DİSKRİMİNAT ANALİZİ VE BUNUNLA İLGİLİ BİR UYGULAMA

Dr. Aydın ÖZTÜRK[1]
Dr. Şaban KARATAŞ[2]

Ö Z E T

Hangi populasyona ait oldukları kesinlikle bilinmeyen fertlerin ayırım ve sınıflandırılmalarını konu edinen diskriminant analizinin teorisi ana hatlarıyla açıklanmıştır.

Sınıflandırma işlerinde kullanılan kriterin seçimine ait genel esaslar izah edilerek, çok değişkenli normal populasyonlardaki distriminant fonksiyonu ve bununla ilgili bazı konular üzerinde durulmuştur. Ayrıca, ele alınan iki akar türü için diskriminant analizinin nasıl yapıldığı gösterilmiştir.

1. GİRİŞ

Üzerinde çalışılan örneğe giren fertler bazan değişik populasyonlardan gelirler. Bu fertlerin hangi populasyona ait olduğu çoğu zaman kolaylıkla tesbit edilemez. Eğer tek bir karakter fertlerin gruplarını ayırdetmede yeterli bir delil olamıyorsa o zaman iki veya daha fazla karakter bakımından bazı ölçülerin alınması gerekir. Böylece, bu ölçülere dayanan herhangi bir indeks değeri fertlerin mensub oldukları populasyonlara sınıflandırılmalarında daha etkili bir ölçü olabilir.

Sınıflandırmada kullanılan indeksin teşkili eldeki problemin amacına göre değişir. Yakın zamana kadar çeşitli ayırım ve sınıflandırma işlemleri için değişik kriterler kullanılmıştır. Burada,

(1) Atatürk Üniversitesi, Ziraat Fakültesi İstatistik Asistanı.
(2) Atatürk Üniversitesi, Ziraat Fakültesi İstatistik Profesörü.
Dergi Komisyonuna geliş tarihi: 6.6.1975 .

uygulama alanı son zamanlarda bir hayli genişleyen diskriminant fonksiyonu üzerinde durulacaktır. Bu fonksiyonun çıkarılışı yanlış sınıflandırma ihtimallerinin minimum yapılması esasına dayanır. Diskriminant fonksiyonu ile yapılan sınıflandırmada fertler riski en az olan populasyona dahil edilirler.

Bu çalışmada önce sınıflandırma işleminde kullanılan kriterin seçimine ait genel esaslar izah edilecek, daha sonra çok değişkenli normal populasyonlarda diskriminant fonksiyonu ve bununla ilgili bazı konular üzerinde durulacaktır.

2. Sınıflandırmada kullanılan kriterin seçimi.

Anderson (1958) sınıflandırma kriterinin seçimine ait genel esasları geniş bir şekilde incelemiştir. Bu bölümde sadece m_1 ve m_2 gibi iki populasyonun mevcut olduğu farzedilecektir. Elde edilen sonuçları ikiden fazla populasyon için de genelleştirilebilir.

Fertlerin m_1 ve m_2 populasyonlarından gelme ihtimalleri sırasiyle q_1 ve q_2 olsun[1]. Dağılışın bir yoğunluk fonksiyonunun olduğunu farzederek m_1 populasyonunkini $f_1(\mathbf{x})$ ve m_2 ninkini de $f_2(\mathbf{x})$, $[\mathbf{x}' = x_1\ x_2 \dots x_p)]$ ile gösterelim. Ayrıca bu populasyonlara ait sınıflandırma bölgeleri R_1 ve R_2 olsun. Buna göre aslında m_1'e mensup bir ferdin gene aynı populasyona doğru olarak sınıflandırılma ihtimali,

$$P(m_1/m_1, R) = \int_{R_1} f_1(\mathbf{x})\, dx,\ dx = dx_1\, dx_2 \dots dx_p,\ R_1,\ R_2\} ER \dots (1)$$

şeklindedir. Aynı ferdin yanlışlıkla m_2 ye dahil edilmesi ihtimali de bulunabilir,

$$P(m_2/m_1, R) = \int_{R_2} f_1(\mathbf{x})\, dx = 1 - \int_{R_1} f_1(\mathbf{x})\, dx \quad (2)$$

