4. AYIRMA (DİSKRİMİNANT) ANALİZİ

4.1 TANIM, AMAÇ ve ÖRNEKLER

Ayırma Analizi (AA) için farklı tanımlamalar yapmak mümkündür.

Tanım 4.1 Her birinde $p \geq 2$ tane değişken bulunan $g \geq 2$ tane grubu ele alalım. Değişkenler vektörü $\underline{X}' = [X_1 \quad X_2 \quad .... \quad X_p]$ olmak üzere $k = 1,2, \dots, g$ iken, k.ncı grup için bu vektörü $\underline{X}'_k = [X_1 \quad X_2 \quad .... \quad X_p]$ ile gösterelim. Bu gruplardan elde edilecek olan ve bu grupları birbirinden ayıran doğrusal fonksiyonlar yardımıyla, hangi gruba ait olduğu bilinmeyen $\underline{X}'_i = [X_{1i} \quad X_{2i} \quad .... \quad X_{pi}]$ gözlemini/gözlemlerini, bu gruplardan birine atamak için uygulanan istatistiksel analize Ayırma Analizi denir.

Bu tanıma göre, önce bağımsız ya da açıklayıcı değişkenlerin $(X_j, j = 1,2, \dots, p)$ doğrusal bir fonksiyonu ya da fonksiyonları elde edilmeli ve bu fonksiyonlar kullanılarak gözlemler bu gruplardan herhangi birisine sınıflandırılmalıdır.

Tanım 4.2 AA, grup ortalamaları arasındaki farklılığı en büyük yapacak şekilde bağımsız veya açıklayıcı değişkenlerin doğrusal fonksiyonlarını bulma işlemidir. Bu doğrusal fonksiyonlara diskriminant (ayırma) fonksiyonları denir.

AA'nde gruplar, özel araştırma grupları olarak önceden belirlenir. Grupları ayırt etmek için gruplardaki beklenen farklılığı ölçen özelliklere ilişkin ayırt edici değişkenler kümesi seçilir ve bu değişkenlerin doğrusal kombinasyonları üretilir. Bu doğrusal kombinasyonlar kullanılarak, yeni gözlemler ait oldukları gruplara sınıflandırılır.

Verilen tanımlar dikkate alındığında AA'nin amaçları ve kullanım yerleri şu şekilde özetlenebilir.

i) Grupları birbirinden ayırmayı sağlayacak olan diskriminant fonksiyonları bulmak

ii) Bulunan diskriminant fonksiyonları yardımıyla, yeni bir gözlemi en az hata ile gruplardan birine (ait olduğu gruba) atamak (grup üyeliklerini kestirmek)

iii) Çalışmaya alınan değişkenlerden hangilerinin grup üyeliğini kestirmekteki katkısının daha fazla olduğunu belirlemek

Kullanım yerleri için örnekler:

Örnek 4.1 Bir anestezi uzmanı, bir anestezi ilacının kalp ameliyatı geçirecek hastalarda güvenli olarak kullanılıp kullanılamayacağını belirlemek istesin. Anestezi uzmanı hastalara ilişkin Yaş, Kan basıncı, Ağırlık vb temel (veya özel) bilgileri (açıklayıcı değişkenlere ait değerleri) elde edebilir. Bu tür bilgiler, farklı özellikteki hastalarda anestezi ilacının ne tür etki göstereceğinin bilinmesi açısından önemlidir. Gönüllüler üzerinde denenmeden önce, böyle bir çalışmanın deney hayvanları (örneğin köpekler) üzerinde yapıldığını ve kalp ameliyatı geçirecek köpeklere ilişkin bazı temel bilgiler elde edildikten sonra, ameliyat sırasında anestezi ilacı açısından gözlenen köpeklerin, güvende olanlar ve olmayanlar olarak iki gruba ayrıldığını kabul edelim.

Buna göre, anestezi ilacı açısından güvende olan ve olmayan gruplarındaki köpeklerin temel (veya özel) değişkenlerinden yararlanarak oluşturulacak doğrusal fonksiyonlar yardımıyla, ameliyat edilecek olan ve ilgili değişken değerleri bilinen yeni bir köpek için anestezi ilacının güvenli olup olmadığı önceden kestirilebilir. Diğer bir ifadeyle, ameliyat edilecek köpek, bu iki gruptan birisine (ait olduğu gruba) sınıflandırılabilir.

Tablo 4.1 Hastalara ait açıklayıcı değişken ölçüm sonuçları

	Anestezi açısından güvende			Anestezi açısından güvende değil		
Birim No	Yaş	Kan basıncı	Ağırlık	Yaş	Kan basıncı	Ağırlık
1	---	---	---	---	---	---
2	---	---	---	---	---	---
.		.			.	
.		.			.	
.		.			.	
N	---	---	---	----	----	----

Örnek 4.2 Üç Farklı dönemde (Yaklaşık MÖ 4000, Yaklaşık MÖ 1850, Yaklaşık MS 150) elde edilen Erkek Mısır kafatasları üzerinde; X_1: Kafatasının genişliği (mm), X_2: Kafatası tabanı ile bregma noktası arasındaki yükseklik(mm), X_3: Kafatasının tabanı ile elveo arasındaki uzunluk(mm) ve X_4: Burun yüksekliği(mm) ölçülerek veri düzenlemesi yapılıyor. Bu veriler kullanılarak grupları ayıran doğrusal fonksiyonlar bulunup, bu doğrusal fonksiyonlar yardımıyla, yeni bulunan bir kafatasının aynı özelliklerini ölçerek hangi döneme ait olduğu istatistiksel olarak belirlenebilir.

Tablo 4.2 Üç farklı döneme ait kafatası ölçüm değerleri

	MÖ 4000				MÖ1850				MS150			
Birim No	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
1	---				---				---			
2	---				---				---			
.	.				.				.			
.	.				.				.			
.	.				.				.			
N	---				---				---			

Örnek 4.3 Bir kurumda çalışan iki farklı yardımcı personel (düz memurlar ve sekreterler) üzerinde yapılan bir çalışmada, personele üç farklı ölçek (X_1: ortam tercihi ölçeği; X_2: neşelilik-yalnızlık tercihi ölçeği ve X_3: tutuculuk-özgürlük tercihi ölçeği) uygulanmış ve her bir personel için bu ölçeklere ait skorlar gözlenmiştir. Bu skorlar yardımıyla, yeni işe alınacak bireylerin hangi grupta çalışmaya daha yatkın olduğunu belirlemek amacıyla AA kullanılabilir. Bu iki personel grubu arasında diskriminant fonksiyonları oluşturulur ve yeni işe alınacak ya da diğer personel arasında bu iki iş alanından birine aday olacaklar, alacakları ölçek puanları yardımıyla bu gruplardan birine sınıflandırılabilir.

