Karadeniz Teknik Üniversitesi

Fen Fakültesi
İstatistik ve Bilgisayar Bilimleri Bölümü

Çok Değişkenli İstatistiksel Yöntemler

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Dr. Öğr. Üyesi Uğur ŞEVİK

Ders İçeriği

- BÖLÜM 3: Uzaklık ve Benzerlik Ölçüleri
  - Veri Türlerine Göre Ölçülerin Seçimi
  - Uzaklık ve Benzerlik Ölçüleri için Veri Yapısı
  - Sürekli Ve Kesikli Sayısal Veriler İçin Uzaklık Ölçüleri
  - Sürekli Ve Kesikli Sayısal Veriler İçin Benzerlik Ölçüleri
  - Sıklık Sayıları İçin Uzaklık Ölçüleri
  - İkili Sınıflı (Binary) Veriler İçin Uzaklık Ve Benzerlik Ölçüleri
  - Verilerin Standartlaştırılması
  - Örnekler
  - Kümeler/Örneklemler/Gruplar Arası Uzaklık Ölçüleri

Ders Hedefleri

- Çok değişkenli analizde değişkenler arasındaki uzaklık ve benzerlik kavramlarının tanımı ve uzaklık-benzerlik ölçülerinin farklı sayı türlerine göre hesaplanması

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

- *Bazı çok değişkenli istatistiksel çözümlemede X veri matrisi yerine, n tane gözlem arasındaki uzaklıklardan (distance) (ya da benzemezliklerden (dissimilarity)) ya da benzerliklerden (similarity) oluşan nxn boyutlu matrislerden ya da pxp boyutlu benzemezlik ve benzerlik matrislerinden yararlanılmaktadır.*
- *Örneğin **kümeleme analizi**, **çok boyutlu ölçekleme** ve **faktör analizi** gibi birçok çok değişkenli istatistiksel yöntemin başlangıç noktası, veriden elde edilen bu tür matrislerdir.*
- *İki gözlem/değişken arasında büyük benzerlik değeri bu gözlemin/değişkenin birbirine **yakın/benzer** olduğunu, **küçük benzerlik değeri** ise bu iki gözlemin/değişkenin birbirinden **uzak/benzemez** olduğunu gösterir. îki değişken arasındaki benzerliğin belirlenmesinde en çok kullanılan ölçü **Pearson ilişki katsayısıdır.***

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

- Örneğin 4 gözlem için 4x4 boyutundaki uzaklık matrisi (D) aşağıdaki gibi verilir. Bu matriste, herhangi d_{ij} elemanı iki gözlem arasındaki mesafeyi verir.

Gözlem	1	2	3	4
1	-	d_{12}	d_{13}	d_{14}
2	d_{21}	-	d_{23}	d_{24}
3	d_{31}	d_{32}	-	d_{34}
4	d_{41}	d_{42}	d_{43}	-

- Yukarıdaki matris simetrik olduğu için sadece alt ya da üst üçgen şeklinde de sunulmakla birlikte sıklıkla aşağıdaki gibi de sunulur.

Gözlem	1	2	3
2	d_{21}		
3	d_{31}	d_{32}	
4	d_{41}	d_{42}	d_{43}

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

- i ve j gibi iki gözleme ilişkin uzaklık olan d_{ij} aşağıdaki özellikleri taşır;
  - $d_{ij} = d_{ji}$ (Simetri özelliği / Gözlem 1 ve 2'nin uzaklığı gözlem 2 ve 1'in uzaklığına eşittir).
  - Eğer $i \neq j$ ise $d_{ij} > 0$ (Negatif olmama özelliği / gözlem 1 ve 2'nin uzaklığı sıfırdan büyük bir değerdir.)
  - Eğer i=j ise $d_{ij} = 0$ (Tanım özelliği / bir gözlemin kendisi ile uzaklığı sıfırdır.)
  - $d_{ik} \leq d_{ij} + d_{jk}$ (Örneğin üç gözlem (i, j, k) ele alındığında, herhangi ikisi arasındaki uzaklık diğer iki çift uzaklığın toplamını geçemez.)
- En bilinen benzemezlik (uzaklık) ölçüsü Öklit (Euclidean) uzaklık ölçüsüdür. Benzemezlik ölçülerinin en küçük değeri 0, en büyük değeri ise sınırsızdır.

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Veri Türlerine Göre Ölçülerin Seçimi

Uzaklık (Dissimilarity) Ölçüleri	*Benzerlik (Similarity) Ölçüleri*
Sayısal Veriler	• Pearson İlişki Katsayısı Kosinüs Benzerlik Ölçüsü
• Öklit Uzaklık Ölçüsü • Kare Öklit Uzaklık Ölçüsü • Chebychev (Çebişev) Uzaklık Ölçüsü • Manhattan City-Block Uzaklık Ölçüsü • Minkowski Uzaklık Ölçüsü • Karl-Pearson Uzaklık Ölçüsü / Standartlaştırılmış Öklit Uzaklığı Ölçüsü • Korelasyon Uzaklığı Ölçüleri	
Sıklık Sayıları	
• Ki-kare Uzaklık Ölçüsü • Phi-Kare Uzaklık Ölçüsü	

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Veri Türlerine Göre Ölçülerin Seçimi

