1+ a_o- a_1- a_2 + a_3)\\
p_1(010) &= f(010)(1+ a_o- a_1+ a_2 - a_3)\\
p_1(011) &= f(011)(1+ a_o- a_1+ a_2 + a_3)\\
p_1(100) &= f(100)(1+ a_o+ a_1 -a_2 - a_3)\\
p_1(101) &= f(101)(1+ a_o+ a_1 -a_2 + a_3)\\
p_1(110) &= f(110)(1+ a_o+ a_1 +a_2 - a_3)\\
p_1(111) &= f(111)(1+ a_o+ a_1 +a_2 + a_3)
\end{aligned} \qquad (4.2.27)$$

Sınıflandırmadaki kaide

$$\begin{aligned}
\Gamma_1 &: h_1(\underline{a}_1, \underline{z}) \geq 0\\
\Gamma_2 &: h_1(\underline{a}_1, \underline{z}) < 0
\end{aligned} \qquad (4.2.28)$$

şeklindedir. $\underline{a}_1$ değerlerini hesaplamak için daha önce izah edilen metodlara göre bir yol takip edilir. Önce (2.6.65) den $\underline{a}_1$ in başlangıç değerleri hesaplanır. Bulunan değerler aşağıdaki gibidir.

$$a_{10} = - 0.009 \qquad a_{12} = 0.008$$

$$a_{11} = 0.037 \qquad a_{13} = 0.107$$

Bunlar kullanılarak (2.683) ün yardımıyla $\hat{\lambda}_1 = -21.838$ bulunur. Muhtelif $(\underline{z})$ değerlerine göre hesaplanan $\hat{f}(\underline{z})$ aşağıdaki gibidir.

$$\begin{aligned} \hat{f}(000) &= 0.136 \qquad \hat{f}(100) = 0.167 \\ \hat{f}(001) &= 0.055 \qquad \hat{f}(101) = 0.027 \\ \hat{f}(010) &= 0.365 \qquad \hat{f}(110) = 0.170 \\ \hat{f}(011) &= 0.075 \qquad \hat{f}(111) = 0.010 \end{aligned} \tag{4.2.29}$$

Böylece $\hat{\lambda}_1$ ve $\hat{f}(\underline{z})$ değerlerini kullanmak suretiyle $\underline{a}_1$ in çözümüne geçilebilir. Fakat (2.6.82) den de anlaşılacağı gibi çözüm uzun bir komputer hesaplamasını gerektirmektedir. $|h_1(\underline{a}_m \underline{z})| < 1$ olduğuna dikkat ederek (2.6.82) seriye açılıp, derecesi iki ve daha fazla olan terimler atılmıştır. (2.6.82) tekrar yazılırsa aşağıdaki eşitlik elde edilir.

$$\sum_{\underline{z} \in \underline{z}_\delta} \frac{c_1(\underline{z})\, \phi_\delta(\underline{z})}{1 + h_1(\underline{a}_1, \underline{z})} - \sum_{\underline{z} \in \underline{z}_\delta} \frac{c_2(\underline{z})\, \phi_\delta(\underline{z})}{1 + h_1(\underline{a}_2, \underline{z})}$$

$$= -\lambda_1 \sum_{\underline{z} \in \underline{z}_\delta} \hat{f}(\underline{z})\, \phi_\delta(\underline{z}) \tag{4.2.30}$$

(2.6.71) e göre $h_1(\underline{a}_1, \underline{z}) = -h_1(\underline{a}_2, \underline{z})$ olup aşağıdaki açınımları yapılırsa

$$\begin{aligned} \frac{1}{1 + h_1(\hat{\underline{a}}_1, \underline{z})} &\cong 1 - h_1(\hat{\underline{a}}_1, \underline{z}) \\ \frac{1}{1 - h_1(\hat{\underline{a}}_1, \underline{z})} &\cong 1 + h_1(\hat{\underline{a}}_1, \underline{z}) \end{aligned} \tag{4.2.31}$$

bulunur. Bunlarda yukarıda yerine koyulursa

$$\sum_{\underline{z} \in \underline{z}_\delta} \phi_\delta(\underline{z}) \left\{ c_1(\underline{z}) - c_2(\underline{z}) - c(\underline{z})\, h_1(\underline{a}_1, \underline{z}) \right\}$$

$$= -\lambda_1 \sum_{\underline{z} \in Z_\beta} \hat{f}(\underline{z})\, \phi_\ell(\underline{z}) \qquad (4.2.32)$$

bulunur. Böylece denklem sistemleri lineer şekle dönüştürülmüş olur. İlgili değerler kullanılırsa aşağıdaki eşitlik elde edilir.

$$\begin{bmatrix} 1005 & -257 & 235 & -671 \\ 74 & -290 & 76 & -20 \\ -68 & 76 & -290 & 66 \\ 192 & -20 & 66 & -290 \end{bmatrix} \begin{bmatrix} \hat{a}_{10} \\ \hat{a}_{11} \\ \hat{a}_{12} \\ \hat{a}_{13} \end{bmatrix} = \begin{bmatrix} 0.000 \\ 74.388 \\ -28.868 \\ 28.653 \end{bmatrix} \qquad (4.2.33)$$

Bu denklem sistemleri çözüldüğü zaman şu değerler bulunur.

$\hat{a}_{10} = 0.0167$ $\qquad$ $\hat{a}_{12} = 0.0635$

$\hat{a}_{11} = -0.2454$ $\qquad$ $\hat{a}_{13} = 0.1412$

Bunların doğruluk derecesi başlangıç değerlerine bağlıdır. İlk değerler mukayese edilirse bunların çok farklı olduğu görülür. Bu bakımdan bulunan en son değerler başlangıç değerleri olarak alınıp işlemlere istenildiği kadar devam edilir.

Bir önceki tahminlerle bir sonrakiler arasındaki ortalama fark takriben 0.0025 oluncaya kadar işlemlere devam edilmiştir. Her safhada bulunan $\hat{a}_1$ değerleri cetvel 4.7. deki gibidir. Görüldüğü

Cetvel 4.7. Parametrelerin Çeşitli Safhalardaki Tahminleri

	Başlangıç	1. Tahmin	2. Tahmin	3. Tahmin
a_{10}	-0.009	0.0167	-0.0537	-0.0575
a_{11}	0.008	-0.2454	-0.2742	-0.2758
a_{12}	0.037	0.0635	0.0683	0.0684
a_{13}	0.107	0.1412	0.0493	0.0436

gibi yaklaşım çok çabuk olmuştur. Parametrelerin tahminlerinden sonra, sınıflandırma kaidesi için $h_1(\underline{a}_1, \underline{z})$ hesaplanması gerekmek-

tedir. Son olarak hesaplanan parametre değerleri her $\underline{z}$ için $h_1(\underline{a}_1,\underline{z})$ yi hesaplamada kullanılırsa cetvel 4.8. deki değerler elde edilir.

Cetvel 4.8. Muhtelif $\underline{z}$'ler İçin Hesaplanan $h_1(\hat{\underline{a}}_1,\underline{z})$ Değerleri

Durum	$h_1(\underline{a}_1,\underline{z})$	$h_1(\hat{\underline{a}}_1,\underline{z})$
000	$a_o - a_1 - a_2 - a_3$	0.1063
001	$a_o - a_1 - a_2 + a_3$	0.1935
010	$a_o - a_1 + a_2 - a_3$	0.2431
011	$a_o - a_1 + a_2 + a_3$	0.3303
100	$a_o + a_1 - a_2 - a_3$	-0.4453
101	$a_o + a_1 - a_2 + a_3$	-0.3581
110	$a_o + a_1 + a_2 - a_3$	-0.3085
111	$a_o + a_1 + a_2 + a_3$	-0.2213

Sınıflandırma kaidesine göre cetvel 4.8. deki 100, 101, 110 ve 111 durumlarına düşen fertlerin ikinci populasyona, diğerlerinin ise birinci populasyona sınıflandırılmaları gerekir.

Martin ve Bradly metodu ilk metodla mukayese edilirse (111) duruma hariç diğerlerinin sınıflandırılışları aynıdır. (111) durumunda birinci populasyondaki fertlerin nisbî frekansları daha yüksek olmasına rağmen ikinci populasyona sınıflandırılmıştır. Aslında bu durumdaki toplam fert sayısı üç olup diğerlerine nazaran çok düşüktür. Bu bakımdan (111) için, her iki metodlada bulunan sınıflandırma kaidelerine fazla güvenilemez. Son metodun en önemli avantajı çok-değişkenli durumların herhangi biri veya birkaçında müşa-

hadelerin olmadığı hallere tatbik edilebilmesidir.