Tablo 4.3 Personel alımı için adaylara ait bazı özellikler

Birim No	Memur			Sekreter		
	X_1	X_2	X_3	X_1	X_2	X_3
1	---			---		
2	---			---		
.	.			.		
.	.			.		
.	.			.		
N	---			---		

Örnek 4.4 Kanser tanısı koymak için kesin tanı koyma yöntemi hastalardan biyopsi yoluyla doku örneği alınmasıdır; ancak biyopsi yaklaşımı invaziv (yayılan) bir yöntemdir. Bu nedenle, daha ucuz ve gerçekleşmesi uzmanlık gerektirmeyen tanı araçlarının kullanılması tıp alanının güncelliğini yitirmeyen konularından biridir. Bu çerçevede yapılan bir çalışmada, prostat kanseri tanısı koymak amacıyla biyopsi alınmış hastaların, ucuz ve gerçekleştirilmesi uzmanlık gerektirmeyen X_1: Total PSA ve X_2: PSA yoğunluk değerleri de ölçülmüştür. Biyopsi sonucunda prostat kanseri tanısı konan (Biyopsi(+)) ve konmayan (Biyopsi(-)) gruplarındaki hastaların Total PSA ve PSA yoğunluk değerleri yardımıyla, bu grupları ayıran doğrusal fonksiyon bulunur. Bu fonksiyon kullanılarak yeni başvuran bir hastanın prostat kanseri olup olmadığı, biyopsi yapılmadan belirlenebilir.

Tablo 4.4 Hastaların biyopsi sonuçları

Birim No	Biyopsi(+)		Biyopsi(-)	
	X_1	X_2	X_1	X_2
1	---		---	
2	---		---	
.	.		.	
.	.		.	
.	.		.	
N	---		---	

4.2 Genel Bilgiler

AA ile ilgili bazı temel kavramlara bu kısımda yer vereceğiz.

4.2.1. Değişkenler: AA'inde değişkenler bağımlı değişken (tek) ve bağımsız (açıklayıcı) değişkenler ($p \geq 2$ tane) olmak üzere ikiye ayrılmaktadır. Bağımlı değişken iki veya daha fazla kategoriye sahip nitel türden sınıflama veya bazen nicel türden fakat sıralama düzeyinde ölçülen bir değişken olabilmekte iken, bağımsız değişkenler nicel türden kesikli ya da sürekli değişken olup ölçme düzeyleri en az eşit aralıklıdır.

Yukarıda Örnek:1.1 de bağımlı değişken (güven durumu: iki kategorili) iken, bağımsız değişkenler (yaş, kan basıncı, ağırlık); Örnek:1.2 de bağımlı değişken (Dönem: üç kategorili) iken, bağımsız değişkenler ($X_1 \quad X_2 \quad X_3 \quad X_4$) dür, v.s.

AA için öncelikle bağımlı değişkene karar verilmelidir. Bağımlı değişkenin kategorileri birbirinden tam bağımsız (mutually exclusive), yani ayrık olmalıdır. Bir diğer ifade ile her bir gözlem sadece bir kategoriye (gruba) yerleştirilebilen bir yapıda olmalıdır. Bağımlı değişken için gruplar oluşturulduktan sonra, gruplardaki bağımsız değişkenler belirlenmelidir. AA'nin amacı gruplar arasındaki farklılığı en büyük yapacak olan değişken ya da değişkenleri belirlemek ve kullanılan bağımsız değişkenlerin farklılığı temelinde grup üyeliklerini kestirmek olduğundan, AA'nin başarısı seçilen bağımsız değişkenler ile yakından ilgilidir. Bağımsız değişken seçim işlemi, hangi değişkenlerin grup üyeliği hakkında daha fazla bilgi sağlayacağı konusundaki teorik bilgiye bağlı olarak yapılabileceği gibi, maliyet, uygunluk ve uygulamada kolaylık sağlayacak özelliklere bağlı olarak da yapılabilir.

4.2.2. Örneklem Büyüklüğü: AA'nde örneklem büyüklüğü için genel bir yaklaşım, bağımlı değişkenin kategorilerini gösteren her bir grupta örneklem büyüklüğü bağımsız değişken sayısının en az 4 veya 5 katı olmasıdır. Gruplarda bağımsız değişken sayısından az gözlem olmamalıdır. Örneğin değişken sayısı 4 ya da 5 iken her bir grup için örneklem büyüklüğü en az 20 olmalıdır. Gruplarda gözlem sayılarının $(n_k, k = 1,2, \dots, g)$ birbirinden farklı olması AA'nde önemli bir problem oluşturmamakla birlikte, çok farklı örnek hacimlerine sahip gruplarla çalışmak tercih edilmemelidir. Böyle durumlarda gözlem sayısı çok fazla olan grup/ gruplardan, diğer grup/gruplarla karşılaştırma yapabilecek düzeyde rsatgele örnekler seçilebilir.

4.2.3. AA'nin Diğer Bazı Çok Değişkenli Analizlerle Benzerlikleri/Farklılıkları: AA ile çok değişkenli varyans analizi (MANOVA) arasında benzerlik/farklılıklar vardır. AA, MANOVA'nın tersi olarak kabul edilebilir. Çünkü; MANOVA'da bağımsız değişkenler gruplar iken her bir gruptaki değişkenler de bağımlı değişkenlerdir. AA ise tam tersi olduğu yukarıdaki açıklamalardan görülmektedir. Ayrıca; MANOVA sonucunda elde edilen istatistiksel önemlilik (gruplar arasında anlamlı bir farkın bulunması), yüksek oranda doğru sınıfla anlamına gelmez. Bu nedenle; anlamlı bir sınıflama ya da ayırım için istatistiksel önemlilik gerekli, fakat yeterli değildir.

AA ile lojistik regresyon analizi(LRA) arasında da benzerlik/farklılıklar vardır. LRA'nde bağımsız değişkenler için veri tipi, varyans-kovaryans matrislerinin homojenliği ve çok değişkenli normallik varsayımı aramazken, AA'nde bağımsız değişkenlerin sürekli (ya da kesikli) nicel veri tipinde olması, grupların homojen varyans-kovaryans matrisli ve gruplardaki bağımsız değişkenler bakımından dağılımların çok değişkenli normal dağılım göstermesi varsayımları aranır. Bu nedenle, özellikle gruplarla ilgili homojenlik varsayımının sağlanmadığı durumlarda LRA, AA'ne iyi bir alternatif olabilmektedir. LRA, bağımsız değişkenlerin bazılarının sürekli, bazılarının da nitel değişken olmaları durumunda da AA yerine kullanılabilmektedir.