Uzaklık (Dissimilarity) Ölçüleri	*Benzerlik (Similarity) Ölçüleri*
İkili Veriler	
Kare Öklit Uzaklık Ölçüsü	Russell ve Rao Benzerlik Ölçüsü
Öklit Uzaklık Ölçüsü	Basit Benzerlik Ölçüsü
Büyüklük Farkları Uzaklık Ölçüsü	Jaccard Benzerlik Ölçüsü
Biçim Farkları Uzaklık Ölçüsü	Parçalı Benzerlik Ölçüsü
Değişim Uzaklık Ölçüsü	Rogers ve Tanimoto Benzerlik Ölçüsü
Durum Uzaklık Ölçüsü	Sokal ve Sneath Benzerlik Ölçüsü 1
Lance ve Williams Uzaklık Ölçüsü	Sokal ve Sneath Benzerlik Ölçüsü 2
	Sokal ve Sneath Benzerlik Ölçüsü 3
	Sokal ve Sneath Benzerlik Ölçüsü 4
	Sokal ve Sneath Benzerlik Ölçüsü 5
	Kulczynski Benzerlik Ölçüsü 1
	Kulczynski Benzerlik Ölçüsü 2
	Hamann Benzerlik Ölçüsü
	Goodman ve Kruskal Lamda Benzerlik Ölçüsü
	Anderberg D Benzerlik Ölçüsü
	Yule Q Benzerlik Ölçüsü
	Yule Y Benzerlik Ölçüsü
	Ochiai Benzerlik Ölçüsü
	Fi 4 nokta Benzerlik Ölçüsü
	Yayılım Benzerlik Ölçüsü

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Uzaklık ve Benzerlik Ölçüleri İçin Veri Yapısı

- Sürekli ve kesikli sayısal verilere (oransal ya da aralık ölçekli) ilişkin ölçüler ile diğer veri tiplerine ilişkin ölçüler bu yapı üzerine kurulmuştur.

Tablo 7.2. Uzaklık ve Benzerlik Ölçüleri Veri Yapısı

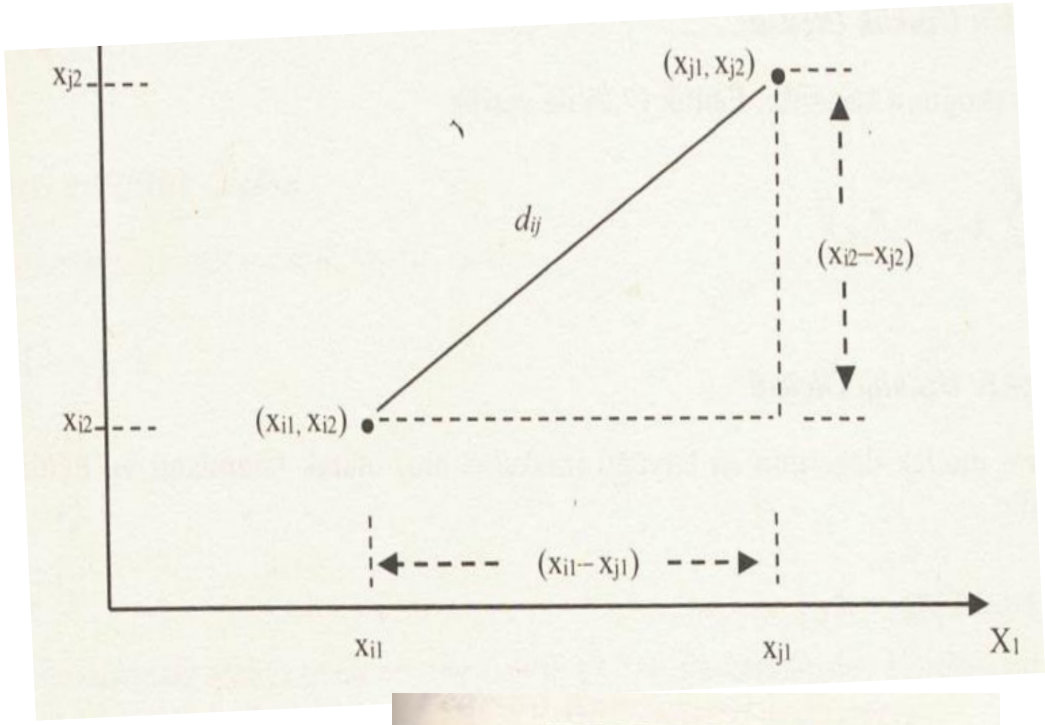
	Değişkenler				
Gözlem	X_1	X_2	.	.	X_P
1	x_{11}	x_{12}	.	.	x_{1p}
.	.	.	.	.	.
.	.	.	.	.	.
i	x_{i1}	x_{i2}	.	.	x_{ip}
j	x_{j1}	x_{j2}	.	.	x_{jp}
.	.	.	.	.	.
.	.	:	.	.	.
n	x_{n1}	x_{n2}	.	.	x_{np}

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Sürekli ve Kesikli Sayısal Veriler İçin Uzaklık Ölçüleri