4.2.3. Normal değişkenlerin kalitatif değişkenlere çevrilmesi

Bundan önce incelenen bölümlerde ele alınan değişkenlerin ya hep sürekli veya hep kesikli olduğu farzedilmişti. Bazan kalitatif ve kantitatif değişkenler bir arada olabilir. Bu gibi hallerde kalitatif değişkenlere 0,1,2 gibi sıra numarası vererek bir dereceye kadar kantitatif şekle sokmak suretiyle analiz yapılabilir. Bununla beraber diğer bir yol, kantitatif değişkenlerin kalitatif hale getirilmeleridir. Her değişken 2,3,4 kısma bölünerek analizleri, önceki bölümlerdeki gibi yapılır.

Bu bölümde çok özel bir hal incelenerek, tek bir normal değişkenin kalitatif hale getirilmesinden dolayı ayırım gücünde hasıl olacak kayıp incelenecektir. y gibi normal bir değişkenin Π_1 populasyonundaki ortalaması 0, varyansı 1; $\Pi_2$deki ortalaması δ ve varyansı Π_1 deki gibi olsun. Ayrıca $|\delta| < \sqrt{2}$ şartını kabul edelim. Bu değişken için Mahalanobis mesafesi

$$\Delta^2 = (\underline{u}^{(1)} - \underline{u}^{(2)})' \underline{V}^{-1} (\underline{u}^{(1)} - \underline{u}^{(2)}) \qquad (4.2.34)$$

$$= \delta^2$$

olur. y'yi iki durumlu kesikli değişken şekline çevirelim. Π_1 ve Π_2 populasyonlarında herhangi bir ferdin i'inci (i=1,2) durumda olması ihtimallerini $P(i|\Pi_1)$ ve $P(i|\Pi_2)$ ile gösterelim. O zaman optimum kaide

$$P(i|\Pi_1) > P(i|\Pi_2) \qquad (4.2.35)$$

şeklindedir. Yani bu şart sağlanıyorsa, i'inci durumdaki fertler Π_1, aksi halde Π_2 populasyonuna dahil edilir.

(4.2.35) in heriki tarafının logaritması alınırsa

$$\ln\left[P(i|\Pi_1)\right] - \ln\left[P(i|\Pi_2)\right] > 0 \qquad (4.2.36)$$

bulunur. Ayrıca,

$$X = \ln \left[P(i \mid \Pi_1) \right] - \ln \left[P(i \mid \Pi_2) \right] \qquad (4.2.37)$$

yazılırsa optimum kaide

$$X > 0 \qquad (4.2.38)$$

şeklini alır.

Sürekli değişkenlerin bölünmesi aşağıdaki şekilde gösterildiği gibidir.

Şekil 4.1. Sürekli Değişkenlerin Bölünmesi

Buna göre, değerleri $\delta/2$ den küçük olan fertler birinci duruma (p_1 ihtimali ile gösterilen), büyük olanlar ikinci duruma dahil edilir. Her populasyondaki ihtimaller ve bunlara ait değişkenler cetvel 4.9. da gösterilmiştir.

Cetvel 4.9. Çeşitli Durumlardaki İhtimaller ve İlgili X- Değişkenleri

İhtimaller \ i	1	2
$P(i \mid \Pi_1)$	p_1	p_2
$P(i \mid \Pi_2)$	p_2	p_1
X	w_1	w_2

cetveldeki $w_1 = \ln \frac{p_1}{p_2}$ ve $w_2 = \ln \frac{p_2}{p_1}$ dir. İlgili değerler kullanılarak X'in Π_1 populasyonundaki beklenen değeri ve variyansı

$$E(X \mid \Pi_1) = w_1 p_1 + w_2 p_2$$
$$\mathrm{Var}(X \mid \Pi_1) = w_1^2 p_1 + w_2^2 p_2 - \left[E(X \mid \Pi_1) \right]^2 \qquad (4.2.39)$$

şeklinde bulunur.

$$p_1 = \int_{-\infty}^{\delta/2} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)\, dt$$
$$= \Phi(\delta/2) \qquad (4.2.40)$$

şeklinde gösterelim. $\Phi(\delta/2)$ sıfır etrafında Taylor serisine göre açılırsa

$$\Phi(\frac{1}{2}\delta) = \Phi(0) + \frac{1}{2}\delta\left[\frac{d\Phi(t)}{dt}\right]_{t=0} + \cdots$$
$$\simeq \frac{1}{2}(1 + \frac{1}{\sqrt{2\pi}}\delta) \qquad (4.2.41)$$

bulunur. O halde

$$p_1 = \frac{1}{2}(1 + \frac{1}{\sqrt{2\pi}}\delta)$$
$$p_2 = \frac{1}{2}(1 - \frac{1}{\sqrt{2\pi}}\delta) \qquad (4.2.42)$$

olur. Diğer taraftan

$$w_1 = \ln\left[\frac{1 + \frac{1}{\sqrt{2\pi}}\delta}{1 - \frac{1}{\sqrt{2\pi}}\delta}\right] \qquad (4.2.43)$$
$$= \ln(1 + \frac{1}{\sqrt{2\pi}}\delta) - \ln(1 - \frac{1}{\sqrt{2\pi}}\delta)$$

şeklinde yazılabilir. Bu da sıfır etrafında Taylor serisine göre açılabilir. Çünkü $\delta < \sqrt{2\pi}$ dir. Böylece

$$w_1 = \frac{2\delta}{\sqrt{2\pi}}(1 - \frac{\delta}{2\sqrt{2\pi}}) \qquad (4.2.44)$$

elde edilir. Aynı şekilde

$$w_2 = -\frac{2\delta}{\sqrt{2\pi}}\left(1-\frac{\delta}{2\sqrt{2\pi}}\right) = -w_1 \qquad (4.2.45)$$

bulunur. X'in her iki populasyondaki beklenen değeri

$$E(X|\pi_1) = (2p_1 - 1)\, w_1$$

$$= \frac{\delta^2}{\pi}\left(1 - \frac{\delta}{2\sqrt{2\pi}}\right) \qquad (4.2.46)$$

$$= -E(X|\pi_2)$$

şeklindedir. Variyansı ise

$$Var(X|\pi_1) = w_1^2 - \left[E(X|\pi_1)\right]^2$$

$$= \frac{2\delta^2}{\pi}\left(1-\frac{\delta}{2\sqrt{2\pi}}\right)^2\left(1-\frac{\delta^2}{2\pi}\right) \qquad (4.2.47)$$

$$= Var\,(X|\pi_2)$$

bulunur. X-değişkeni için Mahalonobis mesafesini Δ_1^2 ile gösterirsek

$$\Delta_1^2 = \frac{\left[E(X|\pi_1) - E(X|\pi_2)\right]^2}{Var\,(X)}$$

$$= \frac{2}{\pi} \cdot \frac{\delta^2}{1-\frac{\delta^2}{2\pi}} \qquad (4.2.48)$$

bulunur. Sürekli değişken için $\Delta^2 = \delta^2$ idi. Böylece sürekli değişkenin kesikli değişkene çevrilmesinden meydana gelen kayıp için Δ_1^2/Δ^2 oranı bir ölçü olarak kullanılırsa

$$\frac{\Delta_1^2}{\Delta^2} = \frac{2}{\pi} \cdot \frac{1}{1-\frac{\delta^2}{2\pi}} \qquad (4.2.49)$$

eşitliği elde edilir. Dikkat edilirse $\delta = 0$ için Δ_1^2/Δ^2 oranı

$\frac{2}{\pi}$ = 0.636 olur. δ büyüdükçe bu oran küçülür. Diğer bir ifade ile populasyonlar arasındaki mesafe büyüdükçe sürekli değişkenlerin kesikli hale getirilmeleri halinde ayırım gücü azalır.

4.3. Diskriminant Analizinde Kayıp Müşahadeler

Bazan kayıtların tutulmasındaki hata ve bunun gibi sebeplerden dolayı fertler üzerinde yapılan müşahadelerden bir kısmı kaybedilmiş olabilir. Böyle hallerde aynı fertteki kaybedilmemiş müşahadelerden faydalanarak elde edilecek malûmatı arttırmak için kayıpların tahmini, oldukça karışık bir meseledir. Bu durum diskriminant analizi ve çoklu regresyon gibi konularda daha da karışık bir durum gösterir. Çünkü değişkenlerin dağılışları çoğu zaman farklılık gösterirler.

