

# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

TEMEL KAVRAMLAR

HAZIRLAYAN : Cemile YILDIZÇAKAR



# 1-VERİ TİPLERİ

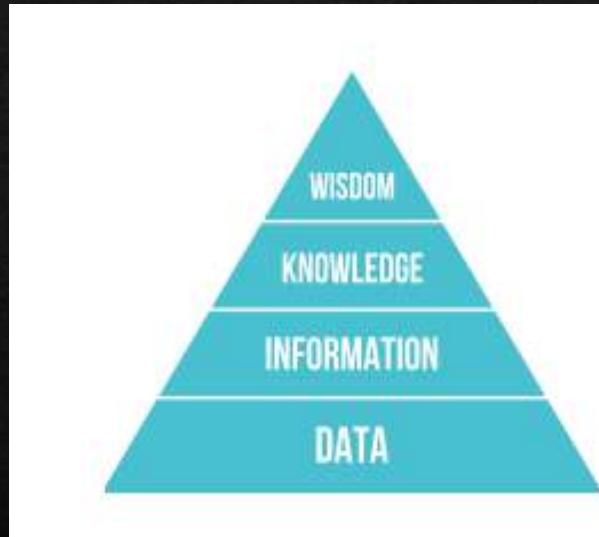
## ❖ Veri nedir??

Araştırma amacı ile toplanan gözlemler veya ölçümlerdir.

Cemile YILDIZÇAKAR

## ❖ NİCEL – NİTEL??

Sayılar, kodlar, harfler , kategoriler ... gibi farklı şekilde ifade edilebilen ancak sayısal değer taşımayan veriler NİTEL, sayısal değer taşıyan veriler ise NİCEL veridir.





# VERİ TIPLERİ ÖZELLİKLERİ

## 01 NİTEL VERİ (Qualitative)

- \* Tanımlar ile ifade edilebilir.
- \* Veri gözlemlenebilir ancak ölçülemez.

Cemile YILDIZÇAKAR

## 02 NİCEL VERİ (Quantitative)

- \* Sayılar ile ifade edilebilir.
- \* Veri ölçülebilir.
  - \*\*KESİKLİ VERİ (SAYILARBİLİR TÜRDEN VERİLER)
  - \*\*SÜREKLİ VERİ (BİR ARALIKTA DEĞER ALAN VERİLER)

# Örnekler;

- ❖ Medeni Durum
- ❖ Kan grubu
- ❖ İlçe Nüfusu
- ❖ Hava sıcaklığı
- ❖ Kütüphanedeki kitap sayısı.



Cemile YILDIZÇAKAR

- ❖ Aracın Hızı
- ❖ Forma Numarası
- ❖ Kahve Acılık Derecesi
- ❖ Yarış Bitirme Derecesi
- ❖ Maaş bilgisi





# Nicel Veri Türleri

## KESİKLİ VERİ (Discrete)

- ❖ Sayılabilen türdeki verilerdir.

Tam sayılar ile ifade edilirler.

\*Sokaktaki insan sayısı.

\*Çokgendeki kenar sayısı.

## SÜREKLİ VERİ (Continuous)

- ❖ Bir aralıktaki tüm değerleri alabilen verilerdir.

Reel Sayılar ile ifade edilir.

\*Hava Basıncı.

\*Paketteki pirinç miktarı.

Cemile YILDIZÇAKAR



# 2-ÖLÇEK TÜRLERİ

- ❖ Sınıflama(Nominal)
- ❖ Sıralama(Ordinal)
- ❖ Eşit Aralıklı (Interval)
- ❖ Eşit Oranlı(Ratio)



Cemile YILDIZÇAKAR



## SINIFLAMA (Nominal)

- ❖ Kategoriler
- ❖ Aralarında sıralama ilişkisi bulunmamaktadır.
- Cinsiyet
- Sıcak- soğuk içecekler.



Cemile YILDIZÇAKAR

## SIRALAMA (Ordinal)

- ❖ Puanlar sıra dizisini göstermektedir.
- ❖ Aralarında daha büyük, daha kötü, daha iyi ... gibi ilişkiler mevcuttur.

- Sınav sonuçları (AA, AB,..)
- Yarış dereceleri



## EŞİT ARALIKLI (Interval)

- ❖ Sayılar bir miktarla karşılık gelir.
  - ❖ SIFIR – İzafî sıfır olarak ifade edilir.
- Sıcaklık
  - Sınav notu

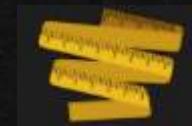


## EŞİT ORANLI (Ratio)

- ❖ Sıfır yokluğu ifade eder.
- ❖ Negatif değer almazlar.

Cemile YILDIZÇAKAR

- Uzunluk
- Ağırlık
- Sahip olunan arkadaş sayısı





# 3- İstatistik Uygulama Türleri

## Tanımlayıcı İstatistik (Descriptive)

- ❖ Veri setinin farklı istatistiksel araçlarla özetlenmesi.
- Tablolar
- Grafikler
- Özeti istatistikler (ortalama, varyans,...)

Cemile YILDIZÇAKAR



## Çıkarsamalı İstatistik (Inferential)

- ❖ Eldeki verileri kullanarak ilgilenilen kitle için çıkışsama yapılması, karar verilmesi, geleceğe yönelik tahminleme ve genelleme yapılması.
- Çıkarsama
- Hipotez testleri

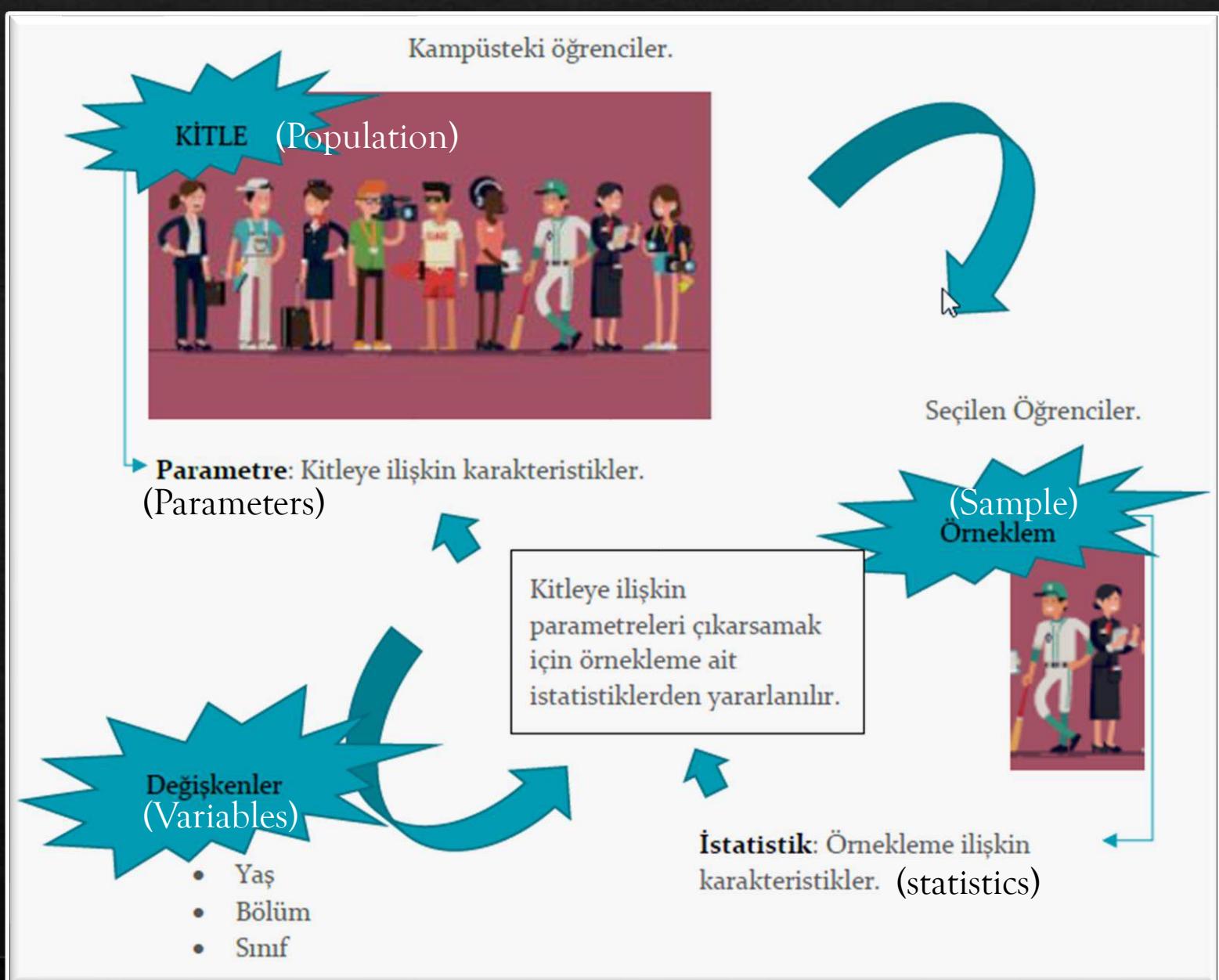




## 4-Temel Kavramlar

Örneğin; satın alınan ya da üretilen malzemelerin her bir parçasına kalite kontrol yapmanız maliyet ve zaman açısından mümkün değildir, kitlenin özelliklerini taşıyan bir grup örnekleme alınarak gerekli kontroller sağlanır ve karar verme işlemi yapılır.

Cemile YILDIZÇAKAR





❖ TEŞEKKÜR EDERİM ...

Cemile YILDIZÇAKAR

9.12.2020



# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-3

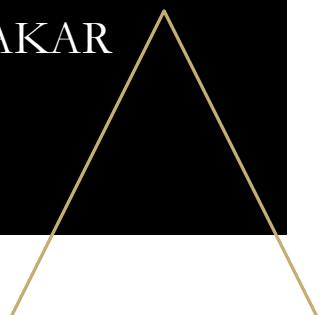
CEMİLE YILDIZÇAKAR

22.12.2020



# İÇİNDEKİLER

Cemile YILDIZÇAKAR

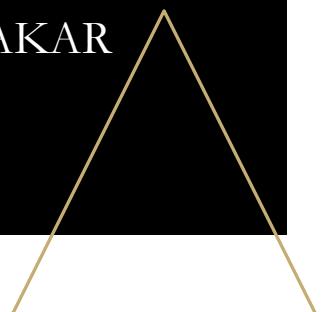


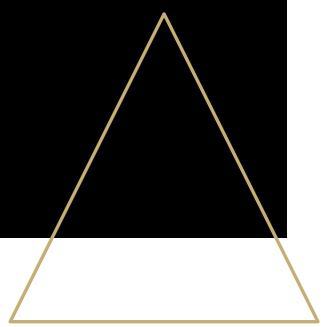
# Değişim Ölçüleri

- **Değişim:**

Bir dağılımda ölçümler arasında gözlenen farklılık ve değişikliğe **değişim**, veriler arasındaki değişimden kaynaklanan farklılıkların istatistiksel ölçülerine ise **değişim ölçüleri** denir.

Cemile YILDIZÇAKAR





## Nitel Veriler için Frekans Tablosu

Örnek;

Bir sağlık meslek yüksek okulunda ilgili programlara 400 öğrenci Hemşirelik, 300 öğrenci ameliyathane hizmetleri, 200 öğrenci Anestezi teknikeri, 100 öğrenci acil servis hizmetlerine kabul edilmiş olsun.

Sınıf	Frekans (fi)	Göreli Frekans Relative Frequency	Yüzde (%)	Kümülatif (%)
Hemşirelik	400	0,333333333	33,33333	33,33333333
Ameliyathane h.	300	0,25	25	58,33333333
Anestezi t.	200	0,166666667	16,66667	75
Acil Servis h.	100	0,083333333	8,33333	83,33333333
Düzen	200	0,166666667	16,66667	100
<b>TOPLAM ::</b>	<b>1200</b>	<b>1</b>	<b>100</b>	

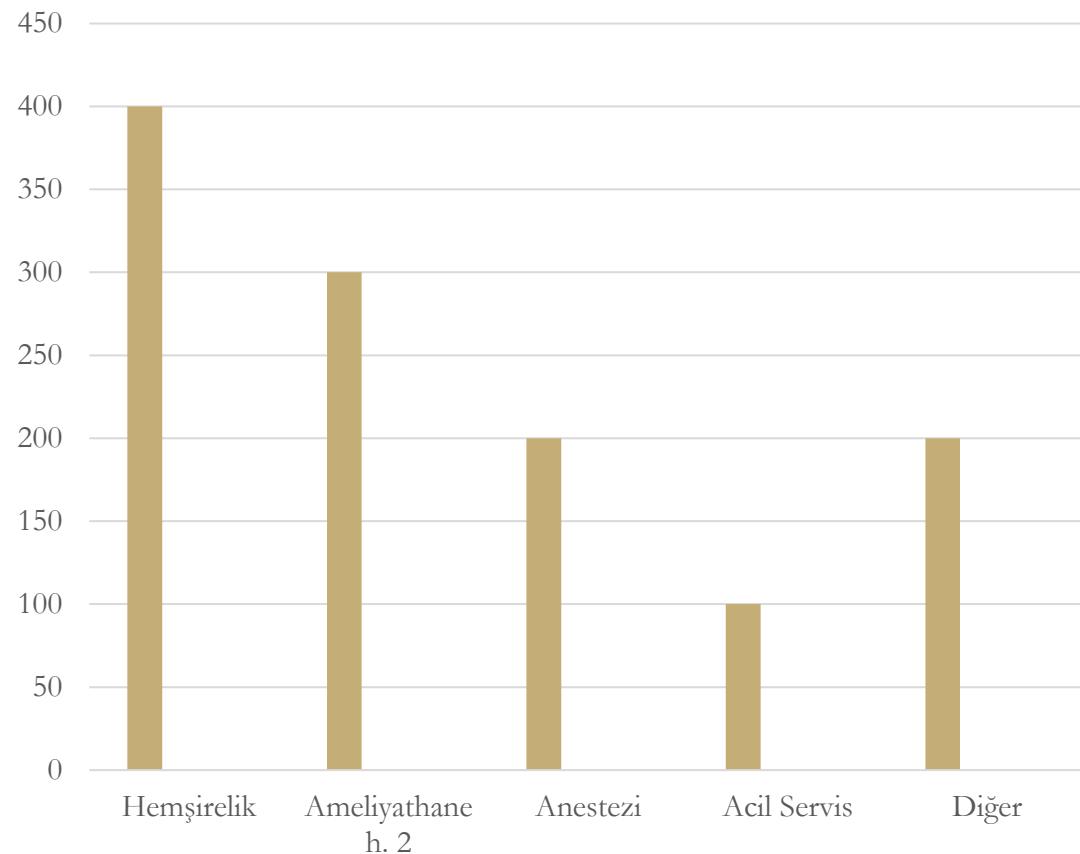
Cemile YILDIZÇAKAR

# Çubuk grafiği (Bar Chart)

Dağılım çubukları grafiği, kesikli nicel verilerde ve nitel verilerde kullanılır. Çubuk grafiğinde sınıflar, tabanları eşit ve birbirine bitişik olmayan dikdörtgenlerle temsil edilir.