Benzer yoldan m_2'ye ait bir ferdin gene m_2 ye doğru sınıflandırılması ihtimali,

$$P(m_2/m_2, R) = \int_{R_2} f_2(\mathbf{x})\, dx \quad (3)$$

ve yanlış sınıflandırılma ihtimali de,

$$P(m_1/m_2, R) = \int_{R_1} f_2(\mathbf{x})\, dx = 1 - \int_{R_2} f_2(\mathbf{x})\, dx \quad (4)$$

olarak bulunur.

m_1 den bir fert çekme ihtimali q_1 olduğundan bu populasyona ait bir ferdi doğru sınıflandırma ihtimali $q_1 P(m_1/m_1, R)$ yanlış sınıflandırma ihtimali ise $q_1 P(m_2/m_1, R)$ dir. Aynı şekilde m_2 den alınan bir ferdin doğru sınıflandırılma ihtimali $q_2 P(m_2/m_1, R)$ ve yanlış sınıflandırma ihtimali de $q_2 P(m_1/m_2, R)$ dir.

Fertlerin yanlış sınıflandırılmalarından ileri gelen kayıpları her iki populasyon için de aynı olmayabilir. m_1'e yapılan yanlış sınıflandırma m_2 ye yapılandan daha fazla kayba yol açabilir. Bu

(1) q_1 ve q_2 ye a priori ihtimaller de denmektedir. Bunlar populasyonların nisbî fonksiyonlarını ifade ederler.

kayıpların meydana getirdiği zamana genellikle yanlış sınıflandırma maliyeti denir. Bu maliyetleri C_{12} (m_2 deki ferdi yanlışlıkla m_1'e sınıflandırmanın maliyeti) ve C_{21} (m_1 deki ferdi yanlışlıkla m_2 ye sınıflandırmanın maliyeti) ile gösterirsek m_1 ve m_2 ye yapılan yanlış sınıflandırma maliyetleri bulunabilir. Bunlar,

m_1 için $C_{21}q_1 \; P(m_2/m_1, R)$

m_2 için $C_{12}q_2 \; P(m_1/m_2, R)$ ... (5)

şeklindedir. Böylece ortalama maliyet bu iki maliyetin toplamına eşitti .Yani,

$$G(C, q) = C_{21}q_1 \; P(m_2/m_1, R) + C_{12}q_2 \; P(m_1/m_2, R) \quad \text{............} \quad (6)$$

dir. Bu ifade (2) ve (4) no. lu eşitlikler yardımiyle aşağıdaki gibi daha açık şekliyle yazılabilir.

$$G(C, q) = C_{21}q_1 \int_{R_2} f_1(x)\, dx + C_{12}q_2 \int_{R_1} f_2(x)\, dx \quad \text{..............} \quad (7)$$

Sınıflandırmada seçilecek en iyi yol yukarıdaki ifadenin minimum yapılmasına bağlıdır. Kolayca anlaşılacağı gibi R_1 ve R_2 bölgeleri aşağıdaki gibi seçilirse bu ifade minimum olur.

$$R_1 : C_{21} f_1(x) > C_{12}q_2 f_2(x)$$
$$R_2 : C_{21}q_1 f_1(x) < C_{12}q_2 f_2(x) \quad (8)$$

Diğer bir ifadeyle; herhangi bir $\bar{x}$ ferdi ,

$$\frac{f_1(x)}{f_2(x)} > \frac{C_{12}q_2}{C_{21}q_1} \quad (9)$$

ise m_1 populasyonuna,

$$\frac{f_1(x)}{f_2(x)} < \frac{C_{12}q_2}{C_{21}q_1} \quad (10)$$

ise, m_2 populasyonuna sınıflandırılırsa (7) no. lu ifadenin değeri en küçük olur. Eşitliğin sağlanması halinde sınıflandırma rastgele yapılır.

3. Çok Değişkenli İki Normal Populasyonda Sınıflandırma

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

p- elementli bir x vektörü her iki populasyonda da çok değişkenli normal dağılışlar göstersin. Bunun birinci ve ikinci populasyonlardaki yoğunluk fonksiyonu,

$$f_1(x) = \frac{1}{(2\pi)^{p/2} \,|V^{(1)}|^{1/2}} \exp\left[- \frac{1}{2} (x-\mu^{(1)})' \, (V^{(1)})^{-1} \, (x-\mu^{(1)}) \right]$$

$$f_2(x) = \frac{1}{(2\pi)^{p/2} \,|V^{(2)}|^{1/2}} \exp\left[- \frac{1}{2} (x-\mu^{(2)})' \, (V^{(2)})^{-1} \, (x-\mu^{(2)}) \right] \quad \text{..............} \quad (11)$$