A) Öklit (Euclidean) Uzaklık Ölçüsü

- Uzaklık ölçüleri arasında en çok kullanılanlardan biridir. X1 ve X2 gibi iki değişken olması durumunda herhangi iki gözlem x-y uzayında yandaki grafik ile gösterilebilir.
- Grafik'deki iki nokta (gözlem) arasındaki uzaklık (d_{ij}), Pisagor bağıntısı yardımıyla kolayca hesaplanabilir.
- (X_{i1},X_{i2}) ve (X_{j1},X_{j2}) gibi iki nokta arasındaki uzaklık aşağıdaki gibi hesaplanır ve bu uzaklığa Öklit uzaklığı denir.



$$d_{ij}=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2}$$

Üç değişken olduğunda Öklit uzaklığı aşağıdaki gibi yazılacaktır;

$$d_{ij}=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+(x_{i3}-x_{j3})^2}$$

P değişkenli bir yapı için genelleştirilmiş Öklit uzaklığı

$$d_{ij}=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$$

Burada,

x_{ik} : i. gözlemin k. değişken değeri,

x_{jk} : j. gözlemin k. değişken değeri,

p : değişken sayısıdır.

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Sürekli ve Kesikli Sayısal Veriler İçin Uzaklık Ölçüleri

B) Kare Öklit Uzaklık Ölçüsü

Öklit uzaklığının karesidir.

$$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

C) Chebychev Uzaklık Ölçüsü

- Farkların mutlak değerinin en büyüğü (maksimumu) olarak tanımlanır

$$d_{ij} = \max_k^p \left| x_{ik} - x_{jk} \right|$$

D) Manhattan City-Block Uzaklık Ölçüsü

- Farkların mutlak değerlerinin toplamı olarak tanımlanır

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|$$

E) Minkowski Uzaklık Ölçüsü

- m=1 için Manhattan City-Block uzaklık ölçüsünü, m=2 için Öklit uzaklık ölçüsünü verir, m artarken uzaklık Chebychev uzaklık ölçüsüne yaklaşır.

$$d_{ij} = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^m \right]^{1/m}$$

F) Karl Pearson Uzaklık Ölçüsü /Standartlaştırılmış Öklit Uzaklığı Ölçüsü

- Öklid uzaklığındaki farkların $\frac{1}{s_k^2}$ ile düzeltilmesi/standartlaştırılması ile elde edilir

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{1}{s_k^2} (x_{ik} - x_{jk})^2}$$

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Sürekli ve Kesikli Sayısal Veriler İçin Benzerlik Ölçüleri

a) Pearson Korelasyon Katsayısı

- İki rassal değişken arasındaki ilişkinin doğrusal derecesini belirlemek için kullanılmaktadır.
- Bu korelasyon yöntemi iki değişkenin kovaryansının, yine bu değişkenlerin standart sapmalarının çarpımına bölünmesiyle elde edilir.

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (X_i - \overline{Y})^2}}$$

veya

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

b) Kosinüs Benzerlik Ölçüsü

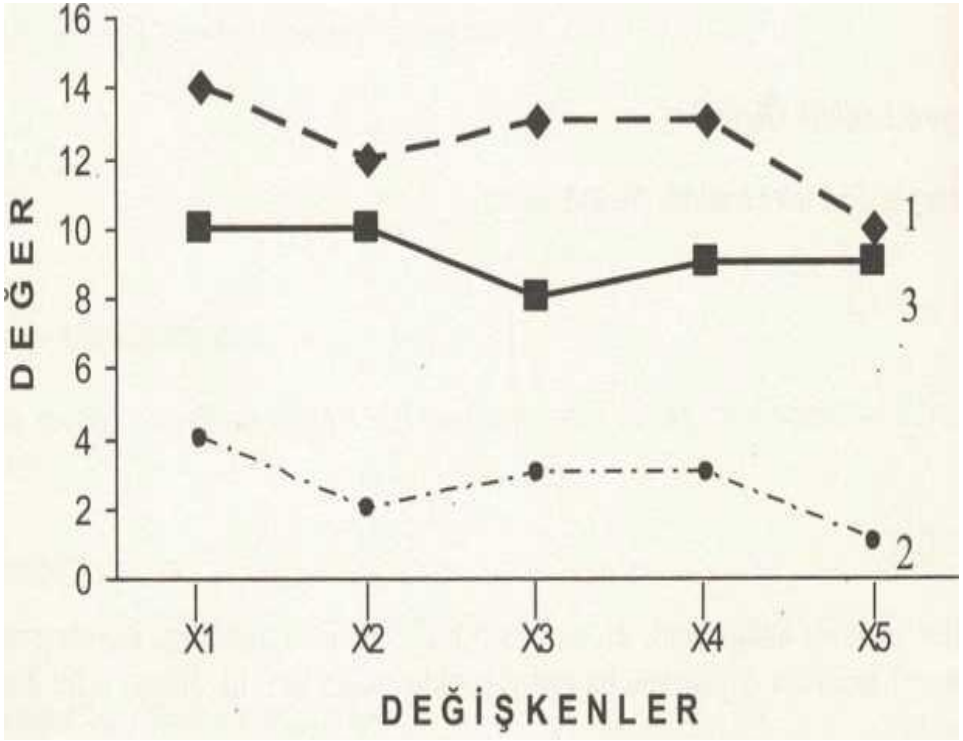
$$Benzerlik_{ij} = \frac{\sum_{i,j}^{p} x_i x_j}{\sqrt{\sum_{i=1}^{p} x_i^2 \sum_{j=1}^{p} x_j^2}}$$