Eğer bağımsız değişkenler bakımından gruplar normal dağılımlı fakat, homojen varyans-kovaryans matrisli değilse, bu durumda **kuadratik (karesel) ayırma analizi** yaklaşımı

kullanılabilmektedir. Bu yaklaşımda kuadratik ayırma fonksiyonu grupları ayırmada ve gözlemleri sınıflandırmada kullanılmaktadır.

4.2.4. Ayırma Analizleri Çeşitleri: AA çeşitleri bağımlı değişkendeki grup sayısına ve grup varyans-kovaryans matrislerinin homojenlik varsayımını sağlayıp sağlamamasına göre belirlenmektedir.

i) Doğrusal Ayırma Analizi: Grup varyans-kovaryans matrisleri homojen olduğunda kullanılır. Bağımlı değişkendeki grup sayısına göre ikiye ayrılır. Eğer $g = 2$ ise iki grup doğrusal ayırım analizi, eğer $g > 2$ ise çoklu grup doğrusal ayırım analizi kullanılır.

ii) Kuadratik (karesel) Ayırma Analizi: Grup varyans-kovaryans matrisleri homojen olmadığında kullanılır.

4.2.5. Ayırma Analizinin Aşamaları: AA iki aşamalı bir işlem olarak ele alınabilir:

i) Ayırma fonksiyonlarının önemlilik testi

ii) Sınıflama (yeni gözlemler için grup üyeliklerinin kestirilmesi)

İlk aşama MANOVA'dakine benzer hesaplamalardan oluşur. Tüm bağımsız değişkenler dikkate alınarak gruplar arasında önemli bir farklılık olup olmadığı ya da incelenen bağımsız değişkenlerin doğrusal fonksiyonu yardımıyla grupların ayırt edilip edilemeyeceği hipotez testi ile belirlenir. Bu amaçla toplam varyans kovaryans matrisi (S) ve ortak grup içi varyans kovaryans matrisi (S_{wg}) yardımıyla elde edilecek çok değişkenli F testlerinden veya Wilks'in Λ istatistiğinin Ki-kare yaklaşımından yararlanılır. Test sonucunda önemli bir farklılık bulunursa (yani ayırıcı fonksiyonların oluşturulabileceği kanısına varıldığında) hangi değişkenlerin modele katkısının daha önemli olduğu da belirlenebilir.

AA'nde elde edilecek fonksiyonlar, grup ortalamaları arasındaki farkın en büyük yapılması prensibine dayanır. Bu kapsamda, kanonik fonksiyonlardan ve Fisher ayırma fonksiyonlarından yararlanılır. Kanonik yaklaşımda, değişkenlerin en uygu kombinasyonu belirlenerek, birinci fonksiyonun gruplar arasında en iyi ayırımı yapması, ikinci fonksiyonun ikinci en iyi ayırımı yapması sağlanır … ve bu şekilde devam edilir. Yani diğer fonksiyonlarda kendisinden öncekilere göre daha düşük, fakat kendisinden sonrakilere göre daha yüksek ayırım yapacak şekilde oluşturulur. Buna ek olarak oluşturulan ayırım fonksiyonları, birbirinden bağımsız, yani birbirine diktir. Böylece fonksiyonların gruplar arasındaki ayırıma katkıları çakışmaz. İlk doğrusal fonksiyon değişkenliğin (varyansın) en büyük kısmını açıklarken, ikinci doğrusal fonksiyon birinci fonksiyon tarafından açıklanamayan değişimin en büyük kısmını açıklar be bu süreç bu şekilde devam eder.Ardışık olarak kanonik fonksiyonları ve kanonik kökleri elde etmek için Kanonik Korelasyon Analizine ihtiyaç duyulur.

Fisher ayırma fonksiyonları da grup ortalamaları arasındaki farkın en büyük yapılması prensibine dayanır. Ancak; bu yaklaşımda grup sayısı kadar ayırma fonksiyonu üretilir.

AA'nin ikinci aşaması olan sınıflama işlemi; kanonik ayırma fonksiyonları, çok değişkenli normal dağılım yoğunluk fonksiyonu ve Fisher'in sınıflama fonksiyonu kullanılarak yapılabilmektedir.

4.2.6. Ayırma Analizinin Varsayımları ve Kısıtlayıcıları: AA'nin iki temel varsayımı ve üç kısıtlayıcısı vardır.

Varsayımlar:

i) Bağımsız Değişkenlerin Çok Değişkenli Normal Dağılım Göstermesi: Bağımlı değişkenin kategorilerinde bağımsız değişkenler vektörüne ait dağılımın çok değişkenli normal dağılım ile uyumlu olması varsayılır. Bu varsayımın sağlanıp sağlanmadığı hipotez testi ile kontrol edilebilir. Ayrıca; bağımsız değişkenlerin herhangi bir doğrusal kombinasyonunun örnekleme dağılımının da normal dağıldığı varsayılır. Çok değişkenli normal dağılım varsayımının sağlanabilmesi için gerekirse bağımsız değişkenler üzerine dönüşümler uygulanarak bu varsayımın dönüşümler yardımıyla sağlanılmasına çalışılır. Eğer her şeye rağmen bağımsız değişkenlerin çok değişkenli normal dağılım varsayımı sağlanmıyorsa, ayırım fonksiyonları olarak kullanılacak olan doğrusal fonksiyonların kestiriminde sorunlar ortaya çıkar. Çünkü bu doğrusal fonksiyonlar bağımsız değişkenlerin fonksiyonu olacağından, bu durumda doğrusal fonksiyonların örnekleme dağılımı da normal olmaz. Böyle bir durumun ortaya çıkması halinde birimlerin sınıflandırılmasında ayırma analizi yerine lojistik regresyon analizi tercih edilmelidir.

