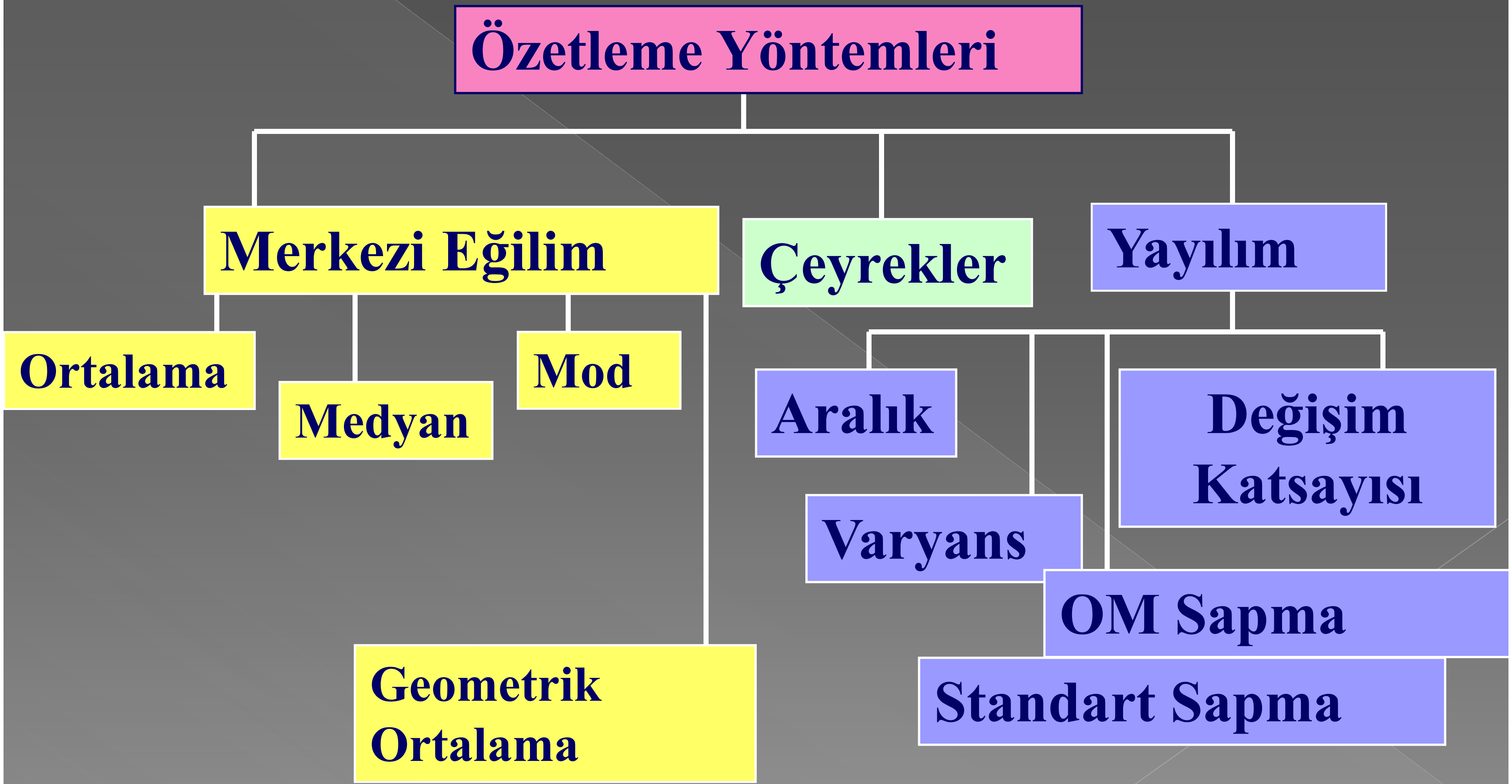
İstatistik 2
Bölüm 3

Genel Tekrar 3

Sayısal Özetleme Yöntemleri



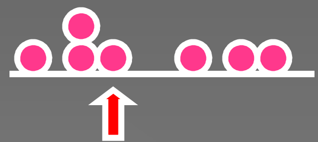
Merkezi Eğilim Ölçüleri

Merkezi Eğilim

- **Ortalama**
- **Medyan**
- **Mod**
- **Geometrik Ortalama**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$



$$\bar{X}^G = (X^1 x X^2 x \ldots x X^n)^{1/n}$$

Ortalama (Aritmetik)

- Ortalama
  - Örnek Ortalaması

Örnek gözlem sayısı

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

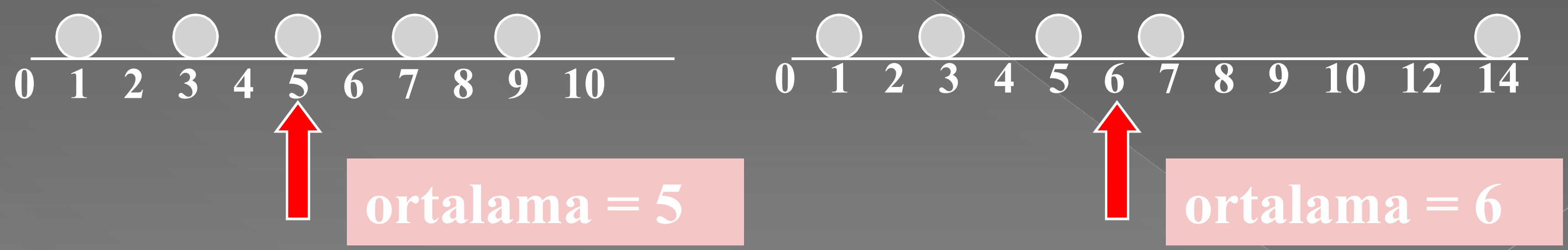
  - Populasyon Ortalaması

Populasyon gözlem sayısı

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

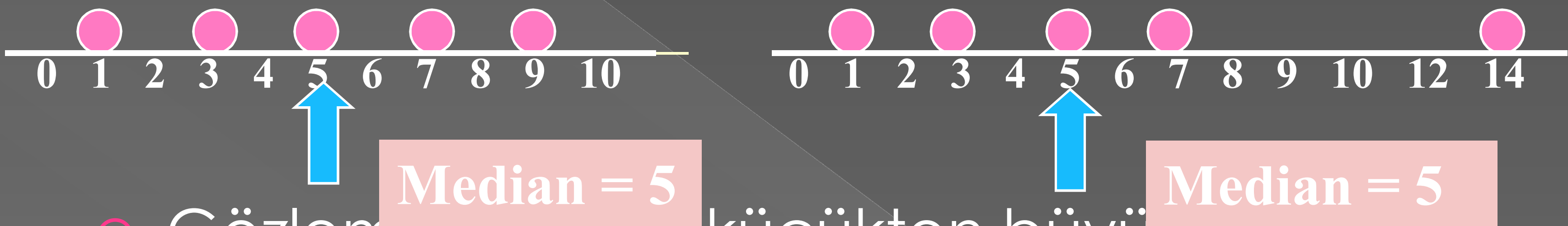
Ortalama

- En yaygın kullanılan merkezi eğilim ölçüsüdür.
- Veri setinde aşırı uçlar varsa bu ölçü etkilenir.



Medyan

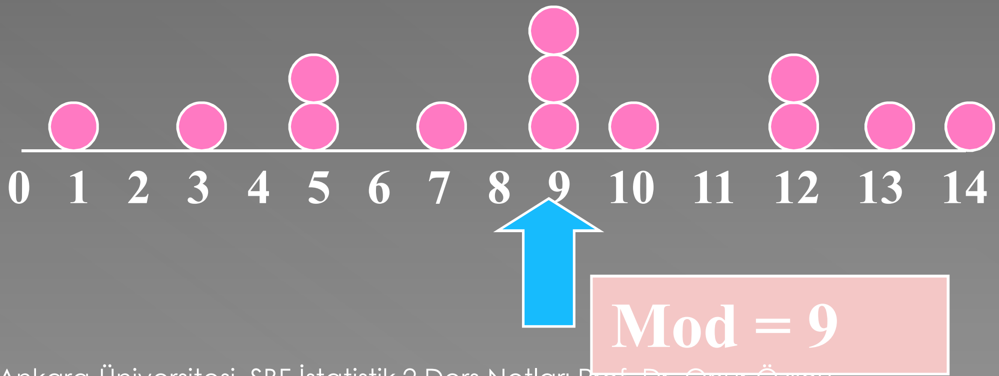
- Aşırı uç değerler etkilemez.
- Daha sağlam bir ölçüdür.