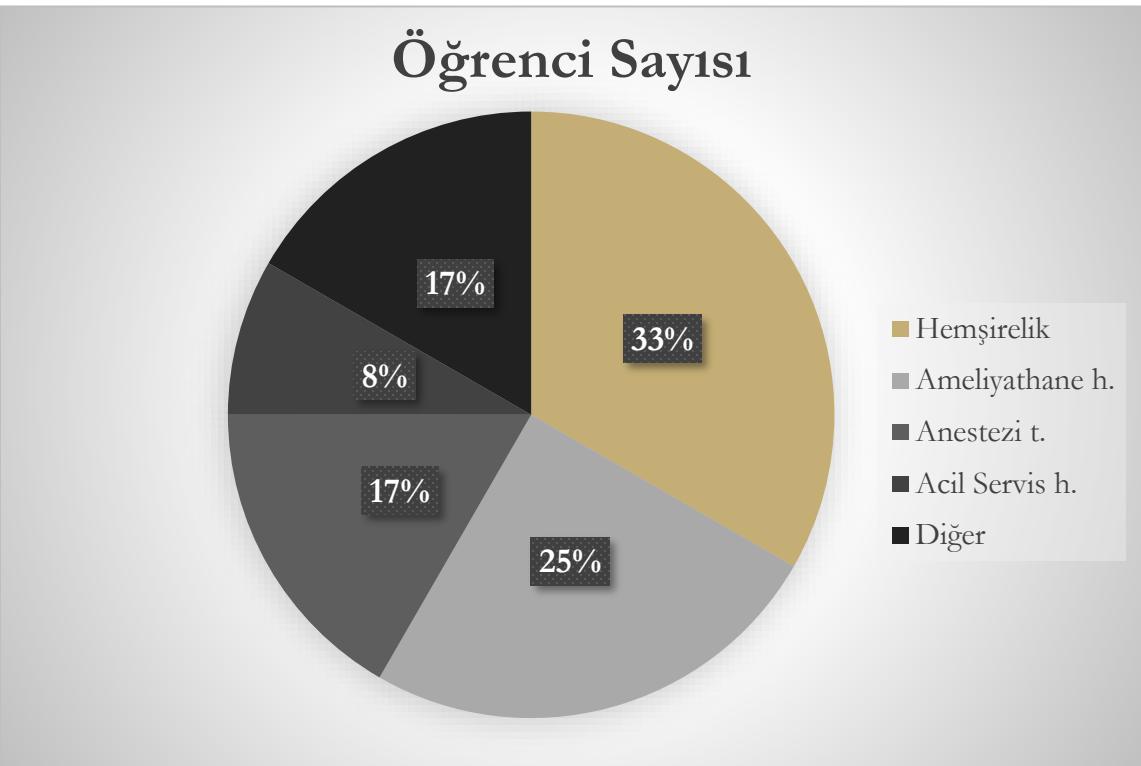
Cemile YILDIZÇAKAR

Bar Chart



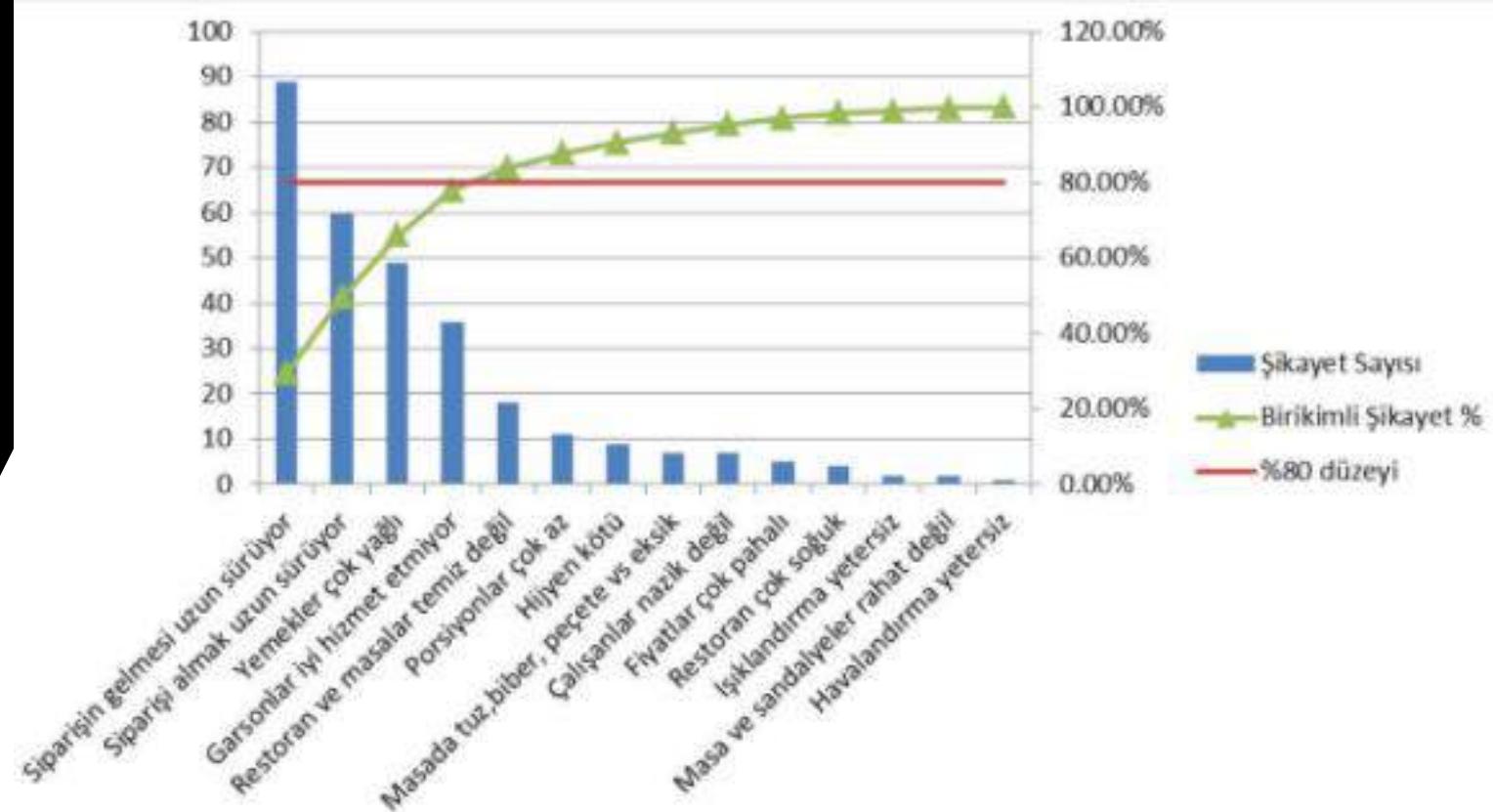
# Paste Grafigi (Pie Chart)

Her sınıfı düşen frekansın bir dairenin parçası ile gösterildiği grafik türüdür. Bu grafiği çizebilmek için görelî frekanslar hesaplanır. Her sınıfı ilişkin görelî frekans 3600 ile çarpılarak o sınıfı ilişkin daire dilimleri bulunur. Tüm sınıflar için yapıldığında daire tamamlanmış olur. Daha çok sınıflandırılabilen verilerde kullanılır.



# Pareto Diagramı

Soldan sağa doğru azalan sıradan yükseliğe göre düzenlenmiş nitel değişken kategorilerini içeren bir çubuk grafiğidir.



# Nicel Veriler İçin Tanımlayıcı Yöntemler

## NOKTA GARFİĞİ (DOT PLOT)

- Her bir verinin tek bir nokta ile gösterildiği grafiktir.
- Veri setindeki boşluklar, kümeler, verinin yayılımı net olarak görülebilir.

Cemile YILDIZÇAKAR

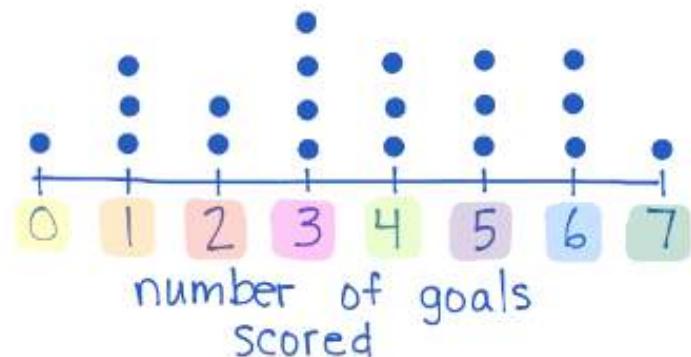
data set:

6	3	6	3	5
7	4	6	5	3
4	4	5	1	0
3	2	2	1	1

number of goals	frequency
0	1
1	3
2	2
3	4
4	3
5	3
6	3
7	1

data set:

6	3	6	3	5
7	4	6	5	3
4	4	5	1	0
3	2	2	1	1



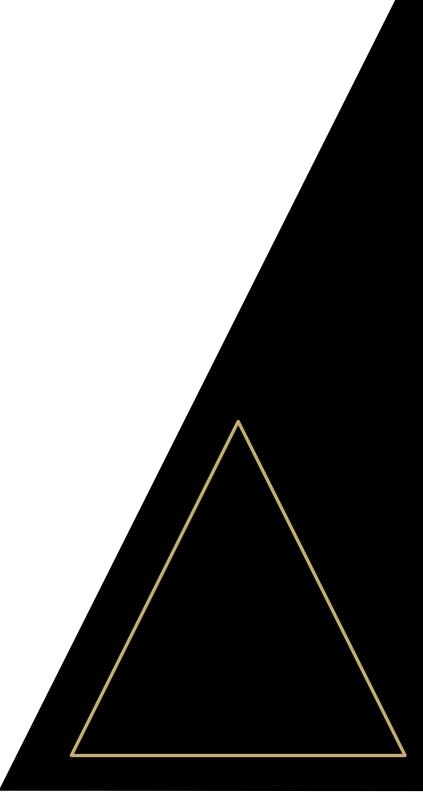


## Gövde Yaprak Grafiği (Steam and Leaf display)

44, 46, 47, 49, 63, 64, 66, 68, 68, 72, 72, 75, 76, 81, 84, 88, 106

Stem	Leaf
4	4 6 7 9
5	
6	3 4 6 8 8
7	2 2 5 6
8	1 4 8
9	
10	6

- Veri setinde yer alan bir değerin bir kısmının gövde diğer kısmının yaprak olarak ayrılarak gösterilmesidir.



Cemile YILDIZÇAKAR



# Gruplandırılmış Frekans Dağılımı Oluşturma

- Veri Seti (S):** 20 kişilik bir sınıfın öğrencilerinin matematik test puanları

Ham Veri																			
96	90	80	67	60	51	40	30	51	60	60	67	80	90	51	60	67	60	60	51
Büyükten Küçüğe Sıralanmış Veri																			
96	90	90	80	80	67	67	67	60	60	60	60	60	60	51	51	51	51	40	30

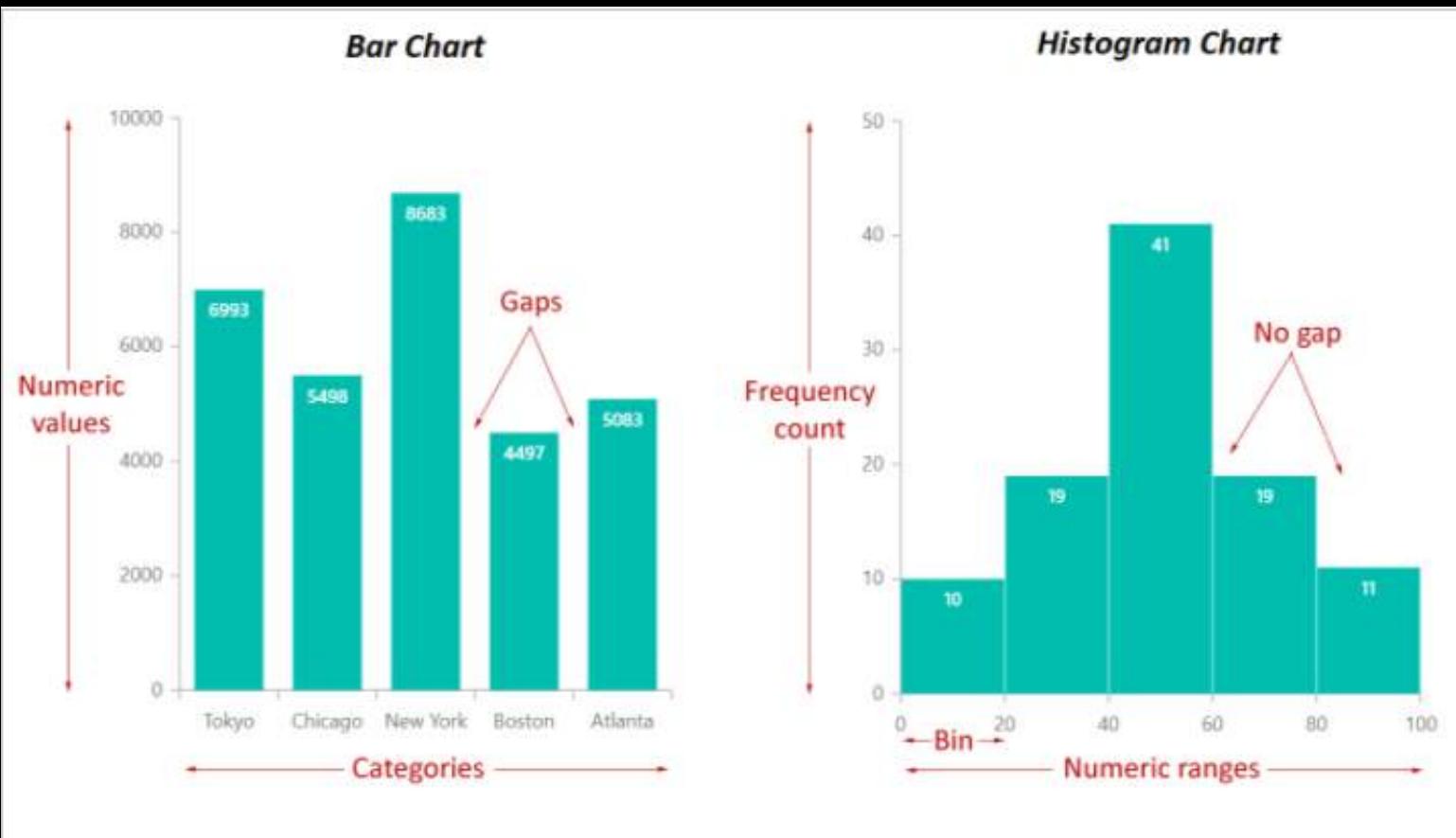
Interval	Frequency	Kümülatif Frekans	Relative Frequency	Percentage
30 - 47	2	2	0,1	10
48 - 65	10	12	0,5	50
66 - 83	5	17	0,25	25
84 - 101	3	20	0,15	15
<b>total::</b>	<b>20</b>		<b>1</b>	<b>100</b>

□ Sınıf sayısı:  $\lceil \frac{17}{4} \rceil \approx \sqrt{\text{number of observations}}$

□ Sınıf genişliği:  $\lceil \frac{(96-30)}{4} \rceil = \frac{\text{highest score} - \text{lowest score}}{\text{number of classes}}$

\*\*Sınıflar ayrık, sınıf genişliği sabit olmalı.