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Sürekli ve Kesikli Sayısal Veriler İçin Benzerlik Ölçüleri

Örnek:

	Değişkenler				
Gözlem	X1	X2	X3	X4	X5
1	14	12	13	13	10
2	4	2	3	3	1
3	10	10	8	9	9



Grafik'de görüldüğü gibi 1. ve 2. gözlemin profili birbirine benzerken ilişki katsayısı da yüksektir (r=0,98). Dolayısıyla 1 ve 2 no'lu gözlemler benzer yapıdadır; ancak bu iki gözlem en uzak iki gözlemdir (Öklit uzaklığı=21,932)

	Korelasyon Katsayıları			Öklit Uzaklığı		
Gözlem	1	2	3	1	2	3
1	1,000			-		
2	0,983	1,000		21,932	-	
3	0,118	0,105	1,000	7,814	15,000	-

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Sıklık Sayıları için Uzaklık Ölçüleri

a) Ki-Kare Uzaklık Ölçüsü

- İki gözlem arasındaki ki-kare istatistiğinin karekökü uzaklık ölçüsü olarak kullanılır (Burada G: Gözlenen sıklık, B: Beklenen sıklık).

$$\left(\sqrt{(G-B)^2/B}\right)$$

b) Phi-Kare Uzaklık Ölçüsü

İki gözlem arasındaki ki-kare uzaklık ölçüsünün toplam gözlem sayısına bölünüp kare- kökünün alınması ile elde edilir

$$d_{ij} = \sqrt{Ki-kare_{ij}/(n_i+n_j)}$$

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
İkili Sınıflı (Bınary) Veriler İçin Uzaklık Ve Benzerlik Ölçüleri

- İki sınıflı 10 değişken için bir veri örneği aşağıdaki Tablo'da verilmiştir. Bu tür değişkenlerde genellikle 1: varlığı, 0: yokluğu belirtir.

Tablo 7.5. İki Sınıflı Değişkenler İçin Uzaklık ve Benzerlik Ölçüleri Veri Örneği

	Değişkenler									
Gözlem	1	2	3	4	5	6	7	8	9	10
i	0	0	1	1	0	1	1	0	1	1
j	1	0	0	0	1	0	1	1	0	1
.	.	.	.	.	.	.	.	.	.	.

- İki sınıflı veriler için geliştirilen uzaklık ve benzerlik ölçülerinin hesaplanmasında 2x2 boyutlarındaki çapraz tablolardan yararlanılır.

Tablo 7.6. 2×2 Boyutlu Genel Tablo

	Gözlem j		
Gözlem i	**1**	**0**	Toplam
1	*a*	*b*	*a+b*
0	*c*	*d*	*c+d*
Toplam	*a+c*	*b+d*	*a+b+c+d=p*

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
İki Sınıflı Veriler için Uzaklık ve Benzerlik Ölçüleri

a) Öklid Uzaklık Ölçüsü

$$d_{ij} = \sqrt{b+c}$$

b) Büyüklük Farkları Uzaklık Ölçüsü

En küçük değeri 0, en büyük değeri sınırsızdır.

$$d_{ij} = \frac{(b-c)^2}{(a+b+c+d)^2}$$

c) Biçim Farkları Uzaklık Ölçüsü

En küçük değeri 0, en büyük değeri 1 'dir.

$$d_{ij} = \frac{bc}{(a+b+c+d)^2}$$

a) Russell ve Rao Benzerlik Ölçüsü

- Sadece "var" olarak uyuşan çiftlerin (1-1) toplam içindeki payını verir. "Yok" olarak uyuşan çiftleri (0-0) dikkate almaz. Dolayısıyla sadece 1-1 olarak uyuşan çiftlere önem verilmek istendiğinde tercih edilir.
- Örneğin T.C. vatandaşları 1, T.C. vatandaşı olmayanların 0 olarak dikkate alındığı bir çalışmada, sadece T.C. vatandaşları ile ilgileniliyorsa uygun bir benzerlik ölçüsüdür. 0 ile 1 arasında değişir.

$$Benzerlik_{ij} = \frac{a}{a+b+c+d} = \frac{a}{p}$$

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
İki Sınıflı Veriler için Benzerlik Ölçüleri

b) Basit Benzerlik Ölçüsü (Simple Similarity Measure)

Toplam içerisinde birlikte "var" ve "yok" olarak uyuşan çiftlerin oranıdır, 1-1 ve 0-0 uyuşumlarına eşit ağırlık verir. Dolayısıyla 1-1 ve 0-0 uyuşumlarına daha fazla önem verildiğinde kullanılır. 0 (benzerlik yok) ile 1 (tam benzerlik) arasında değişim gösterir.

$$Benzerlik_{ij} = \frac{a+d}{a+b+c+d} = \frac{a+d}{p}$$

c) Jaccard Benzerlik Ölçüsü

Benzerlik oranı da denir. Birlikte "yok" olarak uyuşan çiftleri pay ve payda da dikkate almaz. Diğer bir deyişle 0-0 olarak uyuşan çiftler önemsiz olarak dikkate alınır. 0 ile 1 arasında değişim gösterir.