şeklinde yazılabilir. Burada,

$$\mu^{(i)} = (\mu_1^{(i)}, \mu_2^{(i)}, \ldots\ldots \mu_p^{(i)}) \quad (12)$$

ortalama vektörü,

$$V^{(i)} = \begin{bmatrix} \sigma_{11}^{(i)} & \sigma_{12}^{(i)} & \ldots\ldots & \sigma_{1p}^{(i)} \\ \sigma_{21}^{(i)} & \sigma_{22}^{(i)} & \ldots\ldots & \sigma_{2p}^{(i)} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{p1}^{(i)} & \sigma_{p2}^{(i)} & \ldots\ldots & \sigma_{pp}^{(i)} \end{bmatrix} \qquad (13)$$

variyans kovariyans matrisidir.

(9) ve (10) numaralı eşitsizlikler çözülüp her iki tarafın tabii logaritması alınırsa R_1 ve R_2 sınıflandırma bölgeleri şöyle bulunur:

$$R_1 : W(x) = \frac{f_1(x)}{f_2(x)} > \ln \frac{C_{12}q_2}{C_{21}q_1}$$

$$R_2 : W(x) = \frac{f_1(x)}{f_2(x)} < \ln \frac{C_{12}q_2}{C_{21}q_1} \qquad (14)$$

$f_1(x)$ ve $f_2(x)$ in (11) deki değeri kullanılırsa,

$$W(x) = \frac{1}{2} \ln \frac{|V^{(2)}|}{|V^{(1)}|} - \frac{1}{2} [(x-\mu^{(1)})' (V^{(1)})^{-1} (x-\mu^{(1)}) - (x-\mu^{(2)}) (V^{(2)})^{-1} (x-\mu^{(2)})] \qquad (15)$$

fonksiyonu elde edilir. Görüldüğü gibi bu ikinci dereceden p-değişkenli bir polinomdur. Bu polinoma çok değişkenli istatistikte kareli diskriminant fonksiyonu denmektedir. Pratikte ayırım ve sınıflandırma işlemlerinde kullanılan bu gibi fonksiyonların mümkün olduğu kadar daha basit formda olmaları istenir. m_1 ve m_2 populasyonlarının variyans matrislerinin,

$$V = V^{(1)} = V^{(2)} \qquad (16)$$

olduğu kabul edilirse (15) deki fonksiyon aşağıdaki basit şekle dönüşür,

$$W(x) = x' V^{-1} (\mu^{(1)} - \mu^{(2)}) - \frac{1}{2} (\mu^{(1)} + \mu^{(2)}) V^{-1} (\mu^{(1)} - \mu^{(2)}) \qquad (17)$$

Böylece $\mathbf{x}$ ferdini yoğunluk fonksiyonları $f_1(x)$ ve $f_2(x)$ olan m_1 ve m_2 populasyonlarına sınıflandırmak için $W(x)$ değeri hesaplanarak (14) deki R_1 ve R_2 bölgeleri bulunur. Eğer,

$$W(x) > \ln \frac{C_{21}q_1}{C_{12}q_2} = C \qquad (18)$$

ise ilgili fert m_1'e aksi halde m_2'ye sınıflandırılır.

$$W(\mathbf{x}) = \ln \frac{C_{12}q_2}{C_{21}q_1} \qquad (19)$$

olduğu zaman sınıflandırma rastgele yapılır.

$q_1 = q_2$ ve $C_{12} = C_{21}$ özel hali için,

$$\ln \frac{C_{12}q_2}{C_{21}q_1} = 0 \qquad (20)$$

olduğundan sınıflandırma bölgeleri,

$$R_1 : W(\mathbf{x}) > 0$$
$$R_2 : W(\mathbf{x}) < 0 \qquad (21)$$

şeklini alır. Buna göre **x** ferdinin fonksiyon değeri pozitifse sınıflandırma m_1'e, negatifse m_2'ye yapılır.

4. A Priori İhtimallerinin Bilinmediği Durumlarda Sınıflandırma

q_1 ve q_2 ihtimalleri bilinmediği zaman R_1 ve R_2 bölgelerini birbirinden ayıran sınır noktası daha farklı bir şekilde bulunur.