Cemile YILDIZÇAKAR



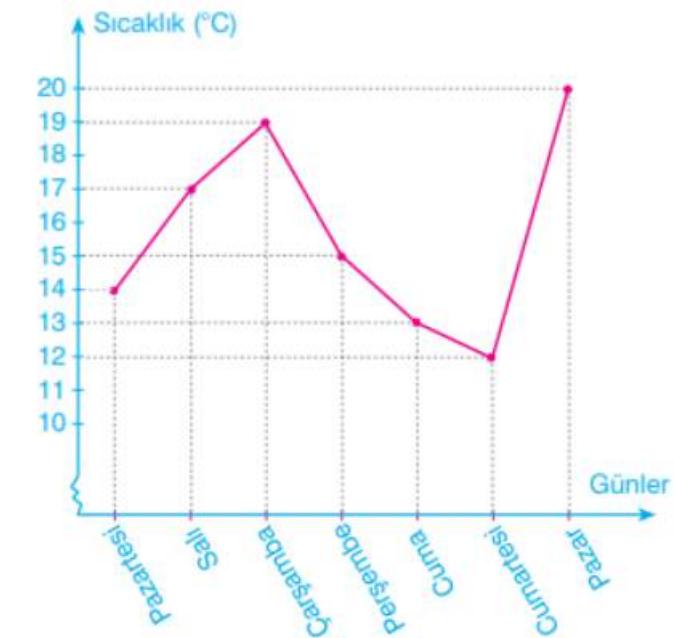
Cemile YILDIZÇAKAR

# Diğer Grafikler

Cemile YILDIZÇAKAR

## Çizgi Grafiği (Line Chart)

Aşağıdaki grafikte bir yerleşim biriminde bir hafta boyunca ölçülen hava sıcaklık değerleri gösterilmiştir.

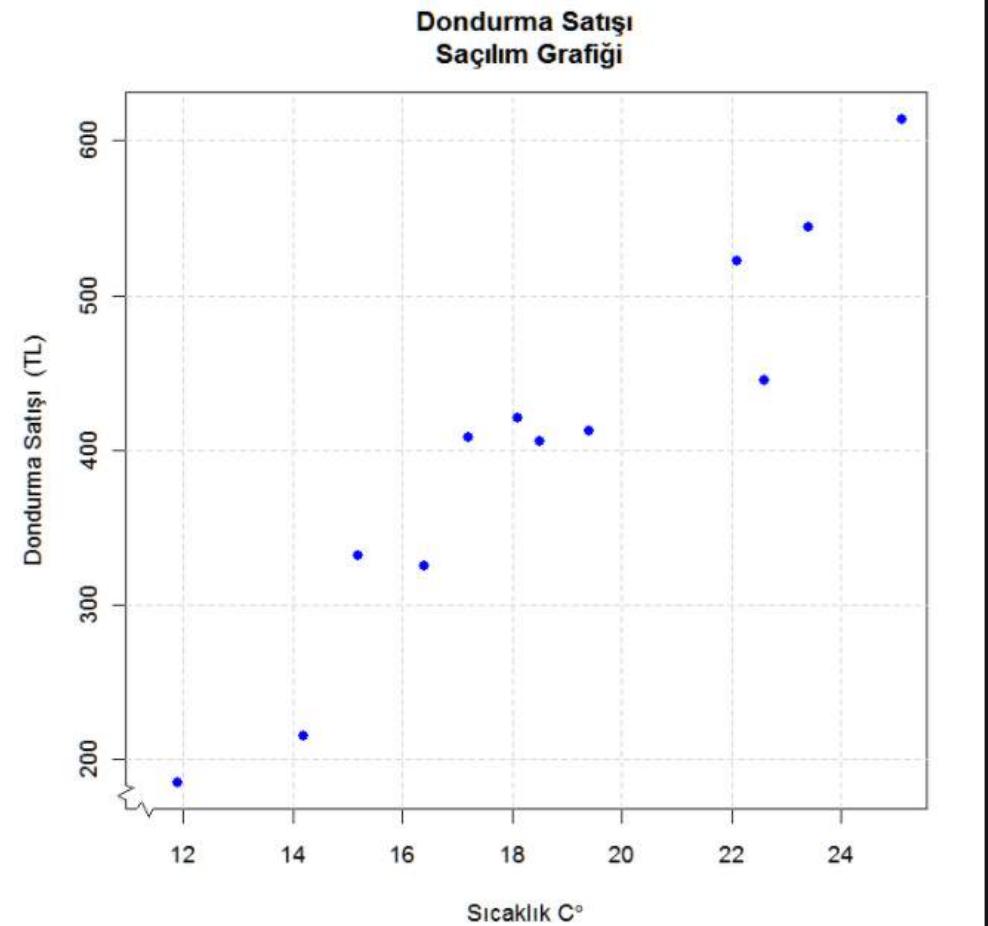


# Diger Grafikler

Sıcaklık (C°)	Dondurma Satışı (TL)
14,2	215
16,4	325
11,9	185
15,2	332
18,5	406
22,1	522
19,4	412
25,1	614
23,4	544
18,1	421
22,6	445
17,2	408

Cemile YILDIZÇAKAR

## Serpme Diagramı (Scatter Diagram)



# Çapraz Tablo

Cemile YILDIZÇAKAR

Türkiye AB'ye üye olamalı \* Cinsiyet Crosstabulation

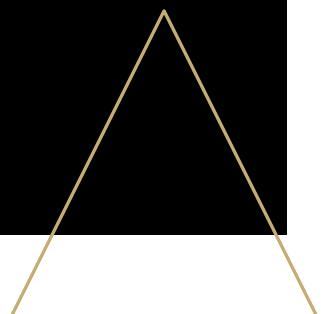
	Türkiye AB'ye üye olamalı		Cinsiyet		Total
			Kadın	Erkek	
Kesinlikle Katılıyorum	Kesinlikle Katılıyorum	Count	15	28	43
		% within Cinsiyet	11,5%	16,1%	14,1%
	Katılıyorum	Count	36	45	81
		% within Cinsiyet	27,7%	25,9%	26,6%
	Kararsızım	Count	38	40	78
		% within Cinsiyet	29,2%	23,0%	25,7%
Katılmıyorum	Katılmıyorum	Count	24	19	43
		% within Cinsiyet	18,5%	10,9%	14,1%
	Kesinlikle Katılmıyorum	Count	17	42	59
		% within Cinsiyet	13,1%	24,1%	19,4%
	Total	Count	130	174	304
		% within Cinsiyet	100,0%	100,0%	100,0%

# Merkezi Eğilim Ölçüleri

## (Measures of central tendency)

- Merkezi Eğilim Ölçüleri, belli bir özelliğe ya da değişkene ilişkin ölçme sonuçlarının, hangi değer etrafında toplandığını gösteren ve veri grubunu özetleyen ölçülerdir.
- Konum ölçüleri olarak da bilinir.

Cemile YILDIZÇAKAR



# ORTALAMA (MEAN)

- Üzerinde inceleme yapılan veri setindeki elemanların toplanıp incelenen eleman sayısına bölünmesiyle elde edilen yer ölçüsüne aritmetik ortalama denir.
- Halk dilinde ortalama ifadesi kullanıldığında ilk akla gelen kavram aritmetik ortalamadır.

Sample mean

$$\bar{X}$$

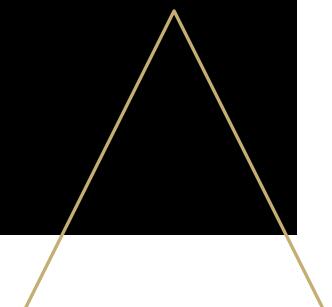
X-bar

Population mean

$$\mu$$

Greek letter (mu)

Cemile YILDIZÇAKAR



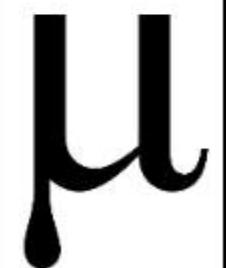
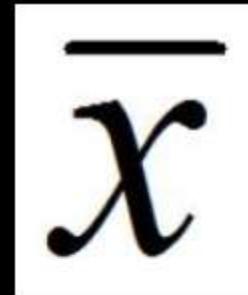


The **sample mean** is represented by *x bar*  $\bar{x}$ . It is given by the formula

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Örneklemde  
gözlenen veriler

Örneklem  
veri sayısı



The **population mean** is represented by the Greek letter *mu* ( $\mu$ ). It is given by the formula

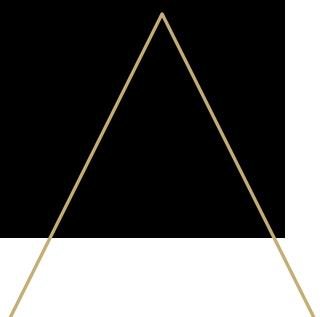
$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N}$$

Kitledeki veriler

Kitle veri sayısı

Veri setindeki her bir değişken hesaba katıldığı için bir verinin değişmesi tüm ortalamayı etkiler.

Cemile YILDIZÇAKAR



## Gruplanmış Seriler İçin Aritmetik Ortalama

$$\bar{x} = \frac{\sum_{i=1}^k x_i f_i}{\sum_{i=1}^k f_i}$$

$$\sum_{i=1}^k f_i = n$$

f : frekans  
k: grup sayısı  
i = 1,2,3,.....,k

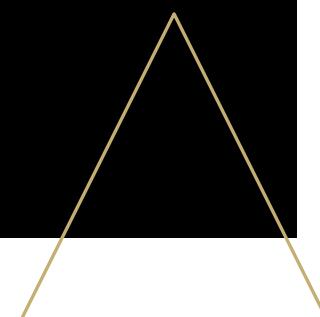
Grup	Frekans	$x_i f_i$
51	1	51
66	3	198
72	4	288
82	5	410
94	7	658
$\sum f_i = 20$		1605

Örnek: Yandaki tabloda bir Samsung bayisindeki LCD televizyonların ekran boyutlarına göre satış miktarları verilmiştir. Frekans dağılımının aritmetik ortalamasını hesaplayınız.

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^k x_i f_i}{\sum_{i=1}^k f_i} = \frac{51(1) + 66(3) + \dots + 94(7)}{1+3+4+5+7} \\ &= \frac{1605}{20} = 80,25\end{aligned}$$

Kaynak: Hamdi Emeç

Cemile YILDIZÇAKAR



## Sınıflanmış Seriler İçin Aritmetik Ortalama

$$\bar{x} = \frac{\sum_{i=1}^k m_i f_i}{\sum_{i=1}^k f_i}$$

f : frekans

k : sınıf sayısı

i = 1, 2, 3, ..., k

m : sınıf orta noktası

$$\sum_{i=1}^k f_i = n$$

- Sınıflanmış serilerde her bir sınıf içindeki değerlerin neler olduğu bilinmediğinden dolayı ve yalnızca her bir sınıfın frekans değerleri bilindiğinden dolayı sınıfı temsil etmek üzere sınıf orta noktaları hesaplamada kullanılır.

