Bölüm 3

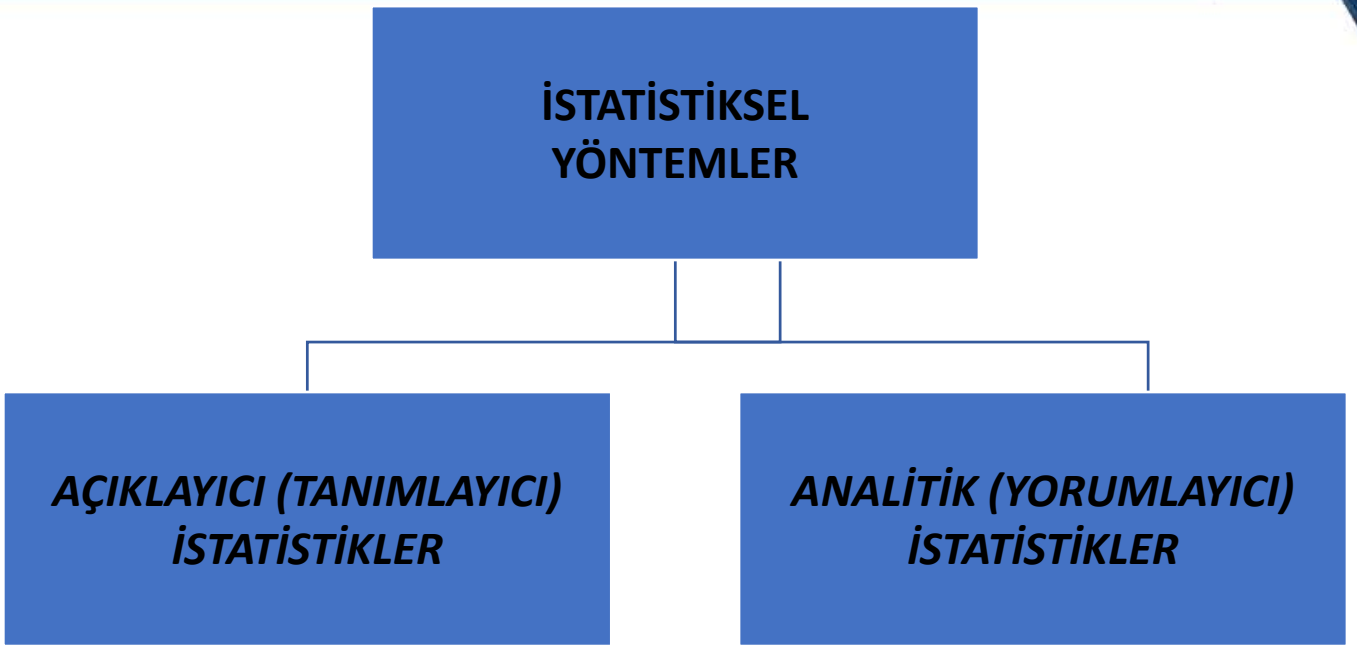
Sayısal Tanımlayıcı Ölçütler

Öğrenme hedefleri

Bu bölümde aşağıdaki konular planlanmıştır:

- Sayısal verilerin merkezi eğilim, değişim (variation) ve şekil özelliklerinin tanımlanması
- Bir popülasyon için tanımlayıcı özet ölçülerinin hesaplanması
- Bir kutu diyagramının (box-plot) oluşturulması ve yorumlanması
- Kovaryans ve korelasyon katsayılarının hesaplanması

Giriş



Eğer sadece verileri özetlemek, sunmak, gerçekleri kataloglamak istiyorsak tanımlayıcı istatistikleri kullanırız.

Ancak örnek verisine dayalı olarak yorumlama yapmak istendiğinde veya belirsizlik altında karar verileceğinde yorumlayıcı istatistik metotlarından faydalanırız.

Tanımlayıcı istatistikler

- Bir veya birden fazla dağılışı karşılaştırmak için kullanılan ve ayrıca örnek verilerinden hareket ile frekans dağılışlarını sayısal olarak özetleyen değerlere tanımlayıcı istatistikler denir.
- Verilerin özetlenmesinde sayısal metotlarla birlikte birçok güçlü grafik tabanlı tekniklerin olduğunu bir önceki dersimizde gördük. Grafik tabanlı teknikler özellikle önemlidir. Herhangi iyi bir istatistiksel veri analizi daima verilerin **grafik çizimi** (plotting the data) ile başlamalıdır.

Tanımlayıcı istatistikler

Tanımlayıcı İstatistikler

Yer (Mer.Eğilim) Ölçüleri

1)Aritmetik ort.
2)Geometrik ort.
3)Harmonik ort.
4)Mod
5)Medyan

Değişkenlik Ölçüleri

1) Range (Değişim Aralığı)
2) Varyans
3) Standart Sapma
4) Değişkenlik(Varyasyon) Katsayısı

Çarpıklık Ölçüleri

Basıklık Ölçüleri

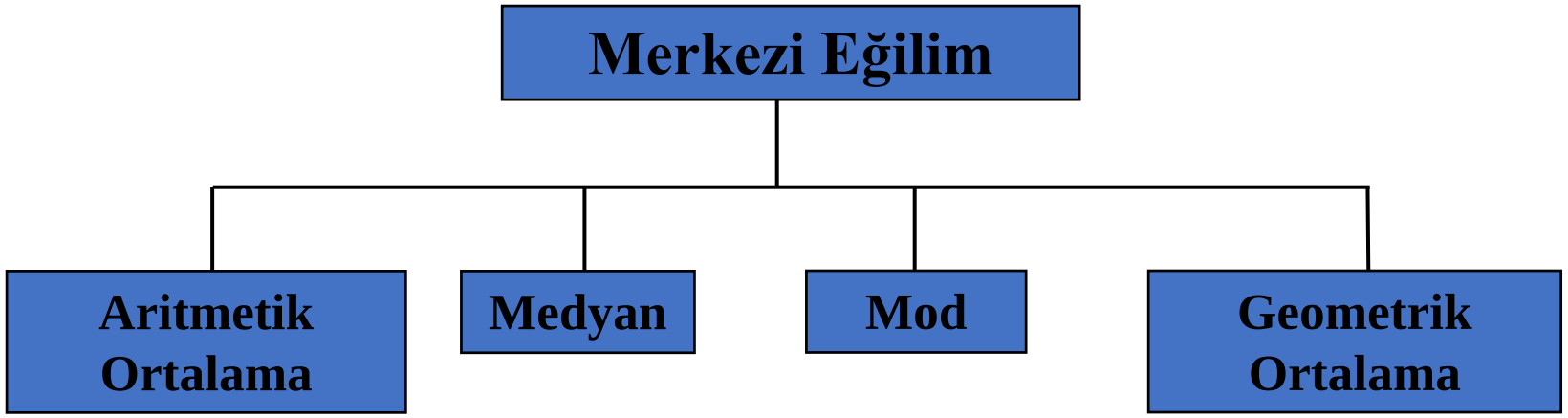
Özet tanımlar

- **Merkezi eğilim**, tüm veri değerlerinin tipik veya merkezi bir değer etrafında dağılımını gösterir.

- **Değişim (variation)**, değerlerin dağılma veya saçılma miktarıdır.

- **Şekil (shape)**, değerlerin en düşük değerden en yüksek değere dağılım motifidir.

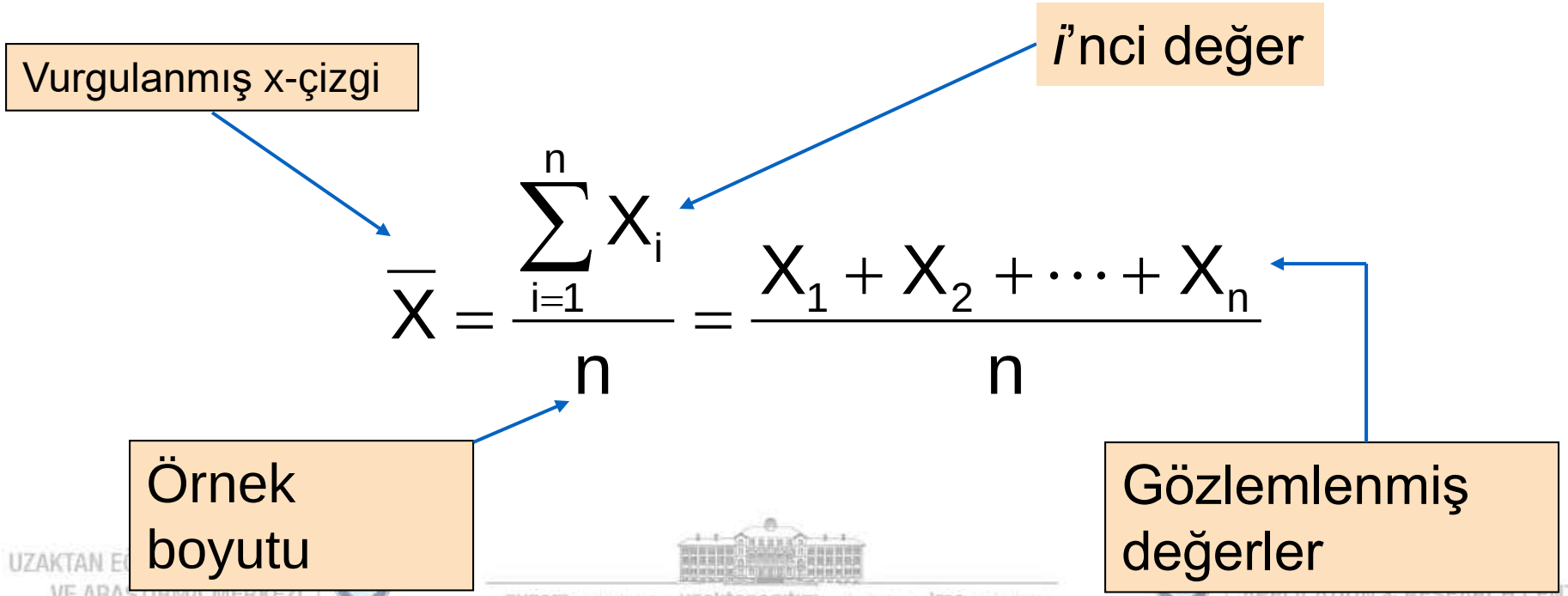
Merkezi eğilim ölçüleri



Merkezi eğilim ölçüleri: Ortalama

- Aritmetik ortalama (çoğunlukla "ortalama" olarak adlandırılır) merkezi eğilimin en yaygın ölçütüdür

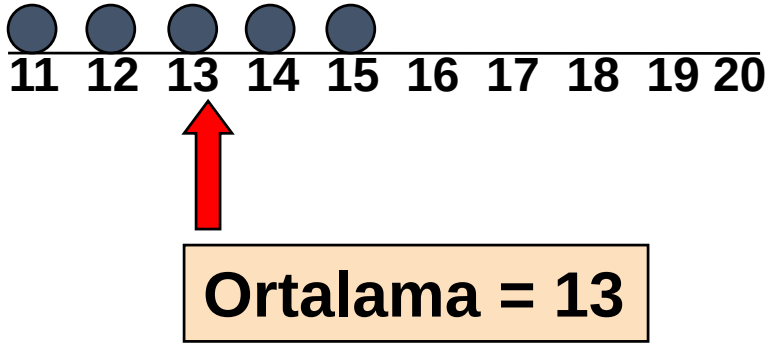
  - n boyutundaki bir örnek için:



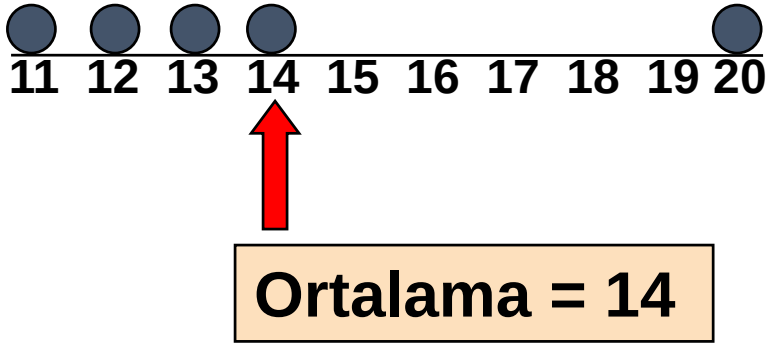
guzem.gazi.edu.tr • uzaktanegitim.gazi.edu.tr • lms.gazi.edu.tr

Merkezi eğilim ölçüleri: Ortalama

- Merkezi eğilim ölçülerinin en yaygınıdır
- Ortalama = değerlerin toplamının değerlerin sayısına bölünmesi
- Uç değerlerinden etkilenir (sınırdışı değerler)



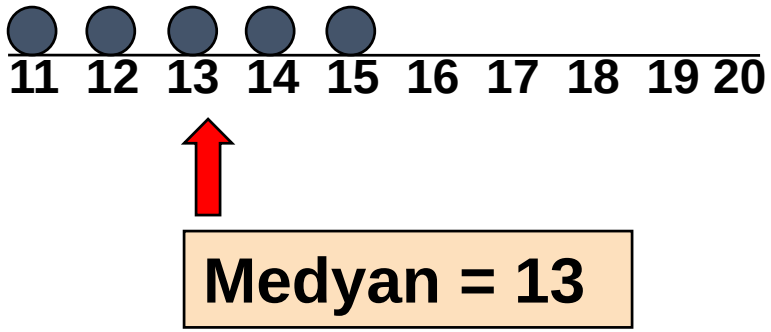
$$\frac{11+12+13+14+15}{5}=\frac{65}{5}=13$$

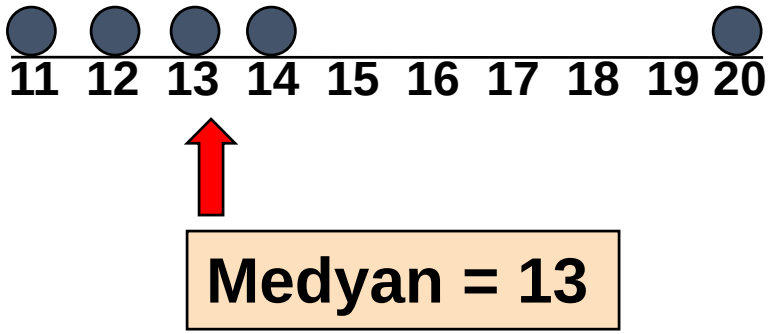


$$\frac{11+12+13+14+20}{5}=\frac{70}{5}=14$$

Merkezi eğilim ölçüleri:
Medyan (ortanca- orta değer)

- Sıralı bir seride, medyan "orta" sayıdır. (% 50 yukarıda,% 50 aşağıda)





- Uç değerlerden etkilenmez

Merkezi eğilim ölçüleri: Medyan'ın yönlendirilmesi

- Değerler sayısal sırayla (küçükten büyüğe) olduğunda ortanın konumu:

$$Medyan\ konumu = sıral\ verinin\ \frac{n+1}{2}\ inci\ konumunda$$

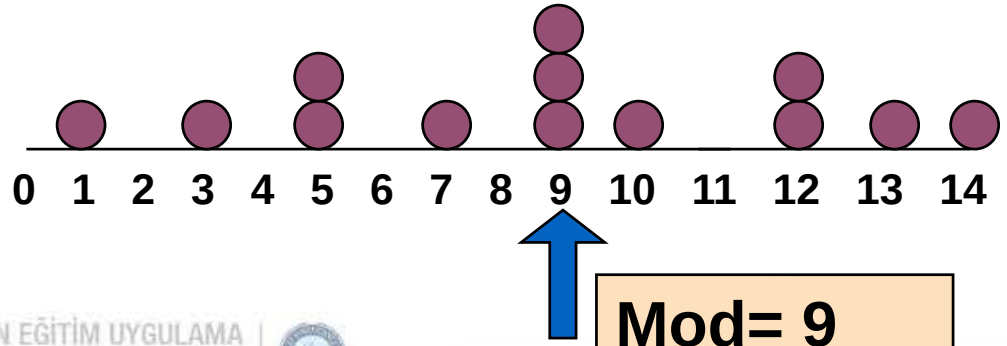
- Eğer değerlerin sayısı tekse, medyan ortadaki sayıdır.