$W(\mathbf{x})$ in m_1 ve m_2 populasyonlarındaki beklenen değerleri sırasiyle A_1 ve A_2 olsun. **x** vektörünün her iki populasyonda da ortalaması $\mu^{(i)}$ variyansı **V** olan çok değişkenli normal yapılış gösterdiği farzedilirse,

$$A_1 = E\,[\,W(\mathbf{x}) \mid (m_1)\,]$$

$$= \frac{1}{2}\,(\mu^{(1)}-\mu^{(2)})'\,V^{-1}\,(\mu^{(1)}-\mu^{(2)})$$

$$= \frac{1}{2}\,\Delta^2$$

$$A_2 = E\,[\,W(\mathbf{x} \mid m_2)\,]$$

$$= -\frac{1}{2}\,(\mu^{(1)}-\mu^{(2)})'\,V^{-1}\,(\mu^{(1)}-\mu^{(2)})$$

$$= -\frac{1}{2}\,\Delta^2 \qquad (22)$$

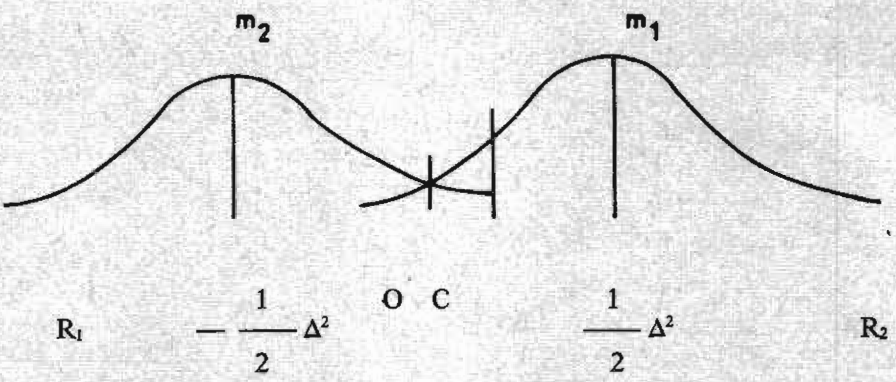
olduğu kolayca bulunur. Burada Δ iki populasyon arasındaki standardize edilmiş mesafeyi göstermektedir.

x'in variyansı m_1 ve m_2 de eşit olduğundan bunun doğrusal bir kombinasyonu olan $W(\bar{x})$ in variyansı da iki populasyonda aynıdır.

$$\sigma^2_W = \mathrm{Var}\,[W(\mathbf{x}) \mid m_1]$$
$$= \mathrm{Var}\,[W(\mathbf{x}) \mid m_2]$$
$$= (\mu^{(1)}-\mu^{(2)})'$$
$$\mathbf{V}^{-1}\,(\mu^{(1)}-\mu^{(2)}) \qquad (23)$$
$$= \Delta^2$$

dir. Özetleyecek olursak, örnekteki fert sayısı yeter derecede fazla ise $W(\mathbf{x})$, m_1 de ortalaması $\Delta^2/_2$ variyansı Δ^2, m_2'de ise ortalaması $-\Delta^2/_2$ variyansı Δ^2 olan normal dağılış gösterir (Şekil 1).

R_1 ve R_2 bölgelerini birbirinden ayıran sınır noktası C olsun. Her bir populasyona yapılan yanlış sınıflandırma ihtimalleri aşağıdaki gibi bulunur.



Şekil : 1. m_1 ve m_2 poplasyonlarına ait R_1 ve R_2 bölgeleri.

$$P(m_1/m_2, R) = \int_C \frac{1}{\sqrt{2\pi}\,\Delta} \exp\left\{-\frac{1}{2\Delta^2}\left[W(\mathbf{x}) + \frac{1}{2}\Delta^2\right]^2\right\} dW(\mathbf{x})$$

$$= 1 - \Phi\left(\frac{C + \Delta^2/2}{\Delta}\right)$$

$$P(m_2/m_1, R) = \int_{-\infty}^{C} \frac{1}{\sqrt{2\pi}\,\Delta} \exp\left\{-\frac{1}{\Delta^2}\left[W(\mathbf{x}) - \frac{1}{2}\Delta^2\right]\right\} dW(\mathbf{x})$$