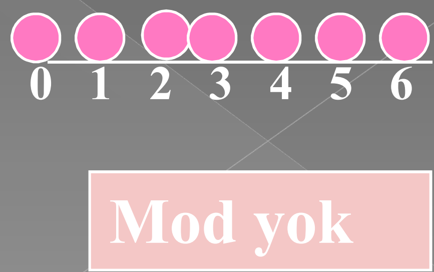


- Gözlem değerleri küçükten büyüğe doğru sıralandığında Medyan ortada kalan gözlemdir.
  - N yada n tek sayı ise ortada kalan gözlem
  - N yada n çift sayı ise ortadaki iki gözlemin ortalaması

Mod

- Başka bir merkezi eğilim ölçüsüdür.
- En sık gözlemlenen değerdir.
- Aşırı uçlardan etkilenmez.
- Sayılabilen yada sayılamayan veriler için kullanılabilir.
- Veri setinde bazen mod olmayabilir.
- Veri setinde birden fazla mod olabilir.





Geometrik Ortalama

- Bir değişkende zaman içinde meydana gelen değişim oranını hesaplamada kullanılır.

$$\bar{X}_G = \left(X_1 \times X_2 \times \text{L} \times X_n \right)^{1/n}$$

- Zaman içinde bir yatırımın getirisini hesaplar
  - Yatırımın zaman içinde durumunun ne olduğunu hesaplar.

$$\bar{R}_G = \left[\left(1 + R_1\right) \times \left(1 + R_2\right) \times \text{L} \times \left(1 + R_n\right) \right]^{1/n} - 1$$

Örnek

$100,000 bir yatırımın değeri 1. sene sonunda $50,000 düşmüş ve ikinci yılın sonunda tekrar $100,000 yükselmiştir. Ortalama getiri ve geometrik ortalama nedir?

$$X_1 = \$100,000 \qquad X_2 = \$50,000 \qquad X_3 = \$100,000$$

Ortalama getiri

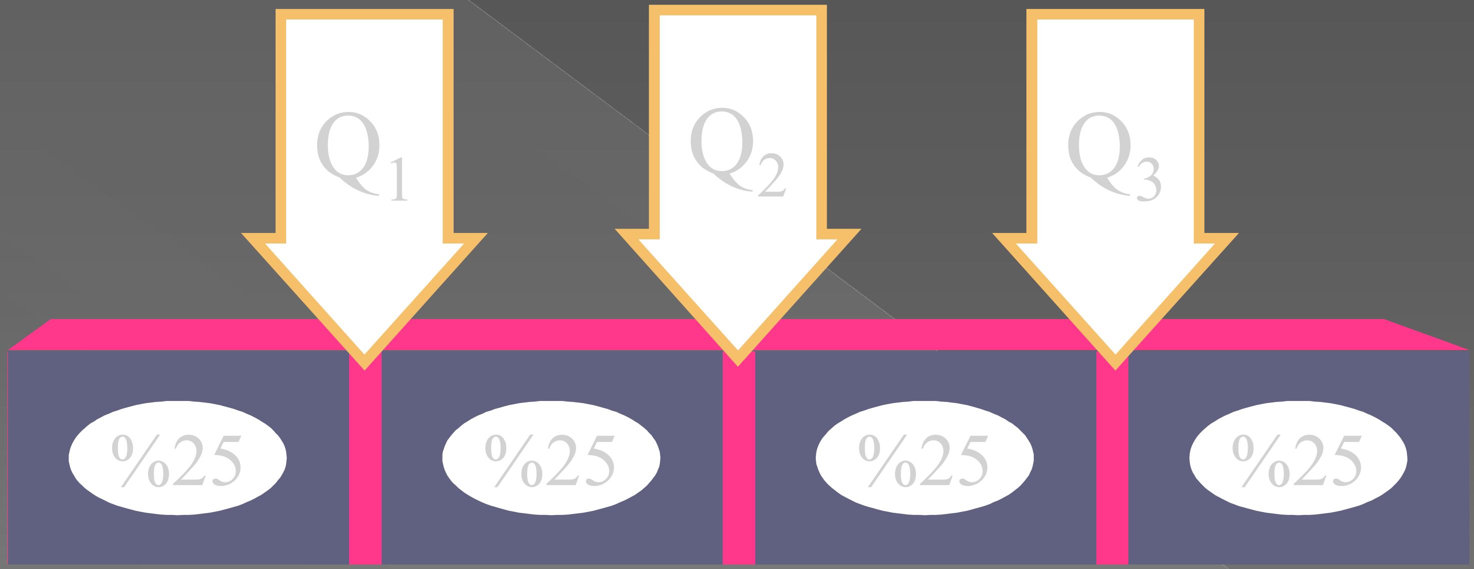
$$\overline{X} = \frac{(-\%50) + (\%100)}{2} = \%25$$

Geometrik getiri

$$\overline{R}_G = \left[(1 + (-\%50))x(1 + (\%100))\right]^{1/2} - 1$$

$$= \left[(0{,}50)\mathrm{x}(2)\right]^{1/2} - 1 = 1^{1/2} - 1 = \%0$$

Çeyreklerin Pozisyonları



Çeyrekler

- Veri setini dörde böler
- i inci çeyreğin pozisyonu $$\left(Q_i\right) = \frac{i\left(n+1\right)}{4}$$
- 1. ve 3. çeyrekler merkezi eğilim ölçüsü değildir.
- Medyan ise bir merkezi eğilim ölçüsüdür.

Çeyrekler: Örnek

- Küçükten… büyüğe doğru sıralı: 106, 109, 114, 116, 121, 122, 125, 129
- Q_1:

$$i=\frac{25}{100}(8)=2 \qquad Q_1=\frac{109+114}{2}=111.5$$

- Q_2:

$$i=\frac{50}{100}(8)=4 \qquad Q_2=\frac{116+121}{2}=118.5$$

- Q_3:

$$i=\frac{75}{100}(8)=6 \qquad Q_3=\frac{122+125}{2}=123.5$$

Çeyrekler

Veri seti: küçük...büyük: 11 12 13 16 16 17 18 21 22

1. Çeyrek: Q1= $\frac{n+1}{4}$ (9+1)/4 =2,5. gözlem. Böyle bir gözlem olmadığı

için 2 ve 3. gözlemlerin ortalamasını alarak 1. gözlemi elde edebiliriz. (12+13)/2 = 12.5

3. Çeyrek: Q3= 3(n+1)/4'üncü gözlemdir. 3 (9+1)/4 =7,5. gözlem. Böyle bir gözlem olmadığı

için 7 ve 8. gözlemlerin ortalamasını alarak 3. gözlemi elde edebiliriz. (18+21)/2 =1 9,5.

Çeyrekler

Yüzdeliğin yeri $$Y_y = (n+1)\frac{y}{100}$$

Y_{89} 89. yüzdeliği ifade etmektedir.