$$Benzerlik_{ij} = \frac{a}{a+b+c}$$

e) Parçalı Benzerlik Ölçüsü (Dice Similarity Measure)

Bu ölçü Czekanowski ölçüsü veya Sorensen ölçüsü olarak da bilinir. Birlikte "var" olan çiftlere iki kat ağırlık verir. Birlikte "yok" olarak uyuşan çiftlerin sayısı pay ve paydada yer almaz. 0-1 arasında değişim gösterir.

$$Benzerlik_{ij} = \frac{2a}{2a+b+c}$$

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Verilerin Standartlaştırılması

- Değişkenlerin farklı ölçüm birimlerine sahip olduğu durumlarda ya da değerlerin ortalama ve standart sapmalarının birbirinden çok farklı olduğu durumlarda uzaklık ve benzerlik ölçüleri hesaplanmadan önce verilerin standartlaştırılması gerekmektedir.
- **Örnek:** Dört ülkeye ilişkin **N**üfus **A**rtış **H**ızı (yüzde), **B**ebek **Ö**lüm **H**ızı (binde), **D**oğuşta **Y**aşam **B**eklentisi (yıl), **T**oplam **D**oğurganlık **H**ızı (kişi), **K**entsel **N**üfus **O**ranı (yüzde), Kişi Başına **G**ayri **S**afı **M**illi **G**elir (Dolar), **İ**lk **Ö**ğretime **K**ayıt Oranı (yüzde), Toplam Yetişkin **O**kur-**Y**azarlık **O**ranı (yüzde) verileri (2006-2007) Tablo'da verilmiştir.

Dört Ülkeye İlişkin Bazı Temel Göstergeler (2006-2007)

Ülke	NAH	BÖH	DYB	TDH	KNO	GSMH	İÖK	OYO
Fransa	0,4	4	80	1,9	77	34810	99	99
Türkiye	1,6	26	69	2,4	67	4710	87	87
Macaristan	-0,2	7	73	1,3	66	10030	89	99
Mısır	1,9	28	70	3,1	42	1250	95	71

z ile Standartlaştırılmış $z = \dfrac{x - \mu}{\sigma}$

Ülke	NAH	BÖH	DYB	TDH	KNO	GSMH	İÖK	OYO
Fransa	-0,5297	-0,9800	1,4094	-0,3603	0,9425	1,4569	1,1802	0,7538
Türkiye	0,681	0,7800	-0,8054	0,2948	0,2693	-0,5265	-0,9986	-0,1508
Macaristan	-1,135	-0,7400	0,0000	-1,1465	0,2020	-0,1759	-0,6355	0,7538
Mısır	0,9836	0,9400	-0,6040	1,2120	-1,4137	-0,7545	0,4539	-1,3568

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Verilerin Standartlaştırılması

* Standartlaştırma yöntemleri *

1. Z-standartlaştırması : orjinal verilerin ortalaması 0, standart sapması 1 olan yeni bir skora dönüştürülmesidir. ortalama 0 olduğu için verilen bir puanın ortalamanın altında ya da üstünde olduğu kolaylıkla söylenebilir. * $z_i = \frac{x_i - \bar{x}}{s}$

2. Dağılım Aralığı (DA) 0 ile 1 Arasında olacak şekilde standartlaştırma : Herbir değerden en küçük değer çıkartılıp dağılım aralığına bölünmesiyle hesaplanır.

* $s_i = \frac{x_i - x_{min}}{DA}$

3. En Büyük Değer 1 olacak şekilde standartlaştırma : Herbir gözlemin dağılımdaki en büyük değere bölünmesiyle elde edilir.

* $s_i = \frac{x_i}{x_{max}}$

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Verilerin Standartlaştırılması

4. Aritmetik ortalama 1 olacak şekilde standartlaştırma :

* $SI = \frac{Xi}{\bar{X}}$

5. ortalama 50, standart sapma 10 olacak şekilde standartlaştırma (T standartlaştırılması) :

$T = 10Z + 50$

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Uzaklık Ölçülerinin Hesaplanması

- **Örnek:** Aşağıdaki tabloda standartlaştırılmış veriler dikkate alınarak oransal ve aralık veriler için hesaplanan bazı uzaklık ölçüsü matrisleri verilmiştir.

Ülke	NAH	BÖH	DYB	TDH	KNO	GSMH	İÖK	OYO
Fransa	-0,5297	-0,9800	1,4094	-0,3603	0,9425	1,4569	1,1802	0,7538
Türkiye	0,681	0,7800	-0,8054	0,2948	0,2693	-0,5265	-0,9986	-0,1508
Macaristan	-1,135	-0,7400	0,0000	-1,1465	0,2020	-0,1759	-0,6355	0,7538
Mısır	0,9836	0,9400	-0,6040	1,2120	-1,4137	-0,7545	0,4539	-1,3568

- **a) Öklit Uzaklığı**
- 4x4 boyutlarındaki uzaklık matrisinin ilk elemanı;

$$dFransa,Türkiye = \sqrt{(-0,5297-0,681)^2+(-0,98-0,78)^2+\ldots+(0,7838-(-0,1508))^2} = 4,455$$

- Gözlemler için elde edilen matrisi iki farklı şekilde aşağıda verilmiştir. En yakın iki ülke Türkiye ve Mısır, en uzak iki ülke Fransa ve Mısır'dır.