$$= \Phi\left(\frac{C - \frac{1}{2}\Delta^2}{\Delta}\right) \tag{24}$$

C'yi hesaplamak için yanlış sınıflandırma maliyetlerini eşit kılan denklem çözülür. Yani, bilinen C_{12} ve C_{21} değerleri için,

$$C_{12}\,P(m_1/m_2, R = C_{21}\,P(m_2/m_1, R) \tag{25}$$

eşitliğini sağlayan C değeri iki populasyon arasındaki sınır noktasını verir. $C_{12} = C_{21}$ olduğunda yukardaki eşitlik daha açık olarak,

$$\Phi\left(\frac{C - \frac{1}{2}\Delta^2}{\Delta}\right) = 1 - \Phi\left(\frac{C + \frac{1}{2}\Delta^2}{\Delta}\right) \tag{26}$$

şeklini alır. Normal dağılış eğrisinin simetrik özelliğinden faydalanarak,

$$\frac{C-\frac{1}{2}\Delta^2}{\Delta} = -\frac{C+\frac{1}{2}\Delta^2}{\Delta}$$

veya

$$C = 0$$

bulunur. Böylece yanlış sınıflandırma maliyetleri eşit olduğu zaman R_1 ve R_2 yi birbirinden ayıran sınır noktası m_1 ve m_2 populasyonuna ait W (**x**) in ortalamalarının orta yerinde olur. C'nin bu değeri için (24) no. lu eşitlikte yerine koyulursa her iki populasyondaki yanlış sınıflandırma ihtimallerinin eşit olduğu görülür.

$$P(m_1/m_2, R) = P(m_2/m_1, r) = \Phi\left(-\frac{1}{2}\Delta\right) \qquad (28)$$

5. Parametrelerin Tahmini

Diskriminant fonksiyonunda yer alan $\mu^{(1)}$, $\mu^{(2)}$ ve **V** parametreleri genellikle bilinmediklerinden bunlara ait örnek tahminlerinin kullanılması gerekir. Her bir populasyondan alınan belirli genişliktekiörnek değerlerine göre fonksiyonun parametreleri tahmin edilir. Parametrelerin bu tahmin değerlerine dayanan yeni diskriminant fonksiyonu daha sonraki fertlerin sınıflandırılmalarında kullanılır.

m_1 populasyonundan $X_1, X_2, \ldots, X_{n1}$ ve m_2 den de $X_{n+1}, X_{n1+2}, \ldots, X_{n1+n2}$ müşahedeleri yapılmış olsun. Böyle bir örnekten $\mu^{(1)}$, $\mu^{(2)}$ ve **V** nin tahminleri için sırasıyla,

$$\mathbf{x}^{(1)} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_j$$

$$\mathbf{x}^{(2)} = \frac{1}{n_2}\sum_{j=n_1+1}^{n_1+n_2}\mathbf{x}_j$$

$$\mathbf{S} = \frac{1}{n_1+n_2-2}\left[\sum_{j=1}^{n_1}(\mathbf{x}_j-\mathbf{x}^{(1)})(\mathbf{x}_j-\mathbf{x}^{(1)})' + \sum_{j=n_1+n_2}^{n_1+n_2}(\mathbf{x}_j-\mathbf{x}^{(2)})(\mathbf{x}_j-\mathbf{x}^{(2)}\right] \qquad (29)$$

değerleri kullanılır. Böylece **W (x)** in örnek tahmini olan **W (x)**,

$$W(x) = x'S^{-1}(x^{(1)} - x^{(2)}) - \frac{1}{2}(x^{(1)} + x^{(2)})\,S^{-1}(x^{(1)} - x^{(2)}) \qquad (30)$$

şeklinde yazılabilir (Sitgreaves, 1952)

Yukardaki $x^{(1)}$, $x^{(2)}$ ve **S** nin beklenen değerleri bilindiği gibi sırasiyle $\mu^{(1)}$, $\mu^{(2)}$ ve **V** yi verirler. Bu tahminleri kullanarak elde edilen w (x) şüphesiz **W (x)** in sapmasız (unbiased) bir tahmini değildir. y ve z gibi bir değişken çifti için,

$$E(y \cdot Z) = E(y) \cdot E(z)$$

bağıntısı her zaman geçerli değildir. Benzer şekilde iki şans değişkeninin çarpımı gibi terimleri ihtiva eden **W (x)** in beklenen değeri W (x)'e eşit olmaz. Ancak, parametrelerin bilindiği durumda en iyi kriter olarak bulunan W (x) istatistiğin bu şartlar altında yaklaşık tahmini $x^{(1)}$, $x^{(2)}$ ve **S** yi kullanmak suretiyle bulunur.