Kayıp müşahadelerin olduğu fertleri denemeden ayrı tutmak şüphesiz en basit ve kestirme yoldur. Fakat mevcut verileri kullanmak suretiyle populasyonun parametrelerini en doğru bir şekilde tahmin edilmesi gerekmektedir.

Diskriminant analizi çerçevesinde bu konu şimdiye kadar çok az kimse tarafından incelenmiştir. Jackson (1968), kayıp müşahadelerin tahmini için iki yol tavsiye etmektedir. Birincisi, her kayıp değer için, ilgili değişkenin örnek ortalaması alınır. Kayıp müşahadelerin tahmini olarak kullanılan bu değer muhakkak ki birçok bakımlardan sakıncalıdır.

Jackson'un tavsiye ettiği ikinci metod ise şöyledir.

$$\underline{X}' = (\underline{X}_1, \ldots, \underline{X}_p) \qquad (4.3.50)$$

çoklu müşahadeleri gösteren matris olsun. X, nxp matris, $\underline{X}_1, \ldots \underline{X}_p$ ise kayıp müşahadeleri olan nx1 boyutlarındaki vektörlerdir.

Bunlardan biri (meselâ $\underline{X}_1$) bağımlı değişken, diğerleri bağımsız değişken olarak seçilip çoklu regresyon yapılır. Bağımsız değişkenlerdeki kayıp müşahadeler için, o değişkenin ortalaması koyulur. Böylece bağımlı değişkendeki bütün eksik kısımlar tamamlanır. Bu işlem diğer bütün değişkenler için ayni şekilde yürütülür. Bir önceki tahminlerle sonraki tahminler arasındaki fark

istenilen seviyeye düşünceye kadar işlemlere devam edilir.

Jackson'un dayandığı teorik esas analitik olarak ispatlanmamış fakat ampirik olarak sağlanmıştır. Metodun dayandığı normallik faraziyesinin birçok hallerde bozulduğu çok büyük bir örnek üzerine çalışılarak ayrıca belirtilmiştir.

Araştırmamızda, kayıp müşahadelerin tahmini $\underline{X}$ in çok-değişkenli normal dağılış gösterdiği faraziyesi esas alınarak yapılacaktır.

$$\underline{X} = \underline{A\theta} \tag{4.3.51}$$

lineer modelini alalım. Modeldeki $\underline{\theta}$, mxp parametre matrisi, $\underline{A}$ ise aşağıdaki gibidir.

$$\underline{A} = \begin{vmatrix} 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{vmatrix} \Big\} n_1 \text{ tane} \qquad \begin{vmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \end{vmatrix} \Big\} n_2 \text{ tane} \tag{4.3.52}$$

n_1, n_2 her gruptaki fert sayısını göstermektedir. Kayıp müşahadelerin tahmini bu modele göre yapılacaktır. Önce bir değişkenli modeli ($\underline{X}$, bu takdirde nxl vektördür) olarak hata (residual) kareler toplamını bulalım: $\underline{\beta}$ nın en küçük kareler tahmini bilindiği üzere

$$\underline{A}'\underline{A}\,\hat{\underline{B}} = \underline{A}'\underline{x} \tag{4.3.53}$$

normal eşitliğinden bulunur. Fakat, çoğu zaman $\underline{A}'\underline{A}$ matrisinin tersi yoktur. Onun için $\underline{B}$, amaca uygun bir matris olmak üzere parametreler arasında

$$\underline{B}\,\underline{\beta} = \underline{0} \tag{4.3.54}$$

şeklinde bir sınırlandırma yapılarak, β nın tahmini

$$\hat{\underline{\beta}} = (\underline{A}'\underline{A}+\underline{B}'\underline{B})^{-1}\underline{A}'\underline{x} \tag{4.3.55}$$

bulunur. Hata kareler toplamı ise

$$S = (\underline{x}-\underline{A}\,\hat{\underline{B}})'(\underline{x} - \underline{A}\,\hat{\underline{B}})$$

$$= \underline{x}'\underline{x} - \underline{x}'\ \underline{A}\ \hat{\underline{B}}$$

$$= \underline{x}'\left[\ \underline{I} - \underline{A}\ (\underline{A}'\underline{A} + \underline{B}'\ \underline{B})^{-1}\ \underline{A}'\ \right]\ \underline{x} \qquad (4.3.56)$$

şeklini alır. $\underline{I}$ birim matrisini göstermektedir.

Buna göre kayıp müşahadeler hata kareler toplamını minimum yapmak suretiyle bulunur.

Şimdi, tekrar çok-değişkenli

$$\underline{X} = \underline{A}\ \underline{\Theta} \qquad (4.3.57)$$

lineer modelini inceliyelim. Benzer şekilde, buna ait hata kareler toplamının da

$$\begin{aligned} \underline{S}_e &= (\underline{X} - \underline{A}\ \hat{\underline{\Theta}})'\ (\underline{X} - \underline{A}\ \hat{\underline{\Theta}}) \\ &= \underline{X}'\left[\ \underline{I} - \underline{A}\ (\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1}\ \underline{A}'\ \right]\ \underline{X} \end{aligned} \qquad (4.3.58)$$

olduğu kolayca bulunabilir. $\underline{S}_e$, pxp kare matrisidir. Bir değişkenli halde kayıp müşahadelerin tahmini için S yi minimum yapan değerler seçilir. Fakat değişken sayısı birden fazla olduğu zaman mesele daha da karışık bir şekil alır. Çünkü bir değişkenli halde sadece bir sayıya tekabül eden $\underline{S}_e$ bu taktirde bir matristir.

Çok-değişkenli lineer modellerde kayıp müşahadelerin tahmini için Donald (1971), $\underline{S}_e$ nin izinin yâni köşegen elemanlarının toplamının minimum yapılmasını kriter olarak kullanmıştır. Ayrıca $\underline{S}_e$ nin determinantının alternatif bir kriter olabileceğine de işaret etmiş fakat bununla ilgili herhangi bir çalışma yapmamıştır.

Donald'ın çok-değişkenli lineer modeller için geliştirdiği metodu diskriminant analize tatbik ederek kayıp müşahadelerin nasıl tahmin edilebileceğini izah edelim. $\underline{S}_e$ nin izi $Tr(\underline{S}_e)$ şeklinde gösterilirse

$$\begin{aligned} Tr(\underline{S}_e) &= Tr\ \left\{\ \underline{X}'\left[\ \underline{I} - \underline{A}(\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1}\ \underline{A}'\ \right]\ \underline{X}\ \right\} \\ &= \sum_{j=1}^{p} \underline{X}'_j\left[\ \underline{I} - \underline{A}(\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1}\ \underline{A}'\ \right]\ \underline{X}_j \end{aligned} \qquad (4.3.59)$$

bulunur. Böylece yukarıdaki ifadeyi minimum yapan değerler kayıp

müşahadelerin yerine alınırlar. Dikkat edilirse $[\underline{I} - \underline{A}(\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1}\underline{A}']$ matrisi pozitif yarıbelirli (positive semi definite) dir. Yâni $\underline{X}_j'[\underline{I} - (\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1}\underline{A}']\underline{X}_j$ bütün $\underline{X}_j$ değerleri için sıfırdan büyüktür. Bu sebeple $\underline{X}_j'[\underline{I} - \underline{A}(\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1}\underline{A}']\underline{X}_j$ nin her terimi ayrı ayrı minimum yapılırsa $Tr(\underline{S}_e)$ de minimum olur. Durum her $\underline{X}_j$ için ayrı ayrı incelendiğinde problemin bir değişkenli modeldeki kayıp müşahadelerin tahmininin aynı olduğu görülür.

Donald'ın yukarıda izah edilen kriteri ile, diskriminant analizindeki kayıp müşahadeler için tatbik etmek istediğimiz "genelleştirilmiş variyans" kriteri mukayese edilecektir. Bu kriter, çok-değişkenli lineer model esas alınarak hesaplanan hata matrisi $\underline{S}_e$ nin determinantının minimum yapılması esasına dayanır (Öztürk, 1972). Genelleştirilmiş variyans kriterini tatbik etmeden önce bunun mahiyetini kısaca açıklayalım.