Ülkeler	Fransa	Türkiye	Macaristan	Mısır
Fransa	-	4,455	3,089	5,285
Türkiye	4,455	-	3,068	**2,729**
Macaristan	3,089	3,068	-	4,672
Mısır	5,285	**2,729**	4,672	-

Ülkeler	Fransa	Türkiye	Macaristan
Türkiye	4,455		
Macaristan	3,089	3,068	
Mısır	5,285	**2,729**	4,672

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Uzaklık Ölçülerinin Hesaplanması

- Amaca uygun olarak değişkenlere ilişkin Öklit uzaklığı matrisi de elde edilebilir. Aşağıda verilen değişkenlere ilişkin Öklit uzaklığı matrisine göre, en uzak iki değişken TDH ve OYO iken en yakın iki değişken DYB ve GSMH'dir.

Gösterge	NAH	BÖH	DYB	TDH	KNO	SMH	İÖK	OYO
NAH	0,000	0,609	3,127	0,480	3,142	3,057	2,505	3,374
BÖH	0,609	0,000	3,340	0,927	3,222	3,292	2,842	3,373
DYB	3,127	3,340	0,000	2,992	1,439	0,366	1,270	1,412
TDH	0,480	0,927	2,992	0,000	3,227	2,964	2.209	3,413
KNO	3,142	3,222	1,439	3,227	0,000	1,215	2,419	0,721
GSMH	3,057	3,292	**0,366**	2,964	1,215	0,000	1,404	1,365
İÖK	2,505	2,842	1,270	2,209	2,419	1,404	0,000	2,472
OYO	3,374	3,373	1,412	**3,413**	0,721	1,365	2,472	0,000

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Uzaklık Ölçülerinin Hesaplanması

- **b) Chebychev Uzaklık Ölçüsü**
- 4x4 boyutlarındaki uzaklık matrisinin ilk elemanı;

$$d_{Fransa,\ Türkiye} = \max\left[|-0,5297-0,681|, |-0,98-0,78|, \ldots, |0,7838-(-0,1508|\right] = 2,2148$$

- Gözlemlere ilişkin uzaklık matrisi aşağıdadır. En yakın iki ülke Türkiye ve Mısır'dır.

Ülkeler	Fransa	Türkiye	Macaristan
Türkiye	2,2148		
Macaristan	1,8157	1,8159	
Mısır	2,3561	**1,6829**	2,3584

- **c) Manhattan City-Block Uzaklık Ölçüsü**
- 4x4 boyutlarındaki uzaklık matrisinin ilk elemanı;

$$d_{Fransa,\ Türkiye} = |-0,5297-0,681| + |-0,98-0,78| + \ldots + |0,7838-(-0,1508)| = 11,581$$

- Gözlemlere ilişkin uzaklık matrisi aşağıdadır. En yakın iki ülke Türkiye ve Mısır'dır.

Ülkeler	Fransa	Türkiye	Macaristan
Türkiye	11,581		
Macaristan	7,230	7,268	
Mısır	14,423	**6,151**	12,155

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Uzaklık Ölçülerinin Hesaplanması

- **d) Minkowski Uzaklık Ölçüsü**
- m=4 için elde edilen sonuçlar aşağıda verilmiştir.
- Gözlemlere ilişkin uzaklık matrisi aşağıdadır. En yakın iki ülke Türkiye ve Mısır'dır.

Ülkeler	Fransa	Türkiye	Macaristan
Türkiye	2,942		
Macaristan	2,184	2,157	
Mısır	3,283	**1,978**	3,057

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Benzerlik Ölçülerinin Hesaplanması

- **a) Korelasyon Matrisi**
- Değişkenlere göre standartlaştırılmış verilerden gözlemler için hesaplanan korelasyon matrisi ise aşağıdadır.

ÜLKELER	Fransa	Türkiye	Macaristan	Mısır
Fransa	1,000	-0,876	0,639	-0,738
Türkiye	-0,876	1,000	-0,375	0,389
Macaristan	0,639	-0,375	1,000	-0,951
Mısır	-0,738	0,389	-0,951	1,000

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Sıklık Sayıları için Uzaklık Ölçülerinin Elde Edilmesi

- **Örnek:** Bir ülkedeki 4 büyük ilde bir yıl içinde meydana gelen 5 önemli suça karışan kişi sayıları aşağıda verilmiştir. Bu veriler dikkate alınarak sıklık sayıları için saplanan Ki-kare ve Phi-kare uzaklık ölçüleri matrisleri aşağıda verilmiştir.

Beş Önemli Suça Karışan Kişi Sayıları

İller	Soygun	Saldırı	Tecavüz	Araba Hırsızlığı	Adi Hırsızlık
A	340	242	201	640	808
B	184	293	342	601	1668
C	68	103	89	467	1017
D	85	148	194	483	1206

a) Ki-kare uzaklık ölçüleri

Ki-kare yardımıyla A ve B illeri arasındaki uzaklığı bulmak için ikişerli ki-kare istatistikleri ((G-B)^2/B) hesaplanır ve karekökleri alınır.