6. Diskriminant Fonksiyonu İle Çoklu Regresyon Arasındaki İlişki.

Bir önceki bölümde olduğu gibi m_1 ve m_2 populasyonlarından $n_1 + n_2$ fertli bir örnek çekilmiş olsun. Bunlar için,

$$i = 1, 2$$
$$j = 1, 2, \ldots, n$$

$$y_j^{(i)} = \beta'(x_j^{(i)} - \bar{x})$$

şeklinde bir çoklu regresyon denklemi yazılabilir. Burada,

$$\beta' = (\beta_1, \beta_2, \ldots \beta_p)$$

çoklu regresyon katsayılarını, **x** ise

$$\bar{x} = \frac{n_1 x^{(1)} + n_2 x^{(2)}}{n_1 + n_2}$$

şeklinde ifade edilebilen genel ortalamayı göstermektedir. Bağımlı değişken olan $y_j^{(i)}$ birinci ve ikinci populasyonlarda sırasıyla,

$$y_j^{(1)} = \frac{n_2}{n_1 + n_2} \qquad j = 1, 2, \ldots, n_1$$

$$y_j^{(2)} = \frac{-n_1}{n_1 + n_2} \qquad j = n_1 + 1, n_1 + 2, \ldots n_1 + n_2 \qquad (31)$$

değerleriyle tarif edilirse, bu değerlere göre elde edilen çoklu regresyon denkleminin katsayıları (30) no.lu eşitlikteki diskriminant fonksiyonunun katsayılarına orantılı olur. (Öztürk, 1973) Böylece $y_j^{(i)}$ lere yukarıdaki gibi uygun değerler verildiği taktirde elde edilen çoklu regresyon denklemi bir çeşit diskriminant fonksiyonu olmaktadır.

Çoklu regresyon denklemi ile diskriminant fonksiyonu arasındaki bu ilişki çok önemlidir. Kayıp müşahedelerin tahmini fonksiyona girecek değişkenlerin seçimi gibi diskriminant analizinde henüz kesin bir açıklığa kavuşturulamamış bazı problemlerin çözümünde çoklu regresyon için geliştirilen metotların kullanılmaları bu benzerlik sayesinde mümkün olmaktadır.

7. İki Akar Türünün Ayırım ve Sınıflandırılması Üzerine Bir Uygulama.

Birbirine çok benzeyen B. Praetiosa ve B. Rubrioculus türlerini ayırdedici en önemli karakterleri böceklerin bacaklarında bulunan sensory (his setaesi) ve tactile (dokunma setaesi) organlarının uzunluklarından yalnız başına sensory uzunluğuna veya tactile uzunluğuna bakmak suretiyle bu iki tür arasında ayırım yapmak oldukça zordur. Bu sebeple iki setae'nin bir arada incelenmesi zorunlu olmaktadır.

Cetvel 1. de B. Praetiosa ve B. Rubrioculus'a ait sensory ve tactile uzunlukları verilmiştir.[1] (Ecevit ve Enns)

CETVEL 1

Praetiosa ve B. Rubrioculus da Sensory ve Tactile Uzunlukları (mikron olarak)

	B. Praetiosa			B. Rubrioculus	
No.	Sensory (x_1)	Tactila (x_2)	No.	Sensory (x_1)	Tactila (x_2)
1	40.95	24.99	21	19.52	19.04
2	42.35	30.47	22	24.99	26.61
3	36.19	18.57	23	20.23	18.57
4	32.14	22.61	24	20.23	19.04
5	27.10	21.42	25	21.42	19.52
6	39.28	21.90	26	23.80	21.90
7	32.14	17.38	27	21.42	20.23
8	28.42	19.04	28	21.42	17.14
9	40.75	28.57	29	30.99	21.90
10	40.47	29.19	30	30.19	22.61
11	39.28	27.38	31	23.80	21.42
12	34.52	22.61	32	23.80	21.42
13	40.47	19.76	33	23.80	22.61
14	35.71	20.23	34	23.80	22.61
15	41.66	28.57	35	23.80	21.42
16	34.52	20.23	36	24.99	22.61
17	43.33	32.14	37	26.19	22.61
18	40.95	27.38	38	24.99	22.61
19	35.71	26.19	39	26.19	23.90
20	42.85	30.95	40	23.80	21.42

(1) Orijinal veriler değiştirilerek alınmıştır.