Çok değişkenli bir kitleden çekilen n birimlik bir örneklemin dağılımının çok değişkenli normal dağılım ile uyumlu olup olmadığını incelemede kullanılan bazı yöntemler:

a) Grafik Yöntem

b) Mardia'nın basıklık testi

şeklinde incelenir.

a) Grafik Yöntem:

Değişken sayısının $p \geq 2$ olduğu çok değişkenli bir kitleden rastgele çekilen n birimlik bir örneklem $\underline{X}_1$, $\underline{X}_2$, … , $\underline{X}_n$ olsun. Test edilecek hipotezler:

H_0: Örneklem, $N_p\left(\underline{\mu}\,,\ \Sigma\right)$ dağılımı ile uyumludur

H_1: Örneklem, $N_p\left(\underline{\mu}\,,\ \Sigma\right)$ dağılımı ile uyumlu değildir (4.1)

şeklinde oluşturulur. Teorik dağılımın parametreleri $\left(\underline{\mu}\,,\ \Sigma\right)$ bilinmediğinden örneklemden tahmin edilmesi gerekir. $\underline{\mu}$ kitle ortalaması parametresinin tahmin edicisi $\overline{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n}\underline{X}_i$ örnek ortalama vektörü iken, Σ kitle varyans-kovaryans matrisi parametresinin tahmin edicisi $S = \frac{1}{n-1}\sum_{i=1}^{n}\left(\underline{X}_i - \overline{\underline{X}}\right)\left(\underline{X}_i - \overline{\underline{X}}\right)'$ örnek varyans-kovaryans matrisidir. $i = 1,2,\ldots,n$ için her bir gözlemin H_0 hipotezi altında teorik normal dağılımın ortalamasına olan Kare Mahalanobis

Uzaklıkları $m_i^2 = \left(\underline{X}_i - \overline{\underline{X}}\right)' S^{-1} \left(\underline{X}_i - \overline{\underline{X}}\right)$ 'ler hesaplanır. Bu uzaklıklar küçükten büyüğe doğru sıralanır, söz konusu sıralı dizi $m_{(1)}^2 \leq m_{(2)}^2 \leq \cdots \leq m_{(n)}^2$ olsun. Daha sonra $i = 1,2,\ldots,n$ için her bir gözleme karşılık $P_i = \frac{i-0,5}{n}$ olasılıkları ve $Pr\left(\chi_p^2 \leq \chi_i^2\right) = P_i$ bu olasılıklara karşılık gelen χ_i^2 değerleri, Ki-Kare dağılımının birikimli olasılık dağılımı yardımıyla hesaplanır. Bunun için paket programlardan yararlanabiliriz (Excell, Matlab, SPSS). Böylece $i = 1,2,\ldots,n$ için $\left(m_{(i)}^2,\ \chi_i^2\right)$ nokta ikilileri koordinat düzlemine yerleştirilir. Bu noktaların dağılımı yaklaşık olarak bir doğru üzerinde dağılım gösteriyorsa, bu takdirde H_0 kabul edilir ve örneklemin dağılımının, çok değişkenli normal dağılım ile uyumlu olduğu söylenir. Aksi takdirde; yani noktaların dağılımı bir doğru etrafında bir dağılım göstermiyorsa H_0 ret edilir ve örneklemin dağılımının, çok değişkenli normal dağılım ile uyumlu olmadığı söylenir.

Tablo 4.5 Çok Değişkenli Normal Dağılıma Uyumluluk Testi İşlem Tablosu

$\boldsymbol{i}$	$\boldsymbol{m_i^2}$	$\boldsymbol{m_{(i)}^2}$	$\boldsymbol{P_i = \frac{i-0,5}{n}}$	$\boldsymbol{\chi_i^2}$
1	m_1^2	$m_{(1)}^2$	P_1	χ_1^2
2	m_2^2	$m_{(2)}^2$	P_2	χ_2^2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	m_n^2	$m_{(n)}^2$	P_n	χ_n^2

b) **Mardia'nın basıklık testi:** Test edilecek hipotezler (4.1) tanımlandığı gibidir. Test istatistiği olarak kullanılacak olan çok değişkenli basıklık katsayısı;

$$\hat{\gamma}_{2p} = \frac{1}{n}\sum_{i=1}^{n} m_i^4 \tag{4.2}$$

dir. Burada $m_i^2 = \left(\underline{X}_i - \overline{\underline{X}}\right)' S^{-1} \left(\underline{X}_i - \overline{\underline{X}}\right)$ dir.

H_0 doğru iken ; test istatistiği olan çok değişkenli basıklık katsayısının örnekleme dağılımı için;

$$\hat{\gamma}_{2p} \sim N\left(p(p+2);\ \frac{8p(p+2)}{n}\right) \tag{4.3}$$

olup, test istatistiği,

$$Z = \frac{\hat{\gamma}_{2p} - p(p+2)}{\sqrt{\frac{8p(p+2)}{n}}} \sim N(0,1) \tag{4.4}$$

standart normal dağılım olacaktır. Eşitlik (4.4) ile hesaplanan değer Z_h olsun. $Pr(Z \leq Z_h) = p$ ve α önem seviyesi olmak üzere H_1-çift yönlü iken karar kuralı, eğer; $2p < \alpha$ ise H_0 ret edilir ve örneklemin dağılımının, çok değişkenli normal dağılım ile uyumlu olmadığı söylenir. Eğer;

$2p \geq \alpha$ ise H_0 kabul edilir ve örneklemin dağılımının, çok değişkenli normal dağılım ile uyumlu olduğu söylenir.

Örnek 4.5 En az bir canlı doğum yapmış kadınlar üzerinde X_1: Kadının hemoglobin düzeyi ve X_2: Kadının yaşı değişkenlerine ait ölçümler aşağıdaki gibi düzenlenmiştir. Bu iki değişkenli verinin $N_2\left(\underline{\mu}, \Sigma\right)$ dağılımı ile uyumlu olup olmadığına %5 önem seviyesinde karar veriniz?

Birim (Kadın)	X_1	X_2	**Birim (Kadın)**	X_1	X_2
1	13	20	9	10	42
2	14	25	10	12	30
3	12,5	40	11	10,5	35
4	12	22	12	10	28
5	12,5	33	13	11	25
6	12	35	14	9	40
7	11	21	15	11,5	33
8	10	25			

Çözüm Hipotezler;

H_0: Çok değişkenli örneklemin dağılımı, $N_2\left(\underline{\mu}\ ,\ \Sigma\right)$ dağılımı ile uyumludur.