Kaynak: Hamdi Emeç

Örnek: Aşağıdaki tabloda 30 günlük süre içinde bir restoranın kullandığı et miktarının dağılımı verilmiştir. Günlük kullanılan et miktarlarının aritmetik ortalamasını hesaplayınız:

Sınıflar	$f_i$	$m_i$	$m_i f_i$
30-36'dan az	2	33	66
36-42'den az	6	39	234
42-48'den az	10	45	450
48-54'dan az	7	51	357
54-60'den az	4	57	228
60-66'den az	1	63	63
Toplam	30		1398

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^k m_i f_i}{\sum_{i=1}^k f_i} = \frac{33(2) + 39(6) + \dots + 63(1)}{30} \\ &= \frac{1398}{30} = 46,6 \text{ kg.}\end{aligned}$$

Cemile YILDIZÇAKAR

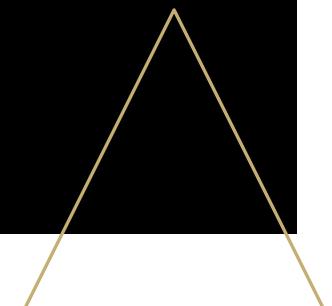
# ORTANCA (MEDIAN)

- Bir veri setindeki tüm değerlerin büyükten küçüğe sıralandığında orta noktasıdır.
- Ortanca veri setini iki eşit parçaya ayırır.
- Eşit aralıklı, oran ve sıralama ölçme düzeyinde ölçülen değişkenler için kullanılır.
- Ortancanın konumunu belirlemek için;

$$ortanca(OR) = \begin{cases} x_j & , \quad j = \frac{n+1}{2} \text{ } n \text{ tek} \\ \frac{x_j + x_{j+1}}{2} & , \quad j = \frac{n}{2} \text{ } n \text{ çift} \end{cases}$$

$$\frac{\text{M E D} + \text{I A N}}{2}$$

Cemile YILDIZÇAKAR

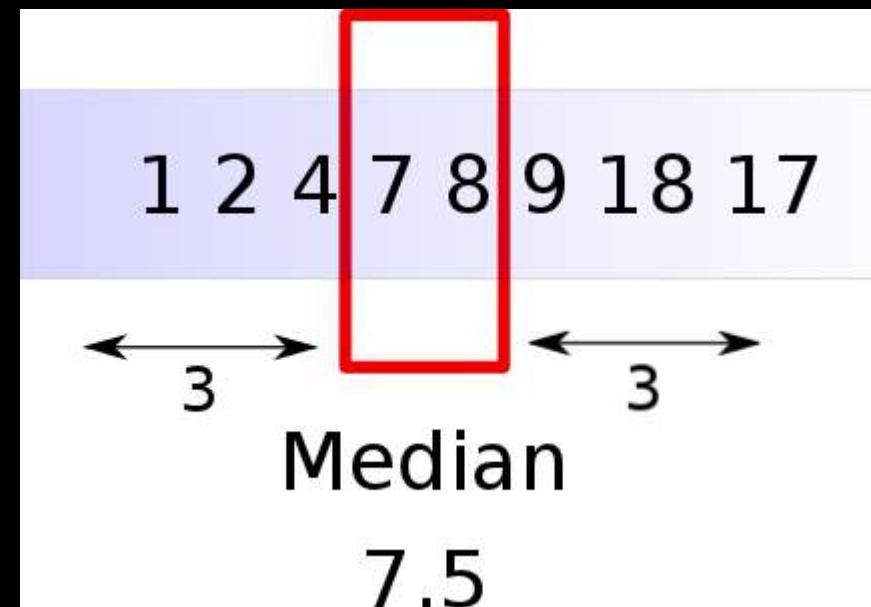


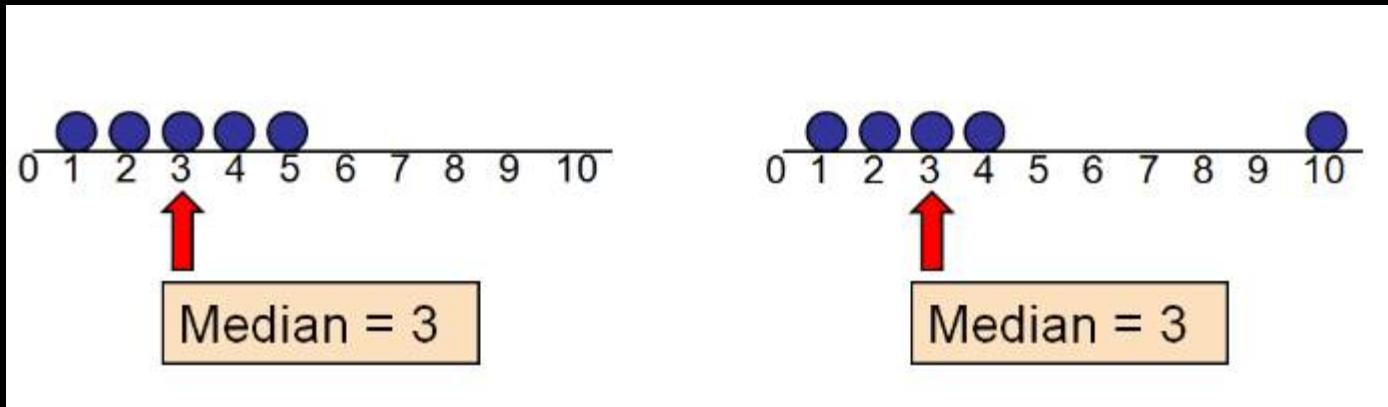
5, 13, 9, 7, 1, 9, 2, 9, and 11

put in  
ascending order

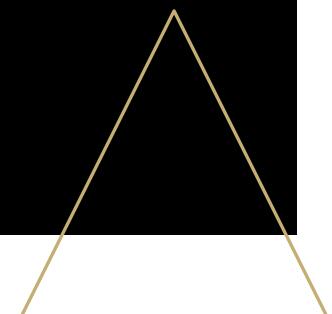
1, 2, 5, 7, **9**, 9, 9, 11, 13

Median  
(middle value)





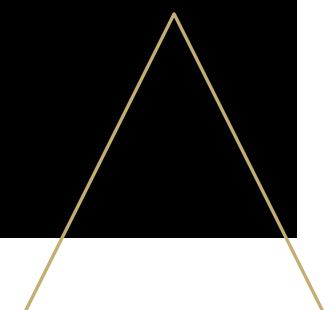
Cemile YILDIZÇAKAR

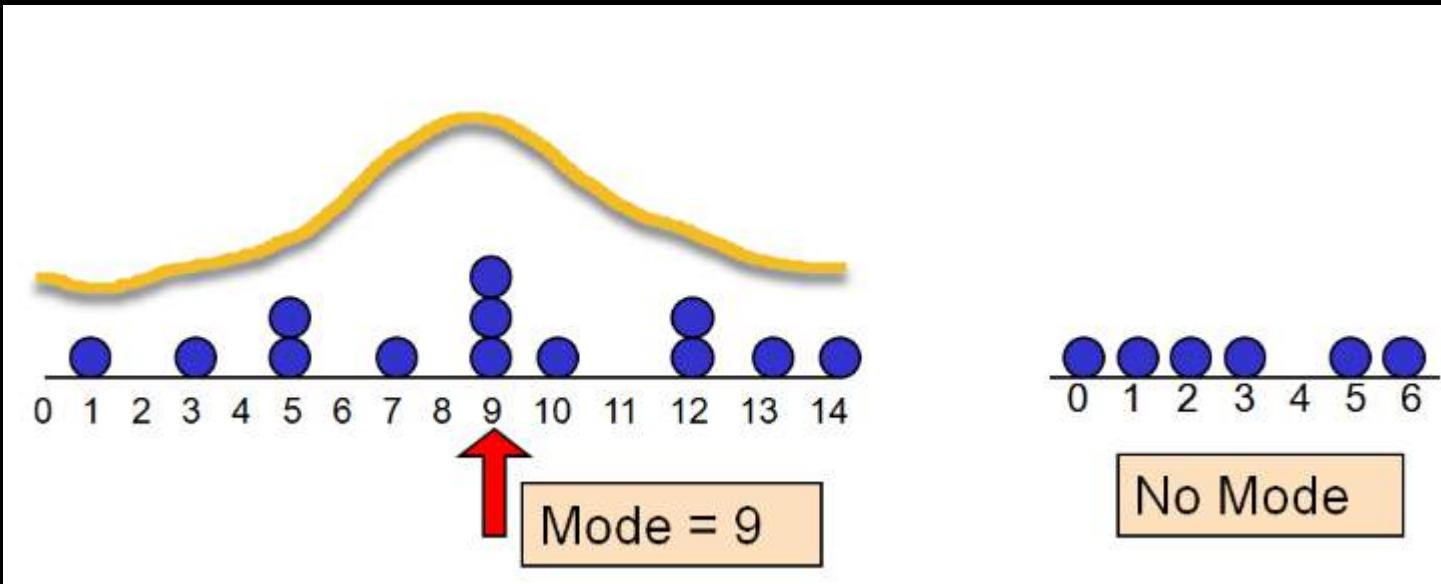


# TEPE DEĞER (MODE)

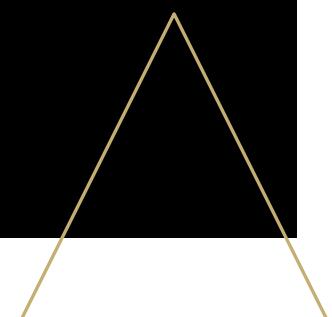
- Bir veri grubunda en çok tekrar eden değere mod denir. Yani en fazla frekansa sahip değer olarak tanımlanır.
- Hiçbir aritmetik işlem gerektirmez.
- Bazı durumlarda, en yüksek frekansa sahip değer iki veya daha fazla sayıda olabilir. Bu durumda veri setinin tek tepe değeri olmaz.
- Bir veri setinde frekanslar eşit ise tepe değeri yoktur.

Cemile YILDIZÇAKAR

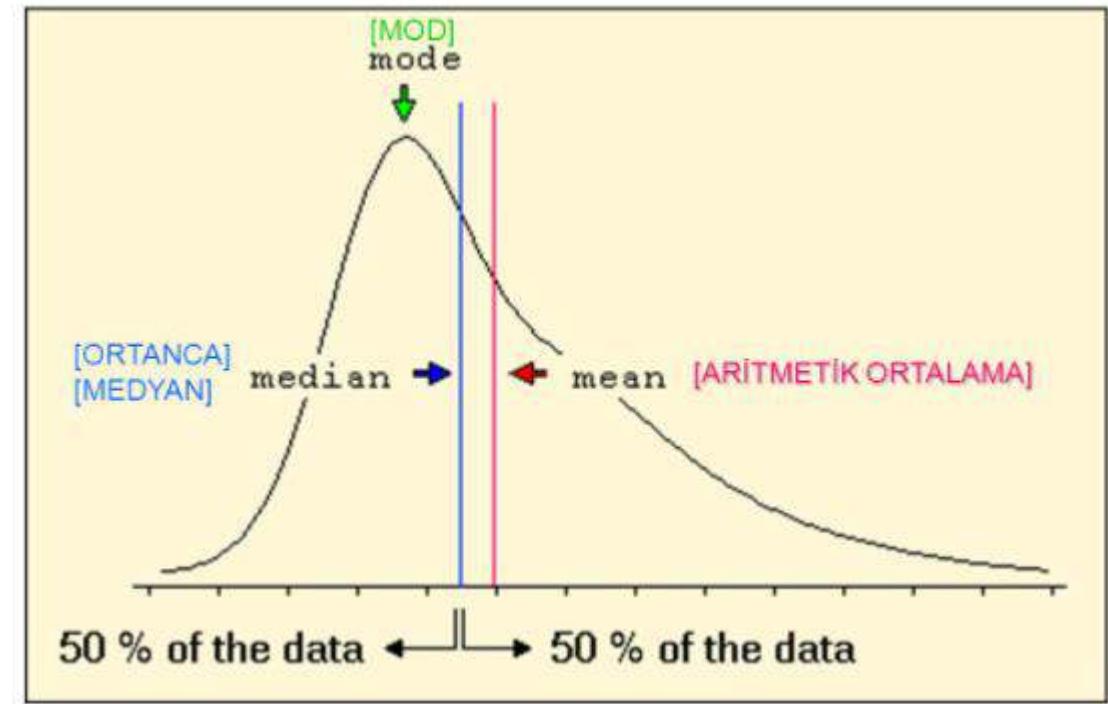
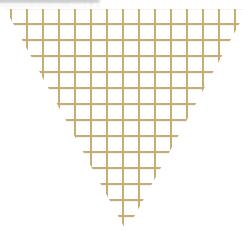
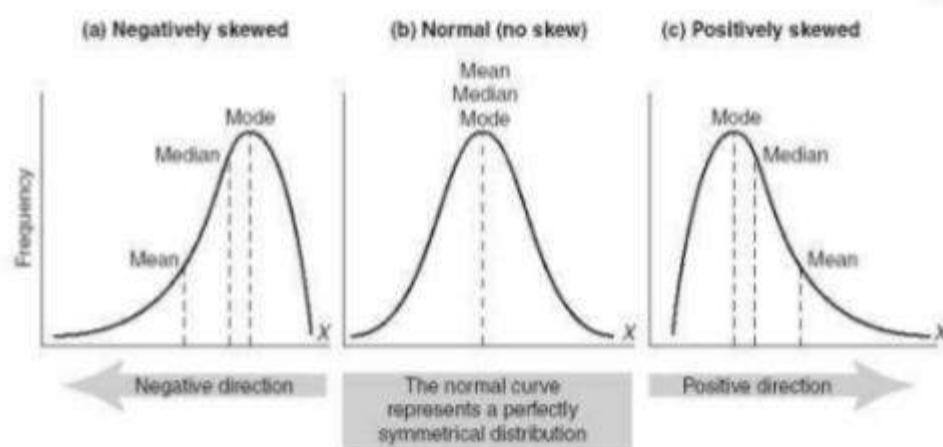




Cemile YILDIZÇAKAR



# Dağılım Şekli



Cemile YILDIZÇAKAR

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB

# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-3

CEMİLE YILDIZÇAKAR

22.12.2020

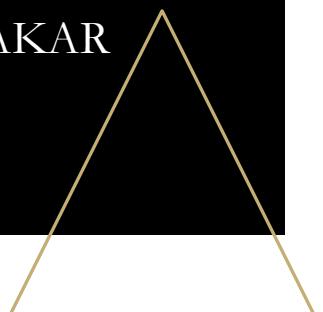


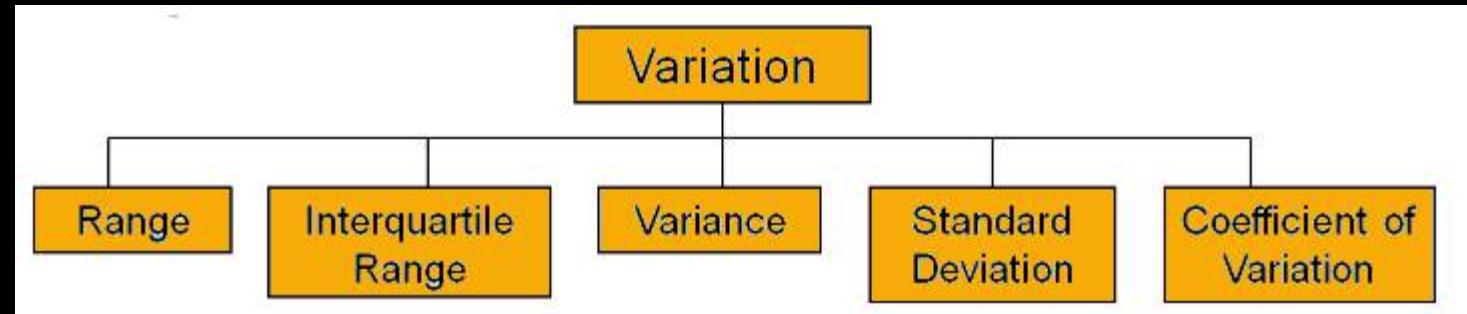
# İçindekiler

Değişim Ölçüleri ( Measures of Variability)