Cetvelde görüldüğü gibi B. Praetiosa türünde sensory ve tactila uzunluklarına ait değişim genişlikleri sırasıyle,

$27.10 < x_1 < 43.33$
$18.57 < x_2 < 32.14$

dür. B. Rubrioculus türünde ise bu değişim genişlikleri,

$19.57 < x_1 < 30.99$
$17.14 < x_2 < 26.61$

olmaktadır. Bütün karakterler birbirlerinin üzerine girmiş durumda olduğundan böcek türlerinin ayırımında karakterlerin hiçbiri tek başına yeterli değildir. Bu sebeple x_1 ve x_2 yi birlikte ihtiva eden diskriminant fonksiyonunun hesaplanması gerekir.

$q_1 = q_2$ ve $C_{12} = C_{21}$ olduğunu farzederek (30) da verilen diskriminant fonksiyonunu elde edelim. Bunun için önce $\bar{x}^{(1)}$, $\bar{x}^{(2)}$ ve S nin hesaplanması gerekir. Cetvel 1. deki verileri kullanarak (29) no. lu formüller yardımiyle bu değerler kolayca bulunabilir. $n_1 = n_2 = 20$ olduğundan,

$$\bar{x}_1^{(1)} = \frac{1}{20} \sum_{j=1}^{20} x_{1j}^{(1)}$$

$$= \frac{1}{20} (40.95 + 42.35 + \ldots + 42.85)$$

$$= 37.4495$$

$$\bar{x}_2^{(2)} = \frac{1}{20} \sum_{j=1}^{20} x_{2j}^{(2)}$$

$$= \frac{1}{20} (24.99 + 30.47 + \ldots + 30.95)$$

$$= 24.4790$$

$$x_2^{(2)} = \frac{1}{20} \sum_{j=21}^{40} x_{1j}^{(2)}$$

$$= \frac{1}{20} (19.52 + 24.99 + \ldots 23.80)$$

$$= 23.9685$$

$$\bar{x}_2^{(2)} = \frac{1}{20} \sum_{j=21}^{40} x_{2j}^{(2)}$$

$$= \frac{1}{20} \ (19.04 + 26.61 + \ldots + 21.42)$$

$$= 21.4545 \qquad (34)$$

değerleri elde edilir. Bunları kullanarak,

$$\bar{x}^{(1)} = (x_1^{(1)} \ x_2^{(1)})$$

$$= (37.4495 \ \ 24.4790)$$

$$\bar{x}^{(2)} = (x_1^{(2)} \ x_2^{(2)})$$

$$= (23.9685 \ \ 21.4545) \qquad (35)$$

bulunur.

S yi hesaplayabilmek için (29) da verilen formülü daha açık bir şekilde aşağıdaki gibi yazalım.

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2} \begin{bmatrix} S_{11}^{(1)} & S_{12}^{(1)} \\ S_{21}^{(1)} & S_{22}^{(1)} \end{bmatrix} + \begin{bmatrix} S_{11}^{(2)} & S_{12}^{(2)} \\ S_{21}^{(2)} & S_{22}^{(2)} \end{bmatrix} \qquad (36)$$

Burada,

$$S_j^{(1)} = \sum_{k=1}^{n_1} (x_{jk}^{(1)} - \bar{x}_j^{(1)}) \ (x_{lk}^{(1)} - \bar{x}_l^{(1)})$$

$$S_j^{(2)} = \sum_{k=n_1+1}^{n_1} (x_{jk}^{(2)} - \bar{x}_j^{(2)}) \ (x_{lk} - \bar{x}_2^{(2)}) \qquad (37)$$

dir.