H_1: Çok değişkenli örneklemin dağılımı, $N_2\left(\underline{\mu}\ ,\ \Sigma\right)$ dağılımı ile uyumlu değildir.

i) Grafik yöntemini uygulayalım. Her bir gözlemin örnek ortalama vektörüne olan kare Mahalanobis uzaklıklarını hesaplayabilmek için, önce örneklemin ortalama vektörü ile varyans kovaryans matrisini ve bu matrisin tersini bulmalıyız.

$$\overline{\underline{X}} = \begin{bmatrix} 11{,}400 \\ 30{,}267 \end{bmatrix} \text{ ve } S = \begin{bmatrix} 1{,}829 & -3{,}293 \\ -3{,}293 & 52{,}495 \end{bmatrix}, S^{-1} = \begin{bmatrix} 0{,}616360 & 0{,}038664 \\ 0{,}038664 & 0{,}021475 \end{bmatrix}$$

i.nci gözlemin örnek ortalamasına olan kare Mahalanobis uzaklığı:

$$m_i^2 = \left(\underline{X}_i - \overline{\underline{X}}\right)' S^{-1} \left(\underline{X}_i - \overline{\underline{X}}\right),\ i = 1,2,\dots,n$$

olup, $i = 1$ için $\underline{X}_1 = \begin{bmatrix} 13 \\ 20 \end{bmatrix}$, $\underline{X}_1 - \overline{\underline{X}} = \begin{bmatrix} 1{,}6 \\ -10{,}267 \end{bmatrix}$ ve

$m_1^2 = [1{,}6 \quad -10{,}267] \begin{bmatrix} 0{,}616360 & 0{,}038664 \\ 0{,}038664 & 0{,}021475 \end{bmatrix} \begin{bmatrix} 1{,}6 \\ -10{,}267 \end{bmatrix} = 2{,}571$ bulunur. Diğer gözlemler için de benzer şekilde uzaklıklar hesaplanabilir. Her bir birim için söz konusu uzaklıklar, bu uzaklıkların sıralı uzaklıkları, $P_i = \frac{i-0{,}5}{n}$ olasılıkları ve $Pr\left(\chi_p^2 \leq \chi_i^2\right) = P_i$ eşitliği ile bulunan χ_i^2 değerleri aşağıdaki Tablo 1.6'da verilmiştir.

$i = 1$ için χ_1^2 değeri $Pr(\chi_2^2 \leq \chi_1^2) = P_1 \implies Pr(\chi_2^2 \leq \chi_1^2) = 0{,}03 \implies \chi_1^2 = 0{,}061$ olup, diğerleri de benzer şekilde hesaplanabilir. $i = 1,2,\dots,15$ için $\left(\boldsymbol{m}_{(i)}^{\boldsymbol{2}}, \boldsymbol{\chi}_{\boldsymbol{i}}^{\boldsymbol{2}}\right)$ nokta ikilileri için serpme diyagramı (saçılım grafiği) Grafik 1.1'de oluşturuldu. Bu grafiğe göre noktaların genellikle bir doğru üzerinde dağılım gösterdiği söylenebilir, sadece 15.nci gözlem doğrudan önemli derecede sapma göstermektedir.. Bu sebeple H_0 hipotezinin ret veya kabul edilme kararı

kuşkuludur, diğer yöntemlere de bakmak ve birlikte değerlendirmekte yarar vardır. Böylece örneklemin dağılımının, $N_2\left(\underline{\mu}\ ,\ \Sigma\right)$ dağılımı ile uyumlu olduğunu söylemek şüphelidir.

SPSS ile hesaplamalar algoritması:

Adım:1 Değişkenleri (bağımsız) tanımla, özelliklerini gir ve verileri gir

Adım:2 Mahalanobis kare uzaklıklarını hesaplayabilmek için bir kukla bağımlı değişken tanımla ve keyfi veri gir

Adım:3 P_i olasılıklarını hesaplayabilmek için gözlem sıra no'ları (1'den n'e kadar) birim değişkeni olarak girilir

Adım:4 Örneklem ortalama vektörü ile varyans-kovaryans matrislerini hesaplamak için **Analyze > correlate > Bivariate** yolunu izleyerek açılan ekranda, bağımsız değişkenleri listeden seçip **Variables** işlem kutusuna aktar ve **Options** penceresini açarak **statistics** bölümünde sunulan iki tercihi de işaretle. **Continue** ve **Ok** tuşları ile istenilenleri hesaplat ve çıktı **(Output**) sayfasında sunulan bilgileri (ortalama vektörleri ve varyans-kovaryans matrislerini) düzenle.

Adım:5 Mahalanobis kare uzaklıklarını $\left(\boldsymbol{m_i^2}\right)$ hesaplamak için **Analyze > regression > Linear** yolunu izleyerek açılan ekranda, bağımlı değişkeni **dependent** işlem kutusuna, bağımsız değişkenleri de **independents** işlem kutusuna aktar. **Save** penceresini açarak **distances** bölümünde **Mahalanobis** seçeneğini işaretle. **Continue** ve **Ok** tuşlarına tıkla, sonuçlar veri sayfasına yeni bir değişken olarak en son sütuna eklenir.

Adım:6 Sıralı Mahalanobis kare uzaklıklarını $\left(\boldsymbol{m_{(i)}^2}\right)$ elde etmek için, veri giriş sayfasında verilen Mahalanobis kare uzaklıklarını $\left(\boldsymbol{m_i^2}\right)$ sütunun en üst satırı üzerine **Mouse** ile gelip sağ tıklayınca açılan seçeneklerden **sort ascending** tıklanır ve bu sütundaki uzaklıkların küçükten büyüğe sıralandığı görülür.

Adım:7 Her bir birime karşılık gelen P_i olasılıklarını elde etmek için **Transform > Compute variable** yolu izlenerek açılan ekranda **target variable** kutusuna olasılık (P_i) yazılır, **Nümeric expression** işlem kutusuna $\left((\boldsymbol{Birim}-\mathbf{0,5})/\boldsymbol{n}\right)$ fonksiyon tanımlaması yapılır ve **Ok** tıklanarak olasılıklar hesaplatılır. Sonuçlar veri giriş sayfasında en son sütunda yeni bir değişken olarak sunulur.