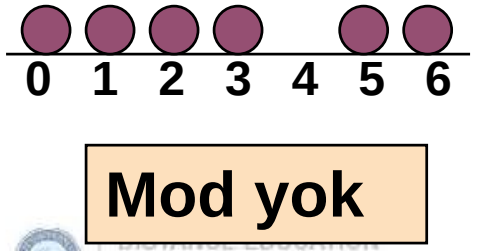
- Eğer değerlerin sayısı çiftse, medyan ortadaki iki sayının ortalamasıdır (n/2 ve (n+2)/2).

$\frac{n+1}{2}$'nin medyan'ın *değeri* olmadığını, yalnızca sıralanan verideki medyan'ın *konumunun* olduğunu belirtmek gerekir.

Merkezi eğilim ölçüleri: Mod (tepe değeri)

- En sık meydana gelen değer
- Uç değerlerden etkilenmez
- Hem sayısal hem de kategorik (nominal) veriler için kullanılır
- Hiç mod olmayabilir
- Birden çok mod mevcut olabilir





Merkezi eğilim ölçüleri: Örnek

Konut Fiyatları (TL):
2,000,000
500,000
300,000
100,000
100,000
Toplam **3,000,000**

- **Ortalama:** (3,000,000/5)
  = **600,000**
- **Medyan:** sıralı verinin orta değeri
  = **300,000**
- **Mod:** en sık tekrarlanan değer
  = **100,000**

Merkezi eğilim ölçüleri: Hangi ölçü seçilmeli?

- **Ortalama,** genellikle, aşırı değerler (sınırdışı değerler) mevcut olmadığı sürece kullanılır
- **Medyan**, aşırı değerlere duyarlı olmadığı için sıklıkla kullanılır. Örneğin, bir bölge için medyan ev fiyatları söylenebilir; aşırı uç değerlere karşı daha az duyarlıdır.
- Bazı durumlarda, hem **ortalamanın** hem de **medyanın** bildirilmesi mantıklıdır.

Bir değişkenin zaman içindeki değişkenlik oranı için merkezi eğilim ölçüsü : Geometrik ortalama & Geometrik getiri oranı

- Geometrik ortalama
  - Zaman içinde bir değişkenin değişim oranını ölçmek için kullanılır

$$\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

- Geometrik ortalama getiri oranı
  - R_i zaman periyodundaki i getiri oranını göstermek üzere bir yatırımın zaman içindeki durumunu ölçer

$$\overline{R}_G = [(1+R_1) \times (1+R_2) \times \cdots \times (1+R_n)]^{1/n} - 1$$

Geometrik ortalama getiri oranı: Örnek

100.000 dolarlık bir yatırımın, birinci yılın sonunda 50.000 dolara gerilediği ve ikinci yılın sonunda 100.000 dolara geri döndüğü gözlemlenmiş :

$$X_1 = \$100{,}000 \quad X_2 = \$50{,}000 \quad X_3 = \$100{,}000$$

50% düşüş 100% artış

Tüm iki yıllık sürecin geri dönüş miktarı sıfırdır, çünkü aynı seviyede başlamış ve bitmiştir.

Geometrik ortalama getiri oranı: Örnek

Aritmetik ortalama ve geometrik ortalama hesaplamak için 1 yıllık getirileri kullanalım :

Aritmetik ortalama getiri oranı:

$$\overline{X} = \frac{(-.5) + (1)}{2} = .25 = 25\%$$

Yanıltıcı sonuç

Geometrik Ortalama getiri oranı:

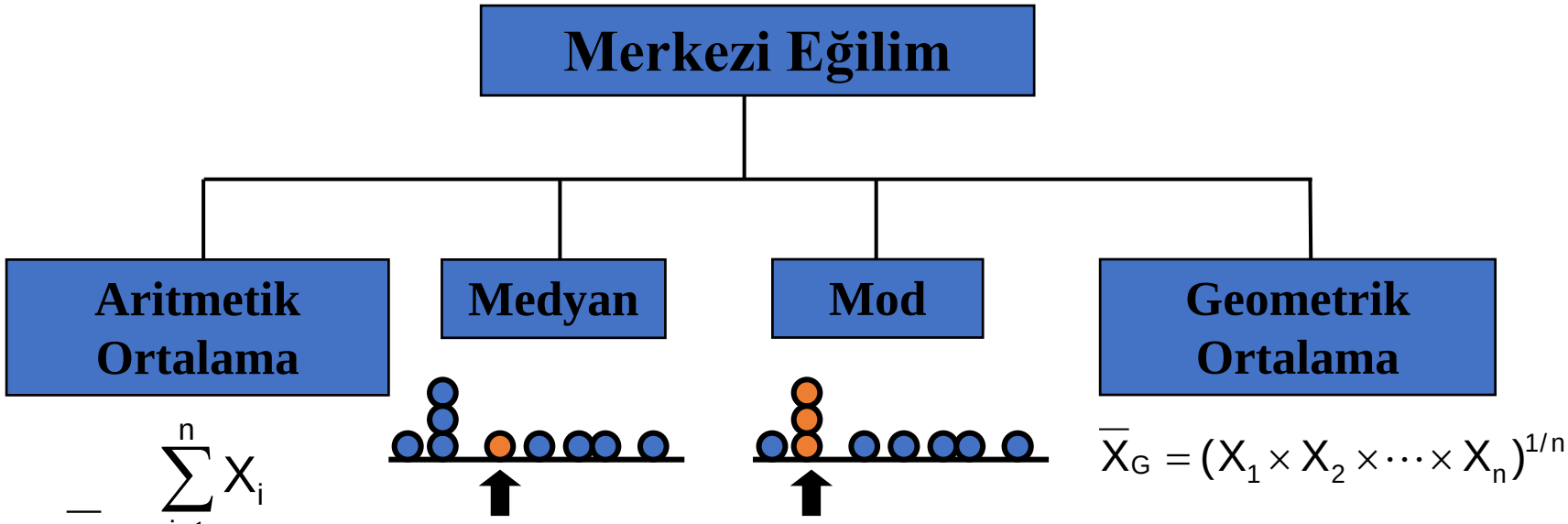
$$\overline{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$

$$= [(1+(-.5))\times(1+(1))]^{1/2} - 1$$

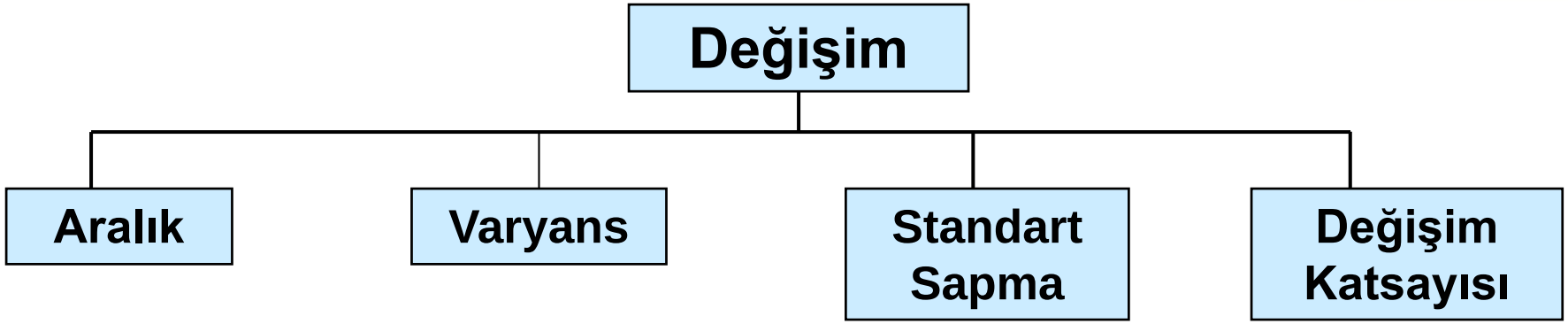
$$= [(.50)\times(2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%$$

Daha çok temsil eden sonuç

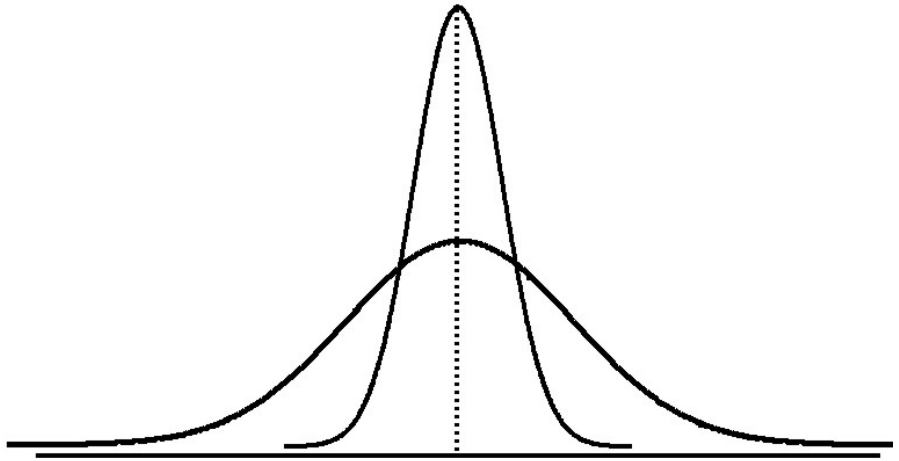
Merkezi eğilim ölçüleri: Özet



Değişim ölçüleri



- Değişim ölçüleri, veri değerlerinin **yayılımı** veya **değişkenliği** veya **dağılımı** hakkında bilgi verir.



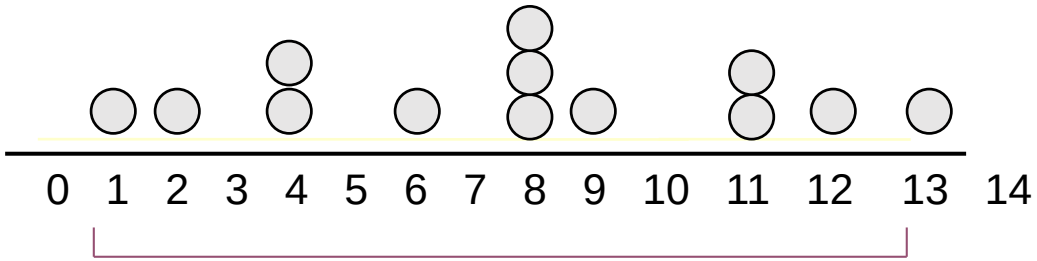
Aynı merkez,
Farklı değişim (sapma)

Değişim ölçüleri: Aralık

- En basit değişim ölçüsüdür.
- En büyük ve en küçük değerler arasındaki farktır:

$$\text{Aralık} = X_{\text{en büyük}} - X_{\text{en küçük}}$$

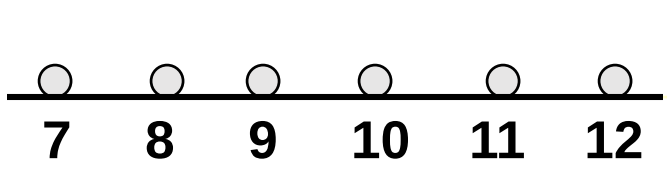
Örnek:



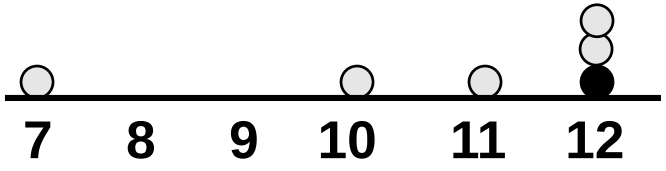
Aralık = 13 - 1 = 12

Değişim ölçüleri:
Neden aralık yanıltıcı olabilmektedir?

- Verilerin nasıl dağılım gösterdiğini ihmal eder.



Aralık = 12 - 7 = 5



Aralık = 12 - 7 = 5

- Uç noktadaki değerlere duyarlıdır.

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Aralık = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Aralık = 120 - 1 = 119

Değişim ölçüleri: Örnek varyansı

- Değerlerin ortalamadan sapmalarının karelerinin (yaklaşık) ortalaması
  - Örnek varyansı:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

$\overline{X}$ = aritmetik ortalama

n = örnek büyüklüğü

X_i = X değişkeninin i'nci değeri

Değişim ölçüleri:
Örnek standart sapması

- En çok kullanılan değişim ölçüsüdür.
- Ortalama çevresindeki değişim miktarını gösterir.
- Varyansın kare köküdür
- Orijinal veri ile aynı birimlere sahiptir.

  - Örnek Standart Sapması:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Değişim ölçüleri: Standart sapma

Standart Sapmanın Hesaplanması İçin Adımlar

1. Her bir değer ile ortalama arasındaki farkı hesapla
2. Her bir farkın karesini al.
3. Fark karelerini topla.
4. Örnek varyansını elde etmek için toplamı (n-1) ile böl.
5. Örnek Standart sapmayı elde etmek için örnek varyansının kare kökünü al.