Çeyrekler

1. Çeyrek → veri setinde yer alan gözlemlerin yaklaşık %25'ini kapsar
2. Çeyrek → veri setinde yer alan gözlemlerin yaklaşık %50'sini kapsar
3. Çeyrek → veri setinde yer alan gözlemlerin yaklaşık %75'ini kapsar
4. Çeyrek → veri setinde yer alan gözlemlerin yaklaşık %100'ünü kapsar

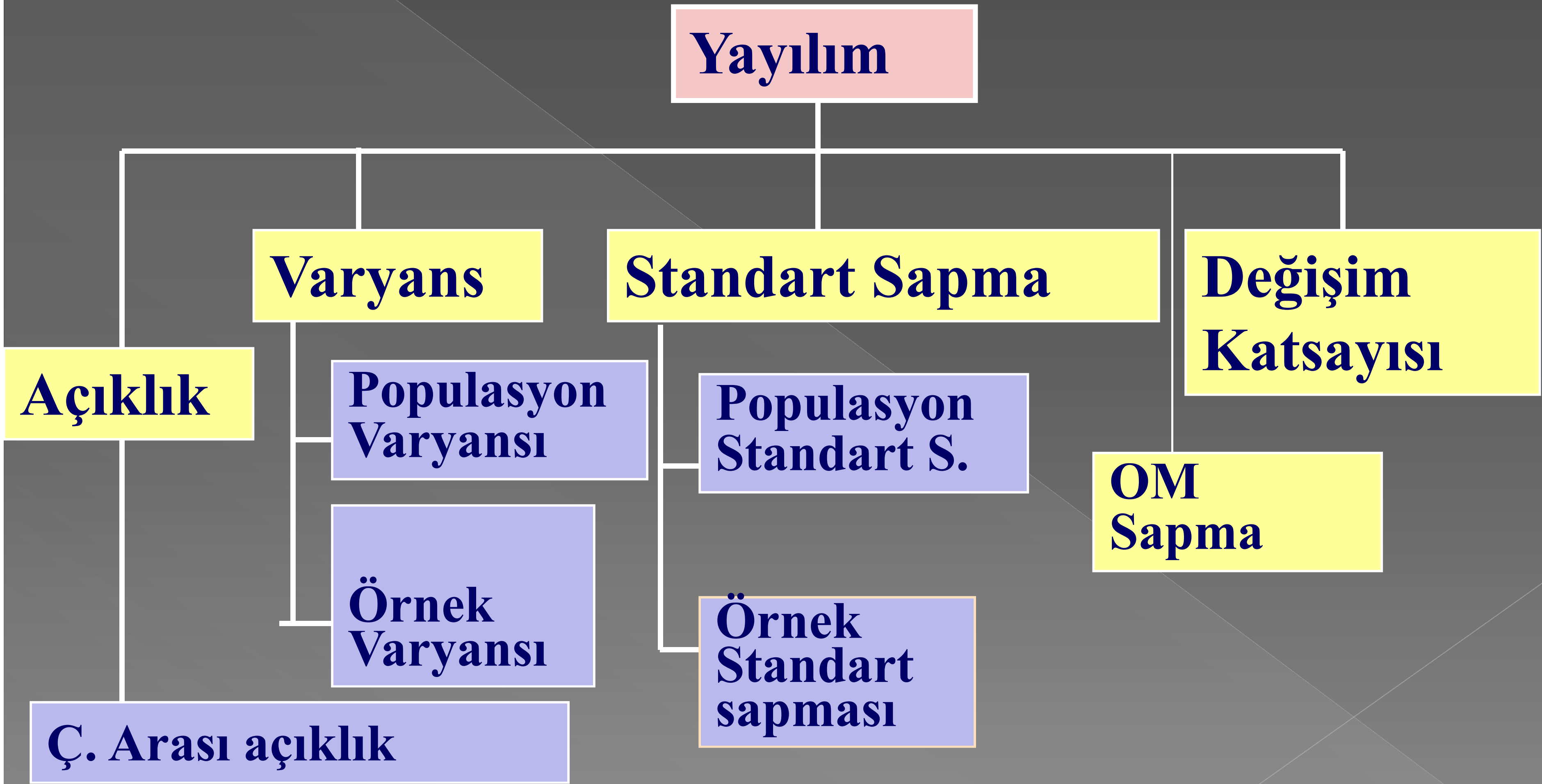
Çeyrekler

1. Çeyrek : $\frac{n+1}{4}$ üncü gözlem,

2. Çeyrek: $\frac{n+1}{2}$ inci gözlem (bu değer aynı zamanda medyan olarak da bilinmektedir)

3. Çeyrek: 3(n+1)/4'üncü gözlemdir (bu aynı zamanda 75. yüzdelik olarak da bilinmektedir.

Yayılım Ölçüleri

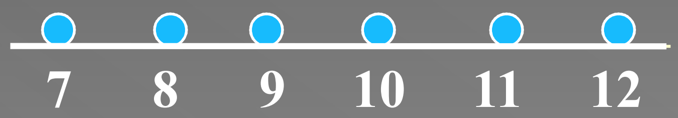


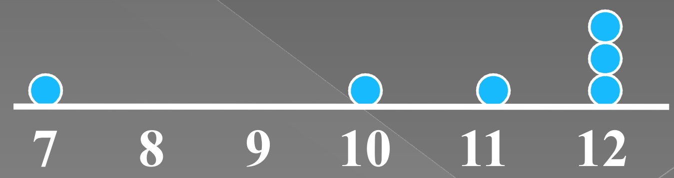
Açıklık

- Bir yayılım ölçüsüdür.
- En büyük gözlem ile en küçük gözlem arasındaki farktır:

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

- Datanın dağılımını göz ardı eder





Açıklık = 12 - 7 = 5

Açıklık:Örnek

- Açıklık iki uç değer arasında kalanları göz ardı eder.
- Açıklık=48 - 35 = 13

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

Çeyrekler arası Açıklık

$$CAA = Q_3 - Q_1$$

- Yayılım ölçüsüdür.
- Aynı zamanda orta yayılım olarak da bilinir
  - %50’lik yayılım
- 1. Ve 3. Çeyrekler arasındaki farktır.
- Aşırı uçlardan etkilenmez

Data büyükten küçüğe sıralı: 11 12 13 16 16 17 17 18 21

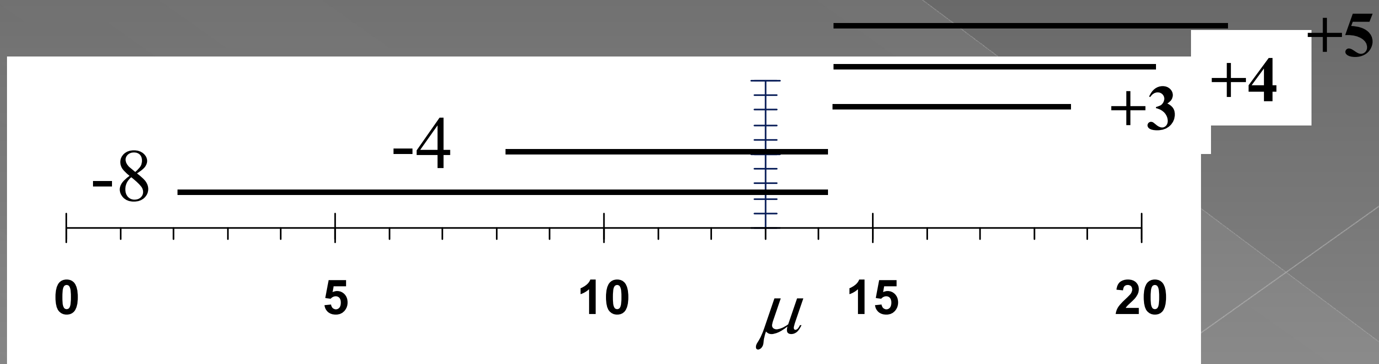
Çeyrekler arası açıklık= 17,5 – 12,5 =5

Ortalamadan Sapmalar

- Veri seti: 5, 9, 16, 17, 18
- Ortalama

$$\mu = \frac{\sum X}{N} = \frac{65}{5} = 13$$

- Ortalamadan sapmalar: -8, -4, 3, 4, 5



Varyans

- En önemli yayılım ölçüsü
- Ortalamadan sapmaları gösterir-ölçümler
  - Örnek varyansı:

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

  - Populasyon varyansı:

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

Standart Sapma

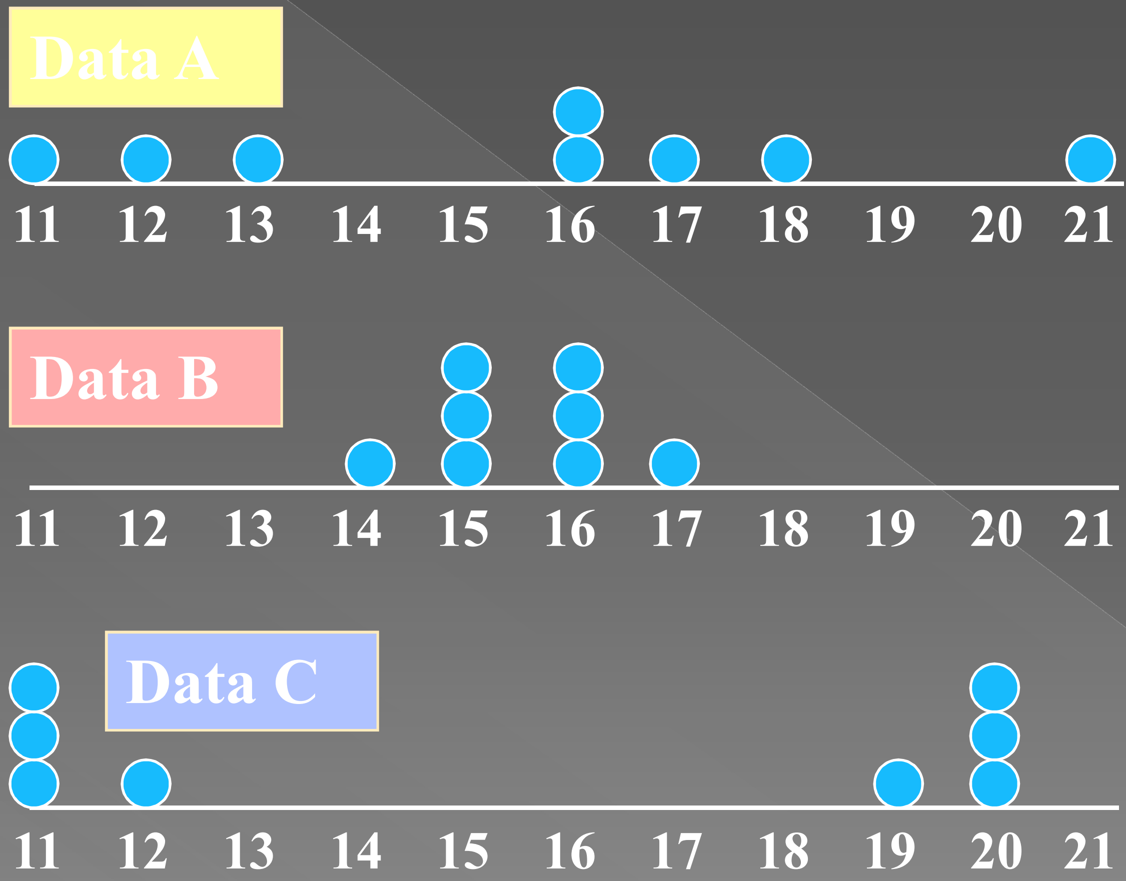
- En önemli yayılım ölçüsü
- Ortalamadan sapmaları gösterir-ölçümler
- Orijinal veri seti ile aynı birim cinsinden
  - Örnek standart sapması:

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

  - Populasyon Standart sapması:

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

Standart Sapmaların Karşılaştırılması



Data A: Ortalama = 15.5
s = 3.338

Data B: Ortalama = 15.5
s = .9258

Data C: Ortalama = 15.5
s = 4.57

Populasyon Varyansı

- Ortalamadan sapmaların karelerinin ortalaması

X	$X-\mu$	$(X-\mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{130}{5} = 26.0$$

Populasyon Standart Sapması

- Varyansın karekōküdür

X	$X-\mu$	$(X-\mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{130}{5} = 26.0$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{26.0} = 5.1$$

Örnek Varyansı

- Ortalamadan sapmaların karelerinin ortalamasıdır

X	$X-\overline{X}$	$(X-\overline{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

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

$$= \frac{663,866}{3}$$

$$= 221,288.67$$

Örnek Standart Sapması

- Varyansın kareköküdür.

X	$X-\overline{X}$	$(X-\overline{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$S^2 = \frac{\sum (X-\overline{X})^2}{n-1} = \frac{663{,}866}{3} = 221{,}288.67$$

$$S = \sqrt{S^2} = \sqrt{221{,}288.67} = 470.41$$

Değişim Katsayısı

- Nisbi yayılımı gösterir-ölçümler
- Her zaman (%) formundadır
- Ortalamaya oranla yayılımı gösterir
- Farklı iki veya daha fazla sayıda veri setinin karşılaştırılmasına yarar
- $$CV = \left(\frac{S}{\overline{\overline{X}}} \right) 100\%$$

CV

$$\mu_1 = 29$$

$$\sigma_1 = 4.6$$

$$C.V._1 = \frac{\sigma_1}{\mu_1}(100)$$

$$= \frac{4.6}{29}(100)$$

$$= 15.86$$

$$\mu_2 = 84$$

$$\sigma_2 = 10$$

$$C.V._2 = \frac{\sigma_2}{\mu_2}(100)$$

$$= \frac{10}{84}(100)$$

$$= 11.90$$

Değişim Katsayılarının Karşılaştırılması

- Hisse S. A:
  - Geçen yılki ortalama fiyat = $50
  - Standart sapma = $5
- Hisse S. B:
  - Geçen yılki ortalama fiyat = $100
  - Standart sapma = $5
- Değişim Katsayısı:
  - Hisse S. A: $CV = \left(\frac{S}{\overline{\overline{X}}}\right)100\% = \left(\frac{\$5}{\$50}\right)100\% = 10\%$
  - Hisse S. B: $CV = \left(\frac{S}{\overline{\overline{X}}}\right)100\% = \left(\frac{\$5}{\$100}\right)100\% = 5\%$

Ortalama Mutlak Sapma (OMS)

- OMS ortalamadan sapmaların mutlak değerlerinin ortalaması
- Yayılım ölçüsü
- Populasyon için

$$\frac{\sum_{i=1}^{N} \left| \mathrm{x}_{\mathrm{i}} - \mu \right|}{N}$$

- Örnek için

$$\frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right|}{n}$$

Ortalama Mutlak Sapma

- Ortalamadan sapmaların mutlak değerlerinin ortalamasıdır.

X	$X-\mu$	$\lvert X-\mu \rvert$
5	-8	+8
9	-4	+4
16	+3	+3
17	+4	+4
18	+5	+5
	0	24

$$M.A.D. = \frac{\sum \lvert X-\mu \rvert}{N} = \frac{24}{5} = 4.8$$

Ortalama-Grup Data

- En yaygın kullanılan merkezi eğilim ölçüsüdür.
- Veri setinde aşırı uçlar varsa bu ölçü etkilenir.

Populasyon: $$\mu = \frac{\sum_{i=1}^{K} f_i m_i}{N}$$

Örnek: $$\overline{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n}$$

Medyan-Grup Data

$$\text{Medyan} = L + \frac{n/2 - F}{f_m} c$$

L: medyan sınıfında yer alan en düşük değer

n: gözlem sayası

F: medyan sınıfına kadar olan sıklıkların toplamı

f_m: medyan sınıfına ait sıklık

c: sınıf aralığının genişliği

Mod-Grup Data

$$Mod = L + \frac{d_1}{d_1 + d_2} c$$

L: Mod sınıfının en düşük değeri

d_1: Mod sınıfı ile bir önceki sınıf arasındaki sınıf sıklıkğı farkı

d_2: Mod sınıfı ile bir sonraki sınıf arasındaki sınıf sıklığı farkı

c: sınıf aralığı genişliği

Varyans-Grup Data

Örnek için varyans formülü:

$$S^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \bar{x})^2}{n-1} \; veya \; \frac{\sum_{i=1}^{K} f_i m_i^2 - n\bar{x}^2}{n-1}$$

Populasyon için varyans formülü:

$$\sigma^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \mu)^2}{N} \; veya \; \frac{\sum_{i=1}^{K} f_i m_i^2}{N} - \mu^2$$

Ortalama Mutlak Sapma (OMS) Grup Data

Populasyon için OMS

$$\frac{\sum_{i=1}^{K} f_i \left| m_i - \mu \right|}{N}$$

Örnek için OMS

$$OMS = \frac{\sum_{i=1}^{n} f_i \left| m_i - \bar{x} \right|}{n} = \frac{500.71}{51} = 9.81$$

Gruplandırılmış Veri Setleri İçin Yüzdelikler, Ondalıklar ve Çeyrekler

- Bu konuyu bir örnek aracılığı ile açıklayalım

Sınıf Aralığı	*Sınıf orta noktası*	Sınıf sıklığı
1 – 5	3	14
6 – 10	8	7
11 – 15	13	8
16 – 20	18	8
21 – 25	23	10
26 – 31	28.5	4

Gruplandırılmış Veri Setleri İçin Yüzdelikler, Ondalıklar ve Çeyrekler

- 1. Çeyrek: $$\frac{n+1}{4}=\frac{52}{4}=13$$

- Bu sonuç veri setinde yer alan gözlem değerleri en küçükten en büyüğe doğru sıralandığında 13. değerin 1. çeyrek olduğunu göstermektedir. Yukarıdaki tablo incelendiğinde 13. gözlemin 1. sınıfta olduğu görülmektedir. Veri seti gruplandırılmış olarak verildiği için 13. gözlemin 1. sınıfta tam olarak hangi değere denk geldiğini bilmek imkansızdır. Ancak bu değeri yaklaşık olarak tahmin etmek mümkündür.