Cetvel 1. deki değerler bu formüllerde yerine koyulursa şu sonuçlar elde edilir.

$$S_{11}^{(1)} = \sum_{k=1}^{20} (x_{1k}^{(1)} - \bar{x}_1^{(1)})^2$$

$$= (40.95 - 37.45)^2 + (42.35 - 37.45)^2 + \ldots + (42.85 - 37.45)^2$$
$$= 437.0253$$

$$S_{12}^{(1)} = \sum_{k=1}^{40} (x_{1k}^{(1)} - \bar{x}_1^{(1)}) \ (x_{2k} - \bar{x}_2^{(1)})$$

$$= (40.95 - 37.45) \ (24.99 - 24.48) + \ldots + (42.85 - 37.45)$$
$$30.95 - 24.48)$$

$= 311.5768$

$$S_{22}^{(1)} = \sum_{k=1}^{20} (x_{2k}^{(1)} - x_2^{(1)})^2$$

$$= (24.99-24.48)^2 + (30.47-24.48)^2 + \ldots + (3.95-24.48)^2$$

$$= 408.4100 \tag{38}$$

Benzer yoldan $S_{11}^{(2)}$, $S_{12}^{(2)}$ ve $S_{22}^{(2)}$ de bulunabilir.

$$S_{11}^{(2)} = 168.4343$$
$$S_{12}^{(2)} = 74.8776$$
$$S_{22}^{(2)} = 84.3299 \tag{39}$$

Böylece,

$$\mathbf{S} = \frac{1}{38} \left[\begin{array}{cc} 437.0253 & 311.5768 \\ 311.5768 & 408.4100 \end{array} \right] + \left[\begin{array}{cc} 168.4343 & 74.8776 \\ 74.8776 & 84.3299 \end{array} \right]$$

$$= \left[\begin{array}{cc} 15.9331 & 10.1698 \\ 10.1698 & 12.9668 \end{array} \right]$$

olduğu kolayca bulunabilir. Bu matris x_1 ve x_2 değişkenlerinin variyans kovariyans matrisidir. Köşegen elemanlar grup içi variyanslar, diğerleri ise kovariyansları göstermektedir.

Variyans - kovariyans matrisinin boyutları 2×2 olduğundan bunun tersi kolayca bulunabilir.

$$\mathbf{S}^{-1} = \left[\begin{array}{cc} 0.125676 & -0.098567 \\ -0.098567 & 0.154426 \end{array} \right] \tag{41}$$

(35) ve (41) deki değerler (29) no. lu denklemde yerine konulursa,

$$W(x) = 1.396122\, x_1 - 0.861720\, x_2 - 23.082645$$

diskriminant fonksiyonu elde edilir. x_1 ve x_2 için belirli değerler için W (x) nin değeri 0'dan büyükse bu ölçülerin alındığı akar B. Praetiosa türüne, aksi halde B. Rubrioculus türüne sınıflandırılır. Cetvel 2. de x_1 ve x_2 ölçüleri verilen akarlara ait fonksiyon değerleri ve sınıflandırıldıkları türler gösterilmiştir.

Fert No.	x_1	x_2	W (x)	Tür
1	23.80	21.42	— 8.312983	B. Rubrioculus
2	26.19	23.80	— 7.027145	»
3	36.19	18.57	11.440869	B. Praetiosa
4	32.14	17.85	6.407014	»
5	30.99	21.90	1.31150778	»

Δ^2 nin örnek tahmini olan D^2 yi

$$D^2 = (\bar{x}^{(1)} - \bar{x}^{(2)})' \, S^{-1} \, (\bar{x}^{(1)} - \bar{x}^{(2)})$$

formülüne göre hesaplayarak yanlış sınıflandırma ihtimali bulunabilir (35) ve (41) deki değerler yukarıda yerine koyulursa,

$$D^2 = 16.214848$$

bulunur. Böylece yanlış sınıflandırma ihtimali,

$$\Phi\,(-D/2) = \int_{-\infty}^{-2.013} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) dz$$

$$= 0.0529$$

olur.

KAYNAKLAR

Anderson, TW. (1958) An introduction to Multivariate Statistical Analysis. John Wiley and Sons Inc. USA.

Ecevit, O. and Enns, W. R. A Statistical Analysis of Two Species of the Genus Bryobia (Acarina: Tetranychidae) Separating B. praetiosa from B. rubrioculus Utilizing Dublex Stae of Tarsi III and IV (Basımda)

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

Öztürk, A. (1973) Diskriminant Analizi ve Bununla İlgili Bazı Problemler Üzerinde bir araştırma. Atatürk Üniversitesi Ziraat Fakültesi, Basılmamış Doktora Tezi.

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

Sitgreaves (1952) On the Distribution of two Random Matrices Used in Classification procedures. Annals of Mat. Statist.