Değişim ölçüleri:
Örnek standart sapması:
hesaplama örneği

Örnek Veri (X_i) :

10	12	14	15	17	18	18	24

n = 8 Ortalama = $\overline{X}$ = 16

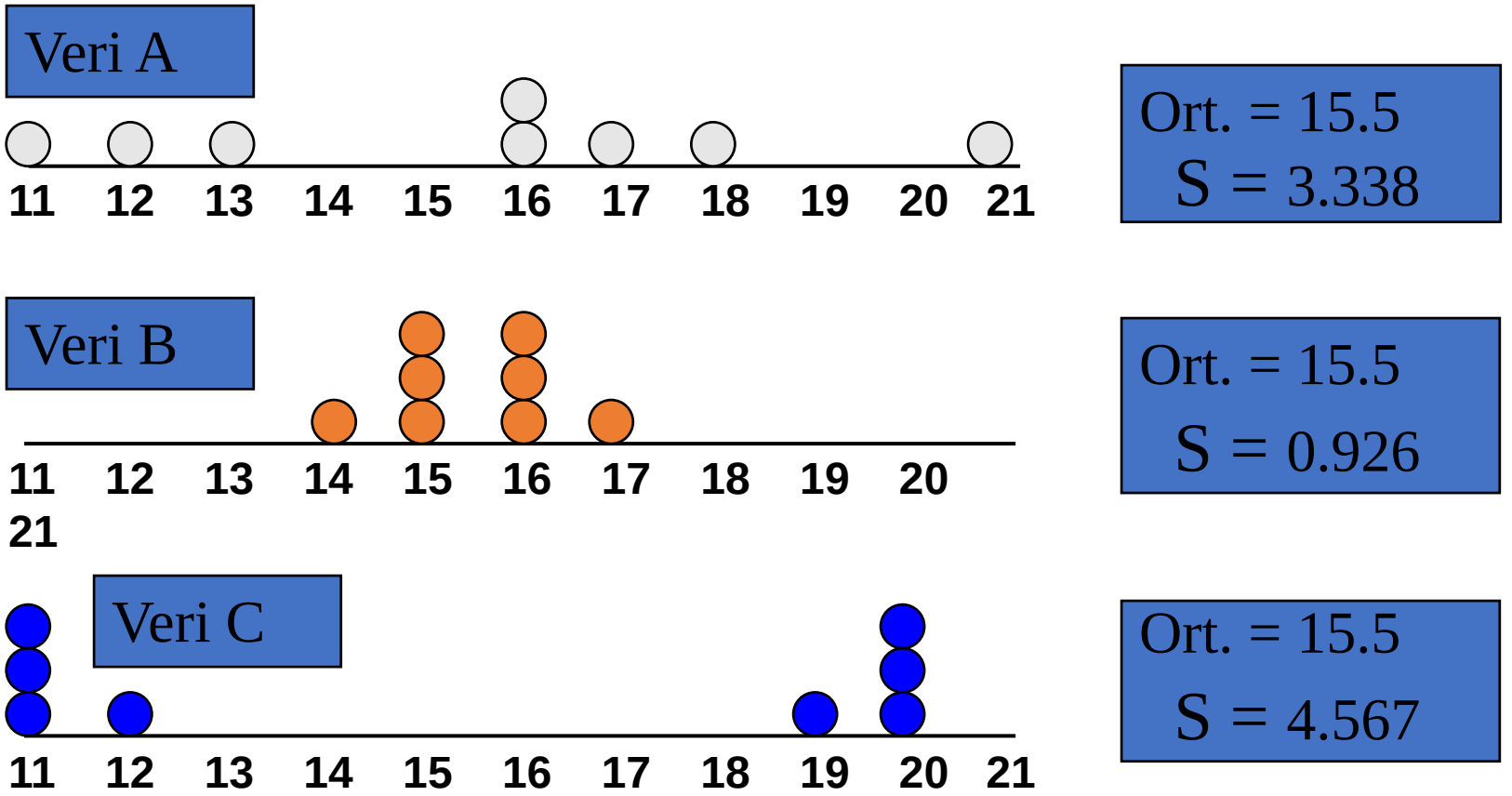
$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \cdots + (24-\overline{X})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{130}{7}} = 4.3095$$

Ortalama çevresindeki "ortalama" saçılımın bir ölçüsü

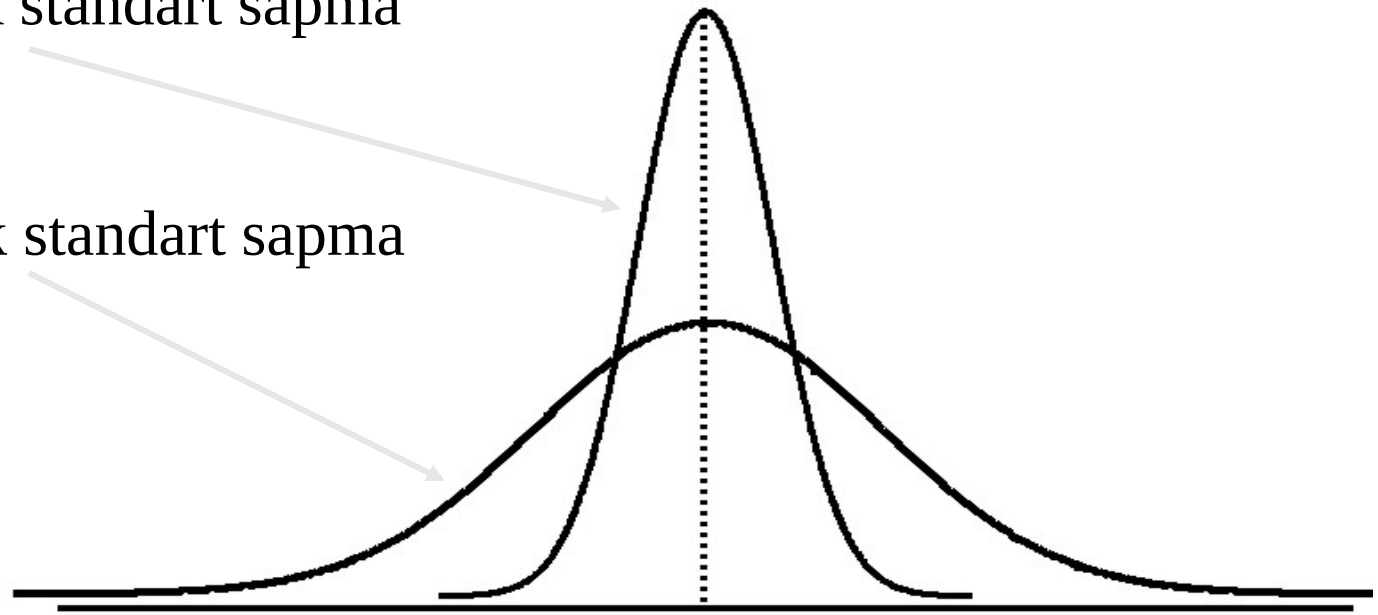
Değişim ölçüleri:
Standart sapmaların kıyaslanması



Değişim ölçüleri:
Standart sapmaların kıyaslanması

Daha küçük standart sapma

Daha büyük standart sapma

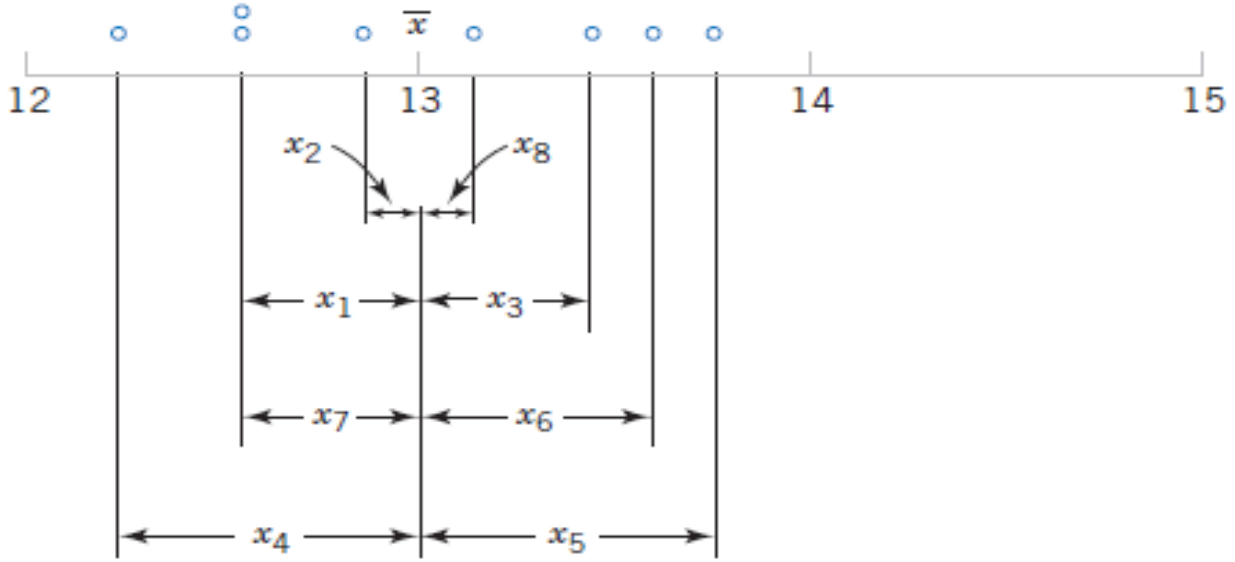


Değişim ölçüleri: Özet özellikler

- Veriler ne kadar çok yayılırsa aralık, varyans ve standart sapma o kadar büyük olacaktır.
- Veriler yoğunlaştıkça aralık, varyans ve standart sapma o kadar az olur.
- Eğer tüm değerler aynı ise (değişim yok), tüm bu ölçüler sıfır olacaktır.
- Bu ölçülerin hiçbiri hiçbir zaman negatif olamazlar.

Örneklem varyansı değişkenliği nasıl ölçer?

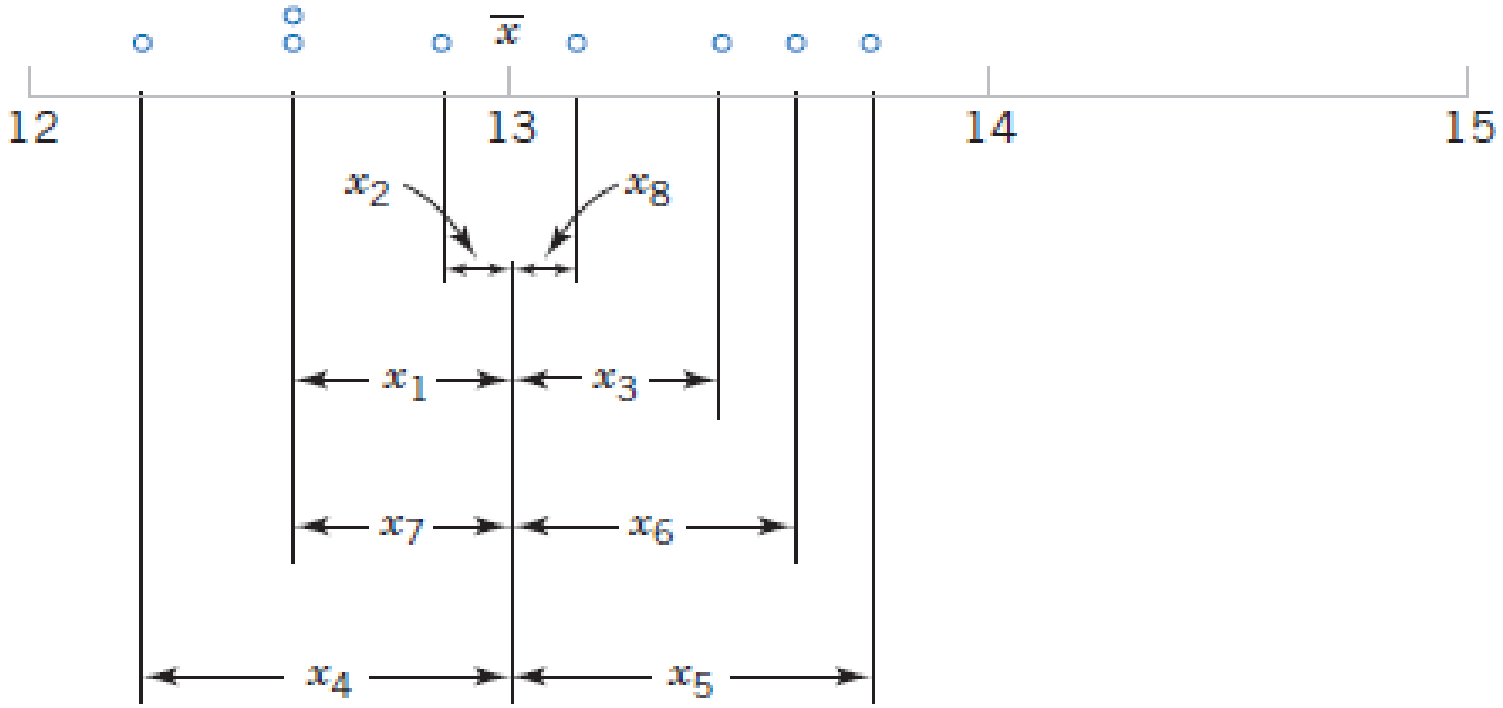
- Örneklem varyansının değişkenliği nasıl ölçtüğünü görebilmek için konnektör çekme kuvveti ile ilgili veriler için $x_i - \overline{x}$ sapmalarını gösteren aşağıdaki şekli inceleyelim.
- Çekme kuvveti verilerindeki değişkenlik miktarı arttıkça $x_i - \overline{x}$ sapmalarının bazılarının mutlak büyüklüğü de artacaktır. $x_i - \overline{x}$ sapmalarının toplamı daima sıfır olacağından negatif sapmaları pozitif miktarlara dönüştüren bir değişkenlik ölçütü kullanmamız gerekmektedir.



UZAKTAN EĞİTİM UYGULAMA VE ARAŞTIRMA MERKEZİ
CENTER

Örneklem varyansı değişkenliği nasıl ölçer?

- Sapmaların karesini almak örneklem varyansında kullanılan yaklaşımdır. Sonuç olarak eğer s^2 küçükse verilerdeki değişkenlik miktarı göreli olarak düşüktür ancak s^2 büyükse değişkenlikte göreceli olarak büyüktür.



Değişim ölçüleri: Değişim katsayısı

- Bağıl değişimi ölçer
- Daima yüzdeli olarak ifade edilir (%)
- Ortalamayla bağlantılı olarak değişimi gösterir.
- Farklı birimlerde ölçülen iki veya daha fazla veri kümesinin değişkenliğini karşılaştırmak için kullanılabilir

$$CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$$

Değişim ölçüleri: Değişim katsayılarının kıyaslanması

- A Hisse senedi:
  - Geçmiş yıl ortalama fiyatı = $50
  - Standart Sapma= $5

$$CV_A = \left(\frac{S}{\overline{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- B Hisse senedi:
  - Geçmiş yıl ortalama fiyatı= $100
  - Standart Sapma= $5

$$CV_B = \left(\frac{S}{\overline{X}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Her iki hisse senedi aynı standart sapmaya sahiptir, fakat B hisse senedi fiyatına göre daha az değişkenlik göstermektedir.