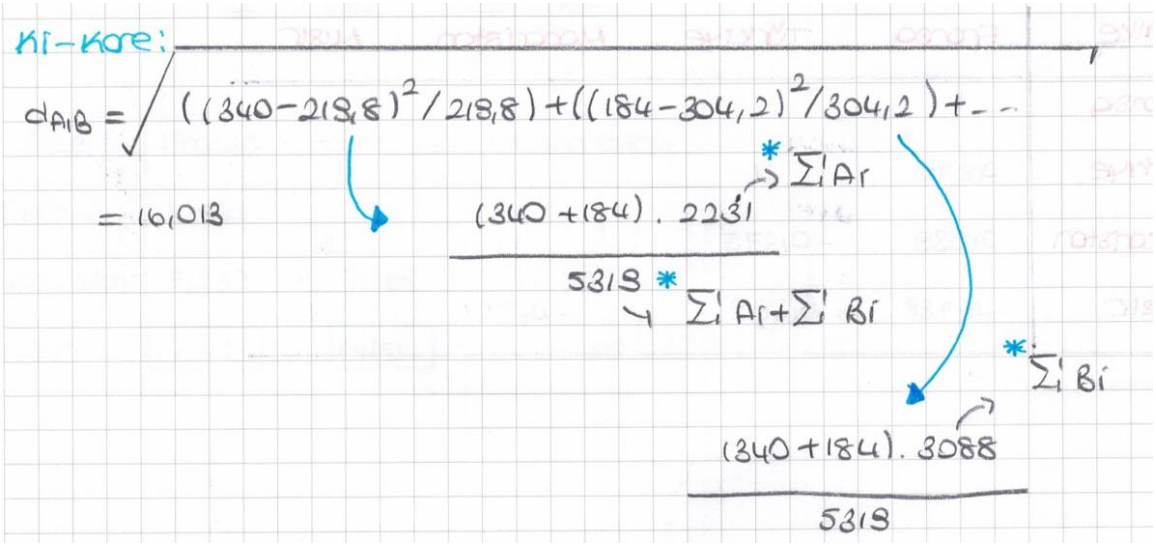
$$d_{A,B} = \sqrt{(340-219{,}8)^2/219{,}8 + (184-304{,}2)^2/304{,}2 + \ldots + (808-1038{,}5)^2/1038{,}5 + (1668-1437{,}5)^2/476{,}5}$$

Gözlemlere ilişkin uzaklık matrisi aşağıda verilmiştir. Buna göre uzak iki il A ve C illeridir.

İller	A	B	C
B	16,013		
C	16,614	10,088	
D	16,539	5,625	5,448

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

Sıklık Sayıları için Uzaklık Ölçülerinin Elde Edilmesi



b) Phi-Kare Uzaklık Ölçüsü

$$d_{A,B} = \sqrt{256{,}427/5319} = 0{,}2195 \cong 0{,}22$$

En uzak iki il A ve C illeridir.

İller	A	B	C
B	0,220		
C	0,264	0,145	
D	0,251	0,078	0,088

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
İkili (Binary) Veriler için Uzaklık ve Benzerlik Ölçülerinin Elde Edilmesi

Örnek : 5 bireyin 8 değişkene ilişkin verileri aşağıda verilmiştir. Uzaklık ve benzerlik ölçülerini bulunuz.

Kişi	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	0	0	1	1	0	0	0	0
2	1	1	1	0	1	1	1	1
3	1	0	0	1	0	1	0	0
4	0	0	0	0	0	1	0	1
5	1	0	0	1	1	0	0	0

X_1 : Düzenli spor — 1 : yapmıyor, 0 : yapıyor

X_2 : Sigara — 1 : içiyor, 0 : içmiyor

X_3 : Cinsiyet — 1 : Kadın, 0 : Erkek

X_4 : Yaş (yıl) — 1 : ≥45, 0 : <44

X_5 : Ağırlık (kg) — 1 : ≥80, 0 : <80

X_6 : Yüksek Tansiyon — 1 : var, 0 : yok

X_7 : Eğitim Düzeyi — 1 : ≥Lise, 0 : <Lise

X_8 : Doktora Düzenli Başvuru — 1 : Hayır, 0 : Evet

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ

İkili (Binary) Veriler için Uzaklık ve Benzerlik Ölçülerinin Elde Edilmesi

Genel Tablo

		Kişi 2		
		1	0	Toplam
Kişi 1	1	1	1	2
	0	6	0	6
Toplam		7	1	8

a) Öklid uzaklığı :

$$d_{rj} = \sqrt{b+c}$$

$$d_{1,2} = \sqrt{1+6} = 2{,}646$$

$$d_{1,2} = \sqrt{(0-1)^2 + (0-1)^2 + (1-1)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2}$$

$$d_{ii} = \sqrt{7} = 2{,}646$$

Kişi	1	2	3	4
2	2,646			
3	1,732	2,449		
4	2,000	2,236	1,732	
5	1,732	2,449	1,414	2,236

b) Benzerlik farkları uzaklık ölçüsü :

$$d_{1,2} = \frac{(1-6)^2}{(1+1+6+0)^2} = \frac{25}{64} = 0{,}391$$

Kişi	1	2	3	4
2	0,391			
3	0,016	0,250		
4	0,000	0,391	0,016	
5	0,016	0,250	0,000	0,016

Buna göre en uzak iki kişi 1. ve 2'inci kişi iken, en yakın iki kişi 1. ve 4'üncü kişiler ile 3. ve 5'inci kişilerdir.