4.3.1. Genelleştirilmiş variyans

Çok-değişkenli istatistikteki variyans-kovariyans matrisi ($\underline{V}$) bir değişkenli istatistikteki variyansa (σ^2) tekabül eder. Diğer taraftan genelleştirilmiş variyans olarak adlandırılan $\underline{V}$ nin determinantı, çok-değişkenli istatistikte variyasyonun başka bir ölçüsü olarak kullanılır (Anderson, 1958; Wilks 1932). Aynı şekilde örnek vektörleri olan $\underline{X}_1, \ldots, \underline{X}_n$ in genelleştirilmiş variyansı

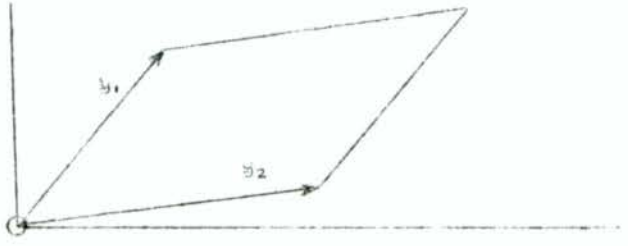
$$|\underline{S}| = \left| \frac{1}{n-1} \sum_{\alpha=1}^{n} (\underline{X}_\alpha - \bar{\underline{X}})(\underline{X}_\alpha - \bar{\underline{X}})' \right| \qquad (4.3.60)$$

şeklindedir. Bunu n-boyutlu uzaydaki p tane nokta cinsinden düşünerek geometrik izahını yapalım :

$$\begin{bmatrix} \underline{y}_1' \\ \vdots \\ \underline{y}_p' \end{bmatrix} = (\underline{Z}_1, \ldots, \underline{Z}_n), \qquad \underline{Z}_\alpha = \underline{X}_\alpha - \bar{\underline{X}} \qquad (4.3.61)$$

olsun. O zaman $\underline{y}_i' \underline{y}_i = a_{ii}$, i'nci vektörün uzunluğunun karesi, $\underline{y}_i' \underline{y}_j = a_{ij}, (i \neq j)$ ise i'nci ve j'nci vektörlerin uzunluklarıyla

bunlar arasındaki açının kosünüsünün çarpımlarına eşit olur.



Şekil 4.2. p= 2 Olduğu Zaman $\underline{y}_1$ ve $\underline{y}_2$ Vektörlerinin Şematik Olarak Gösterilişi

Yukarıdaki şekli inceliyelim: p= 2 olduğu zaman şekil bir paralel kenardır. Eğer p= 3 ise bu sefer kenarları $\underline{y}_1$, $\underline{y}_2$, $\underline{y}_3$ vektörleri olan bir paralel kenar pirizma elde edilir. Genel olarak p-tane vektör için paralel kenar, p-boyutlu uzayda bir şekildir. Bu (p-1) boyutlu paralel düzlem çiftlerinin kesişmesinden meydana gelmiştir.

Şimdi yukarıda izah edilen şeklin hacminin $\underline{A} = (n-1)\,\underline{S}$ nin determinantının kareköküne eşit olduğunu gösterelim.

p= 1 için $|\underline{Y}'\underline{Y}| = |\underline{A}| = \underline{y}_1'\underline{y}_1$ olup, determinantın karekökü $\underline{y}_1$ vektörünün uzunluğunu verir. Aynı şeyin p= k-1 için de doğru olacağını düşünerek p= k olduğu durumu inceliyelim.

k-boyutlu şeklin tabanı k-1 boyutludur. $\underline{y}_1,\ldots,\underline{y}_{k-1}$ vektörlerinin lineer kombinasyonu olan $c_1\underline{y}_1 + c_2\underline{y}_2 + \ldots + c_{k-1}\underline{y}_{k-1}$ vektörü k-boyutlu şeklin tabanında yer alır. En küçük mesafeyi $\underline{v} = \underline{y}_k - (c_1\underline{y}_1 + \ldots + c_{k-1}\underline{y}_{k-1})$ şeklinde tarif edersek bu k-boyutlu şeklin yüksekliği olur. $\underline{v}$ vektörünün minimum olabilmesi için $\underline{y}_1,\ldots,\underline{y}_{k-1}$ e ortogonal olması gerekir. Yâni, $c_1,\ldots,c_{k-1}$ katsayılarının

$$\underline{y}_j'\,\underline{v} = \underline{y}_j'\,\underline{y}_k - (c_1\underline{y}_j'\,\underline{y}_1 + \ldots + c_{k-1}\underline{y}_j'\,\underline{y}_{k-1}) \qquad (4.3.62)$$

$$= 0 \qquad j=1,\ldots,k-1$$

şartını yerine getirmesi lâzımdır. $\underline{Y}_{k-1}$, $\underline{Y}_k$ ve $\underline{C}$ matrisleri aşağı-

daki gibi olsun.

$$\underline{Y}_{k-1} = (\underline{y}_1 \cdots \underline{y}_{k-1})$$

$$\underline{Y}_k = (\underline{y}_1 \cdots \underline{y}_k)$$

$$\underline{C} = \begin{pmatrix} 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_{k-1} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \qquad (4.3.63)$$

$|\underline{C}| = 1$ ve $\left[\underline{Y}_{k-1}\ \underline{v}\right] = \underline{Y}_k\ \underline{C}$ olduğundan

$$|\underline{Y}'_k\ \underline{Y}_k| = |\underline{C}'||\underline{Y}'_k\ \underline{Y}_k||\underline{C}|$$

$$= |\underline{C}'\ \underline{Y}'_k\ \underline{Y}_k\ \underline{C}|$$

$$= \left| \begin{pmatrix} \underline{Y}'_{k-1} \\ \underline{v}' \end{pmatrix} (\underline{Y}_{k-1} \quad \underline{v}) \right| \qquad (4.3.64)$$

$$= \begin{vmatrix} \underline{Y}'_{k-1}\underline{Y}_{k-1} & \underline{0} \\ \underline{0} & \underline{v}'\underline{v} \end{vmatrix}$$

$$= |\underline{Y}'_{k-1}\underline{Y}_{k-1}|\underline{v}'\underline{v}$$

bulunur. Daha önce $\underline{Y}'_{k-1}\ \underline{Y}_{k-1}$ in k-1 boyutlu şeklin hacminin karesi olduğunu kabul etmiştik. Diğer taraftan tabanı k-1 boyutlu şekil olan k-boyutlu şeklin yüksekliği $\underline{v}$ nin uzunluğuna eşit olduğundan bunun hacmi

$$H = (|\underline{Y}'_{k-1}\ \underline{Y}_{k-1}|\ \underline{v}'\underline{v})^{\frac{1}{2}} \qquad (4.3.65)$$

şeklinde bulunur. O halde $|\underline{Y}'\ \underline{Y}| = |\underline{A}|$ nın karekökü p-boyutlu şeklin hacmine eşittir.

4.3.2. Metodların mukayesesi

Cetvel 3.2. deki verileri alarak heriki metodu da kullanıp, kayıp müşahadeleri tahmin edelim. Bu maksatla X_1 değişkenine ait birinci populasyondaki üçüncü müşahade ($y_1 = 1.946$) ile X_2değişkenine ait ikinci populasyondaki dokuzuncu müşahadenin ($y_2 : 1.411$) kayıp olduğu farzedilecektir.

a) Donald'ın kriterine göre kayıp müşahadelerin tahmini :

Deneme matrisi diye adlandırılan 20x3 boyutundaki $\underline{A}$,

$$\underline{A} = \begin{pmatrix} 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \end{pmatrix} \begin{matrix} \} \text{ 10 tane} \\ \\ \} \text{ 10 tane} \end{matrix} \qquad (4.3.66)$$

şeklindedir. Bunda son iki sütunun toplamı birinci sütuna eşit olduğundan $\underline{A}'\underline{A}$ matrisinin tersi yoktur. (4.3.54) de tarif edilen $\underline{B}$ matrisi

$$\underline{B} = (0 \quad 1 \quad 1) \qquad (4.3.67)$$

şeklinde seçilirse

$$\underline{A}'\underline{A} + \underline{B}'\underline{B} = \begin{pmatrix} 20 & 10 & 10 \\ 10 & 11 & 1 \\ 10 & 1 & 11 \end{pmatrix} \qquad (4.3.68)$$

elde edilir. Bazı işlemlerden sonra

$$\underline{A}(\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1}\underline{A}' = \frac{1}{10} \begin{pmatrix} \underline{J} & \vdots & \underline{0} \\ \cdots & \vdots & \cdots \\ \underline{0} & \vdots & \underline{J} \end{pmatrix} \qquad (4.3.69)$$

bulunur. Eşitliğin sağında 10x10 boyutundaki $\underline{J}$ ve $\underline{0}$ matrisleri aşağıdaki gibidir.

$$\underline{J} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & & 1 \end{pmatrix}$$

$$\underline{0} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 0 \end{pmatrix} \qquad (4.3.70)$$