Değişim ölçüleri: Değişim katsayılarının kıyaslanması

- A Hisse Senedi:
  - Geçmiş yıl ortalama fiyatı= $50
  - Standart sapma = $5

$$CV_A = \left(\frac{S}{\overline{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

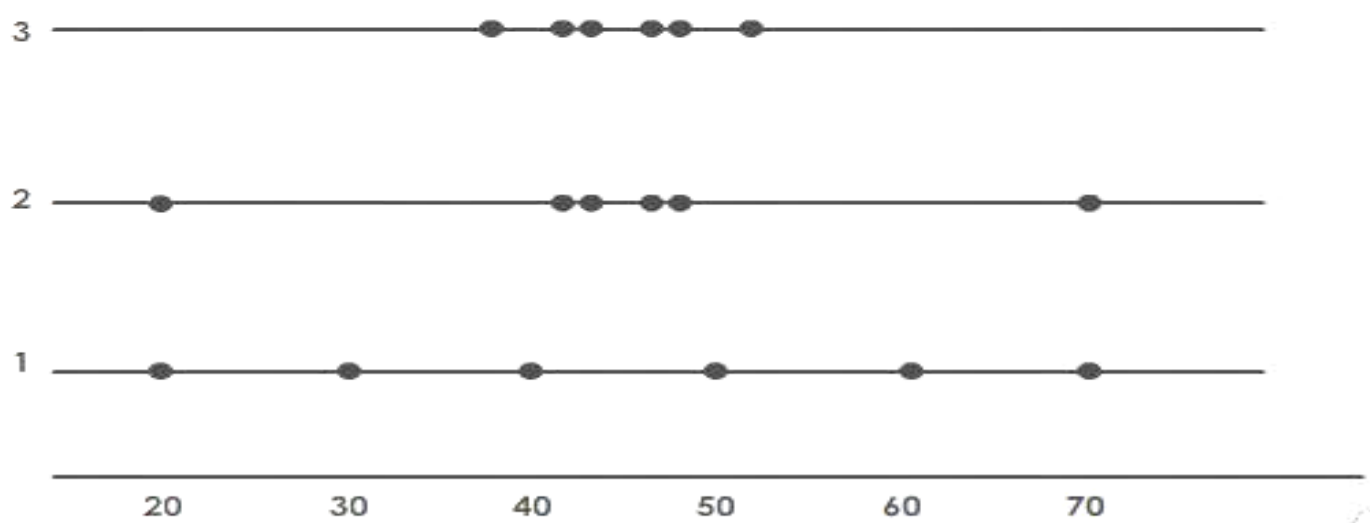
- C Hisse Senedi:
  - Geçmiş yıl ortalama fiyatı= $8
  - Standart sapma= $2

$$CV_C = \left(\frac{S}{\overline{X}} \right) \cdot 100\% = \frac{\$2}{\$8} \cdot 100\% = 25\%$$

C hisse senedinin daha küçük bir standart sapması vardır fakat daha yüksek bir değişim katsayısı vardır

Örnek

20	40	50	30	60	70
47	43	44	46	20	70
44	43	40	50	47	46



UZAKTAN EĞİ
VE ARAŞ
RCH CENTER

Aşırı uç noktaların tespiti: Z-değeri

- Bir veri değerinin **Z-değerini** hesaplamak için, bu veriden ortalama çıkarılır ve standart sapma ile bölünür.
- Z-değeri, bir veri değerinin ortalamalardan olan standart sapmaların sayısıdır.
- Bir veri değeri, Z-değeri -3.0'dan küçük veya +3.0'dan büyükse, aşırı bir uç nokta olarak kabul edilir.
- Z-değerinin mutlak değeri ne kadar büyük olursa, veri değeri ortalamadan o kadar uzaktır.

Aşırı uç noktaların tespiti: Z-değeri

$$Z = \frac{X - \overline{X}}{S}$$

X veri değerini

X örnek ortalamasını

S örnek standart sapmasını göstermektedir

Uç noktaların tespiti: Z-değeri

- SAT sınavı matematik ortalama puanının 490, standart sapmasının 100 olduğunu varsayalım
- 620 test puanı için Z- değerini hesaplayalım.

$$Z = \frac{X - \overline{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

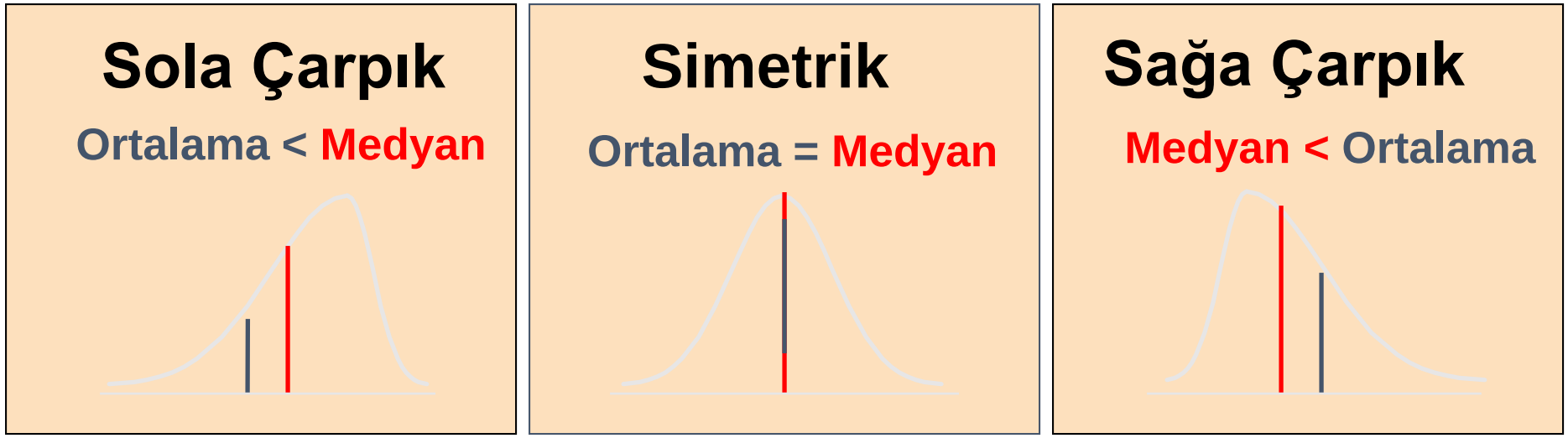
620 puanı, ortalamanın 1.3 üstünde standart sapma göstermekte ve bir aşırı uç nokta olarak görülmemektedir.

Bir dağılımın şekli

- Verinin nasıl bir dağılım gösterdiğini ifade eder
- Şekle ilişkin iki yararlı istatistik şu şekildedir:
  - Çarpıklık
    - ✓Veri değerlerinin ne miktarda simetrik olmadığını ölçer
  - Basıklık (Kurtosis)
    - ✓Basıklık, dağılım eğrisinin zirve yapma durumunu etkilemektedir-yani, eğrinin dağılım merkezine yaklaştıkça ne kadar keskin bir şekilde yükseldiğini gösterir.

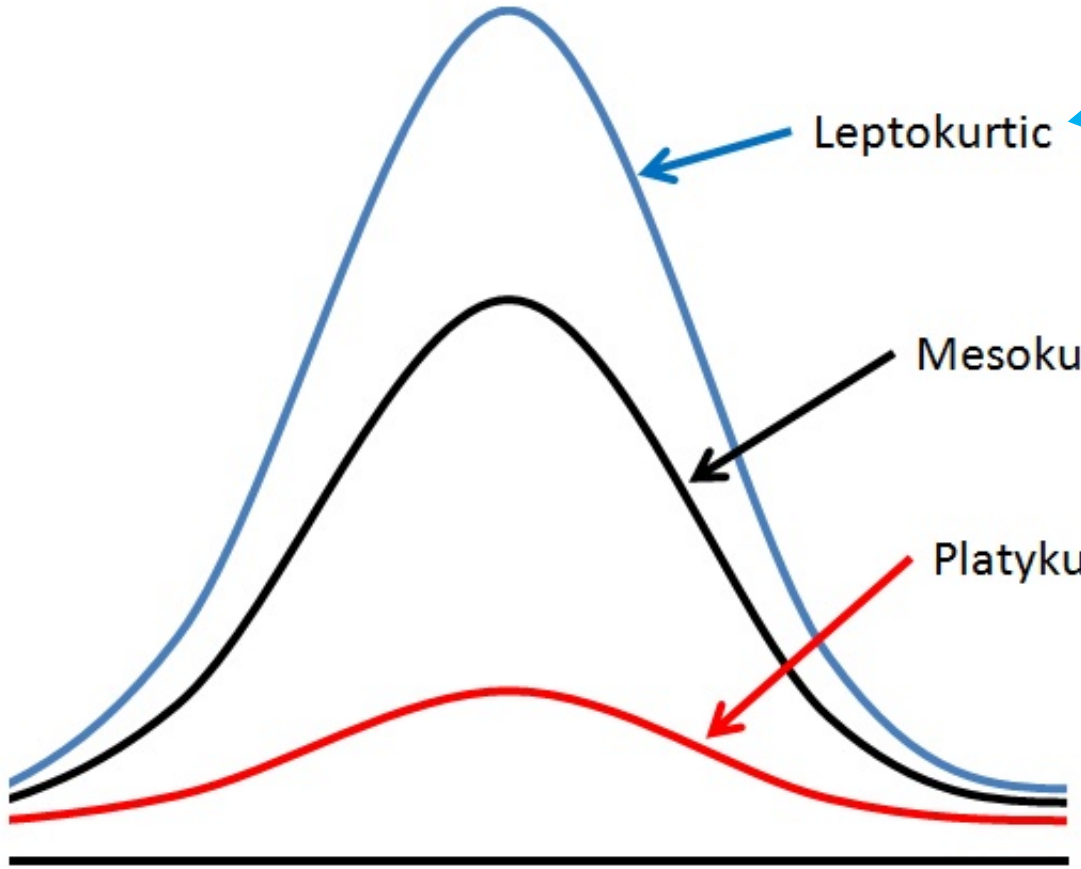
Bir dağılımın şekli (Çarpıklık)

- Verilerin hangi oranda simetrik olmadığını ölçer



Çarpıklık İstatistiği	< 0	0	>0

Bir dağılımın şekli -- Basıklık (Kurtosis) dağılımın merkezine yaklaştığında eğrinin ne kadar keskin bir şekilde yükseldiğini ölçer)



Çan şekline göre Daha Keskin Zirveli (Kurtosis > 0)

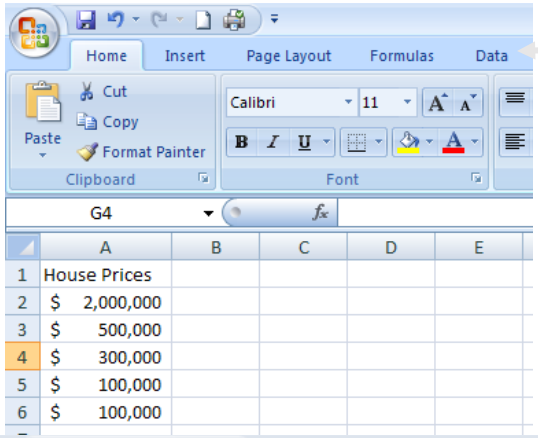
Çan Şekilli (Kurtosis = 0)

Çan şekline göre Daha Düz (Kurtosis < 0)

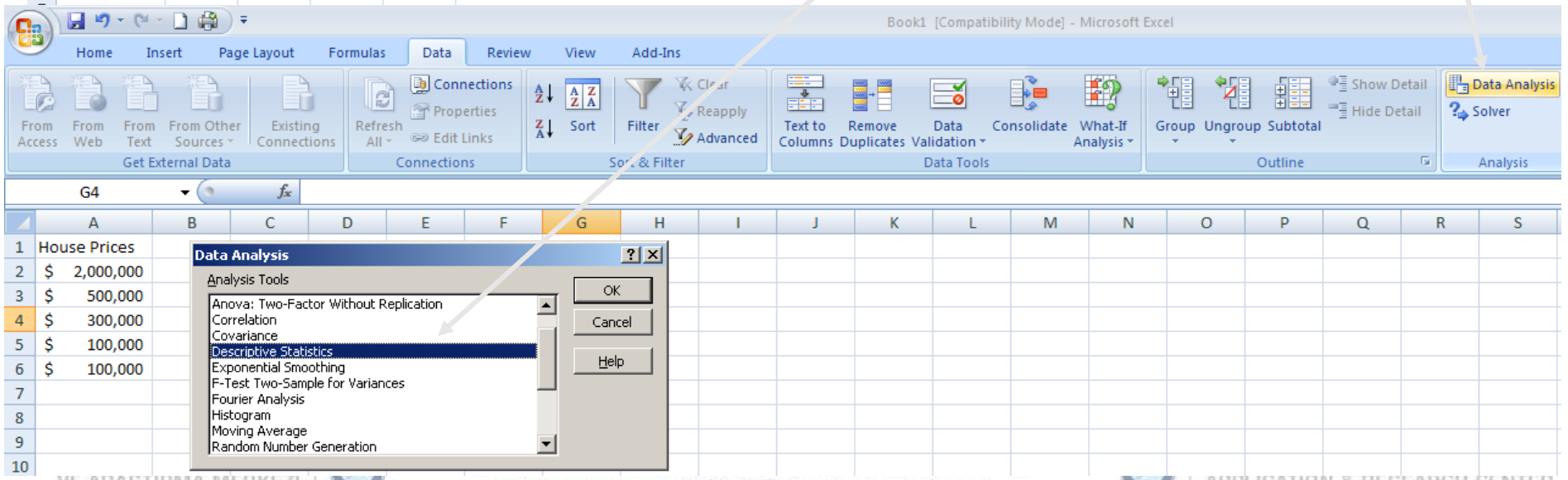
MS Excel fonksiyonlarını kullanarak genel tanımlayıcı istatistiklerin bulunması

Ev Fiyatları		Tanımlayıcı İstatistik		
$ 2.000.000		Ortalama	$ 600.000	=AVERAGE(A2:A6)
$ 500.000		Standart Hata	$ 357.770,88	=D6/SQRT(D14)
$ 300.000		Medyan	$ 300.000	=MEDIAN(A2:A6)
$ 100.000		Mod	$ 100.000,00	=MODE(A2:A6)
$ 100.000		Standart Sapma	$ 800.000	=STDEV(A2:A6)
		Örnek Varyansı	640.000.000.000	=VAR(A2:A6)
		Basıklık (Kurtosis)	4,1301	=KURT(A2:A6)
		Çarpıklık	2,0068	=SKEW(A2:A6)
		Aralık	$ 1.900.000	=D12 - D11
		En Küçük	$ 100.000	=MIN(A2:A6)
		En Büyük	$ 2.000.000	=MAX(A2:A6)
		Toplam	$ 3.000.000	=SUM(A2:A6)
		Sayma	5	=COUNT(A2:A6)

MS Excel veri analizi araçlarını kullanarak genel tanımlayıcı istatistiğin bulunması



1. Veriyi Seç.
2. Veri Analizi butonunu seç.
3. Tanımlayıcı İstatistiği Seç ve Tamam'a bas.

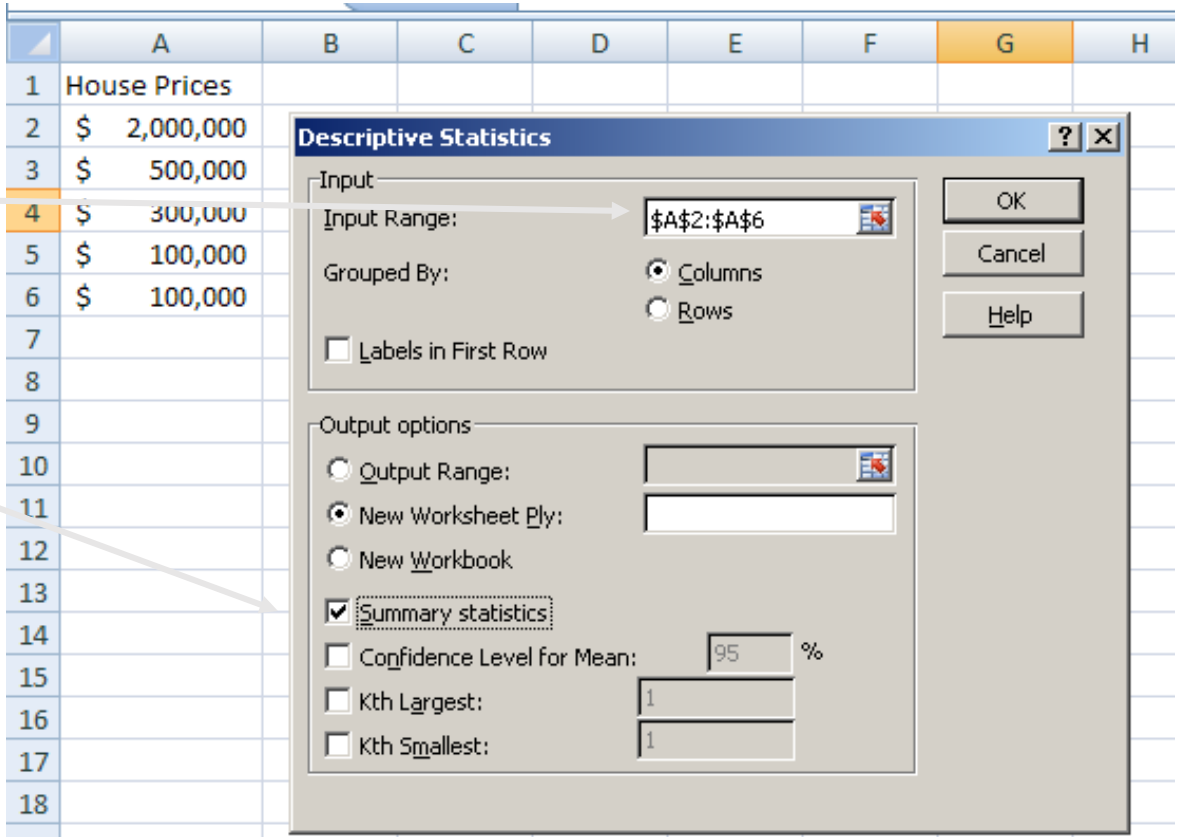


MS Excel veri analizi araçlarını kullanarak genel tanımlayıcı istatistiğin bulunması

4. Hücre aralığını gir.

5. Özet İstatistiği kutucuğunu işaretle.