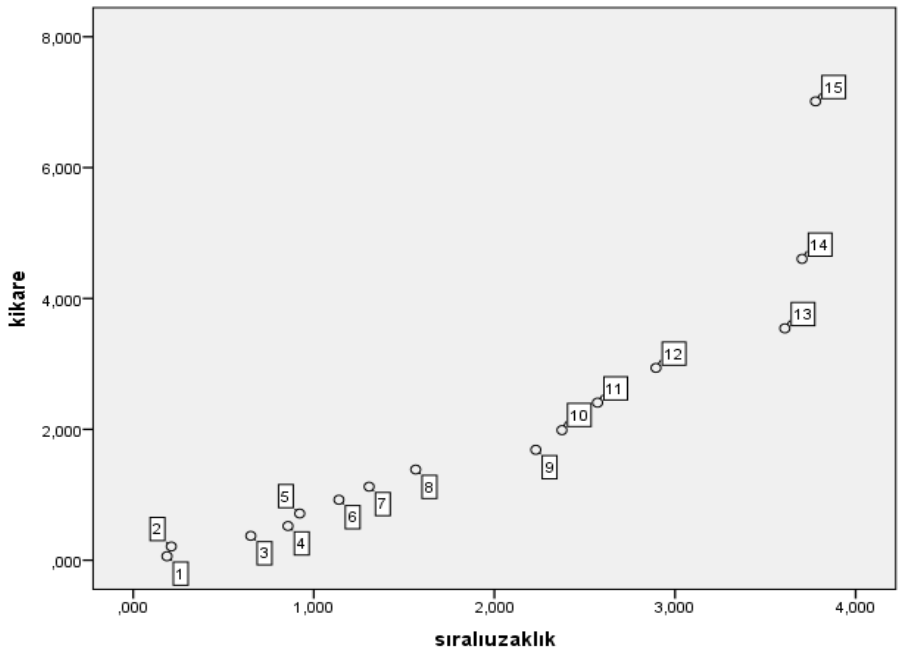
Adım:8 Her bir P_i olasılıklarına karşılık gelen Ki Kare değerlerini hesaplamak için **Transform > Compute variable** yolu izlenerek açılan ekranda **target variable** kutusuna **Ki kare** yazılır. Sonra **Function Groups** içerisinden **Inverse DF** ve **Functions and Special Variables** seçenek kutusundan **Idf.Chisq** seçeneği seçilir ve yandaki aktarma butonu ile **Nümerik Expression** işlem kutusuna aktarılır. Ekranın sol tarafında yer alan kayıtlı değişkenler listesinden olasılık değişkeni seçilerek birinci soru işaretinin yerine ve serbestlik derecesi (**df=p** değişken sayısı) ikinci soru işareti yerine girilir. **Ok** tuşu tıklandığında sonuçlar veri giriş sayfasında yeni değişken **Ki kare** olarak en son sütünde yer alacaktır.

Adım:9 Verilere ait $\left(\boldsymbol{m_{(i)}^2, \chi_i^2}\right)$ nokta ikilileri ile serpme diyagramı çizmek için **Graphs > legacy dialogs > Scatter Dots > Simple scatter > Define** yolunu izleyerek açılan ekranda, Y

Axis işlem kutusuna Ki Kare ve X Axis işlem kutusuna da $\boldsymbol{m_{(i)}^2}$ değişkenleri aktarılır. Ok tuşu ile işlem bitirilir, grafik çıktı sayfasında sunulur.

Tablo 4.6 Çok Değişkenli Normal Dağılıma Uyumluluk Testi İşlem Tablosu

$\boldsymbol{i}$	$\boldsymbol{m_i^2}$	$\boldsymbol{m_{(i)}^2}$	$\boldsymbol{P_i = \frac{i - 0,5}{n}}$	$\boldsymbol{\chi_i^2}$	$\boldsymbol{m_i^4}$
1	2,571	0,187	$[(1 - 0,5)/15] = 0,03$	,061	6,610
2	3,703	0,211	0,10	,211	13,712
3	3,608	0,651	0,17	,373	13,018
4	1,306	0,857	0,23	,523	1,706
5	1,139	0,922	0,30	,713	1,297
6	0,922	1,139	0,37	,924	,850
7	2,229	1,306	0,43	1,124	4,968
8	2,374	1,564	0,50	1,386	5,636
9	2,894	2,229	0,57	1,688	8,375
10	0,211	2,374	0,63	1,989	,045
11	0,651	2,571	0,70	2,408	,424
12	1,564	2,894	0,77	2,939	2,446
13	0,857	3,608	0,83	3,544	,734
14	3,778	3,703	0,90	4,605	14,273
15	0,187	3,778	0,97	7,013	,035
					$\boldsymbol{\sum_{i=1}^{15} m_i^4 = 74,13}$



Grafik 4.1 Verilere ait $\left(m_{(i)}^2, \chi_i^2\right)$ nokta ikilileri için serpme diyagramı

iii) Mardia'nın basıklık testi: ile inceleyelim. Çok değişkenli basıklık katsayısı $\hat{\gamma}_{2p} = \frac{1}{n}\sum_{i=1}^{n} m_i^4$ olup, H_0 doğru iken örnekleme dağılımı $\hat{\gamma}_{2p} \sim N\left(p(p+2); \frac{8p(p+2)}{n}\right)$ olduğundan, test istatistiği;

$$Z = \frac{\hat{\gamma}_{2p} - p(p+2)}{\sqrt{\frac{8p(p+2)}{n}}} \sim N(0,1)$$

dir. Her bir gözlem için m_i^4 değerleri Tablo 1.6'da verilmiştir. Buna göre $\hat{\gamma}_{2p} = \frac{74,13}{15} = 4,942$ ve $p = 2$ için test istatistiğinin H_0 hipotezi altında alabileceği değer;

$$Z_h = \frac{4,942 - 2*4}{\sqrt{\frac{8*2*4}{15}}} = -1,48$$

olarak bulunur. $p = Pr(Z \leq Z_h) = Pr(Z \leq -1,48) = 0,5000 - 0,4306 = 0,0694$ olup, H_1 çift yönlü olduğundan $2p = 2 * 0,0694 = 0,1388$ ve $\alpha = 0,05$ iken, $2p > \alpha$ olduğundan H_0 hipotezi ret edilemez ve böylece örneklemin dağılımı, $N_2\left(\underline{\mu}\ ,\ \Sigma\right)$ dağılımı ile uyumludur.

ii) Bağımsız Değişkenlerin Varyans-Kovaryans Matrislerinin Homojenliği: AA, varyans kovaryans matrislerinin homojenliğine karşı çok duyarlıdır. Bu duyarlılığı etkileyebilecek durumlar:

i) Bağımlı değişkenin kategorilerinde örneklem büyüklüğü yetersiz olduğunda ve varyans-kovaryans matrisleri homojen olmadığında, sınıflandırma fonksiyonlarının kestirim işlemlerinin istatistiksel önemliliği olumsuz olarak etkilenir.

ii) Bağımlı değişkenin kategorilerinde örneklem büyüklüğünün yeterli olması, ancak; varyans-kovaryans matrislerinin homojen olmaması durumunda, gözlemler daha büyük kovaryansa sahip olan gruplara yanlışlıkla sınıflandırılabilir. Bu sebeple AA öncesinde grup içi varyans-kovaryans matrislerinin homojenliği Box M testi ile incelenmelidir.

iii) Varyans kovaryans matrislerinin homojen olmaması durumunda dönüşümlerden yararlanarak homojenlik sağlanabilir. Bu amaçla gruplara göre her bir bağımsız değişkenin incelenmesi gerekir. Ya da, gruplara göre verilerin çok değişkenli normal dağılım gösterdiği, ancak varyans kovaryans matrislerinin benzer olmadığı durumlarda karesel (kuadratik) ayırma analizi tercih edilir.