- Açıklık
- Çeyreklikler
- Varyans
- Standart Sapma
- Değişim Katsayısı

Cemile YILDIZÇAKAR

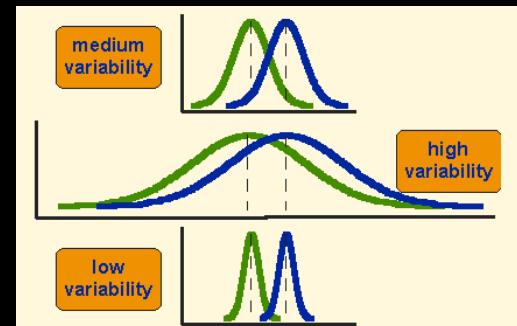
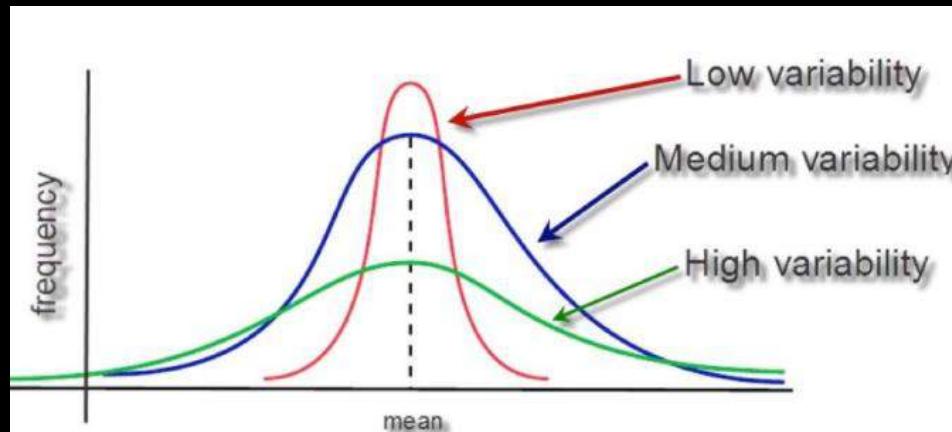




# Değişim Ölçüleri ( Measures of Variability)

- Değişim:

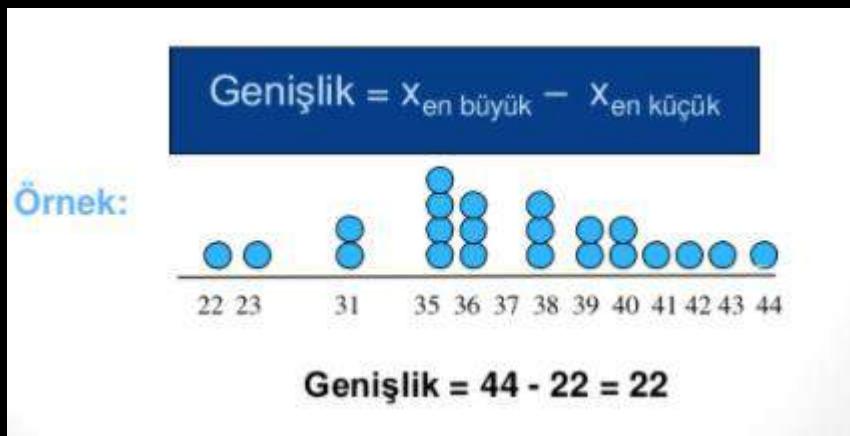
Bir dağılımda ölçümler arasında gözlenen farklılık ve değişikliğe değişim, bir serideki gözlemlerin (birimlerin) birbirinden ya da herhangi bir ortalama değerden uzaklıklarının çeşitli ölçümllerine merkezi değişim ölçülerini denir.



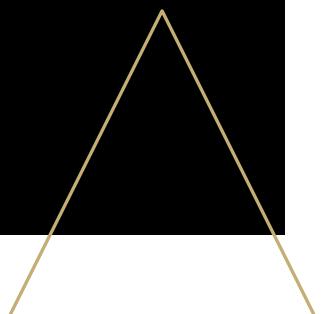
Cemile YILDIZÇAKAR

# ARALIK/AÇIKLIK (RANGE)

- En basit ölçüm ölçüsüdür.
- Ölçümlerin en büyüğü ile en küçüğü arasındaki farktır.

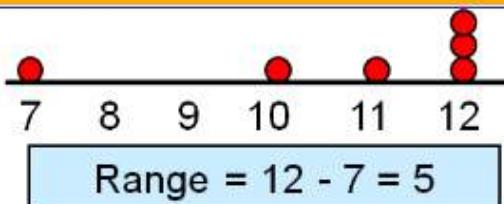
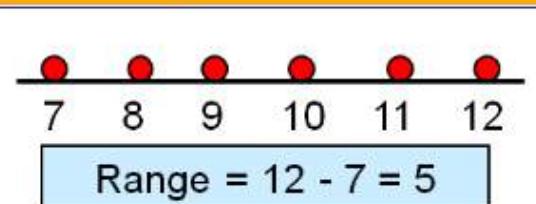


Cemile YILDIZÇAKAR



# Dezavantajları

- Dağılımın şeklini dikkate almaz.



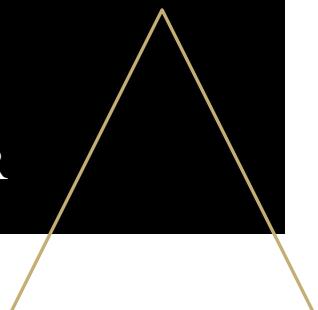
- Sapan değerlerden etkilenir.

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

$$\text{Range} = 5 - 1 = 4$$

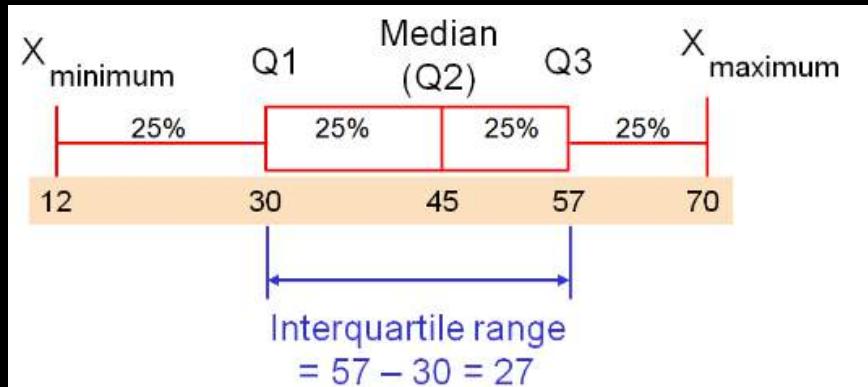
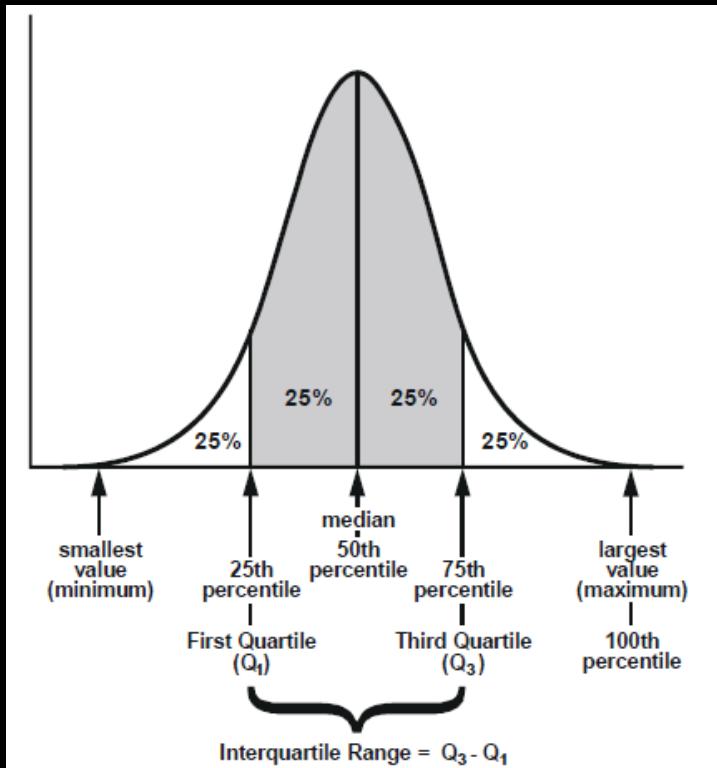
**1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,120**

$$\text{Range} = 120 - 1 = 119$$



# Çeyreklikler

- Sıralanmış veri setini ilk dört parçaya ayıran değerlerdir.



Cemile YILDIZÇAKAR

# Çeyrekliklerin konumunun bulunması



## Quartile Formula

$$\text{Lower Quartile (Q1)} = (N+1) \times 0.25$$

$$\text{Middle Quartile (Q2)} = (N+1) \times 0.50$$

$$\text{Upper Quartile (Q3)} = (N+1) \times 0.75$$

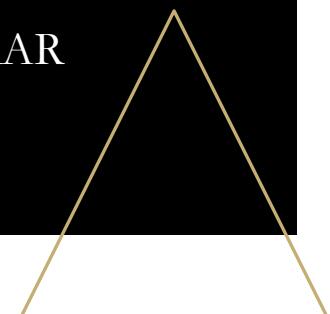
Sample Ranked Data: 11 12 13 16 16 17 18 21 22

(n = 9)

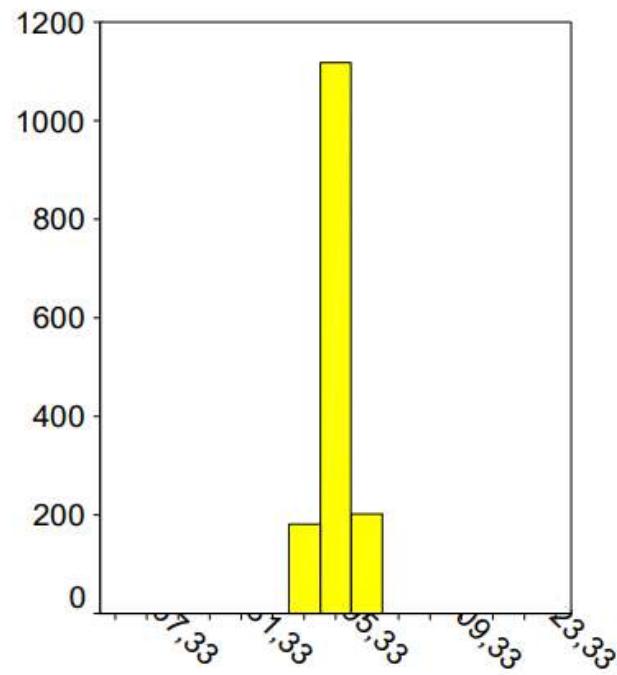
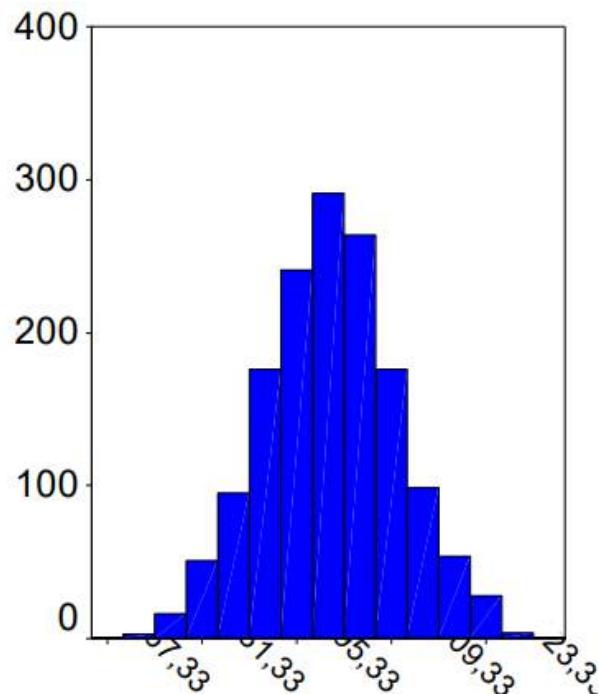
$Q_1$  = is in the  $0.25(9+1) = 2.5$  position of the ranked data  
so use the value half way between the 2<sup>nd</sup> and 3<sup>rd</sup> values,

so  $Q_1 = 12.5$

Cemile YILDIZÇAKAR



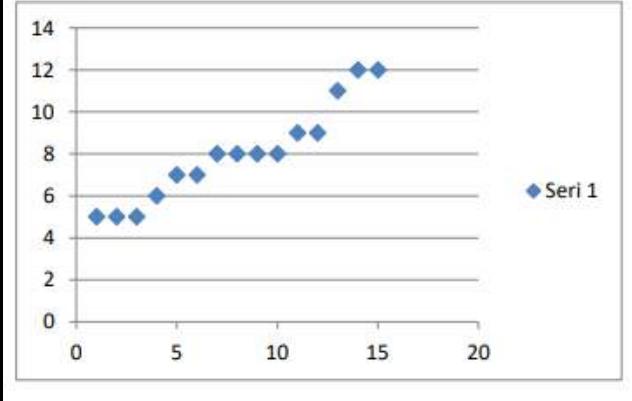
Aşağıdaki iki grafik  $n = 1500$  hacimlik alınan iki farklı örneğin doğrultusunda oluşturulan histogramlardır. Her iki örneğin ortalaması yaklaşık olarak 100 olduğuna göre iki örneğin ayrı anakütlelerden alındığı söylenebilir mi?