6. Tamam'a bas.



Excel çıktısı

Ev fiyatları verilerinin kullanıldığı MS Excel tanımlayıcı istatistik çıktısı:

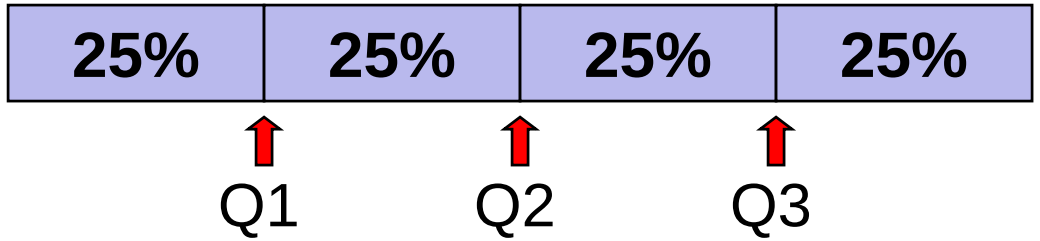
Ev Fiyatları:

$2,000,000
500,000
300,000
100,000
100,000

Ev Fiyatları	
Ortalama	600000
Standart Hata	357770,8764
Medyan	300000
Mod	100000
Standart Sapma	800000
Örnek Varyansı	640.000.000.000
Basıklık (Kurtosis)	4,1301
Çarpıklık	2,0068
Aralık	1900000
En Küçük	100000
En Büyük	2000000
Toplam	3000000
Sayma	5

Çeyrek ölçüleri

- Çeyrekler, sıralanmış veriyi, parça başına eşit sayıda değer düşecek şekilde 4 alt parçaya böler.



- Birinci çeyrek, Q_1, gözlemlerin %25'nin küçük olduğu ve %75'nin büyük olduğu bir değerdir.
- Q_2 medyan ile aynıdır. (gözlemlerin %50'si küçük ve %50'si bu değerden büyüktür)
- Gözlemlerin sadece %25'i üçüncü çeyrekten ,Q_3, büyük değere sahiptir.

Çeyrek ölçüleri:
Çeyreklerin konumunun belirlenmesi

Sıralanmış bir veride uygun bir yere bir çeyreğin yerleştirilmesi şu şekildedir;

Birinci çeyreğin konumu: $\mathbf{Q_1 = (n+1)/4}$ sıralı değeri

İkinci çeyreğin konumu : $\mathbf{Q_2 = (n+1)/2}$ sıralı değeri

Üçüncü çeyreğin konumu : $\mathbf{Q_3 = 3(n+1)/4}$ sıralı değeri

n gözlemlenmiş verilerin sayısı

Çeyrek ölçüleri: Hesaplama kuralları

- Sıralı konum hesaplanırken şu kurallar kullanılır;
  - Eğer sonuç bir tam sayı ise bu sayı sıralı konum için kullanılacak sayıdır.
  - Eğer sonuç, kesirli bir yarımsa, (ör. 2.5, 7.5, 8.5, vb.) karşılık gelen iki veri değerinin ortalaması alınır.
  - Sonuç tam sayı veya kesirli bir yarım değilse, sıralı konumu bulmak için sonuç en yakın tamsayıya çevirilir.

Çeyrek ölçüleri:
Çeyreklerin konumunun belirlenmesi

Sıralı dizi olarak örnek veri: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 , sıralı verinin (9+1)/4 = 2.5 konumundadır

Öyleyse 2'nci ve 3'üncü değerlerin orta noktasındaki değer kullanılmalıdır, **$Q_1 = 12.5$**

Q_1 ve Q_3 merkezi olmayan bir konum ölçütüdür
Q_2 = medyan, merkezi eğilim ölçütüdür.

Çeyrek ölçüleri
çeyreklerin hesaplanması: Örnek

Sıralı dizide Örnek Veri: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 sıralı verinin (9+1)/4 = 2.5 konumunda,

öyleyse **$Q_1 = (12+13)/2 = 12.5$**

Q_2 sıralı verinin (9+1)/2 = 5'inci konumunda,

öyleyse **Q_2 = medyan = 16**

Q_3 sıralı verinin 3(9+1)/4 = 7.5 konumunda,

öyleyse **$Q_3 = (18+21)/2 = 19.5$**

Q_1 ve Q_3 merkezi olmayan bir konum ölçütüdür
Q_2 = medyan, merkezi eğilim ölçütüdür.

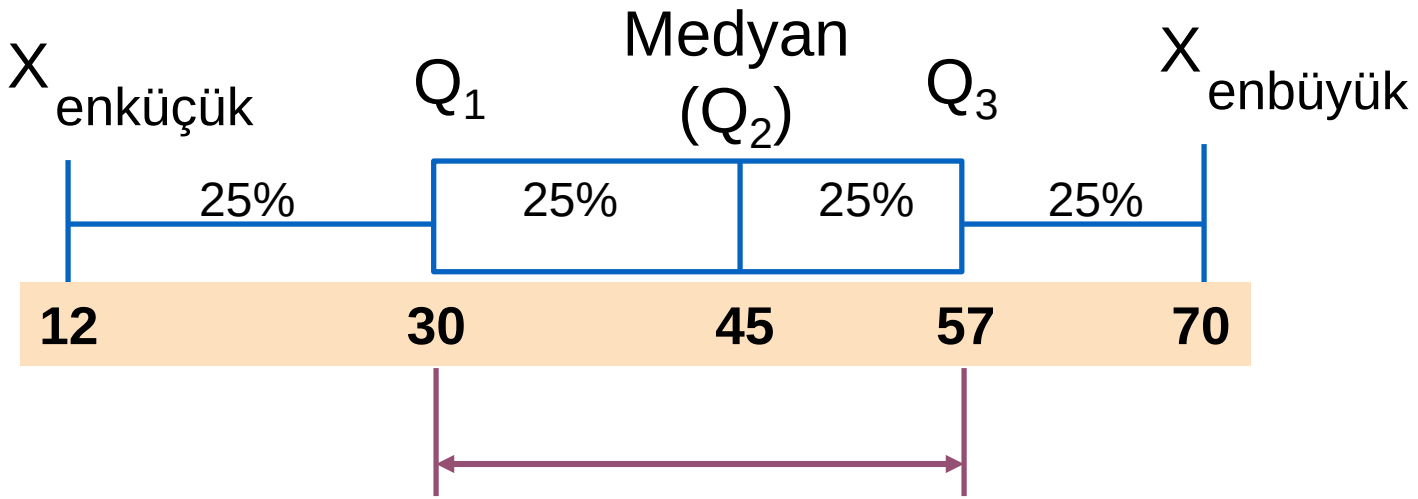
UZAKTAN EĞİTİM UYGULAMA VE ARAŞTIRMA MERKEZİ
guzem.gazi.edu.tr • uzaktanegitim.gazi.edu.tr • lms.gazi.edu.tr
DISTANCE EDUCATION APPLICATION & RESEARCH CENTER

Çeyrek ölçüleri:
Çeyrekler arası aralık (IQR- The Interquartile Range)

- IQR $Q_3 - Q_1$ farkıdır ve verinin ortadaki %50'sinin dağılımını ölçer.
- IQR , verinin ortadaki %50'lik kısmını kapsadığı için orta dağılım olarak da adlandırılır.
- IQR, sıradışı veya aşırı değerlerden etkilenmeyen bir değişkenlik ölçüsüdür.
- Q_1, Q_3, ve IQR gibi sıradışı değerlerden etkilenmeyen ölçütler dirençli ölçütler olarak adlandırılır.

Çeyrekler arası aralığın hesaplanması

Örnek:



Çeyreklerarası Aralık
= 57 – 30 = 27

Beş sayı özeti

Verilerin merkezini, yayılımını ve şeklini tanımlamaya yardımcı olan beş sayı şunlardır:

- $X_{enküçük}$
- Birinci Çeyrek(Q_1)
- Medyan (Q_2)
- Üçüncü Çeyrek(Q_3)
- $X_{enbüyük}$

Beş sayı özeti ve dağılım şekli arasındaki ilişkiler

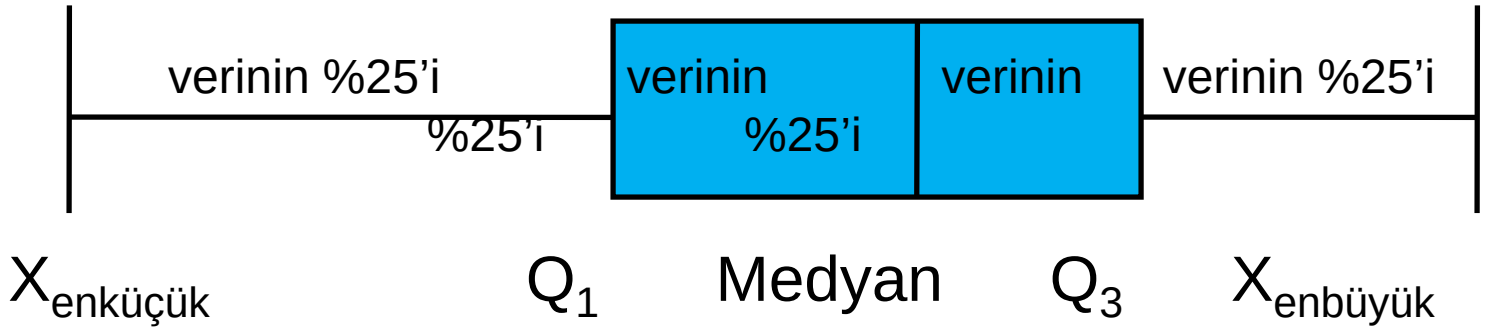
Sola Çarpık	Simetrik	Sağa Çarpık
$\text{Medyan} - X_{\text{enküçük}}$ $>$ $X_{\text{enbüyük}} - \text{Medyan}$	$\text{Medyan} - X_{\text{enküçük}}$ $\approx$ $X_{\text{enbüyük}} - \text{Medyan}$	$\text{Medyan} - X_{\text{enküçük}}$ $<$ $X_{\text{enbüyük}} - \text{Medyan}$
$Q_1 - X_{\text{enküçük}}$ $>$ $X_{\text{enbüyük}} - Q_3$	$Q_1 - X_{\text{enküçük}}$ $\approx$ $X_{\text{enbüyük}} - Q_3$	$Q_1 - X_{\text{enküçük}}$ $<$ $X_{\text{enbüyük}} - Q_3$
$\text{Medyan} - Q_1$ $>$ $Q_3 - \text{Medyan}$	$\text{Medyan} - Q_1$ $\approx$ $Q_3 - \text{Medyan}$	$\text{Medyan} - Q_1$ $<$ $Q_3 - \text{Medyan}$

Beş sayı özeti ve basit kutu diyagramı

- Basit Kutu Diyagramı: Beş sayı özeti temel alınarak verilerin grafiksel gösterimi :

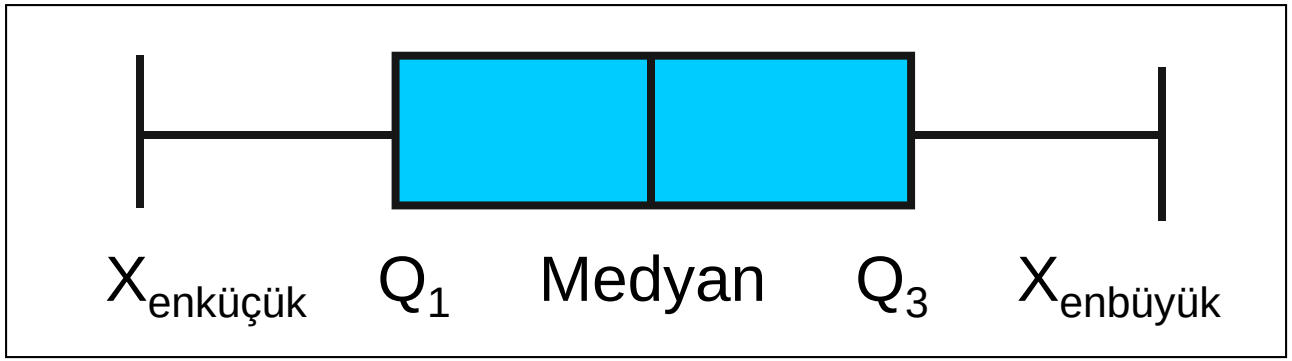
$X_{enküçük}$ -- Q_1 -- Medyan -- Q_3 -- $X_{enbüyük}$

Örnek:



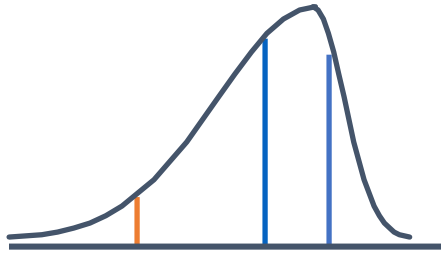
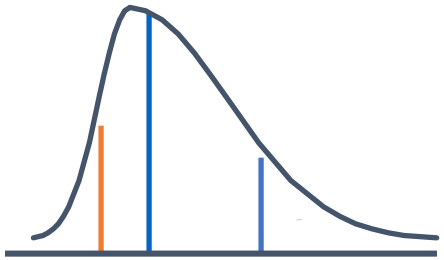
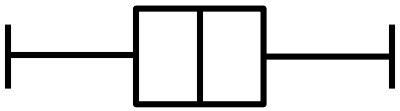
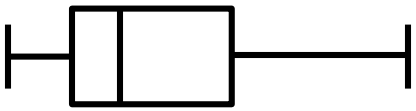
Beş sayı özeti:
Basit kutu diyagramlarının şekli

- Veriler medyan etrafında simetrik ise, kutu ve merkez çizgisi bitiş noktaları arasında ortalanır.