Gruplandırılmış Veri Setleri İçin Yüzdelikler, Ondalıklar ve Çeyrekler

- Çeyreğin hesaplanan değerinin belirlenen sınıfta yaklaşık olarak hangi değere denk geldiğini tahminde kullanılan formül şöyledir:

$$L + (j - \tfrac{1}{2})\frac{U - L}{f}$$

Gruplandırılmış Veri Setleri İçin Yüzdelikler, Ondalıklar ve Çeyrekler

- Hesaplama bize 1. sınıfta 13. gözlemin tahmini değerini göstermektedir. Yani 1. çeyreğe ait tahmini değer 4.57 dir.

$$L + (j - \tfrac{1}{2})\frac{U-L}{f} \longrightarrow 1 + (13 - \tfrac{1}{2})\frac{5-1}{14} = 4.57$$

Gruplandırılmış Veri Setleri İçin Yüzdelikler, Ondalıklar ve Çeyrekler

- 2. Çeyrek (Medyan): $\dfrac{n+1}{2} \Rightarrow \dfrac{51+1}{2} = 26$
- Bu sonuç medyanın 3. sınıfta 26. gözleme karşılık geldiğini göstermektedir.
- Medyanın tahmini değeri:

$$L + (j - \tfrac{1}{2})\frac{U-L}{f} \Rightarrow 11 + (5 - \tfrac{1}{2})\frac{15-11}{8} = 13.25$$

Gruplandırılmış Veri Setleri İçin Yüzdelikler, Ondalıklar ve Çeyrekler

- 3. çeyrek 3(51+1)/4 = 39

$$L+(j-\tfrac{1}{2})\frac{U-L}{f} \Longrightarrow 21+(2-\tfrac{1}{2})\frac{25-21}{10}=21.6$$

Çeyrekler arası açıklık = $3(\frac{n+1}{4}) - \frac{n+1}{4} \Longrightarrow$ 21.6 – 4.85 = 16.75

Populasyon Standart Sapmasının açıklanması

Chebychev (çebişev biçiminde telaffuz edilir) Kuralı

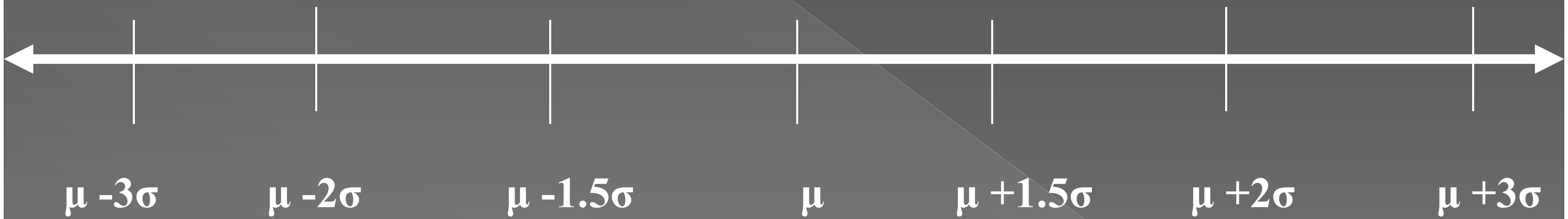
Ortalaması μ ve standart sapması σ olan bir

populasyon üyelerinin en az % 100 (1 -$(\frac{1}{M^2})$' lik

bölümü, M > 1 olması koşulu ile ortalamanın M standart sapması etrafında toplanır.

Populasyon Standart Sapmasının açıklanması

Pratikte Chebychev kuralının nasıl işlemekte olduğunu şu şekilde açıklayabiliriz:



Populasyon Standart Sapmasının açıklanması

Ampirik Kuralı

Ampirik kuralı, Başparmak Kuralı yada Gaus Kuralı olarak da bilinmektedir. Ampirik Kuralı'na göre gözlem değerleri normal dağılma sahip veya çan eğrisi biçiminde ise bu değerlerin yaklaşık olarak

Ampirik Kuralı

- Verilerin normal dağıldığı veya yaklaşık olarak normal dağıldığı varsayılır

Ortalamadan uzaklık	Bu uzaklıkta bulunan değerlerin yüzdesi
$\mu \pm 1\sigma$	68
$\mu \pm 2\sigma$	95
$\mu \pm 3\sigma$	99.7

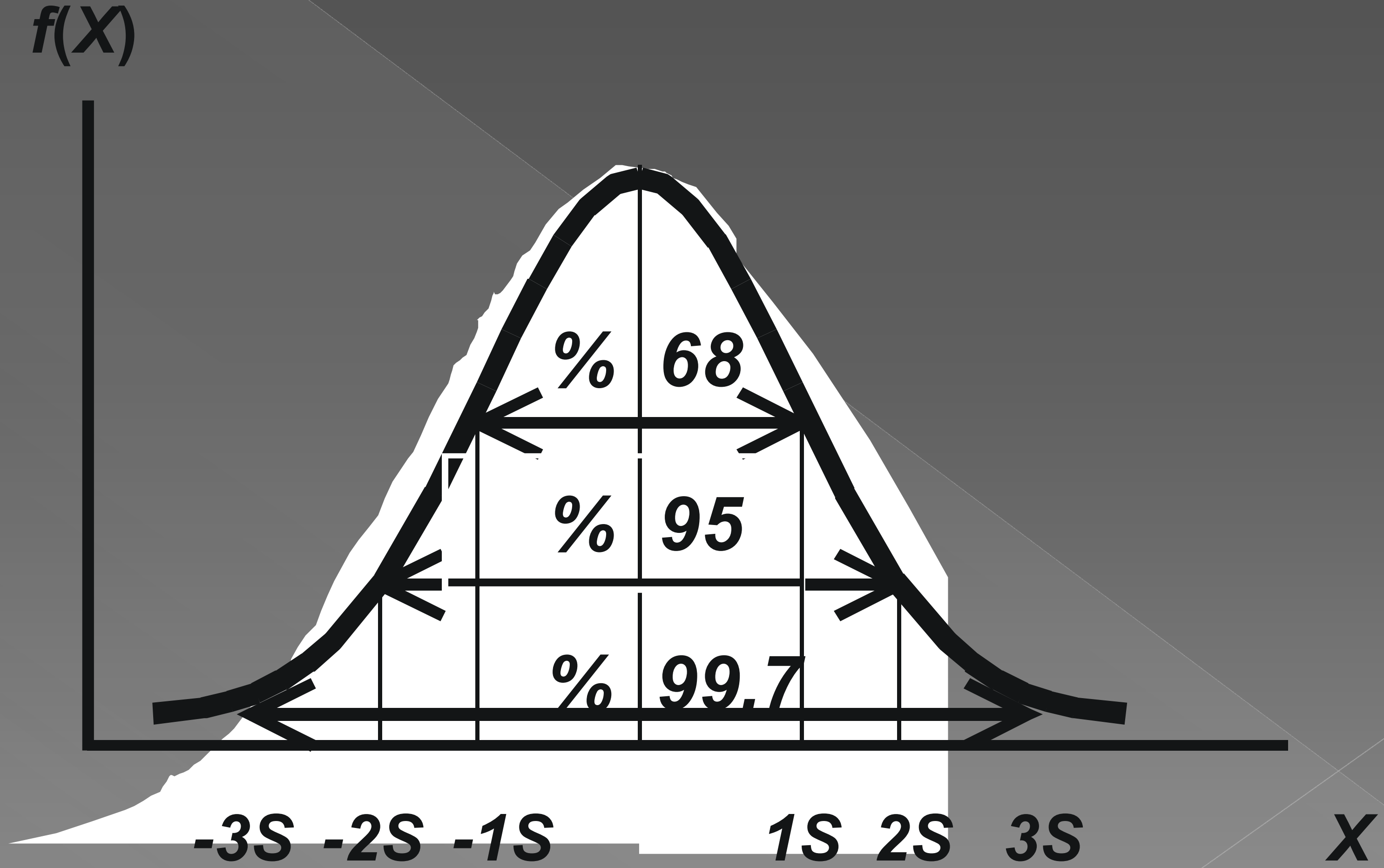
Populasyon Standart Sapmasının açıklanması