Cemile YILDIZÇAKAR

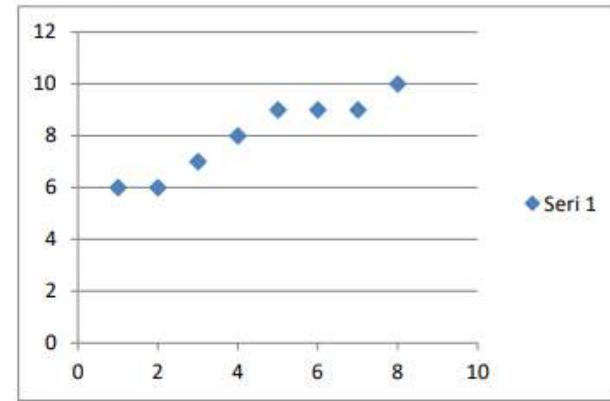
Seri1: 5 5 5 6 7 7 8 8 8 9 9 11 12 12

$$\bar{X}_1 = 8$$



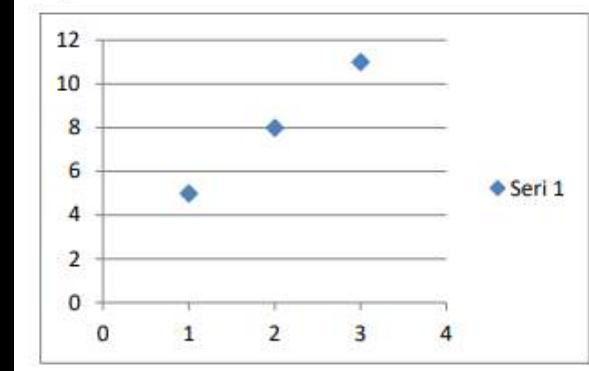
Seri2: 6 6 7 8 9 9 9 10

$$\bar{X}_2 = 8$$



Seri3: 5 8 11

$$\bar{X}_3 = 8$$



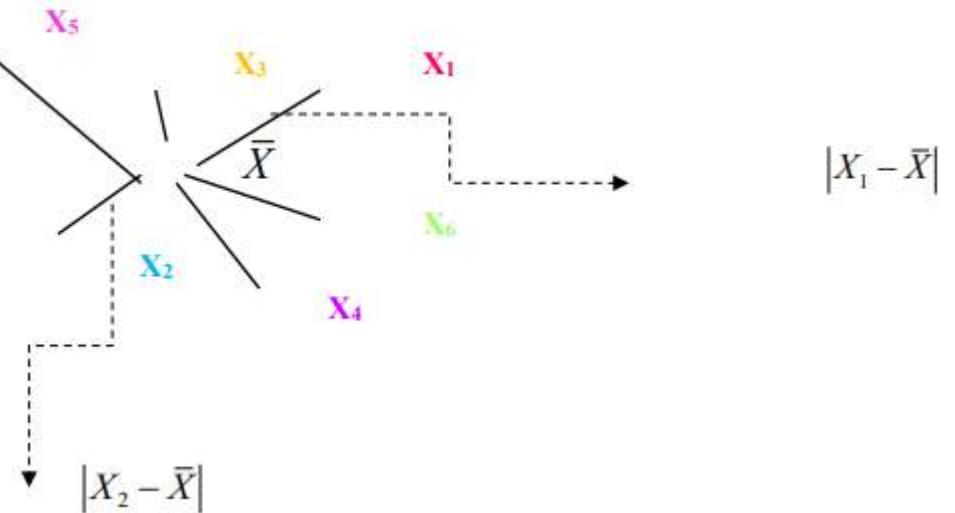
Her üçünün de ortalaması birbiriyle aynı olmasına rağmen, veri sayıları, değişimleri, dağılımları birbirinden farklıdır.

Cemile YILDIZÇAKAR

Verilerin dağılımını ölçmede,  
gözlem değerlerinin  
ortalamadan sapmalarını  $|X_i - \bar{X}|$   
kullanabiliriz. Bu sapmalar yani  
ne kadar büyükse,  $X_i$  gözlemi  
ortalamadan o denli uzakta  
demektir

$$\sum(X_i - \bar{X}) = 0$$

Cemile YILDIZÇAKAR



# VARIANCE /STANDART DEVIATION

- **VARYANS /STANDART SAPMA:**

Kitle setindeki tüm değerleri hesaba katan değişkenlik ölçüsüdür.

Standart sapma varyansın kareköküne eşittir.

Standart sapma ortalama etrafındaki değişimini gösterir.

Standart sapmanın birimi veri setindeki değerlerin birimi ile aynıdır.

$\sigma^2$ : Kitle varyansı

$s^2$  : Örneklem varyansı

Cemile YILDIZÇAKAR



**Basit seriler için:**

Populasyon Varyansı:

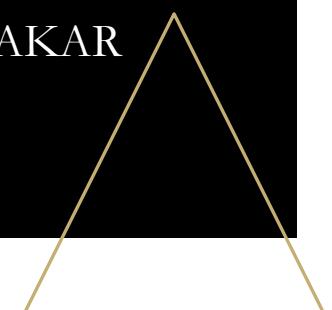
$\mu$  : Populasyon Ortalaması    N : Populasyon Hacmi

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Örnek Varyansı :

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Cemile YILDIZÇAKAR





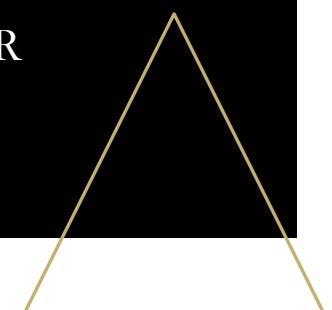
**Gruplanmış Seriler İçin:**

$$s^2 = \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{\sum_{i=1}^k f_i - 1}$$

**Sınıflanmış Seriler İçin :**

$$s^2 = \frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{\sum_{i=1}^k f_i - 1}$$

Cemile YILDIZÇAKAR



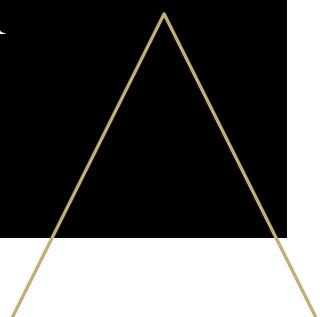
Basit seriler İçin:

Populasyon Standart Sapması:  $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

$\mu$  : Populasyon Standart Sapması  $N$  : Populasyon Hacmi

Örnek Standart Sapması :  $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

Cemile YILDIZÇAKAR





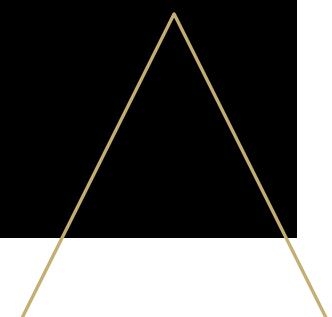
Gruplanmış Seriler İçin:

Sınıflanmış Seriler İçin :

$$s = \sqrt{\frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{\sum_{i=1}^k f_i - 1}}$$

$$s = \sqrt{\frac{\sum_{i=1}^k f_i (m_i - \bar{x})^2}{\sum_{i=1}^k f_i - 1}}$$

Cemile YILDIZÇAKAR



**Örnek:** İstatistik I dersini alan 10 öğrencinin vize notları aşağıdaki gibi sıralanmıştır. Buna göre vize notları için varyans ve standart sapmayı hesaplayınız.

30,41,53,61,68,79,82,88,90,98

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{30 + 41 + \dots + 98}{10} = 69$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(30-69)^2 + (41-69)^2 + \dots + (98-69)^2}{9}$$
$$= \frac{4538}{9} \approx 504,22$$

$$s^2 \approx 504,22 \rightarrow s = \sqrt{s^2} = \sqrt{504,22} \approx 22,45$$

**İstatistik I vizesinden alınan notların ortalama etrafında yaklaşık olarak 22 puan değiştiği görülmektedir.**

Cemile YILDIZÇAKAR



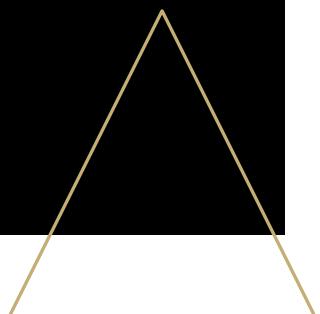
# CHEBYSHEV TEOREMİ

Herhangi bir veri setinde, verilerin ortalamanın  $K$  standart sapma uzağında bulunması oranı  $1-1/K^2$  dir. Burada  $K$ , birden büyük pozitif sayıdır.

**K=2 ve K=3 için;**

- Verilerin en az  $3/4$ ' ü (%75) ortalamanın 2 standart sapma uzağında bulunur.
- Verilerin en az  $8/9$ ' u (%89) ortalamanın 3 standart sapma uzağında bulunur.

Cemile YILDIZÇAKAR



**Örnek:** X değişkeni bir sınıfındaki İstatistik I dersinin başarı notlarını göstermek üzere, örnek ortalamasının 60 varyansının 100 olduğu bilindiğine göre, verilerin  $\frac{3}{4}$ 'ü hagi aralıkta değişir?

$$1 - \frac{1}{k^2} = \frac{3}{4} \rightarrow k = 2$$

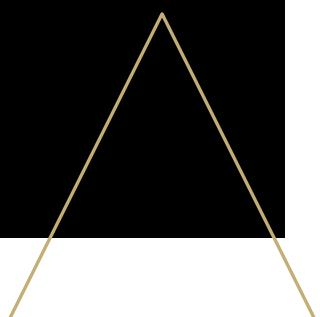
$$(\bar{x} \mp 2s)$$

$$(60 \mp 2.10)$$

$$(40, 80)$$

Kıyanak: [https://kisi.deu.edu.tr/s.ucdogruk/ist1de%c4%9fi%c5%9fkenlik\\_%c3%b6l%C3%A7a7%c3%bccleri\(Bolum3\).pdf](https://kisi.deu.edu.tr/s.ucdogruk/ist1de%c4%9fi%c5%9fkenlik_%c3%b6l%C3%A7a7%c3%bccleri(Bolum3).pdf)

Cemile YILDIZÇAKAR



Ortalama ve standart sapmanın birlikte kullanıldığı bir durum da, verilerin yüzde kaçının hangi aralıkta bulunduğu ölçmeye yarayan Chebyshev Teoremi'dir. Ortalaması  $\mu$ , standart sapması  $\sigma$  olan bir veri kümesindeki gözlemlerin  $\mu \pm k\sigma$  aralığına düşenlerin oranı

en az  $1 - \frac{1}{k^2}$  'dir. **Burada k, 1'den büyük bir sayıdır.**

**Örnek:** Bir lisedeki öğrencilerin IQ ortalaması 105, standart sapması 9'dur. Öğrencilerin en az % kaçının 85,2 ile 124,8 arasında IQ'ya sahiptir?

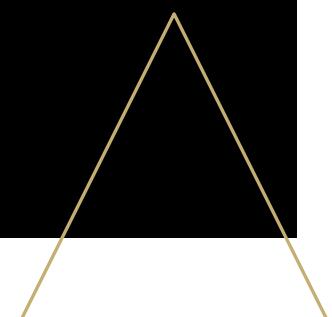
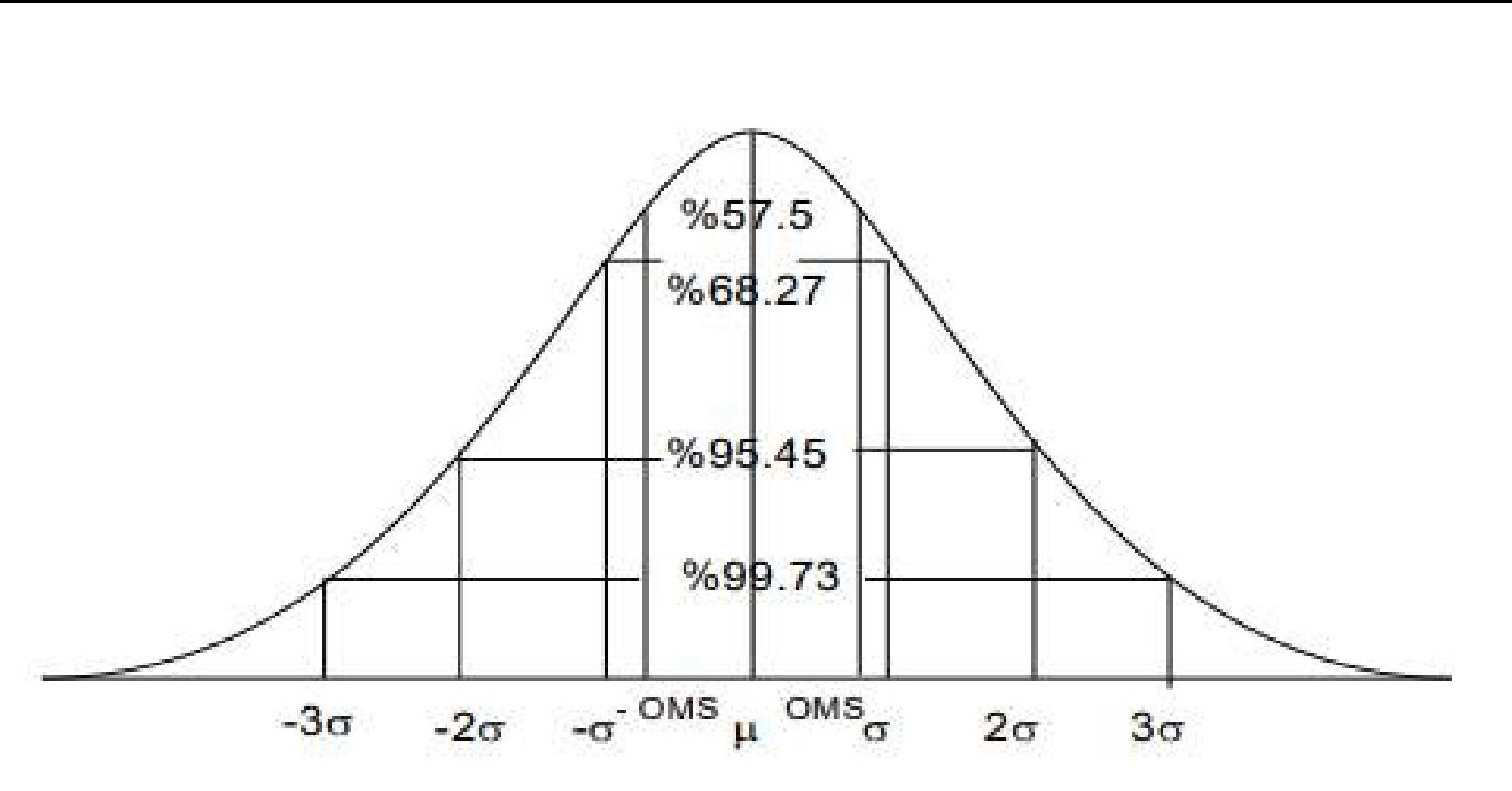
$$\mu - k\sigma = 85,2$$

$$\mu + k\sigma = 124,8$$

Buradan  $k=2,2$  elde edilir.

$$1 - \frac{1}{(2,2)^2} = 0,79$$

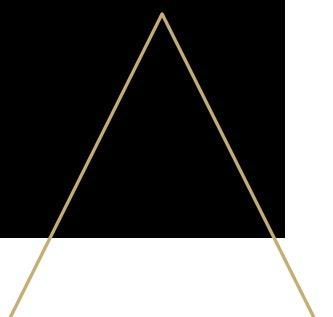
**Yorum:** Öğrencilerin en az %79'u verilen aralıkta IQ skoruna sahiptir.