- Bir kutu diyagramı hem dikey hem de yatay olarak gösterilebilir

Dağılım şekli ve kutu diyagramı

Sola Çarpık	Simetrik	Sağa Çarpık
		
Q_1 Q_2 Q_3	Q_1 Q_2 Q_3	Q_1 Q_2 Q_3
		

Basit kutu diyagramı örneği

- Aşağıdaki veri için altta bir basit kutu diyagramı verilmiştir:

$X_{enküçük}$ Q_1 Q_2 / Medyan Q_3 $X_{enbüyük}$

0 2 2 2 3 3 4 5 5 9 27

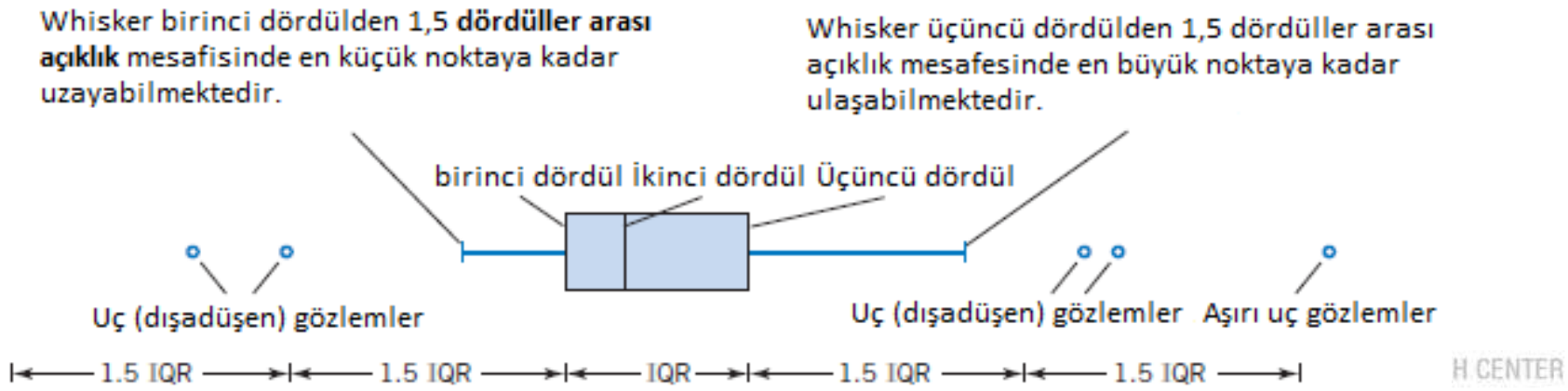
- Basit kutu diyagramından görülebileceği gibi veri sağa çarpıktır

Kutu diyagramları

- Kök-ve-yaprak gösterimi ve histogram veri seti hakkında görsel bir izlenim verirken $\overline{x}$ ve s gibi sayısal değerler verinin yalnızca bir özelliği hakkında bili verirler. Kutu diyagramı merkez, yayılım, simetriden uzaklık ve uç gözlemlerin belirtilmesi gibi veri setinin birçok önemli özelliğini eş zamanlı olarak tanımlayan grafiksel bir gösterimdir.
- Kutu diyagramı yatay veya dikey olarak ayarlanmış bir dikdörtgen bir kutu üzerinde üç tane dördülü, verinin en büyük ve en küçük değerini göstermektedir. Kutu birinci dördülde (q_1) sol (ya da alt) köşeden üçüncü dördülde (q_3) sağ (ya da üst) köşeye bir dördüller arası açıklık mesafesi genişliğindedir.
- İkinci dördülde $q_2 = \overline{x}$ (verilerin %50'sinden büyük nokta veya medyan) kutu boyunca bir çizgi mevcuttur. Bir çizgi ya da whisker kutunun bir ucundan diğerine kadar uzanmaktadır.

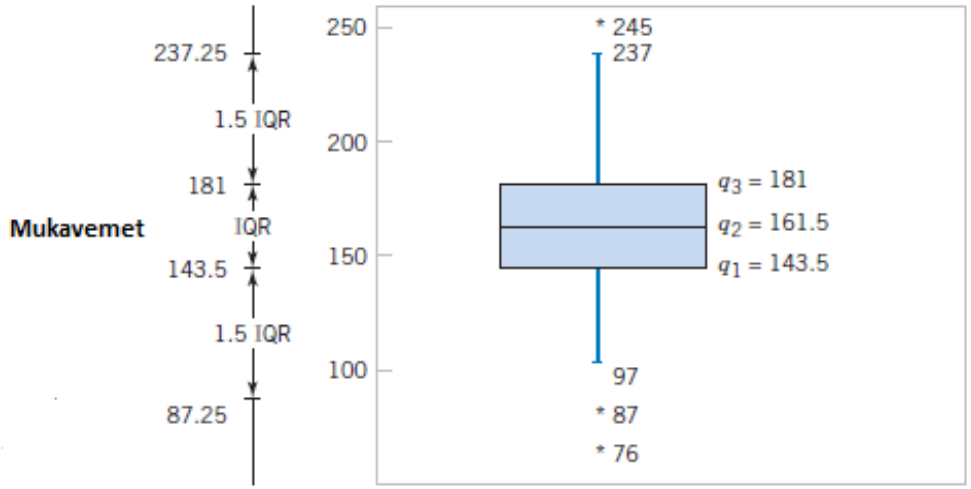
Kutu diyagramları

- Alt whisker birinci dördülden 1,5 dördüller arası açıklık mesafesinde en küçük noktaya giden bir çizgidir. Üst whisker üçüncü dördülden 1,5 dördüller arası açıklık mesafesinde en büyük noktaya giden bir çizgidir.
- Kutuya whisker'lardan daha uzaktaki veriler bireysel noktalar olarak gösterilmektedir. Whisker'dan daha ötede ancak kutunun köşesinden üç dördüller arası mesafeden daha yakın olan noktalar **uç nokta** (outlier) olarak bilinir. Kutunun köşesinden itibaren üç dördüller arası açıklık mesafesinden daha uzaktaki bir nokta **aşırı uç nokta** (extreme outlier) olarak bilinir. Aşağıdaki şekil bir kutu diyagramını göstermektedir.



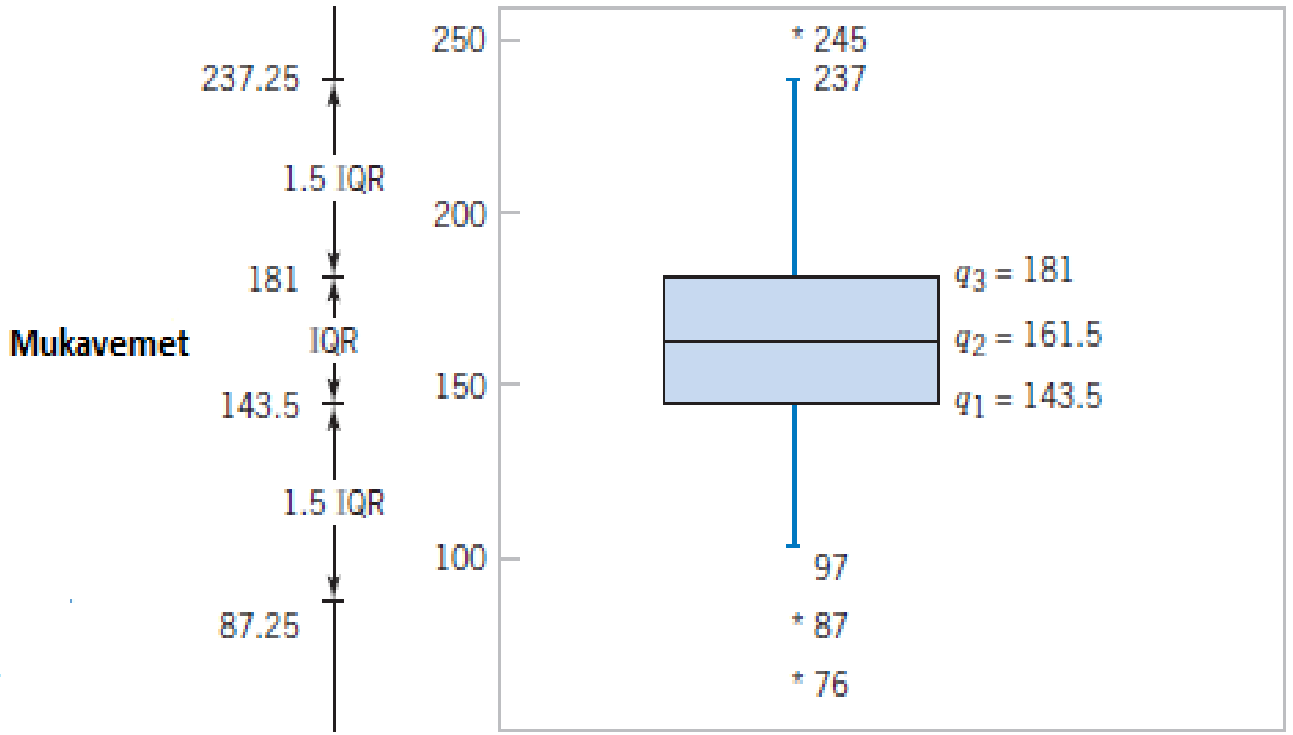
Kutu diyagramları

- Aşağıdaki şekil basınç mukavemeti verileri için kutu diyagramını vermektedir.
- Bu kutu diyagramından basınç mukavemeti verilerinin yaklaşık olarak ortalama etrafında simetrik olduğu görülmektedir.
- Çünkü sağ ve sol whisker'lar ve medyan etrafındaki sağ sol kutuların uzunlukları yaklaşık olarak aynıdır.
- Üst whisker 237 gözlem değerine kadar uzanmaktadır çünkü üst uç noktalar için limitin altındaki en büyük gözlem değeridir. Düşük mukavemet için iki adet ve yüksek mukavemet için bir adet uç nokta bulunmaktadır.



Kutu diyagramları

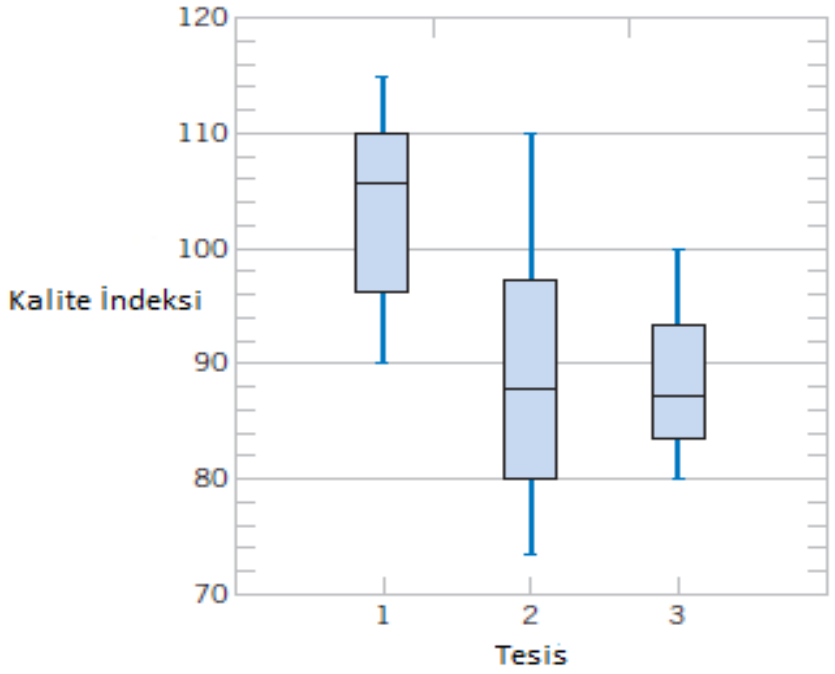
- Bu limit $q_3 + 1{,}5IQR = 181 + 1{,}5(181 - 143{,}5) = 237{,}25$'dir. Alt whisker değeri 97 değerine kadar uzanmaktadır çünkü alt uç noktalar için bu limitin üstündeki en küçük gözlemdir. Bu limit $q_1 - 1{,}5IQR = 143{,}5 - 1{,}5(181 - 143{,}5) = 87{,}25$'dir.



UZAKTAN EĞİTİM UYGUL
VE ARAŞTIRMA MERI
CATION
RESEARCH CENTER

Örnek

- Aşağıdaki şekil üç imalat tesisindeki yarı iletken aygıtlar üzerindeki imalat kalite indeksi için karşılaştırmalı kutu diyagramlarını göstermektedir. Bu grafik incelendiğinde ikinci imalat tesisinde çok büyük miktarda değişkenlik olduğu ve ikinci ve üçüncü tesislerin indeks performanslarını artırması gerektiği görülmektedir.



Bir popülasyon için sayısal tanımlayıcı ölçütler

- Daha önce tartışılan tanımlayıcı istatistikler, *popülasyonu* değil de bir *örneği* tanımlamaktadır.

- **Parametreler** adı verilen, bir popülasyonu tanımlayan özet ölçütler, Yunan harfleriyle belirtilir.

- Önemli popülasyon parametreleri, popülasyon ortalaması, varyansı ve standart sapmasıdır.

Bir popülasyon için sayısal tanımlayıcı ölçüler: Ortalama μ

- Popülasyon ortalaması, popülasyondaki değerlerin toplamının, popülasyon büyüklüğüne, N'e bölünmesiyle elde edilir

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

μ = popülasyon ortalaması

N = popülasyon büyüklüğü

X_i = X değişkeninin i'nci değeri

Bir popülasyon için sayısal tanımlayıcı ölçüler : Varyans σ^2

- Değerlerin ortalamadan sapma karelerinin ortalaması
  - Popülasyon varyansı:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

μ = popülasyon ortalaması

N = popülasyon büyüklüğü

X_i = X değişkeninin i'nci değeri

Not: Serbestlik derecesi

- Dikkat edileceği üzere örneklem varyansının paydası örneklem boyutunun bir eksiği $(n-1)$ iken popülasyon varyansı için payda popülasyon boyutu olan N'dir.
- Eğer popülasyon ortalaması μ'nün gerçek değerini bilseydik örneklem varyansını örneklem gözlem değerlerinin μ civarındaki sapmalarının karelerinin averaj değeri olarak bulabilirdik.
- Pratikte μ değeri neredeyse hiç bilinmediği için gözlem değerlerinin örneklem ortalaması olan $\bar{x}$ civarındaki sapmalarının karelerinin toplamı kullanılmaktadır.
- Ancak x_i değerleri popülasyon ortalaması μ değerinden ziyade kendi averaj değerleri $\bar{x}$'a daha yakın olma eğiliminde olacaktır.

Not: Serbestlik derecesi

- Böylece bunu telafi etmek adına payda da n yerine $n-1$ terimi kullanılmaktadır. Örneklem varyansının hesaplanmasında payda da n kullanılsaydı gerçek popülasyon varyansı σ^2'den daha düşük değişkenlik ölçümü elde edebilirdik.
- Bununla ilgili bir diğer yol örneklem varyansı s^2'nin $n-1$ **serbestlik derecesine** sahip olduğu gerçeğini dikkate almaktır. Serbestlik derecesi terimi n adet sapmanın $x_1 - \overline{x}$, $x_2 - \overline{x}$, ..., $x_n - \overline{x}$ toplamının daima sıfıra eşit olacağı gerçeğidir.
- Böylece $n-1$ adet terime ait sapmalar hesaplandığında geri kalan bir tane sapma değeri otomatik olarak hesaplanabilir. Yani n adet sapmanın $n-1$ tanesi serbest bir şekilde belirlenebilir.