% 68'i($\bar{x}$ -S $\bar{x}$ +S) aralığında

% 95'i $\bar{x}$ (-2S $\bar{x}$ +2S) aralığında

%100'ü $\bar{x}$(-3S $\bar{x}$ +3S) aralığında

Populasyon Standart Sapmasının açıklanması



13	31	22	5	3	6
2	19	17	23	22	17
31	13	31	2	1	5
8	9	7	10	17	6
19	3	19	17	22	23
22	31	13	2	5	22
5	22	13	9	13	13
22	25	13	13	3	2
1	2	17			

Sınıf Aralığı		Sınıf Sıklığı	
1 – 5	m_i	14 f_i	f_i m_i
6 – 10	m_i 3	7	42
11 – 15	8	8	56
16 – 20	13	8	104
21 – 25	18	10	144
26 – 31	23	4	230
	28.5		114

$$\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n} = \frac{690}{51} = 13.53$$

Sınıf Aralığı	m_i	Sınıf Sıklığı f_i	$f_i m_i^2$
1 – 5	3	14	126
6 – 10	8	7	448
11 – 15	13	8	1352
16 – 20	18	8	2592
21 – 25	23	10	5290
26 – 31	28.5	4	3249

Varyans:

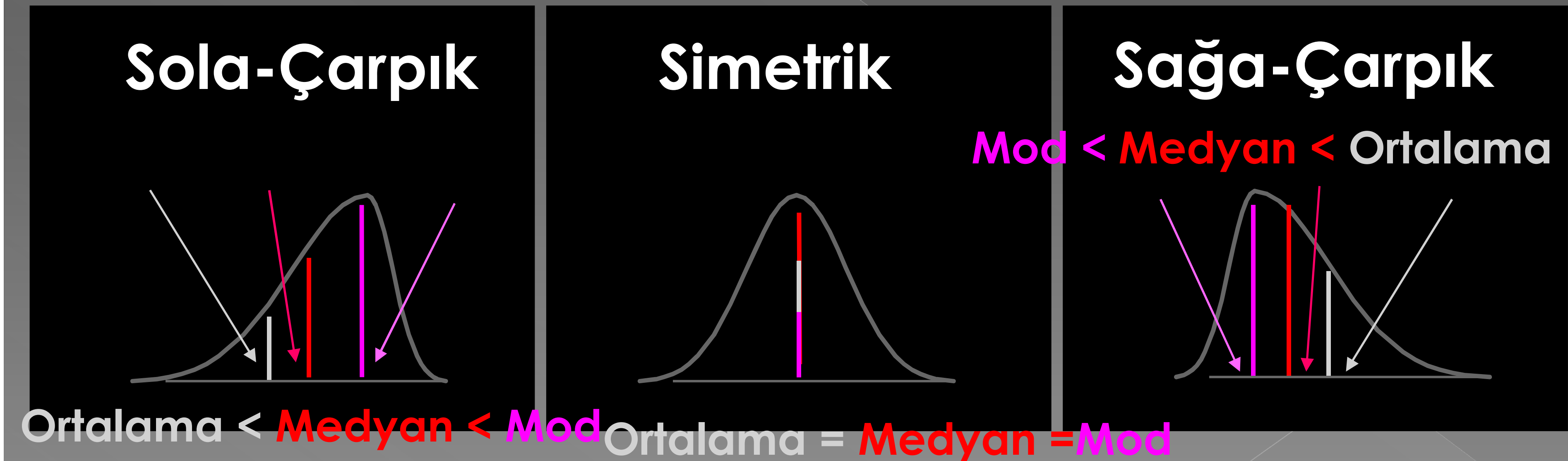
$$S^2 = \frac{\sum_{i=1}^{K} f_i m_i^2 - n\bar{x}^2}{n-1} = \frac{13{,}057 - (51)*(13.53)}{50} = 74.41$$

Standart sapma:

$$S = \sqrt{74.41} = 8.62$$

Dağılımların Şekilleri ve Ölçümleri

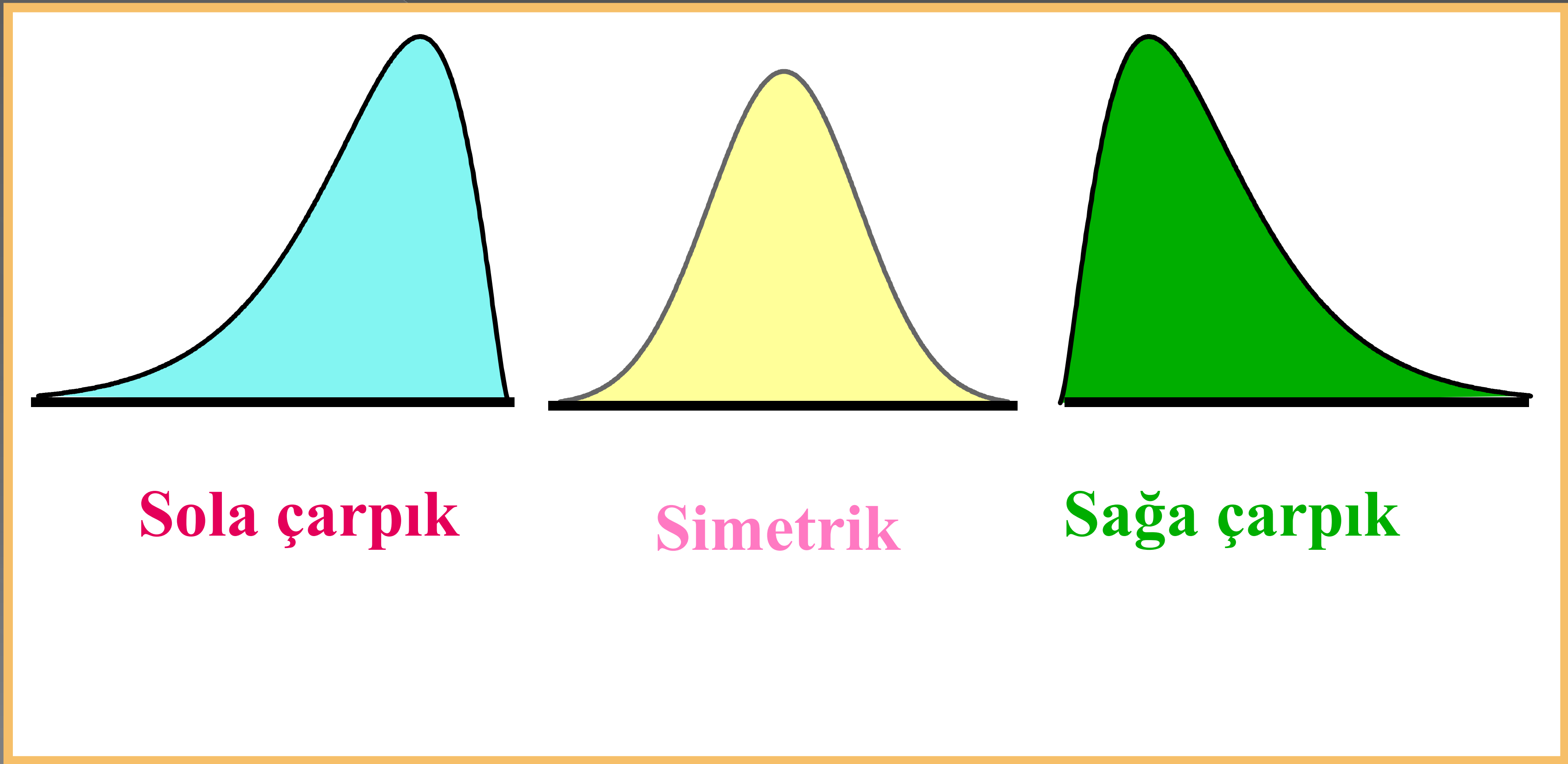
- Datanın dağılımını tanımlar
- Şekil Ölçüleri
  - Simetrik yada Çarpık