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
İkili (Binary) Veriler için Uzaklık ve Benzerlik Ölçülerinin Elde Edilmesi

Benzerlik Ölçüleri

Russell ve Rao Benzerlik Ölçüsü

1. ve 2. kişi arasındaki benzerlik;

$$Benzerlik_{ij} = \frac{1}{1+1+6+0} = \frac{1}{8} = 0{,}125$$

Buna göre, sadece 1-1 gözlerinde uyuşan çiftlere önem verildiğinde en benzer olmayan iki kişi 1- 4'üncü kişiler ile 4-5'inci kişilerdir.

Kişi	1	2	3	4
2	0,125			
3	0,125	0,250		
4	0,000	0,250	0,125	
5	0,125	0,250	0,250	0,000

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Kümeler/Örneklemler/Gruplar Arası Uzaklık Ölçüleri

Verilere ilişkin ortalama, varyans ve kovaryansların bilinmesi durumunda çok değişkenli örneklemlerin (dolayısıyla evrenlerin) arasındaki uzaklıkların belirlenmesine yönelik geliştirilmiş birçok ölçü vardır. Bunlardan en bilinenleri Mahalanobis uzaklık ölçüsü, Hotelling T^2 uzaklık ölçüsü ve Penrose uzaklık ölçüsüdür.

a) Mahalanohis Uzaklık Ölçüsü

Örneklem varyans-kovaryans matrislerinin homojen olduğu varsayımı altında iki grup arasındaki Mahalanobis uzaklığı,

$$D_{ij}^2 = (\mu_i - \mu_j)' \, S^{-1} \, (\mu_i - \mu_j)$$

Burada,

μ_i : i. evrenin (grubun) ortalama vektörü,

μ_i : j. evrenin (grubun) ortalama vektörü,

S^{-1} : Ortak (pooled) varyans-kovaryans matrisinin tersidir.

BÖLÜM 3 : UZAKLIK ve BENZERLİK ÖLÇÜLERİ
Kümeler/Örneklemler/Gruplar Arası Uzaklık Ölçüleri

a) Hotelling T^2 uzaklık ölçüsü

İki grup ya da kümenin ortalama vektörlerinin karşılaştırılmasında kullanılan Hotelling T^2 değeri bir uzaklık ölçüsü olup,

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2} \right) \sum_{k=1}^{p} (\mu_{ik} - \mu_{jk})' S^{-1} (\mu_{ik} - \mu_{jk})$$

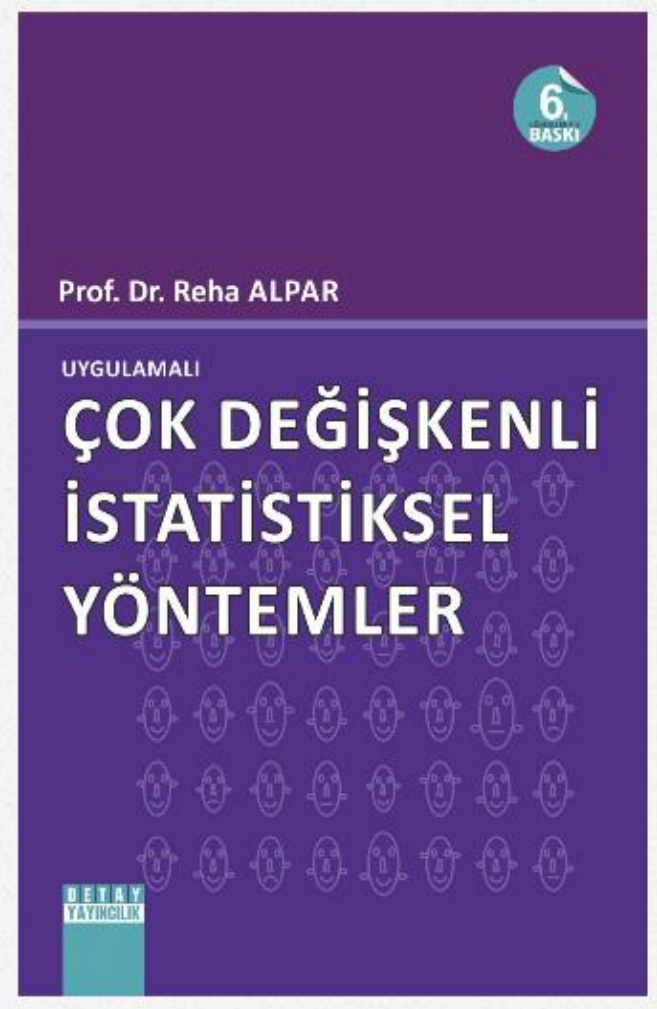
μ_{ik} : i. kümenin k. değişken ortalaması

μ_{jk} : j. kümenin k. değişken ortalaması

S^{-1} : Toplanmış varyans - kovaryans matrisinin tersi.

ÇOK DEĞİŞKENLİ İSTATİSTİKSEL ANALİZ

- Sunum hazırlanırken aşağıdaki kaynaktan yararlanılmıştır.




KARADENİZ TEKNİK ÜNİVERSİTESİ
1955