Bu konu hipotez testlerinde tekrar ele alınacaktır.

Bir popülasyon için sayısal tanımlayıcı ölçüler : Standart sapma σ

- En sık kullanılan değişim ölçüsüdür
- Ortalamadan sapma miktarını gösterir
- Popülasyon varyansının karekökü̈dür
- Orijinal veri ile aynı birimlere sahiptir

  - Popülasyon standart sapması:

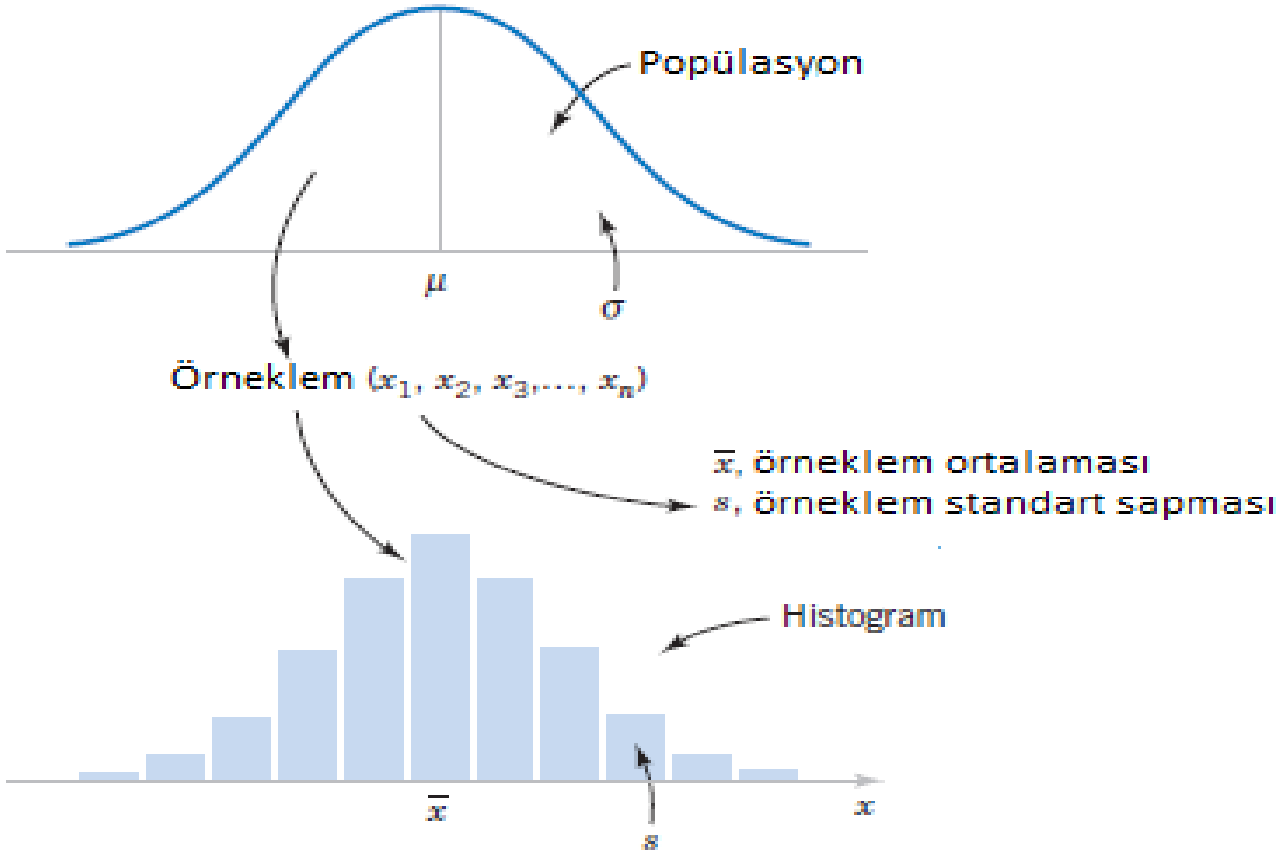
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Örnek istatistiğine karşı popülasyon parametreleri

Ölçüt	Popülasyon Parametresi	Örnek İstatistiği
Ortalama	μ	$\overline{X}$
Varyans	σ^2	S^2
Standart Sapma	σ	S

Popülasyonla örneklem arasındaki ilişki

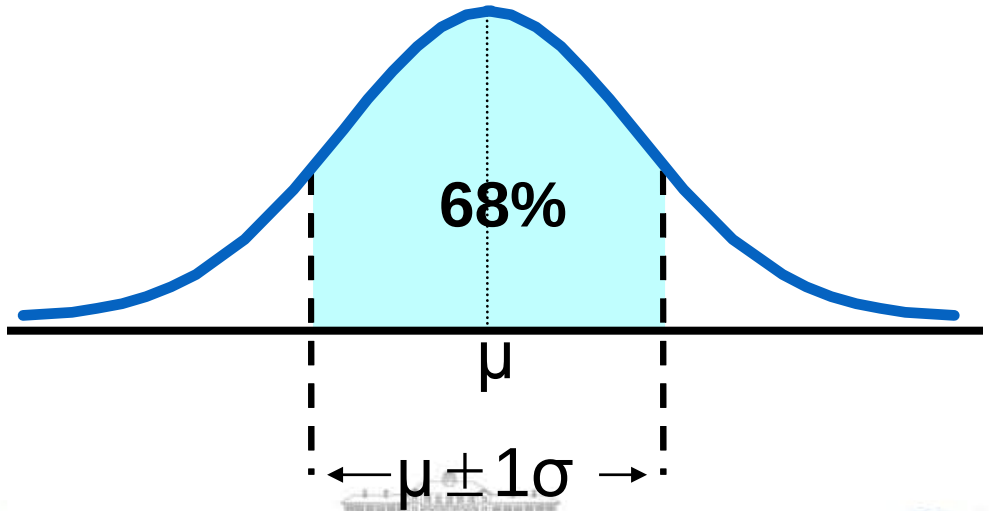
- Popülasyon ve örneklem arasındaki ilişki aşağıdaki şekilde gösterilmektedir.



UZAKTAN EĞİTİM UYG
VE ARAŞTIRMA M
EDUCATION
ON & RESEARCH CENTER

Deneysel (Ampirik) kural

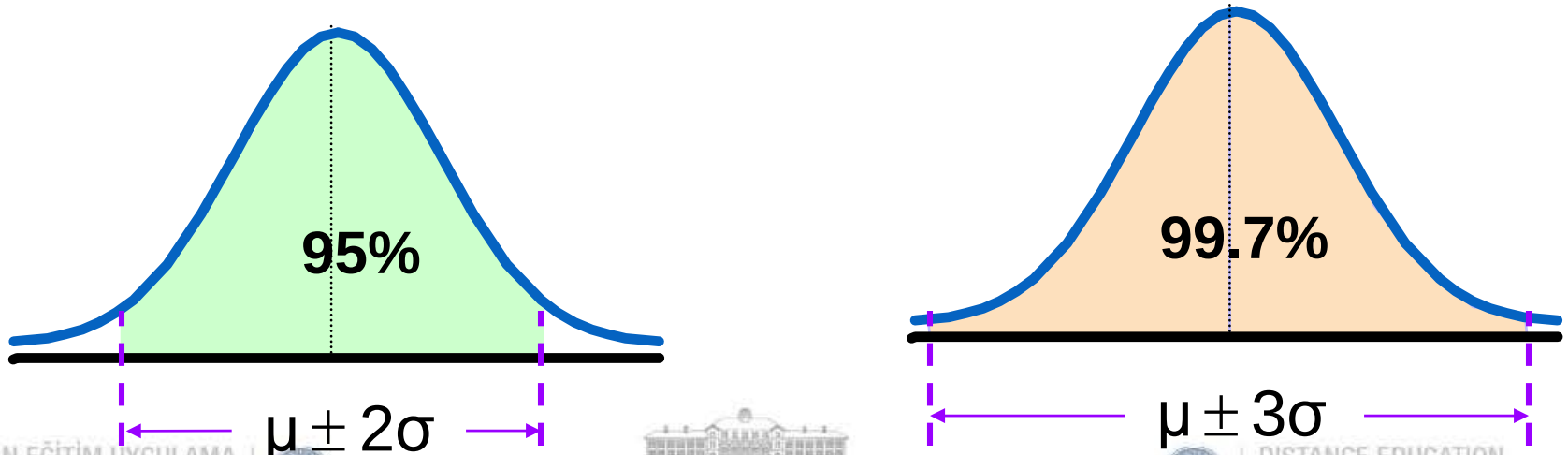
- Ampirik kural, verilerin değişiminin çan şeklindeki bir dağılıma yakınsadığını söyler
- Çan şeklindeki bir dağılımdaki verilerin yaklaşık % 68'i ortalamanın 1 standart sapması içindedir veya $\mu \pm 1\sigma$



Deneysel (Ampirik) kural

- Çan şeklindeki bir dağılımdaki verilerin yaklaşık % 95'i ortalamanın iki standart sapması içindedir veya $\mu \pm 2\sigma$

- Çan şeklindeki bir dağılımdaki verilerin yaklaşık % 99.7'si ortalamanın üç standart sapması içindedir veya $\mu \pm 3\sigma$



UZAKTAN EĞİTİM UYGULAMA VE ARAŞTIRMA MERKEZİ
guzem.gazi.edu.tr • uzaktanegitim.gazi.edu.tr • lms.gazi.edu.tr
DISTANCE EDUCATION APPLICATION & RESEARCH CENTER

Ampirik kuralın kullanımı

- SAT Matematik puanı değişkeninin ortalaması 500 ve standart sapması 90 olacak şekilde çan şeklinde olduğunu varsayalım. Bu takdirde,
  - Teste katılanların %68'i 410 ile 590 arasında bir skor almıştır. (500 ± 90).
  - Teste katılanların %95'i 320 ile 680 arasında bir skor almıştır. (500 ± 180).
  - Teste katılanların % 99.7'si 230 ile 770 arasında bir skor almıştır. (500 ± 270).

Chebyshev kuralı

- Verilerin nasıl dağıtıldığına bakılmaksızın, değerlerin en az $(1 - 1/k^2) \times \%100$ 'ü ortalamanın k standart sapması aralığına girer (k > 1 için)
  - Örnekler:

En az	Aralık
$(1 - 1/2^2) \times 100\% = 75\%$	k=2 $(\mu \pm 2\sigma)$
$(1 - 1/3^2) \times 100\% = 88.89\%$	k=3 $(\mu \pm 3\sigma)$

İki nümerik değişken arasındaki ilişkinin iki ölçütü

- Serpme diyagramları, iki sayısal değişken arasındaki ilişkiyi görsel olarak incelemenize izin verir ve şimdi bu tür ilişkilerin iki niceliksel ölçüsü üzerinde duracağız.

- Kovaryans
- Korelasyon katsayısı

Kovaryans

- Kovaryans, **iki sayısal değişken** (X & Y) arasındaki doğrusal ilişkinin gücünü ölçer.
- Örnek kovaryansı:

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

- Sadece ilişkinin gücü ile ilgilenir
- Hiçbir rasgele etki ima edilmemektedir.

Kovaryansı yorumlama

- İki değişken arasındaki **kovaryans** :

cov(X,Y) > 0 ⟶ X ve Y aynı yönde hareket etme eğilimindedir

cov(X,Y) < 0 ⟶ X ve Y ters yönde hareket etme eğilimindedir

cov(X,Y) = 0 ⟶ X ve Y bağımsızdır

- Kovaryansın büyük bir kusuru mevcuttur:
  - Kovaryansın büyüklüğünden ilişkinin göreceli gücünü belirlemek mümkün değildir

Korelasyon katsayısı

- İki sayısal değişken arasındaki doğrusal ilişkinin göreceli gücünü ölçer

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}$$

$$S_X=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$$

$$S_Y=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-1}}$$

olduğunda

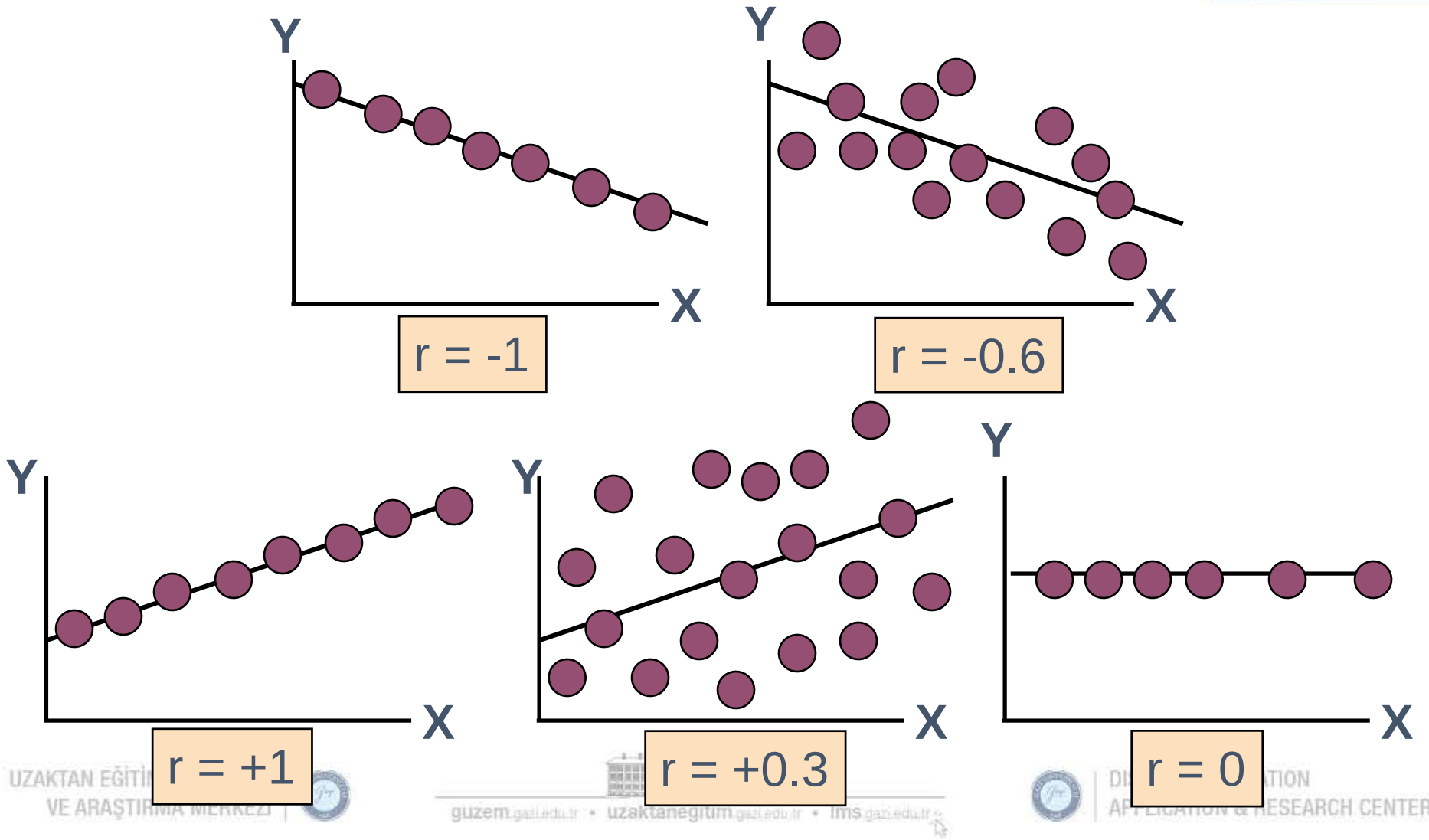
- Örnek korelasyon katsayısı:

$$r=\frac{\operatorname{cov}(X,Y)}{S_X S_Y}$$

Korelasyon katsayısının özellikleri

- Popülasyonun korelasyon katsayısı ρ ile gösterilir.
- Örnek korelasyon katsayısı r ile ifade edilir.
- Hem ρ hem de r aşağıdaki özelliklere sahiptir:
  - Birimden bağımsızdır
  - –1 ile 1 arasında değer alır
  - –1'e ne kadar yakınsa, o kadar güçlü negatif doğrusal bir ilişki mevcuttur.
  - 1'e ne kadar yakınsa, o kadar güçlü pozitif doğrusal bir ilişki mevcuttur.
  - 0'a ne kadar yakınsa, o kadar zayıf doğrusal bir ilişki vardır.