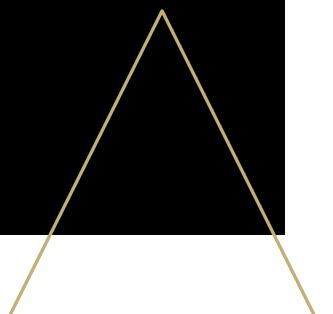
# Değişim Katsayısı (Coefficient of Variation)

- Standart sapmayı ortalamanın bir yüzdesi olarak ifade eden ve iki veya daha fazla populasyondaki varyasyonu (değişkenliği) karşılaştırmada kullanılan ölçüye varyasyon(değişkenlik) katsayısı denir.
- Yüzdelik değerler alır (%).
- Farklı birimlerde ölçülmüş veri setlerinin değişkenliğini karşılaştırmak için kullanılabilir.

$$CV = \left( \frac{s}{\bar{x}} \right) \cdot 100\%$$



DK'sı küçük olan serilerin, diğerlerine göre daha az değişken olduğu (daha homojen yani ortalamaya yakın dağılıma sahip) birimlerden oluştuğu söylenir. Çünkü standart sapma küçüldükçe DK da küçülür. Standart sapmanın küçük olması ise ancak ortalama etrafındaki saçılımın  $\bar{X}$  'ya doğru çekilmesiyle mümkünür birimler  $\bar{X}$  'dan uzaklaşıkça standart sapma da büyümektedir.



**Örnek:** İki hisse senedi olsun. Bu hisselerin 1 aylık ortalama getirileri ve standart sapmaları aşağıdaki gibi ise, siz bir portföy yöneticisi olarak **riski seven** bir yatırımcıya hangi hisse senedini tavsiye edersiniz?

$$\mu_1=70 \quad \sigma_1=6$$

$$\mu_2=25 \quad \sigma_2=3$$

$$DK_1 = \frac{6}{70} * 100 = 8,57$$

$$DK_2 = \frac{3}{25} * 100 = 12$$

DK sı büyük olanı yani 2. hisse senedini tercih etmelidir. Şayet riski sevmeyen yatırımcı olsaydı, ona değişimi görece daha az olan ilk hisse senedini almasını tavsiye ederdik. Borsada riskli senetlerin getirişi de kaybı da büyüktür. Ama riski seven yatırımcı “risk yoksa getiri de yoktur” mantığına sahip olduğu için bu onun tercihidir. Riski düşük olan senetler her ne kadar güvenli olsa da, maalesef getirişi de düşüktür. Elbette bu bir tercih meselesidir.

- Kaynak: <http://auzefkitap.istanbul.edu.tr/kitap/kok/istatistikiu155.pdf>

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

The background of the slide features a photograph of a tropical landscape with several tall palm trees in the foreground and middle ground, and a range of mountains or hills under a clear blue sky. A large black rectangular overlay covers the right side of the slide, containing the text.

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB



# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-4

CEMİLE YILDIZÇAKAR

29.12.2020

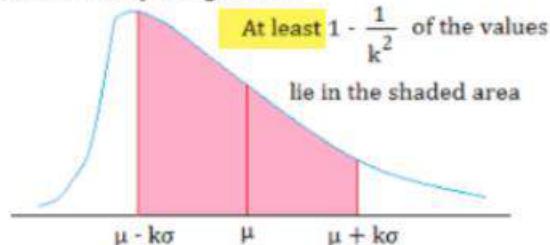


## Chebychev Teoremi

Herhangi bir veri setinde, verilerin ortalamadan  $K$  standart sapma uzakta bulunma oranı  $1 - 1/K^2$  dir. Burada  $K$ , birden büyük pozitif sayıdır.

**K=2 ve K=3 için;**

- Verilerin **en az**  $3/4$ 'ü (%75) ortalamanın 2 standart sapma uzağında bulunur.
- Verilerin **en az**  $8/9$ 'u (%89) ortalamanın 3 standart sapma uzağında bulunur.

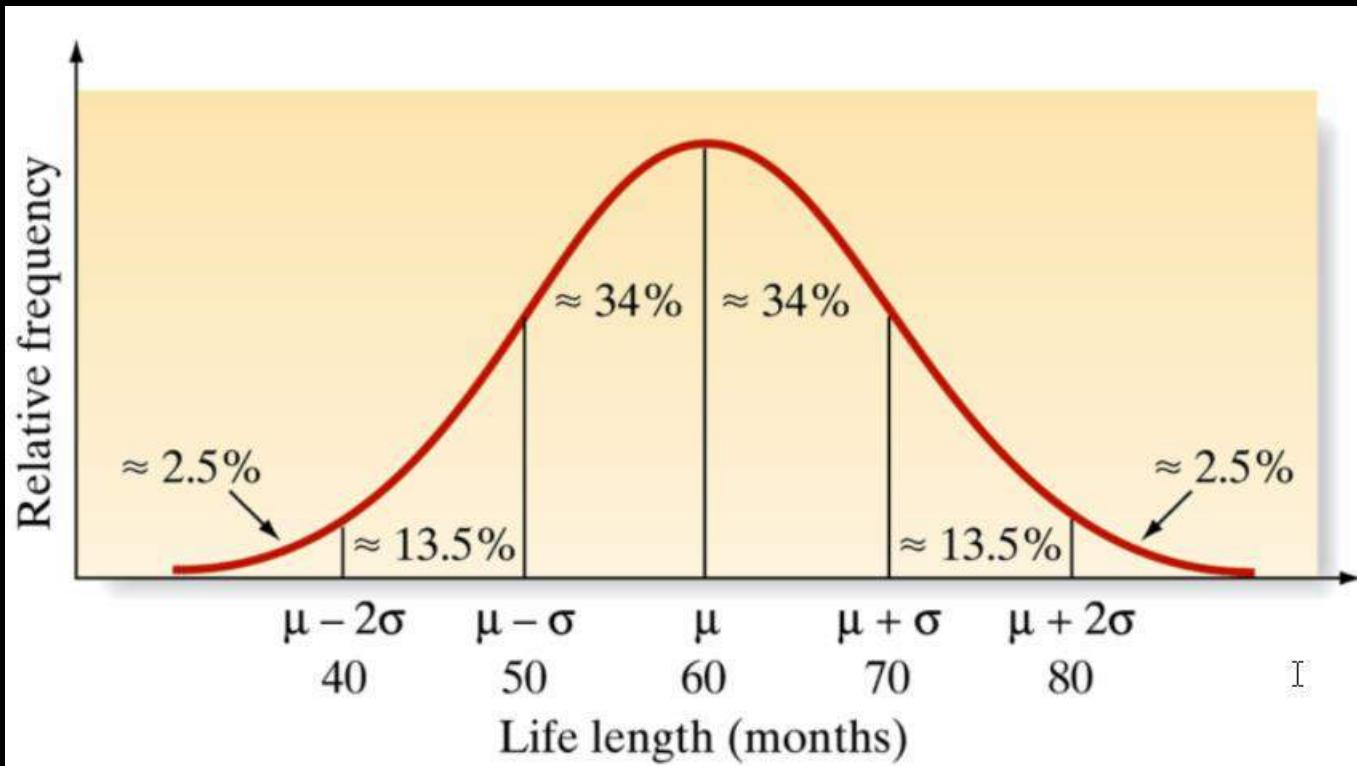


$$\mu \pm \sigma$$



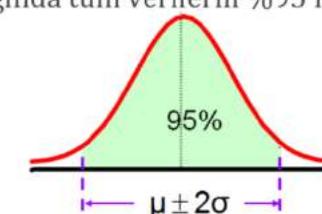
$$100[1 - (1/k^2)]\%$$

Cemile YILDIZÇAKAR

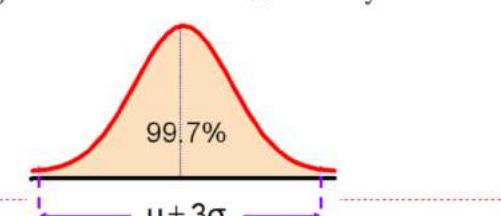


### Deneysel kural (The Empirical Rule)

□  $\mu \pm 2\sigma$  aralığında tüm verilerin %95'i yer alır.

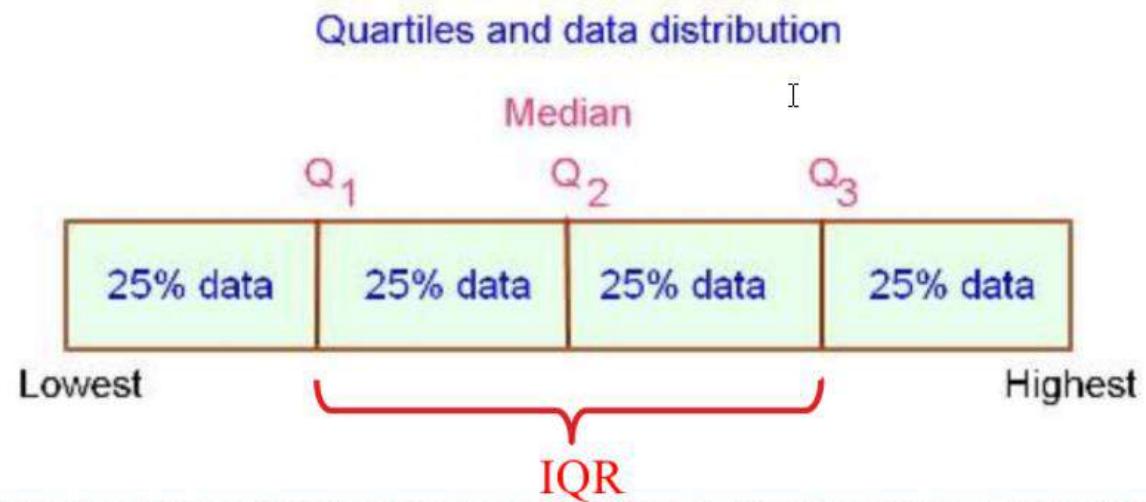


□  $\mu \pm 3\sigma$  aralığında tüm verilerin %99.7'si yer alır.



Cemile YILDIZÇAKAR

A **quartile** divides a sorted data set into 4 equal parts, so that each part represents  $\frac{1}{4}$  of the data set

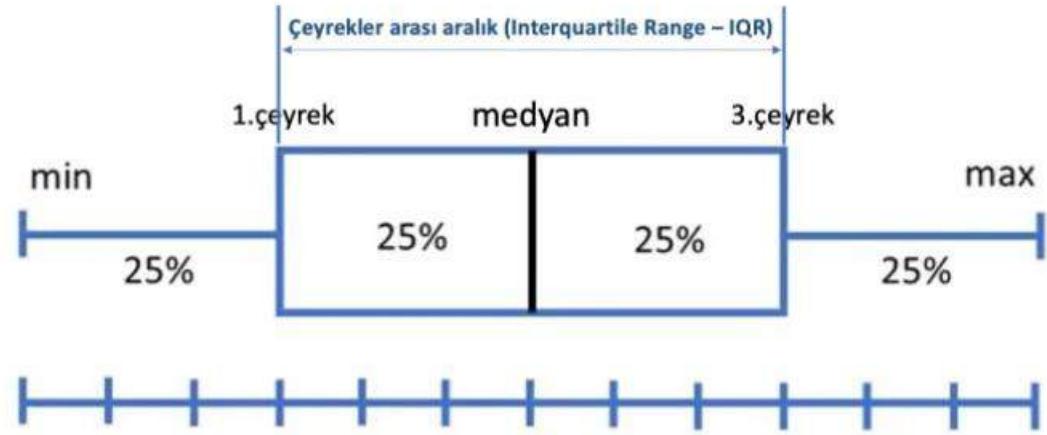


Cemile YILDIZÇAKAR

# Kutu grafigi (Box Plot)

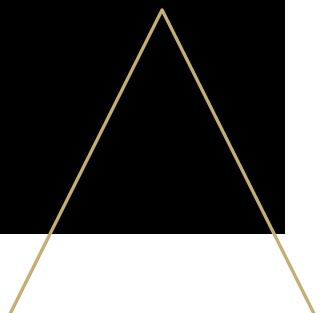
Bir kutu grafigi (Boxplot), veri çeyreklerini (veya yüzdelikleri) ve ortalamaları görüntüleyerek sayısal verilerin ve değişkenliğin görsel olarak dağılımını göstermek için kullanılır. Veri analizinde sıkılıkla kullanılan bir grafik türüdür.

Aşağıdaki görüntü, mükemmel bir normal dağılım olan verileri temsil eder ve çoğu kutu grafiginin bu simetriye uymaz.