4.3 Ayırma Analizine İlişkin Temel Eşitlikler ve Ek Açıklamalar

4.3.1 Değişkenlerin Ayırım Gücü ve Anlamlı Boyut Sayısı

Ayırma fonksiyonları oluşturulmadan önce, bağımsız değişkenlerin oluşturacağı kombinasyonun gözlemleri gruplara atamada başarılı olup olmayacağının belirlenmesi gerekir. Yani, oluşturulacak ayırıcı fonksiyonların ayırıcılık gücü istatistiksel anlamlılık açısından

değerlendirilir. Bağımlı değişkene ait grup (kategori sayısı) g ve her bir gruptaki bağımsız değişken sayısı p olmak üzere, ayırma fonksiyonlarının sayısı;

$$r = min(g-1, p) \tag{4.5}$$

ile bulunur. Değişkenlerin ayırım gücü için test edilecek hipotezler;

H_0 : $Bağımsız\ değişkenler\ grupları\ ayırmada$ önemsizdir

H_1 : $Bağımsız\ değişkenler\ grupları\ ayırmada$ önemlidir (4.6)

şeklinde kurulurken, Anlamlı boyut sayısını belirlemek için hipotezler;

H_0 : $j.nci\ ayırma\ fonksiyonu$ önemsizdir

H_1 : $j.nci\ ayırma\ fonksiyonu$ önemlidir , $j = 1,2,\dots,r$ (4.7)

şeklinde kurulur ve H_0 hipotezi kabul edilinceye kadar ardışık olarak tekrarlanır. (4.6) ile verilen H_0 hipotezini MANOVA analizinde kullanılan Wilks Lambda istatistiği (Λ) ile test edebiliriz. Bu amaçla bağımsız değişkenler kümesine ait toplam varyans (T) gruplar arasındaki farklılığa ilişkin varyans (B) ve gruplar içi farklılığa (hataya) ilişkin varyans(W) şeklinde birbirinden bağımsız iki kaynağa bölünür, öyleki:

$$T = B + W \tag{4.8}$$

dir. Burada:

T: Genel kareler ve çarpımlar toplamı matrisi

$$T = \sum_{k=1}^{g} \sum_{i=1}^{n_k} \left(\underline{X}_{k\,i} - \underline{\overline{X}}_{\cdot\,\cdot}\right)\left(\underline{X}_{k\,i} - \underline{\overline{X}}_{\cdot\,\cdot}\right)' \tag{4.9}$$

B: Gruplar arası kareler ve çarpımlar toplamı matrisi

$$B = \sum_{k=1}^{g} n_k \left(\underline{\overline{X}}_{k\,\cdot} - \underline{\overline{X}}_{\cdot\,\cdot}\right)\left(\underline{\overline{X}}_{k\,\cdot} - \underline{\overline{X}}_{\cdot\,\cdot}\right)' \tag{4.10}$$

W: Hata kareler ve çarpımlar toplamı matrisi

$$W = \sum_{k=1}^{g} \sum_{i=1}^{n_k} \left(\underline{X}_{k\,i} - \underline{\overline{X}}_{k\,\cdot}\right)\left(\underline{X}_{k\,i} - \underline{\overline{X}}_{k\,\cdot}\right)' \tag{4.11}$$

eşitlikleri ile hesaplanırken,

$\underline{\overline{X}}_{k\,\cdot}$: k.ncı gruba ait örnek ortalama vektörü

$$\underline{\overline{X}}_{k\,\cdot} = \frac{1}{n_k} \sum_{i=1}^{n_k} \underline{X}_{k\,i}\ , k = 1,2,\dots,g \tag{4.12}$$

$\underline{\overline{X}}_{\cdot\,\cdot}$: Genel örneklem ortalama vektörü

$$\underline{\overline{X}}_{..} = \frac{1}{\underline{\overline{X}}_{k.}} \sum_{k=1}^{g} \sum_{i=1}^{n_k} \underline{X}_{k\,i} = \frac{1}{N} \sum_{k=1}^{g} n_k \, \underline{\overline{X}}_{k.} \qquad (4.13)$$

eşitlikleri ile hesaplanmaktadır. Burada $N = \sum_{k=1}^{g} n_k$ genel örnek birim sayısı ve n_k: k.ncı gruba ait örnek birim sayısıdır. B ve W matrislerinin serbestlik dereceleri sırası ile $(g-1)$ ve $(N-g)$ dir. Bu bilgilere dayalı olarak, Wilks'in Lambda istatistiği;

$$\Lambda = \frac{|W|}{|B+W|} \;, 0 < \Lambda < 1 \qquad (4.14)$$

şeklinde ifade edilir. Wilks'in Lambda istatistiğinin örnekleme dağılımı F veya χ^2 dağılımlarına yaklaşmaktadır. Bu yaklaşımlar Tablo 4.7 de verilmiştir.

Karar: α önem seviyesi, test istatistiklerinin örnekten hesaplanan değerleri de F istatistikleri için F_h ve χ^2 istatistikleri için de χ_h^2 olsun. Kritik değerler ise ilgili dağılıma ait serbestlik dereceleri ve α önem seviyesi dikkate alınarak belirlenecek olan tablo değeri olmak üzere F_t ve χ_t^2 ile gösterilsin. Eğer, $F_h \leq F_t$ veya $(\chi_h^2 \leq \chi_t^2)$ ya da $p \geq \alpha$ ise (4.6) ile verilen H_0 hipotezi ret edilemez ve böylece bağımsız değişkenlerin grupları ayırmada önemsiz olduğuna karar verilir. Eğer, $F_h > F_t$ veya $(\chi_h^2 > \chi_t^2)$ ya da $p < \alpha$ ise (4.6) ile verilen H_0 hipotezi ret edilir ve böylece bağımsız değişkenlerin grupları ayırmada önemli olduğuna karar verilir. Diğer bir ifadeyle, bağımsız değişkenlerin oluşturacağı doğrusal fonksiyonlar yardımıyla g tane grubun birbirinden ayırt edilebileceğine ve yeni bir birimin gruplardan birisine doğru olarak atanabileceğine karar verilir.