İki değişken için $\underline{S}_e$ nin izi

$$Tr(\underline{S}_e) = \underline{X}_1' \left[\underline{I} - \underline{A}(\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1} \underline{A}' \right] \underline{X}_1 + \underline{X}_2' \left[\underline{I} - \underline{A}(\underline{A}'\underline{A} + \underline{B}'\underline{B})^{-1} \underline{A}' \right] \underline{X}_2 \qquad (4.3.71)$$

olur. Diğer taraftan $\underline{X}_{11}'$, $\underline{X}_{12}'$, $\underline{X}_{21}'$ ve $\underline{X}_{22}'$ 1x10 boyutundaki vektörleri göstermek üzere

$$\underline{X}_1' = (\underline{X}_{11}' : \underline{X}_{12}')$$

$$\underline{X}_2' = (\underline{X}_{21}' : \underline{X}_{22}') \qquad (4.3.72)$$

şeklinde yazılırsa

$$Tr(\underline{S}_e) = \underline{X}_1' \underline{X}_1 + \underline{X}_2' \underline{X}_2 - \frac{1}{10} (\underline{X}_{11}' \underline{J} \underline{X}_{11} + \underline{X}_{12}' \underline{J} \underline{X}_{12} + \underline{X}_{21}' \underline{J} \underline{X}_{21} + \underline{X}_{22}' \underline{J} \underline{X}_{22}) \qquad (4.3.73)$$

$$= \sum_{i=1}^{20} X_{1i}^2 - \frac{1}{10} \left[\left(\sum_{i=1}^{10} X_{11i} \right)^2 + \left(\sum_{i=1}^{10} X_{12i} \right)^2 \right] \qquad (4.3.74)$$

$$+ \sum_{i=1}^{20} X_{2i}^2 - \frac{1}{10} \left[\left(\sum_{i=1}^{10} X_{21i} \right)^2 + \left(\sum_{i=1}^{10} X_{22i} \right)^2 \right]$$

elde edilir.

Birinci ve ikinci gruplardaki y_1 ve y_2 kayıp müşahadelerinin tahmin için $TR(\underline{S}_e)$ nin kısmî türevlerinin bunlara göre alınıp

sıfıra eşitlenmesi gerekir.

$$\frac{\partial\left[\mathrm{Tr}(\underline{S}_e)\right]}{\partial y_1} = 1.8\, y_1 - 2.953 = 0$$

$$\frac{\partial\left[\mathrm{Tr}(\underline{S}_e)\right]}{\partial y_2} = 1.8\, y_1 - 2.048 = 0 \qquad (4.3.75)$$

denklemleri çözülürse $y_1 = 1.640$ ve $y_2 = 1.138$ değerleri bulunur. Bunlar populasyon ortalamaları olan 1.500 ve 1.200 değerleri ile mukayese edildiğinde, gerçek değerlerine çok yakın oldukları anlaşılır.

b) Genelleştirilmiş variyans kriterine göre tahmin: Bir önceki metodda olduğu gibi y_1 ve y_2 müşahadelerinin kayıp olduklarını farzedelim. Hesaplanan $\underline{S}_e$ matrisi aşağıdaki gibidir.

$$\underline{S}_e = \begin{bmatrix} (4.084-2.953y_1+0.9y_1^2) & (-0.560+0.118y_1+0.417y_2) \\ (-0.560+0.118y_1+0.417y_2) & (1.937-2.048y_2+0.9y_2^2) \end{bmatrix} \qquad (4.3.76)$$

Genelleştirilmiş variyans kriteri bunun determinantı olup, kayıp müşahadeleri tahmin etmek için $\underline{S}_e$ nin minimum yapılması gerekir. Bazı işlemlerden sonra

$$\frac{\partial|\underline{S}_e|}{y_1} = -5.588 + 3.458y_1 + 5.950\, y_2 - 3.686y_1y_2$$

$$-2.658\, y_2^2 + 1.620\, y_1y_2^2 \qquad (4.3.77)$$

$$\frac{\partial|\underline{S}_e|}{\partial y_2} = -7.897 + 5.950\, y_1 + 7.004\, y_2 - 5.316\, y_1y_2$$

$$-1.843\, y_1^2 + 1.620\, y_1^2y_2$$

bulunur.

Görüldüğü gibi yukarıdaki iki bilinmiyenli denklem sistemleri ikinci dereceden olup çözümleri basit değildir. Bununla beraber çok küçük bir hata ile y_1 ve y_2 aşağıdaki gibi çözülür. Birinci ve ikinci denklemlerden y_1 ve y_2 çekilirse

$$y_1 = \frac{2.658\ (2.102 - 2.239\ y_2 + y_2^2)}{1.620\ (2.134 - 2.275\ y_2 + y_2^2)}$$

$$y_2 = \frac{1.843\ (4.285 - 3.228\ y_1 + y_1^2)}{1.620\ (4.323 - 3.281\ y_1 + y_1^2)} \qquad (4.3.78)$$

bulunur. Her iki eşitliğin sağ tarafındaki pay ve payda daki parantez değerleri y_1 ve y_2 nin belirli değişim sınırları dahilinde eşit olarak alınırsa

$$y_1 = 1.641 \text{ ve } y_2 = 1.138$$

bulunur. Bu değerler Donald'ın kriterini kullanarak elde edilen değerlerin hemen hemen aynıdır.

Ele aldığımız misalde çeşitli müşahadelerin kaybedildiğini farzederek yapılan hesaplamalar yukarıdaki gibi neticeler vermiştir. Bu tip çalışmalar ikiden fazla değişken içinde yapılarak metodların birbirlerine olan durumları hakkında kesin bir hükme varılabilirdi. Fakat iki değişkenden fazla haller için S_e nin determinantını minimum yapan değerleri bulmak oldukça zordur. Bunun için komputer hesaplamalarına dayalı, fonksiyonu minimum yapan tekniklere başvurmak gerekir.

Donald'ın kriterinin diskriminant analizi için kullandığımız hesaplamalarda sağladığı kolaylıklar şüphesiz arzu edilen bir özelliktir. Fakat itiraz ettiğimiz nokta, bu kriterin değişkenler arasındaki kovariyansı dikkate almamasıdır. Halbuki bu durum çok-değişkenli istatistikte oldukça önemlidir.

4.4. Değişkenlerin Seçimi

Biyometri ve diğer bazı konularla ilgili araştırmalarda,

bir ferde ait müşahadeler genellikle mümkün olduğu kadar çok değişken üzerinde yapılır. Ölçülerin bu derece çok sayıda alınmasında en önemli sebep üzerinde çalışılacak ferdin teminindeki güçlüktür. Bütün değişkenler ölçüldükten sonra, araştırıcı bunlar arasından en önemlilerini seçmek ister. Çünkü fazla sayıda değişkenle çalışmak, çözülmek istenen problemin daha karışık bir durum göstermesine yol açar.

Diskriminant analizinde en uygun değişkeni yâni ayırıma konu olan populasyonların mukayesesinde en fazla farklılığı meydana getirenleri seçmek çok geniş ve önemli bir problemdir. Değişkenlerin mümkün olan bütün kombinasyonlarını kullanarak elde edilecek diskriminant fonksiyonlarından

$$\alpha = \Phi\left(-\frac{1}{2}\Delta\right) \qquad (4.4.79)$$

değerleri hesaplanıp, bunlar arasından en küçük α yı veren kombinasyon seçilebilir. Diğer bir ifade ile, en küçük yanlış sınıflandırma ihtimalini veren kombinasyondaki değişkenler, amaç için en uygun olanlarıdır.

Populasyonun parametreleri bilinmediği zaman örnek tahminlerinin kullanılacağını bundan önceki bölümlerde belirtmiştik. Böyle haller için çeşitli metodlar geliştirilmiştir. Bunların dördü ile, ayrı olarak tatbik etmek istediğimiz bir metod mukayese edilecektir.