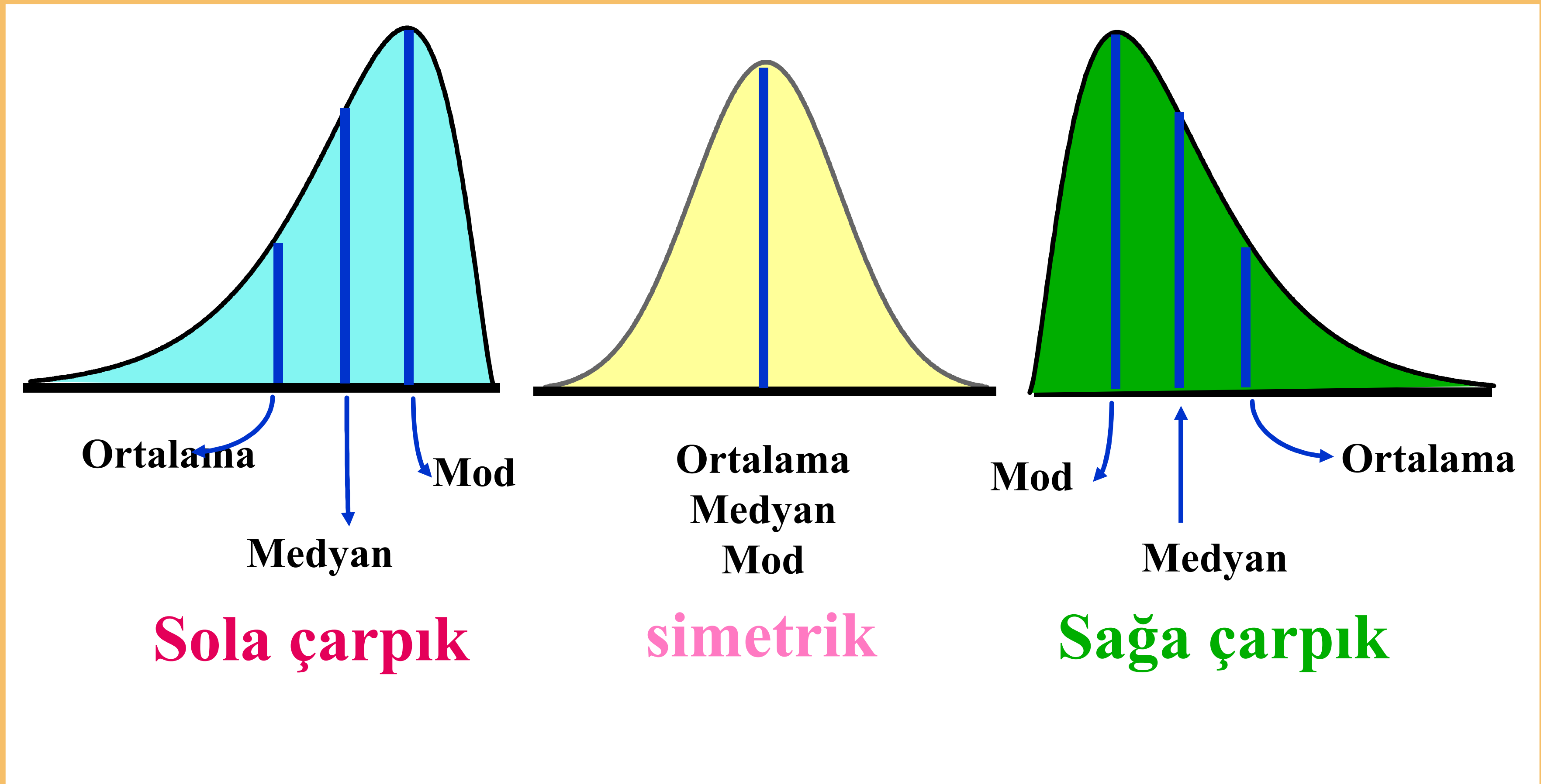
Dağılımların Şekilleri ve Ölçümleri

- **Çarpıklık (Skewness)**
  - Simetrik olmama
  - Uç değerler dağılımın bir yanında toplanmıştır
- **Tepeleme (Kurtosis)**
  - Dağılımın tepelemesi ile ilgili bilgi verir
- **Box and Whisker Grafikleri (Kutu Grafikleri)**
  - Dağılımın grafiksel gösterimidir
  - Çarpıklıkla ilgili bilgi verir

Çarpıklık



Çarpıklık


Ortalama
Mod
Medyan
Sola çarpık
Ortalama
Medyan
Mod
simetrik
Mod
Ortalama
Medyan
Sağa çarpık

Çarpıklık (Skewness):Sınıflandırılmamış data

Çarpıklık (Skewness) Pearson Skewness katsayısı formülü ile hesaplanır

Populasyon için Pearson Skewness katsayısı formülü: Sk = $\dfrac{3(\mu - medyan)}{\sigma}$

Örnek için Pearson Skewness katsayısı formülü:

Sk = $\dfrac{3(\bar{x} - medyan)}{S}$

Çarpıklık (Skewness):Sınıflandırılmış data

Populasyon için Pearson Skewness katsayısı formülü:

$$Sk = \frac{\sum_{i=1}^{N} f_i (x-\mu)^3}{\sigma^3}$$

Örnek için Pearson Skewness katsayısı formülü:

$$Sk = \frac{\sum_{i=1}^{n} f_i (x-\bar{x})^3}{S^3}$$

Çarpıklık

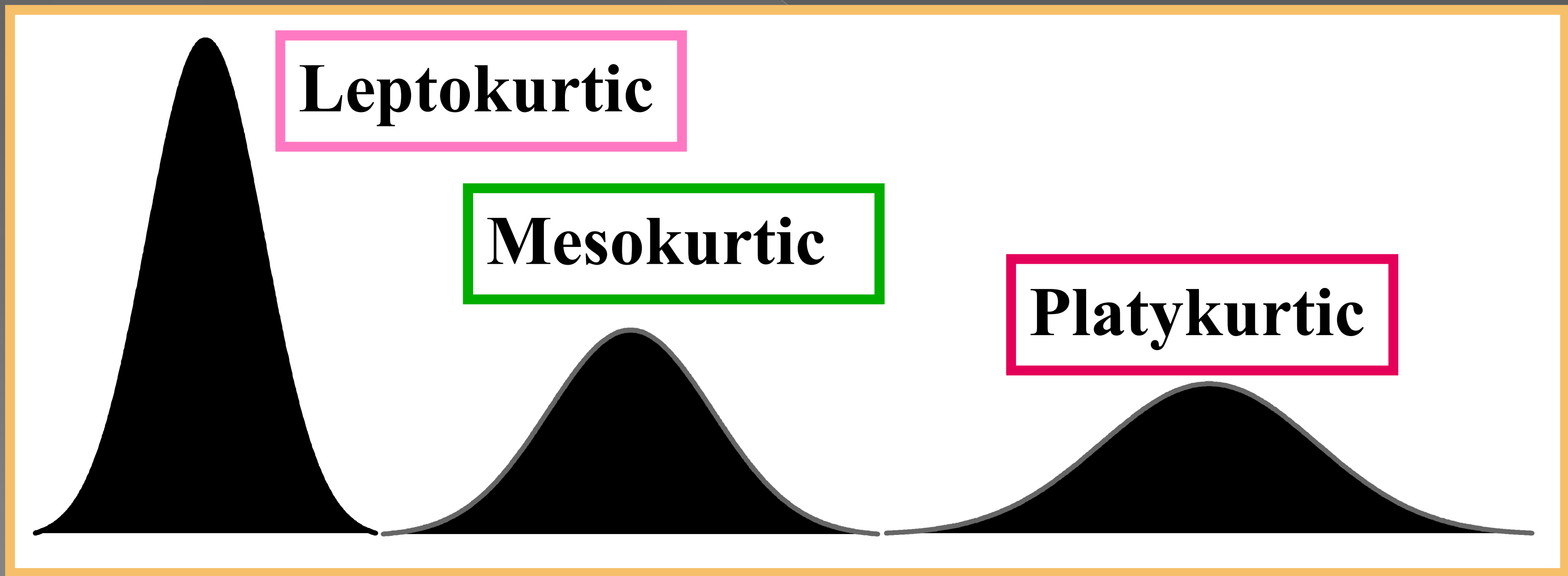
- ÇK < 0, ise dağılım sola çarpıktır.
- ÇK $= 0$, ise dağılım simetrik
- ÇK > 0, ise dağılım sağa çarpıktır.