Tablo 4.7 Wilks'in Lambda istatistiğinin *F* veya χ^2 dağılımlarına yaklaşımı

Değişken sayısı	**Grup sayısı**	***F* dağılımına yaklaşım**	***F* dağılımının s.d.**
$p = 1$	$g \geq 2$	$\left(\frac{N-g}{g-1}\right)\left(\frac{1-\Lambda}{\Lambda}\right)$	$(g-1)\,,(N-g)$
$p = 2$	$g \geq 2$	$\left(\frac{N-g-1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)$	$2(g-1)\,,2(N-g-1)$
$p > 2$	$g = 2$	$\left(\frac{N-p-1}{p}\right)\left(\frac{1-\Lambda}{\Lambda}\right)$	$(p)\,,(N-p-1)$
$p > 2$	$g = 3$	$\left(\frac{N-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)$	$(2p)\,,2(N-p-2)$
χ^2 dağılımına yaklaşım			**χ^2 dağılımının s.d.**
$L = -\left[N - 1 - \frac{(p+g)}{2}\right] ln(\Lambda)$			$p(g-1)$

Diğer taraftan (4.7) ile verilen H_0 hipotezini test etmek için gerekli olan test istatistiği,

$\lambda_1, \lambda_2, \ldots, \lambda_r$ ler $W^{-1}B$ matrisinin sıfırdan farklı özdeğerleri ve m önemli özdeğer veya ayırma fonksiyonu sayısı olmak üzere $(r \leq p \text{ ve } m = 0,1,\ldots,r-1)$

$$V = \left[N - 1 - \frac{(p+g)}{2}\right] \sum_{j=m+1}^{r} ln\left(1+\lambda_j\right) \sim \chi^2_{(p-m)(g-1-m)} \qquad (4.15)$$

şeklinde tanımlıdır. Burada $\lambda_1 > \lambda_2 > \cdots > \lambda_r$ sıralaması mevcut olduğundan ve her bir özdeğer farklı bir ayırma fonksiyonunu tanımlayacağından, bu test işlemi ile ayırma fonksiyonlarının önemliliği ardışık olarak H_0 hipotezi kabul edilinceye kadar tekrarlanır.

1. Adım Önce $j = 1$ alınıp $\lambda_1$özdeğerinin, yani birinci ayırma fonksiyonunun önemli olup olmadığı test edilir. Bu durumda (4.7) ile verilen hipotezler:
H_0: Birinci ayırma fonksiyonu önemsiz $(\lambda_1 = 0)$

H_1: Birinci ayırma fonksiyonu önemli $(\lambda_1 > 0)$

Bu adımda henüz önemli ayırma fonksiyonu olmadığından $\boldsymbol{m = 0}$ dır. V_h : test istatistiğinin örnekten hesaplanan değeri olsun. α önem seviyesi olmak üzere, eğer; $V_h \leq \chi^2_{(p)(g-1);\alpha}$ yani $p \geq \alpha$ ise H_0 ret edilemez. Böylece birinci ayırma fonksiyonunun ve dolayısıyla $\lambda_1 > \lambda_2 > \cdots > \lambda_r$ sıralaması gereğince tüm ayırma fonksiyonlarının önemsiz olduğuna karar verilir.

Eğer; $V_h > \chi^2_{(p)(g-1);\alpha}$ yani $p < \alpha$ ise H_0 ret edilir. Böylece birinci ayırma fonksiyonunun önemli olduğuna karar verilir ve önemli ayırma fonksiyon sayısı bu adım sonunda $\boldsymbol{m = 1}$ alınır, ikinci ayırma fonksiyonunun veya λ_2 özdeğerinin önemli olup olmadığını test etmek için ikinci adıma geçilir.

2. Adım Yeni hipotezler ($\boldsymbol{m = 1}$)

H_0: İkinci ayırma fonksiyonu önemsiz $(\lambda_2 = 0)$

H_1: İkinci ayırma fonksiyonu önemli $(\lambda_2 > 0)$

Eşitlik (4.15) ile hesaplanacak olan test istatistiğinin alabileceği değer V_h olsun. α önem seviyesi olmak üzere, eğer; $V_h \leq \chi^2_{(p-1)(g-2);\alpha}$, yani $p \geq \alpha$ ise H_0 ret edilemez. Böylece ikinci ayırma fonksiyonunun ve dolayısıyla $\lambda_2 > \cdots > \lambda_r$ sıralaması gereğince geriye kalan tüm ayırma fonksiyonlarının önemsiz olduğuna karar verilir.

Eğer; $V_h > \chi^2_{(p-1)(g-2);\alpha}$, yani $p < \alpha$ ise H_0 ret edilir. Böylece ikinci ayırma fonksiyonunun önemli olduğuna karar verilir ve önemli ayırma fonksiyon sayısı bu adım sonunda $\boldsymbol{m = 2}$ alınır, üçüncü ayırma fonksiyonunun veya λ_3 özdeğerinin önemli olup olmadığını test etmek için üçüncü adıma geçilir.

Bu adımsal süreç H_0 hipotezinin kabul edildiği aşamaya kadar devam eder. Bu aşamada test edilen ayırma fonksiyonu ve buna karşılık gelen özdeğer ile birlikte özdeğerler arasındaki büyüklük sıralaması gereğince, kendisinden küçük olan diğer özdeğerler ve bunlara karşılık gelen ayırma fonksiyonları önemsiz olacaktır. Bu sebeple bu fonksiyonlar ihmal edilirler.

Önemli olduğuna karar verilen ayırma fonksiyonları yardımı ile gruplar arasında ayırım yapılabilir. Özdeğerler arasındaki büyüklük sıralaması gereğince gruplar arasında en iyi ayırımı birinci ayırma fonksiyonu yapacaktır, diğer bir ifade ile gruplar arası varyansın (değişimin) en büyük kısmını birinci ayırma fonksiyonu açıklayacaktır. İkinci ayırma fonksiyonu, birinci

ayırma fonksiyonun gruplar arası varyansın açıklayamadığı kısmının en büyük kısmını açıklayacağından birinci ayırma fonksiyonundan sonra grupları en iyi ayıran ayırma fonksiyonu olacaktır. Diğerleri de bu şekilde değerlendirilir. Yani diğer önemli ayırma fonksiyonlarından herhangi birisi grupları ayırmada kendisinden öncekilere göre daha kötü bir performans gösterirken, kendisinden sonrakilere göre daha iyi performans gösterecektir.