4.4.1. Değişkenlerin seçiminde kullanılan metodlar

Bu konudaki metodları kısaca izah ettikten sonra bir misal alınarak mukayese yapılacaktır.

a) t-değerine göre seçim (Weiner ve Dunn 1966): İki gruptaki ortalamaların farklarını kullanmak suretiyle elde edilen Student'in t-değeri her değişken için hesaplanır. Buna göre

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, S} \qquad (4.4.80)$$

hesaplanır. Formüldeki $\bar{X}_1$, $\bar{X}_2$, birinci ve ikinci gruplardaki ortalamayı, S^2 ise iki grubun karıştırılmış variyansını yani,

$$S^2 = \frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_1 + n_2 - 2} \qquad (4.4.81)$$

formülüne göre elde edilen değeri göstermektedir. Böylece en büyük t-değeri olan değişkenler seçilir.

b) Diskriminant fonksiyonunun katsayıları (Weiner ve Dunn, 1966): Bu metodda diskriminant fonksiyonunun katsayısı

$$b_j = \sum_{k=1}^{p} r^{jk} \frac{\bar{X}_{1k} - \bar{X}_{2k}}{S_k} \qquad (4.4.82)$$

kriter olarak kullanılır. S_k^2, birinci metodda olduğu gibi karıştırılmış variyansı ve r^{jk} ise korelasyon matrisinin tersinin j'nci sıra ve k'ncı sütunundaki elemanı göstermektedir. Buna göre mutlak değeri en büyük b_j değerlerine sahip değişkenler seçilir.

c) İki grup arasındaki mesafe (Lebischew, 1962) : Mahalonobis mesafesinin örnekteki tahmini bir değişken için yazılırsa

$$D^2 = \frac{(\bar{X}_1 - \bar{X}_2)^2}{S^2} \qquad (4.4.83)$$

bulunur. Kullanılan bu değer t ye benziyorsa da örnekteki fert sayısına bağlı olmayışı bakımından farklıdır. İki grup arasındaki standardize edilmiş mesafeleri en büyük olan değişkenler ayırım için en etkili karakterler olarak seçilirler.

d) Ayırım gücündeki değişiklik (Urbakh, 1971): Bir değişkenin atılmasiyle D^2 de meydana gelecek değişiklik kriter olarak alınır. b_k ve r^{kk} (4.4.82) deki gibi olsun. Urbakh'ın ileri sürdüğü kriter

$$\frac{b_k^2}{r^{kk}} < \frac{D_p^2 + (n_1 + n_2 - 2p - 2)(\frac{1}{n_1} + \frac{1}{n_2})}{n_1 + n_2 - p - 2} \qquad (4.4.84)$$

şeklindedir. Yukardaki D_p^2, p-değişkene ait D^2 yi göstermektedir. Buna göre b_k^2/r^{kk} değerleri, eşitliğin sağ tarafında hesaplanan değerden büyük olan değişkenler seçilir.

e) Hata kareler toplamındaki (residual sum of squares) değişiklik : Kayıp müşahadelerin tahmininde, genelleştirilmiş variyans kriterini ileri sürerken, çok-değişkenli

$$\underline{X} = \underline{A}\ \underline{\Theta}$$

lineer modeli esas olarak alınmıştı. Şimdi ise, değişkenlerin seçiminde

$$Y = (\underline{X} - \bar{\underline{X}})\ \underline{\beta} \qquad (4.4.85)$$

çoklu regresyon denkleminden hareket ederek, hesaplanan hata kareler toplamındaki değişiklik kriter olarak alınacaktır. Bu kriteri izah etmeden önce, çoklu regresyon ile diskriminant fonksiyonu arasındaki ilişkiyi gösterelim.

$$y_\alpha^{(1)} = \frac{n_2}{n_1 + n_2} \qquad (\alpha = 1, \ldots, n_1)$$

$$y_\alpha^{(2)} = \frac{-n_1}{n_1 + n_2} \qquad (\alpha = 1, \ldots, n_2) \qquad (4.4.86)$$

olsun. Böylece tarif edilen y değişkeni sabit olup her örnekte farklı değerler alır. Gruplar arasındaki farkı dikkate almadan y'yi bağımlı değişken olarak düşünüp çoklu regresyon analizi yapabiliriz. X_i ve y_i lerin çarpımlarının toplamı

$$\frac{n_1\ n_2}{n_1 + n_2}\ (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) \qquad (4.4.87)$$

dir. Bu ise örnek ortalamaları arasındaki farka orantılıdır. Bu şartlar altında bulunan çoklu regresyon katsayılarının diskriminant fonksiyonunun katsayılarından farklı olmadıklarını kolayca ispat edebiliriz.

$$\underline{y}_{\alpha}^{(i)} = \hat{\underline{\beta}}'(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}}) \qquad (4.4.88)$$

modelini alalım. Yukardaki

$$\bar{\underline{X}} = \frac{n_1\bar{\underline{X}}^{(1)} + n_2\bar{\underline{X}}^{(2)}}{n_1 + n_2} \qquad (4.4.89)$$

dir. Normal eşitlikler ise

$$\sum_{i=1}^{2}\sum_{\alpha=1}^{n_i}(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}})(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}})'\hat{\underline{\beta}} = \sum_{i=1}^{2}\sum_{\alpha=1}^{n_i} y_{\alpha}^{(i)}(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}})$$

$$= \frac{n_1 n_2}{n_1+n_2}(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})$$

(4.4.90)

şeklini alır. Diğer taraftan

$$\underline{A} = \sum_{i=1}^{2}\sum_{\alpha=1}^{n_i}(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}})(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}})'$$

$$= \sum_{i=1}^{2}\sum_{\alpha=1}^{n_i}(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}}^{(i)})(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}}^{(i)})'$$

$$+ n_1(\bar{\underline{X}}^{(1)} - \bar{\underline{X}})(\bar{\underline{X}}^{(1)} - \bar{\underline{X}})' + n_2(\bar{\underline{X}}^{(2)} - \bar{\underline{X}})(\bar{\underline{X}}^{(2)} - \bar{\underline{X}})'$$

(4.4.91)

şeklinde yazılabilir. $\underline{S}_w = \sum_{i=1}^{2}\sum_{\alpha=1}^{n_i}(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}}^{(i)})(\underline{X}_{\alpha}^{(i)} - \bar{\underline{X}}^{(i)})'$ yazılırsa

$$\underline{S}_w\hat{\underline{\beta}} = (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})\left[\frac{n_1 n_2}{n_1+n_2} - \frac{n_1 n_2}{n_1+n_2}(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})'\hat{\underline{\beta}}\right]$$

(4.4.92)

bulunur. $(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})'\hat{\underline{\beta}}$ bir sabite olduğundan

$$\hat{\underline{\beta}} \propto \underline{S}_w^{-1}(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) \qquad (4.4.93)$$

dır. Tek değişkenli istatistikte olduğu gibi muameleler içi (hata) variyans-kovariyans matrisi

$$\underline{S}_w = \underline{A} - \frac{n_1 n_2}{n_1+n_2} \underline{d}\,\underline{d}' \qquad (4.4.94)$$

şeklinde ifade edilebilir. $\underline{d} = \bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}$ dir. Diskriminant katsayısı

$$\underline{b} \propto \underline{S}_w^{-1}\,\underline{d} \qquad (4.4.95)$$

idi. Buradan $\underline{d}$ çözülüp $\underline{A}$ cinsinden ifade edilirse

$$\underline{d} = \underline{S}_w \underline{b} = (\underline{A} - \frac{n_1 n_2}{n_1+n_2} \underline{d}\,\underline{d}')\,\underline{b}$$

$$= \underline{A}\,\underline{b} - \frac{n_1 n_2}{n_1+n_2} \underline{d}\,\underline{d}'\,\underline{b} \qquad (4.4.96)$$

bulunur. Dikkat edilirse $\underline{d}'\,\underline{b}$ bir sabitedir. Bu g ile gösterilirse

$$\underline{d} = \underline{A}\,\underline{b} - g \frac{n_1 n_2}{n_1+n_2} \underline{d} \qquad (4.4.97)$$

olur. Bazı kısaltmalardan sonra

$$\underline{b} = \underline{A}^{-1}\,\underline{d}\,(1 + \frac{n_1 n_2}{n_1+n_2} g) \qquad (4.4.98)$$

bulunur. Diğer taraftan çoklu regresyon katsayısı β da $\underline{A}^{-1}\underline{d}$ ye orantılıdır.

O halde; fertlerin iki normal populasyondan birine sınıflandırılmaları bahis konusu olduğu hallerde, y-değişkenine yukarıda tarif edildiği gibi değerler verilmek suretiyle bulunan çoklu regresyon denklemi ile diskriminant fonksiyonu aynı şeyi ifade ederler.

Bu noktadan hareketle, değişkenlerin seçiminde çoklu regresyon

için geliştirilmiş metodlar diskriminant analizinde de tatbik edilebilir.