Değişik korelasyon katsayıları ile örnek verilerinin serpme diyagramı



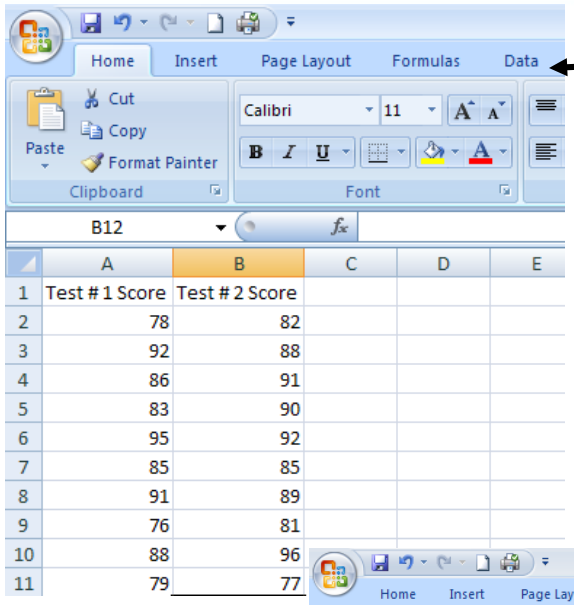
UZAKTAN EĞİTİM VE ARAŞTIRMA MERKEZİ
guzem.gazi.edu.tr • uzaktanegitim.gazi.edu.tr • ims.gazi.edu.tr
DISTANCE EDUCATION APPLICATION & RESEARCH CENTER

MS Excel fonksiyonlarını kullanarak korelasyon katsayısının eldesi

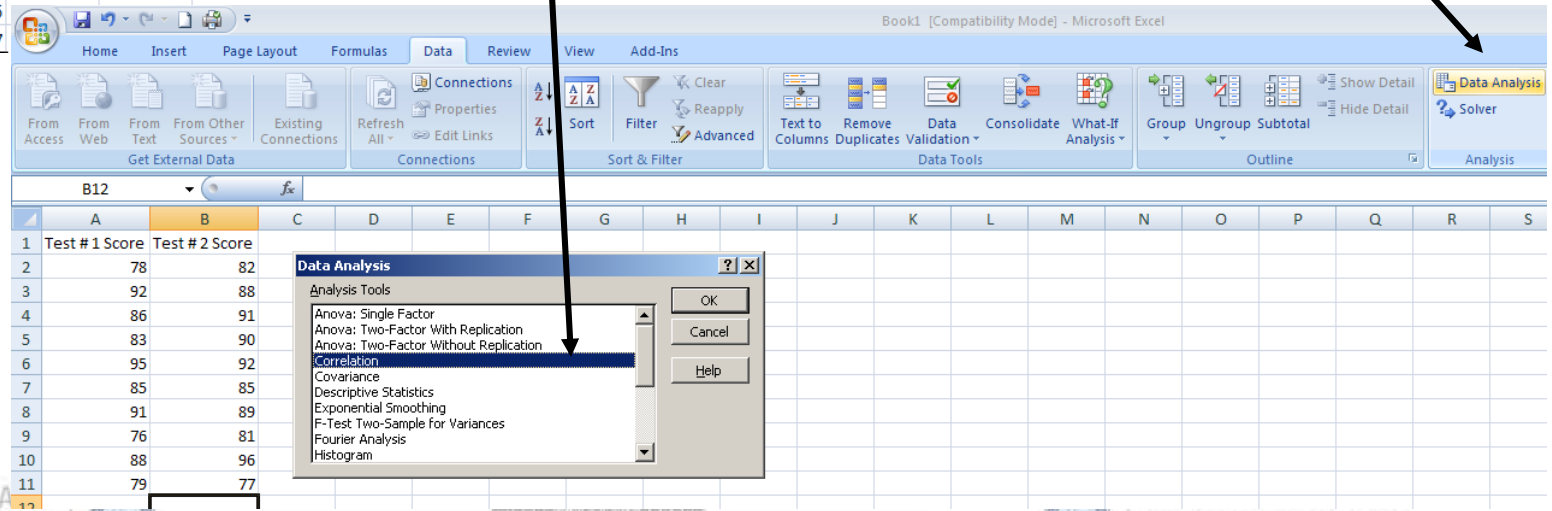
Test #1 Sonucu	Test #2 Sonucu		**Korelasyon Katsayısı**	
78	82		0,7332	=CORREL(A2:A11,B2:B11)
92	88			
86	91			
83	90			
95	92			
85	85			
91	89			
76	81			
88	96			
79	77			

MS Excel veri analizi aracını kullanarak korelasyon katsayısı eldesi

1. Veriyi Seç
2. Veri Analizini seç
3. Korelasyonu seç & Tamam'a bas



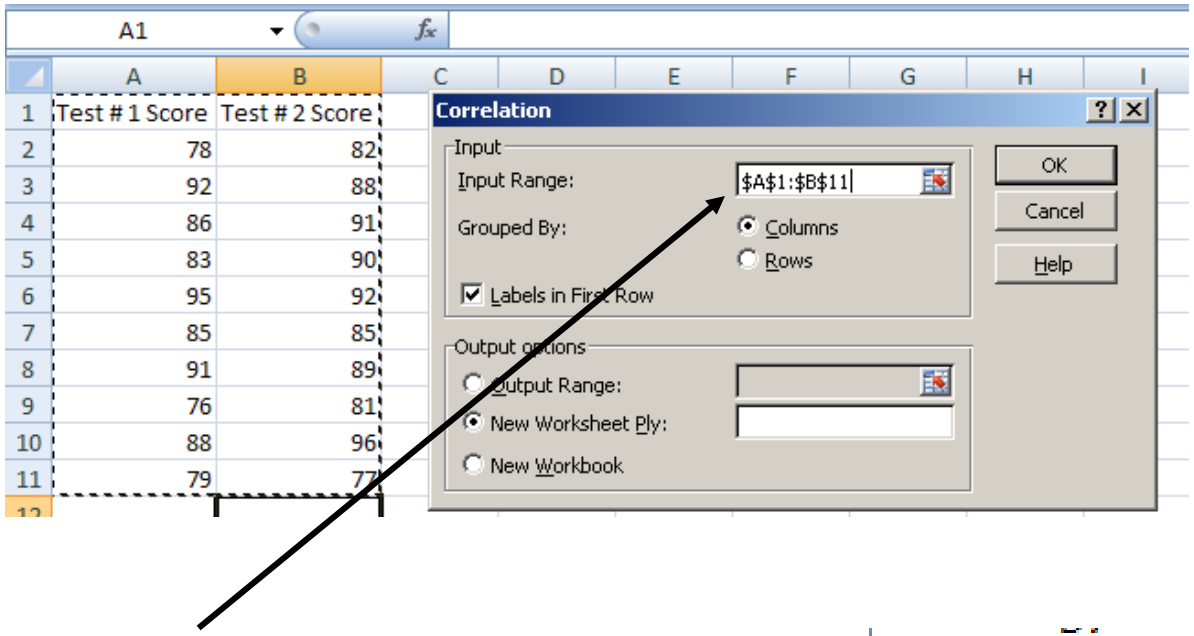
	A	B	C	D	E
1	Test # 1 Score	Test # 2 Score			
2	78	82			
3	92	88			
4	86	91			
5	83	90			
6	95	92			
7	85	85			
8	91	89			
9	76	81			
10	88	96			
11	79	77			



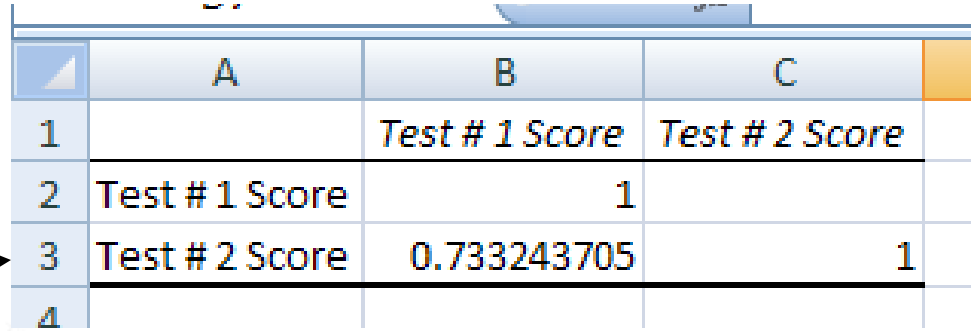
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Test # 1 Score	Test # 2 Score																	
2	78	82																	
3	92	88																	
4	86	91																	
5	83	90																	
6	95	92																	
7	85	85																	
8	91	89																	
9	76	81																	
10	88	96																	
11	79	77																	

UZAKTAN EĞİTİM UYGULAMA VE ARAŞTIRMA MERKEZİ
guzem.gazi.edu.tr • uzaktanegitim.gazi.edu.tr • lms.gazi.edu.tr
APPLICATION & RESEARCH CENTER

MS Excel'i kullanarak korelasyon katsayısı eldesi



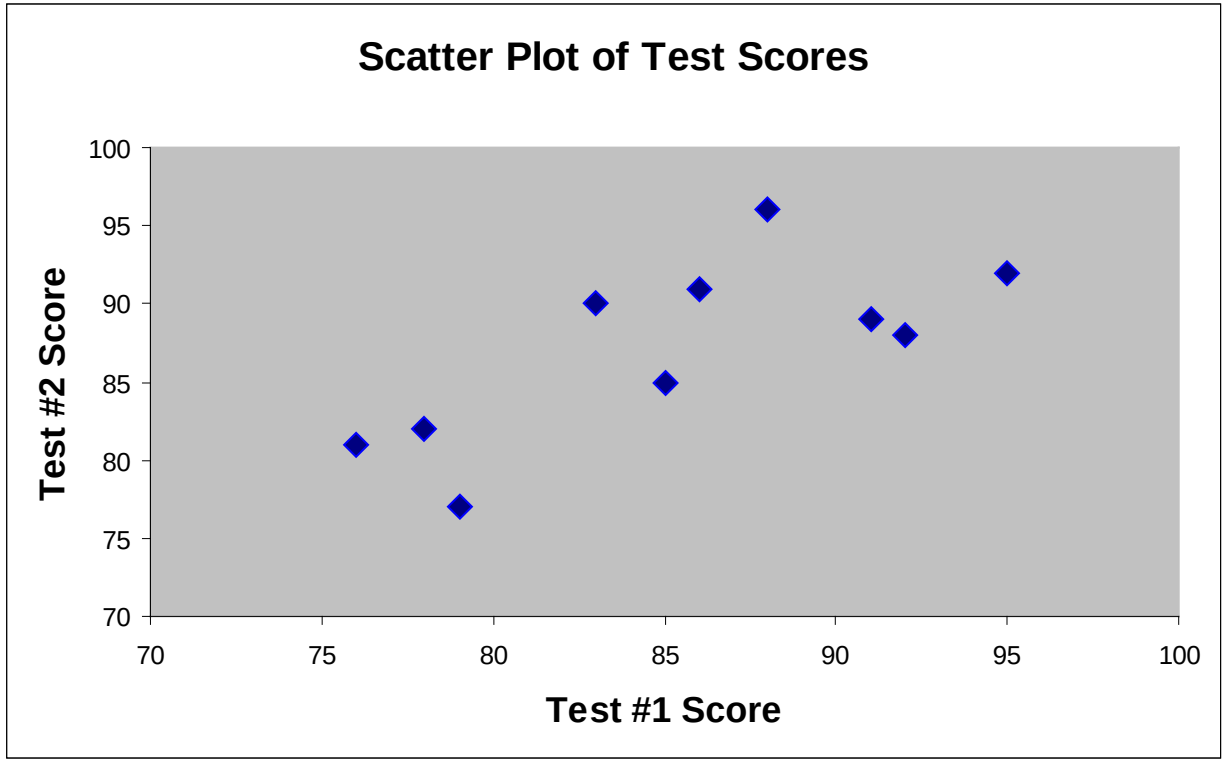
4. Veri aralığını gir ve gerekli seçenekleri seç
5. Çıktıyı almak için Tamam'a bas



	A	B	C
1		*Test # 1 Score*	*Test # 2 Score*
2	Test # 1 Score	1	
3	Test # 2 Score	0.733243705	1

MS Excel'i kullanarak korelasyon katsayısından yorumlama yapma

- r = 0.733
- Test sonucu1 ile test sonucu 2 arasında göreceli olarak güçlü bir pozitif doğrusal ilişki vardır.
- İlk testte yüksek puan alan öğrenciler ikinci testte de yüksek puanlar alma eğilimindedir.



Sayısal tanımlayıcı ölçütlerdeki tuzaklar

- Veri analizi nesneldir
  - Veri setinin önemli özelliklerini en iyi tanımlayan ve ileten özet ölçütleri raporlamalıdır
- Verinin yorumlanması özneldir.
  - Adil, tarafsız ve net bir şekilde yapılmalıdır.

Etik hususlar

Sayısal Tanımlayıcı Ölçütler:

- İyi ve kötü sonuçların her ikisini de raporlamalıdır.
- Adil, nesnel ve tarafsız bir şekilde sunulmalıdır.
- Gerçekleri saptırmak için uygun olmayan özet ölçütlerini kullanmamalıdır.

Bölüm özeti

Bu bölümde

- Merkezi Eğilim Ölçülerini
  - Ortalama, medyan, mod, geometrik ortalama
- Değişim Ölçülerini
  - Aralık, Çeyrekler Arası Aralık, varyans ve standart sapma, değişim katsayısı, Z-değeri
- Dağılımların şekillerini
  - Çarpıklık & Basıklık (Kurtosis)
- 5 sayı Özeti ile verilerin Tanımlanmasını
  - Kutu diyagramları

Bölüm özeti

- Kovaryans ve Korelasyon katsayısını
- Sayısal Tanımlayıcı ölçütlerle ilgili tuzakları ve ahlaki hususları inceledik.