Çarpıklık Katsayısı

$\mu_1 = 23$	$\mu_2 = 26$	$\mu_3 = 29$
$M_{d1} = 26$	$M_{d2} = 26$	$M_{d3} = 26$
$\sigma_1 = 12.3$	$\sigma_2 = 12.3$	$\sigma_3 = 12.3$
$S_1 = \dfrac{3(\mu_1 - M_{d1})}{\sigma_1}$	$S_2 = \dfrac{3(\mu_2 - M_{d2})}{\sigma_2}$	$S_3 = \dfrac{3(\mu_3 - M_{d3})}{\sigma_3}$
$= \dfrac{3(23-26)}{12.3}$	$= \dfrac{3(26-26)}{12.3}$	$= \dfrac{3(29-26)}{12.3}$
$= -0.73$	$= 0$	$= +0.73$

Tepeleme

- Dağılımın dikliği ile ilgili bilgi verir
  - Leptokurtic: yüksek ve ince
  - Mesokurtic: normal
  - Platykurtic: yatık ve yayvan



Tepeleme Katsayısı (Kurtosis): Gruplandırılmamış Data

Populasyon için Tepeleme =

$$\frac{(x-\mu)^4}{\sigma^4}$$

Örnek için Tepeleme = $$\frac{(x-\bar{x})^4}{S^4}$$

Tepeleme Katsayısı (Kurtosis): Gruplandırılmış Data

Populasyon için Tepeleme =

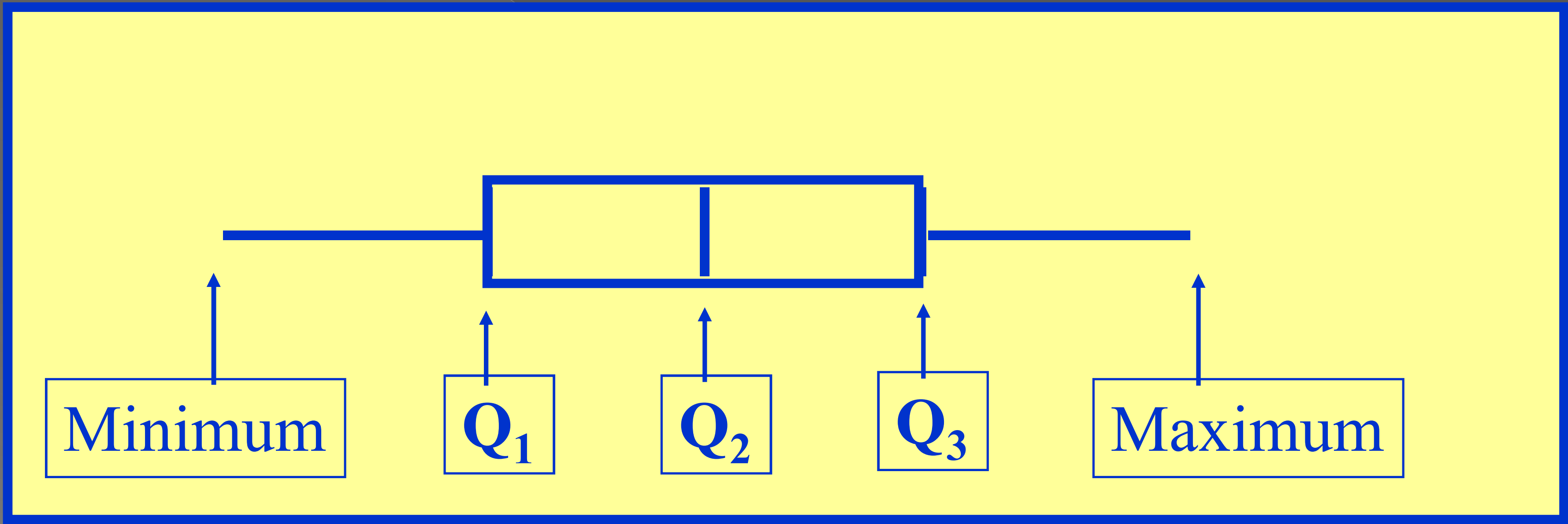
$$\frac{\sum_{i=1}^{N} f_i (x-\mu)^4}{\sigma^4}$$

Örnek için Tepeleme = $\dfrac{\sum_{i=1}^{n} f_i (x-\bar{x})^4}{S^4}$

Box ve Whisker Grafikleri

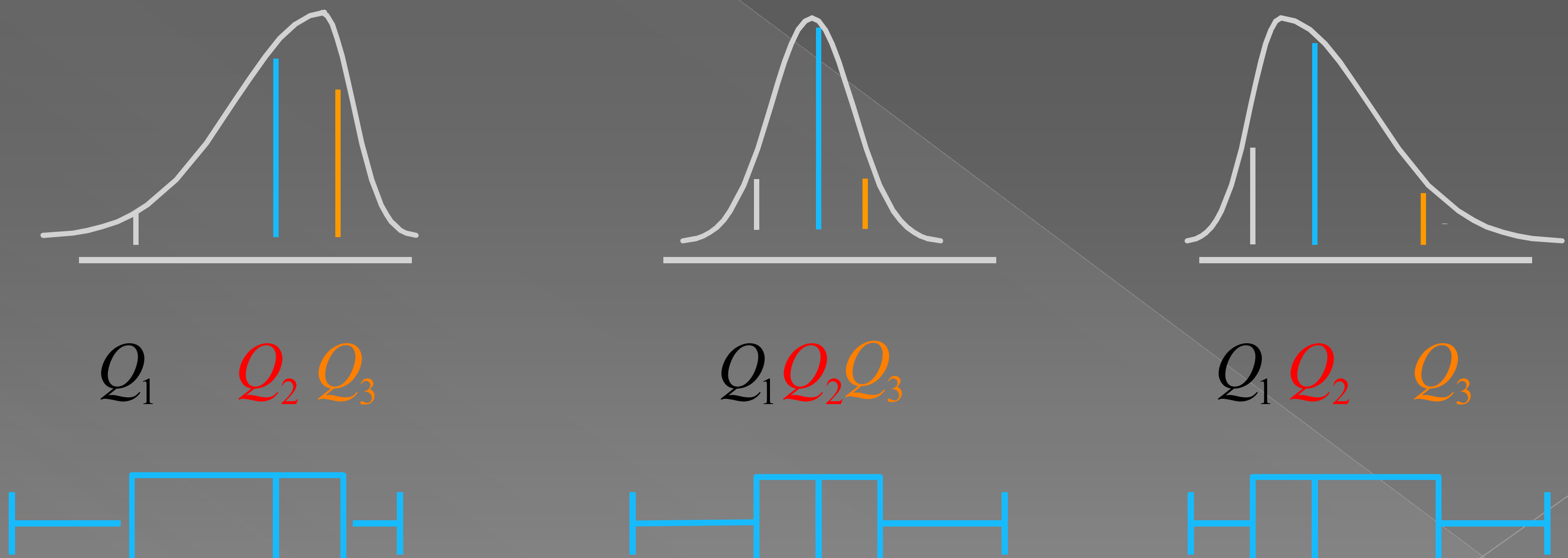
- **B/W grafiklerinde 5 adet sayısal özetleme gösterilir:**
  - **Medyan, Q_2**
  - **1. çeyrek, Q_1**
  - **3. çeyrek, Q_3**
  - **Data setindeki minimum ve maksimum değerler**

Box ve Whisker Grafikleri



Dağılım Şekilleri ve Box-and-Whisker Grafikleri

Sola-Çarpık **Simetrik** **Sağa-Çarpık**



Korelasyon Katsayısı

- İki adet sayılabilen değişken arasındaki ilişkiyi ölçümlemede kullanılır ve r ile gösterilir.

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Korelasyon Katsayısının Özellikleri

- -1 ile 1 arasında bir değer alır
- -1 mükemmel negatif, 1 ise mükemmel pozitif ilişkiyi gösterir.
- 0 yakın olduğunda ise ilişki çok zayıftır.

Farklı Korelasyonlar