Şimdi,

$$\underline{Y} = (\underline{X} - \bar{\underline{X}})\,\underline{\beta} \qquad (4.4.99)$$

modelini ele alalım. $\underline{Y}' = (y_1^{(1)}, \ldots, y_{n_1}^{(1)}, y_1^{(2)}, \ldots, y_{n_2}^{(2)})$ ve $\underline{\beta}' = (\beta_1, \ldots, \beta_p)$ dir. $(\underline{X} - \bar{\underline{X}})$ matrisi ise aşağıdaki gibidir.

$$(\underline{X} - \bar{\underline{X}}) = \begin{pmatrix} x_{11}^{(1)} - \bar{x}_1 & \cdots & x_{p1}^{(1)} - \bar{x}_p \\ \vdots & & \vdots \\ x_{1n_1}^{(1)} - \bar{x}_1 & \cdots & x_{pn_1}^{(1)} - \bar{x}_p \\ x_{11}^{(2)} - \bar{x}_1 & \cdots & x_{p1}^{(2)} - \bar{x}_p \\ \vdots & & \vdots \\ x_{1n_2}^{(2)} - \bar{x}_1 & \cdots & x_{pn_2}^{(2)} - \bar{x}_p \end{pmatrix} \qquad (4.4.100)$$

Buna göre $\underline{\beta}$ nın en küçük kareler tahmini

$$\hat{\underline{\beta}} = \left[(\underline{X} - \bar{\underline{X}})'(\underline{X} - \bar{\underline{X}}) \right]^{-1} (\underline{X} - \bar{\underline{X}})'\, \underline{Y} \qquad (4.4.101)$$

olarak bulunur. Hata kareler toplamı ise

$$\begin{aligned} S_r &= \underline{Y}'\underline{Y} - \underline{Y}'\,(\underline{X}-\bar{\underline{X}})\,\hat{\underline{\beta}} \\ &= \underline{Y}' \left\{ \underline{I} - (\underline{X}-\bar{\underline{X}}) \left[(\underline{X}-\bar{\underline{X}})'(\underline{X}-\bar{\underline{X}}) \right]^{-1} (\underline{X}-\bar{\underline{X}})' \right\} \underline{Y} \end{aligned} \qquad (4.4.102)$$

şeklindedir. Çoklu regresyon denkleminde değişkenler seçilirken, hata kareler toplamında meydana gelen değişikliğe bakılır. Bir değişken hariç tutulduğunda hata kareler toplamındaki artış en küçük ise, ilgili değişkenin diğerlerine nazaran daha az öneme sahip olduğu anlaşılır. Böylece en az farklılığı meydana getiren

değişkenler seçilerek regresyon denklemi dışında bırakılır.

Diskriminant fonksiyonunda da aynı durumun mevcut olması mümkün değildir. Çoklu regresyonda y_i lerin normal dağılış gösterdikleri kabul edilir. Halbuki diskriminant fonksiyonundaki y_i ler $-\frac{1}{2}$ ve $\frac{1}{2}$ değerlerini alan sun'i değişkenlerdir. Faraziyedeki bu farklılık sebebiyle hata kareler toplamında meydana gelecek artışın önemli olup olmadığının kontrolü çok güç bir durum göstermektedir. Bununla beraber, seçimdeki amacımız değişkenleri önemlilik sırasına göre dizmek olduğundan, hata kareler toplamında meydana gelecek artışın önemliliği bahis konusu değildir.

4.4.2. Metodların mukayesesi

Tatbik etmek istediğimiz yukarıdaki metodu, cetvel 3.3. deki verileri kullanarak diğerleriyle mukayese edelim.

$[(\underline{X} - \bar{\underline{X}})'(\underline{X}-\bar{\underline{X}})]$ matrisinin tersi aşağıdaki gibi bulunmuştur.

$$[(\underline{X} - \bar{\underline{X}})'(\underline{X}-\bar{\underline{X}})]^{-1} = \begin{pmatrix} 0.2881 & -0.0867 & -0.1243 & -0.0721 \\ -0.0867 & 0.2465 & -0.0562 & -0.0349 \\ -0.1243 & -0.0562 & 0.2476 & 0.0211 \\ -0.0721 & -0.0349 & 0.0211 & 0.2202 \end{pmatrix} \quad (4.4.103)$$

Diğer taraftan

$$\underline{Y}'(\underline{X}-\bar{\underline{X}}) = (7.296 \quad 4.139 \quad 4.293 \quad 0.435) \quad (4.4.104)$$

olarak hesaplanmıştır. Bunları kullanarak (4.4.102) den $S_r = 3.8603$ bulunur. Bundan sonra en iyi üç değişkeni seçmek için (X_1, X_2, X_3), (X_1, X_3, X_4), (X_1, X_2, X_4) ve (X_2, X_3, X_4) kombinasyonları için S_r değerleri ayrı ayrı hesaplanır. Misal olarak (X_1, X_2, X_3) kombinasyonunu alalım. 3x3 boyutundaki $(\underline{X}-\bar{\underline{X}})'(\underline{X}-\bar{\underline{X}})$ matrisinin tersi aşağıdaki gibidir.

$$[(\underline{X}-\bar{\underline{X}})'(\underline{X}-\bar{\underline{X}})]^{-1} = \begin{pmatrix} 0.2645 & -0.0981 & -0.1174 \\ -0.0981 & 0.2410 & -0.0529 \\ -0.1174 & -0.0529 & 0.2456 \end{pmatrix} \quad (4.4.104)$$

Ayrıca

$$\underline{Y}'(\underline{X}-\bar{\underline{X}}) = (7.296 \quad 4.139 \quad 4.293) \qquad (4.4.105)$$

değerleri kullanılırsa, dördüncü değişkenin atılmasından sonra, hesaplanan hata kareler toplamı $S_{r4} = 4.9237$ olarak bulunmuştur. Aynı şekilde

(X_1,X_2,X_4) kombinasyonu için $S_{r3} = 3.8788$

(X_1,X_3,X_4) kombinasyonu için $S_{r2} = 3.9335$

(X_2,X_3,X_4) kombinasyonu için $S_{r1} = 8.6784$

değerleri bulunur. Bu neticelerden, hata kareler toplamındaki en az artışın (X_1,X_2,X_4) kombinasyonunun kullanılmasıyle meydana geldiği anlaşılır. O halde diskriminant fonksiyonu için X_1,X_2 ve X_4 değişkenleri seçilebilir. Bu üç değişkenden hangisinin daha az önemli olduğunu bulmak için işlemlere aynı şekilde devam edilir. Çeşitli kombinasyonlara ait $S_{r'}$ değerleri aşağıdaki gibi bulunmuştur.

(X_1,X_2) kombinasyonu için $S_{r'4} = 4.9435$

(X_1,X_4) kombinasyonu için $S_{r'2} = 3.9384$

(X_2,X_4) kombinasyonu için $S_{r'1} = 9.6825$

Böylece X_2 ye nazaran X_1 ve X_4 değişkenlerinin daha önemli oldukları anlaşılır.

X_1 ve X_4 ayrı ayrı kullanılarak hesaplanan $S_{r''}$ değerleri

X_1 için $S_{r''4} = 4.9517$

X_4 için $S_{r''1} = 12.4635$

şeklinde bulunarak en önemli değişkenin X_1 olduğu sonucuna varılır.

Değişkenlerin seçimi için bundan önce izah edilen dört ayrı metoda göre elde edilen neticelerle birlikte bütün sonuçlar cetvel 4.10. da gösterilmiştir.

Cetvel 4.10. Çeşitli Metodlara Göre Seçilen Değişkenlerin Önemlilik Derecelerine Göre Sıralanışları

Metodlar	Değişkenlerin sırası
t-değeri	X_1, X_2, X_3, X_4
Diskriminant katsayıları	X_1, X_4, X_2, X_3
Gruplar arasındaki mesafe	X_1, X_2, X_3, X_4
Ayırım gücündeki değişiklik	X_1, X_4, X_2, X_3
S_r deki değişiklik	X_1, X_4, X_2, X_3

Yukarıdaki cetvelde görüldüğü gibi birinci ve üçüncü metodlar aynı neticeyi vermiştir. Ayrıca ikinci, dördüncü ve beşinci metodların da birbirlerinden farksız oldukları anlaşılmaktadır. Birinci grupta olan metodlardaki değişkenlerin sırası X_1, X_2, X_3, X_4, diğerinde ise X_1, X_4, X_2, X_3 şeklindedir. Diskriminant fonksiyonu için bunların hangisinin daha uygun olduğunu populasyona ait hata nispetlerini hesaplamak suretiyle bulabiliriz. (X_1, X_2) ve (X_1, X_2, X_3) kombinasyonlarına ait hata nispetlerinin sırasiyle 0.1056 ve 0.1003 olduğu cetvel 4.2. de gösterilmişti. Aynı şekilde (X_1, X_4) ve (X_1, X_4, X_2) kombinasyonları için de sırasiyle 0.0934 ve 0.0885 bulunur.

Heriki hal için de yeni bulunan 0.0934 ve 0.0885 hata nispetleri öncekilerden daha küçük olduğundan; ikinci, dördüncü ve beşinci metodun daha doğru olduğu sonucuna varılır. Seçilen bu üç metodu birbirlerine göre mukayese edebilmek için fazla sayıda değişken alarak daha geniş bir çalışma yapmak gerekir.

Ö Z E T

Bu araştırmada diskriminant analizinin genel teorisi incelenerek değişkenlerin seçimi, yanlış sınıflandırma ihtimalleri ve kayıp müşahadelerin tahmini gibi konularda bazı çalışmalar yapılmıştır.

Üzerinde çok sayıda ölçüler alınan bir ferdi, bu ölçülere dayanarak mensub olduğu populasyona sınıflandırırken mümkün olduğu kadar az hata yapılması arzu edilir. Bu maksatla, yanlış sınıflandırma maliyetinin beklenen değerini minimum yapan bir sınıflandırma kaidesi bulunmuştur.

Çok-değişkenli iki normal populasyon için yapılan sınıflandırmada Anderson'un ileri sürdüğü aşağıdaki istatistik kullanılır.

$$W(\underline{X})=(\underline{u}^{(1)}-\underline{u}^{(2)})\underline{V}^{-1}\underline{X} - \frac{1}{2}(\underline{u}^{(1)}+\underline{u}^{(2)})'\underline{V}^{-1}(\underline{u}^{(1)}-\underline{u}^{(2)}) \gtrless \ln k$$

formüldeki $\underline{u}^{(1)}$, $\underline{u}^{(2)}$ ortalama vektörlerini $\underline{V}$ ise variyans-kovariyans matrisini gösteren parametrelerdir. Bu parametreler bilinmediği zaman, örnek tahminleri kullanılır. Yukarıdaki istatistik, Fisher'in lineer diskriminant fonksiyonundan sadece sabit bir terim bakımından farklıdır. Populasyon sayısının ikiden fazla olduğu haller için sınıflandırma kaidesi genelleştirilmiştir. Aynı kaide kullanılarak kalitatif verilerle sınıflandırmanın nasıl yapılacağı izah edilmiştir.

Araştırmada öğrenci kayıtlarından ve normal dağılış gösteren şans sayılarından elde edilen veriler kullanılmıştır.

Hata nispetlerini tahmin için geliştirilmiş olan metodlardan beş tanesi alınarak mukayese edilmiş, bunlardan ikisinin en iyi olduğu sonucuna varılmıştır.

Kalitatif verilerle yapılan sınıflandırmada iki ayrı metodun uygulaması gösterilmiştir. Ayrıca bir normal değişkenin kalitatif şekle getirilmesinden dolayı ayırım gücünde meydana gelen değişiklik incelenmiştir.

Kayıp müşahadelerin tahmini için ileri sürülen iki metodun mukayeseleri yapılarak bunların birbirlerine olan üstünlükleri izah edilmiştir.

Araştırmanın son kısmında diskriminant fonksiyonu için kullanılan değişkenlerin seçiminde bazı metodlar incelenmiştir. Diskriminant fonksiyonu ile çoklu regresyon arasındaki ilişki de gösterilerek değişkenlerin seçimi için regresyon analizindeki tekniklerden birinin kullanılmasının mümkün olabileceği belirtilmiştir.

SUMMARY

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

The problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several populations on the basis of these measurements. In constructing a precedure of classification it is desired to minimise the probability of misclassification. For this purpose a classification precedure was obtained by minimising the expected cost of misclassification.

Anderson proposed the following statistics for the classification 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 $\underline{u}^{(1)}$, $u^{(2)}$ and 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 invoking the simulation techniques and gathering observations on the student records.

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

Application of two different methods 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.

SUMMARY

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

The problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several populations on the basis of these measurements. In constructing a precedure of classification it is desired to minimise the probability of misclassification. For this purpose a classification precedure was obtained by minimising the expected cost of misclassification.

Anderson proposed the following statistics for the classification 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 invoking the simulation techniques and gathering observations on the student records.

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

Application of two different methods 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 research some methods which are used to select 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.

Ashton, E.H, Healey, M.C.R ve Lipton, S.(1957) The Descriptive Use of Discriminant Functions in Physical Anthropology. Proc. Roy. Soc B 146, 552.

Bartlett, M.S.(1938), Further Aspects of The Theory of Multiple Regression, Proc. Camb. Phil. Soc. 34,33.

Blackith, R.E.(1960), A Synthesis of Multivariate Techniques to Distinguish Patterns of Growth in Grasshoppers. Biometrics, 16,28.

Burnaby, T.P.(1966), Growth-Invariant Discriminant Functions and Generalised Distances. Biometrics, 22,97.

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

Cramér, H.(1946), Mathematical Methods of Statistics. Princeton University Press, Princeton.

Dixon, J.N.J. ve Massey, F.(1969) Introduction to Statistical Analysis. Mc. Graw-Hill Book Company, New York.

Donald, L.Mc.(1971), On the Estimation of Missing Data in The Multivariate Linear Model. Biometrics, 27,535.

Fisher, R.A.(1936), The Use of Multiple Measurements in Taxonomic Problems. An.Eugen. London, 7,179.

Hills, M.(1966), Allocation Rules and Their Error Rates. Journal of Royal Statist. Soc. B 28,1.

Hotelling, H.(1931) The Generalisation of Student's Ratio. Ann. Math. Statist. 2,360.

Jackson,E.C.(1968), Missing Values in Linear Multiple Discriminant Analysis. Biometrics, 24,835.

Karataş, Ş.(1967), Atatürk Üniversitesi Merinos Sürüsünde Bazı Parametreler ve Tahmin Metodları. Atatürk Üniv. Zir. Fak. Zootekni Böl.

Kendall, M.G.(1957), A Course in Multivariate Analysis. Hafner Publishing Company, New York.

Kendall, M.G. ve Stuart,A.(1969) The Advanced Theory of Statistics. Charles Griffin and Company Limited, London.

Lachenbruch, P.A.(1968), On Expected Probabilities of Misclassification in Discriminant Analysis, Necessary Sample Size, Relation With the Multiple Correlation Coefficient. Biometrics, 24,823.

Lachenbruch, P.A. ve Mickey, M.R.(1968), Estimation of Error Rates in Discriminant Analysis. Technometrics, 10,1.

Lubischew, A.A.(1962), On the Use of Discriminant Functions in Taxonomy. Biometrics, 18,455.

Mahalanobis, P.C.(1948), Historical Note on The D^2-Statistic. Sankhya, 9,237.

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

Okamoto, M.(1963), An Assymptotic Expansion For The Distribution of The Linear Discriminant Function. Ann. of Math. Statist. 34,1286

Öztürk, A.(1972), Experimental Designs For Optimising Discrimination Between Two Simple Response Surface Models. M.Sc. Tezi. Reading Üniversitesi.

Rao, C.R.(1946), Tests With Discriminant Functions in Multivariate Analysis. Sankhya 7,407.

Rao,C.R.(1948), Tests of Signification in multivariate Analysis. Biometrika, 35, 58.

Rao,C.R.(1965), Linear Statistical Inference and Its Applications. John Wiley and Sons Inc. New York

Sezgin,F.(1972), Variyans Analizinin Dayandığı Faraziyeler ve Bunların Bozulmasından Doğan Durumlar. Doktora Tezi, Atatürk Üniversitesi Ziraat Fakültesi Zootekni Bölümü, Erzurum.

Siegel, S.(1956), Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Company Inc. New York.

Sitgreaves,R.(1952), On the Distribution of two Random Matrices Used in Classification Procedures. Annals of Math. Statist.23,263.

Urbakh, V. Yu(1971) Linear Discriminant Analysis: Loss of Discriminating Power When a Variate is Omitted. Biometrics. 27,531.

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.

Warner, H., Toronto, A., Veesey, L ve Stephenson, R.(1961), A Mathematical Approach to Medical Diagnosis. Journal of Amer. Med. Association, 177.

Weiner, J.M. ve Dunn, O.J.(1966), Elimination of Variables in Linear Discrimination Problems. Biometrics, 22,268.

Wilks, S.S.(1932), Certain Generalisation in The Analysis of Variance. Biometrika, 24,471.