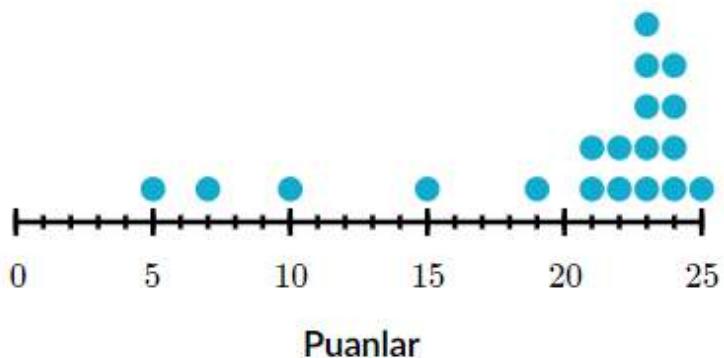


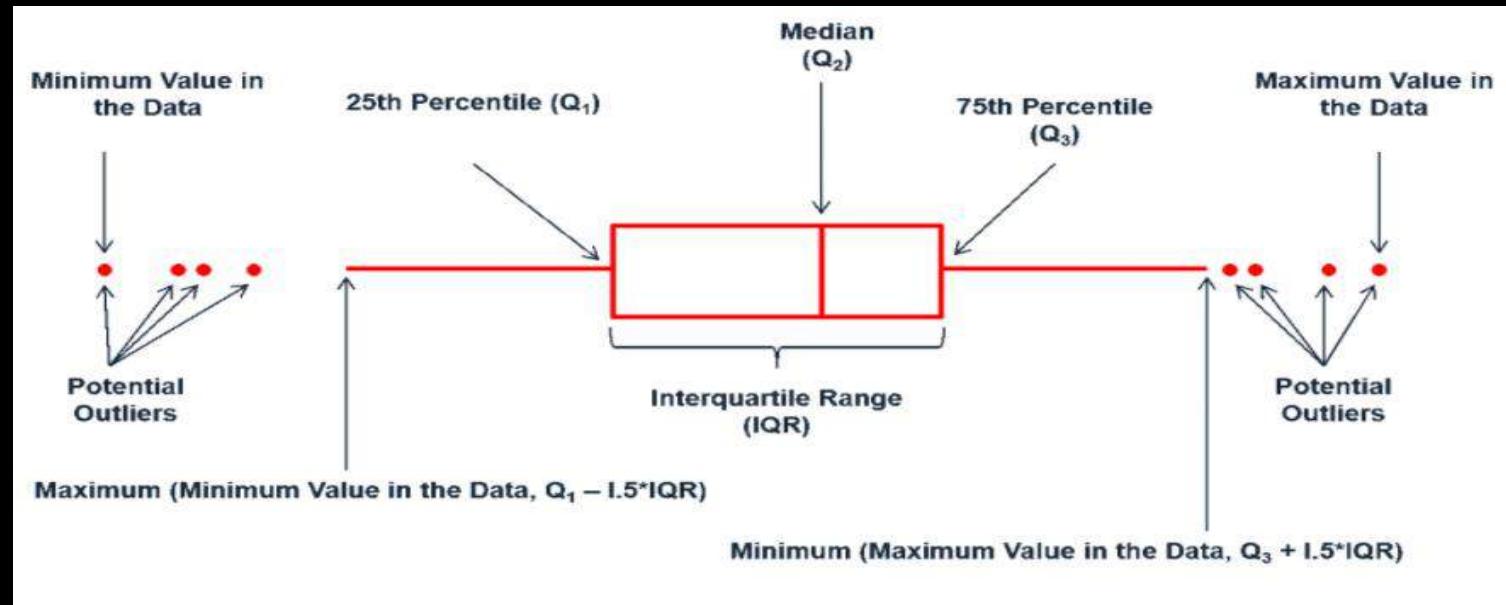
# KUTU GRAFIĞİ (BOXPLOT) NEREDE VE NASIL KULLANILIR?

- Kutu grafiği, verilerdeki değerlerin nasıl yayıldığınn iyi bir göstergesidir. Kutu grafikleri bir histograma göre ilkel gibi görünse de, birçok grup veya veri kümesi arasındaki dağılımları karşılaştırırken yararlıdır.
- Kutu ne kadar uzun olursa veri o kadar dağılmış olur. Kutu ne kadar küçük olursa veri o kadar az dağılmış olur.

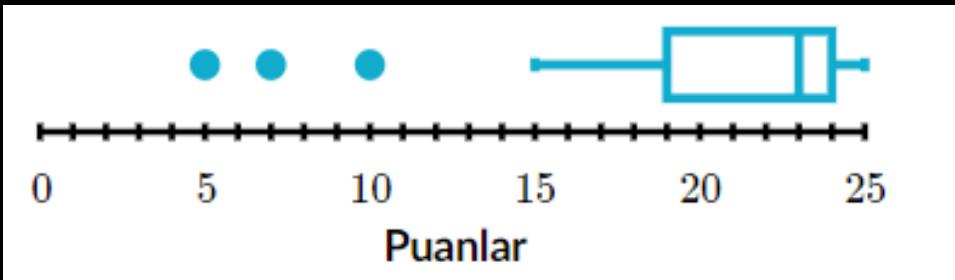
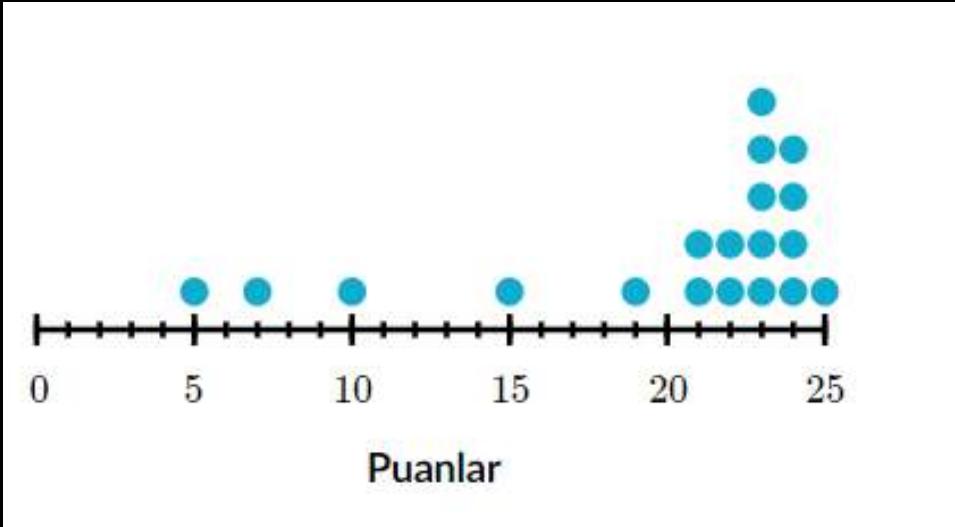


Aşağıdaki dağılım, 19 aday için bir sürücü sınavının puanlarını gösterir. Kaç aykırı değer görünsünüz?





Yayın kullanılan bir kural, eğer bir veri noktası  $1,5 \cdot \text{ÇA}$  üçüncü çeyreğin üstündeyse veya birinci çeyreğin altındaysa, bu veri noktasının bir aykırı değer olduğunu söyler. Farklı şekilde söylesek, düşük aykırı değerler  $Q_1 - 1,5 \cdot \text{ÇA}$ 'nın altındadır ve yüksek aykırı değerler  $Q_3 + 1,5 \cdot \text{ÇA}$ 'nın üstündedir.



## **1- Simetrik – Normal :**

Bir sürecin, normal dağılıma uygun olması verilerin ortalamaya etrafında homojen dağıldığını gösterir.

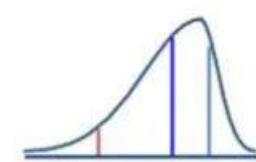
## **2-Eğik – Pozitif :**

Verileriniz alt limite yakın. Bu verilerin, üretilen bir ürünün bir ölçüsü olduğunu kabul edelim. 100 cm nominal değer beklenisi olan bir ölçü var. 0,1 cm de toleransınız olsun. Yani 99 – 101 cm aralığında ürettiğiniz her ürün kabul edilecektir. Ancak yapılan ölçümler sonucu toplanan veriler çoğunlukla 99 cm ve ona yakın değerler (99,4; 99,5; v.b.) çıkıyorsa üretim prosesiniz alt limite yakın üretim yapıyor demektir. İlerleyen dönemde

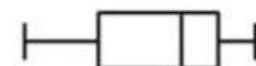
## **3-Eğik – Negatif :**

Verilerinizin üst limite yakın olması demektir. prosesinizin kontrol dışına çıkıp belirtilen alt limite değerin

Eğik - Negatif

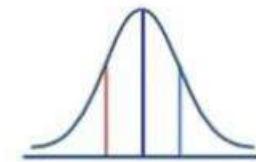


$Q_1$   $Q_2$   $Q_3$

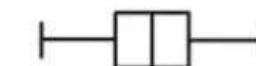


(a) Negatively skewed

Simetrik - Normal

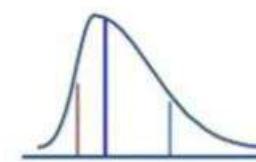


$Q_1$   $Q_2$   $Q_3$

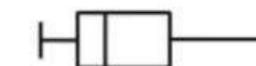


(b) Normal (no skew)

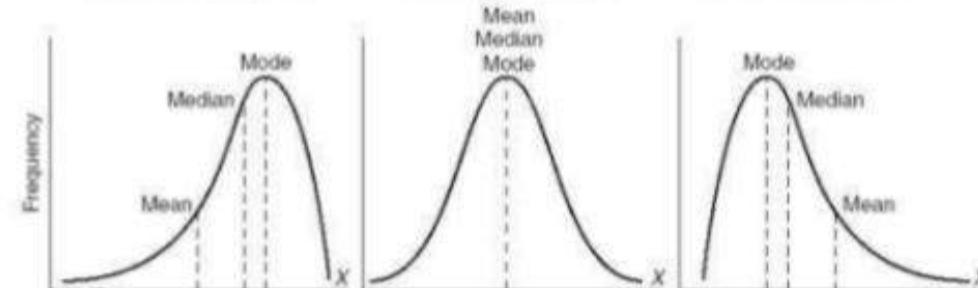
Eğik - Pozitif



$Q_1$   $Q_2$   $Q_3$



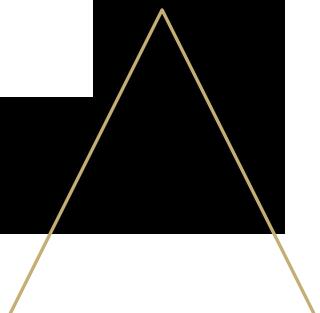
(c) Positively skewed



# Standartlaştırma (z-skoru)

- Standartlaştırma → Her bir değişken değerinden, ortalamanın farkının alınması ve elde edilen farkın standart sapmaya bölünmesidir.
- Böylece ham veriler standart verilere dönüştürülerek, ölçü birimi farklılığı ortadan kaldırılmış olur.

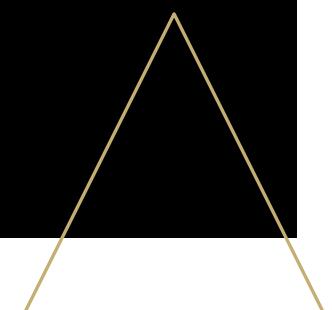
$$z = \frac{x_i - \mu}{\sigma}$$





# Z-skoru

- $z$ -skoru veri setindeki gözlemlerin ortalamaya olan uzaklıklarını gösteren bir ölçüdür.
- $z$ -skor *positif* ya da *negatif* olabilir.
- Bir diğer ifade ile; istatistikte, bir gözlemin  $z$ -skoru (veya standart skor), popülasyon ortalamasının üstünde veya altında standart sapma sayısıdır.
- Yine bir başka ifade ile,  $Z$  skoru yardımıyla elinizde bulunan örnek kümedeki sayısal verilerin, ortalamanın ne kadar altında ya da üstünde olduğunu görebilirsiniz.
- Bir  $z$ -skoru hesaplamak için popülasyon (hesaplanamıyor ise örneklem) ortalamasını ve popülasyon standart sapmasını bilmelisiniz:



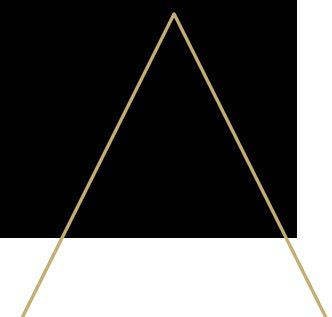
- Birbirinden farklı ölçü birimlerinin karşılaştırılmasında kullanılır.
- Z-score bütün veri yığınlarındaki birimlerin, ortak bir birim aralığına yığılmasını sağlar

$$z = \frac{x_i - \mu}{\sigma}$$

Formülde yer alan  $x(i)$  bizim gözlem değerimizdir.



Aşağıdakilere benzer soruları cevaplamak için bir z-puan görselleştirme oluşturabilirsiniz:

- Değerlerin yüzde kaçı belirli bir değerin altına düşüyor?
  - Hangi değerler olağanüstü sayılabilir? Örneğin, bir IQ testinde hangi puanlar ilk yüzde beşi temsil ediyor?
  - Bir dağıtımın diğerine karşı göreceli toplamı nedir? Örneğin, Michael ortalama bir erkektenden daha uzun ve Emily ortalama bir kadından daha uzun, ancak cinsiyetleri arasında nispeten daha uzun kim?
- 

## Örnek

- Ayşe analiz sınavından 80, istatistik sınavından ise 68 almıştır.
- Analiz sınavında sınıf ortalaması 83, standart sapma ise 10'dur.
- İstatistik sınavında sınıf ortalaması 62, standart sapma ise 6'dır.
- Buna göre Ayşe hangi sınavda daha yüksek performans göstermiştir?

İpucu:

- Negatif bir z skoru, incelenen veri ortalamadan az demektir.
- Pozitif bir z skoru, incelenen veri ortalamadan çok demektir.

## Çözüm

$$\square Z_A = (80-83)/10 = -0.3$$

$$\square Z_I = (68-62)/6 = 1$$

- Ayşe sınıf arkadaşlarına göre istatistik sınavında daha başarılı olmuştur.

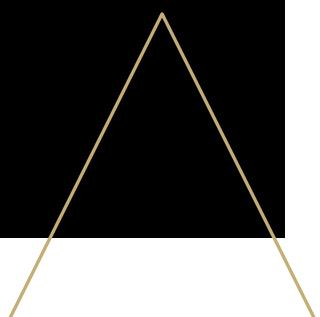
Örneğin ; Bir matematik sınav sonuçlarının olduğu veri setimiz olduğunu düşünelim. Bu sınav sonucunda ortalamanın ( $\mu$ ) 60 olduğu ve standart sapmanın ( $\sigma$ ) ise 10 olduğu tespit edilmiştir. Eğer 49 ve altında puan alan oranını bulmak istersek standardizasyon işlemi sonrası z-puan tablosunu kullanabiliriz.

$$Z = \frac{49 - 60}{10} = -1.1$$

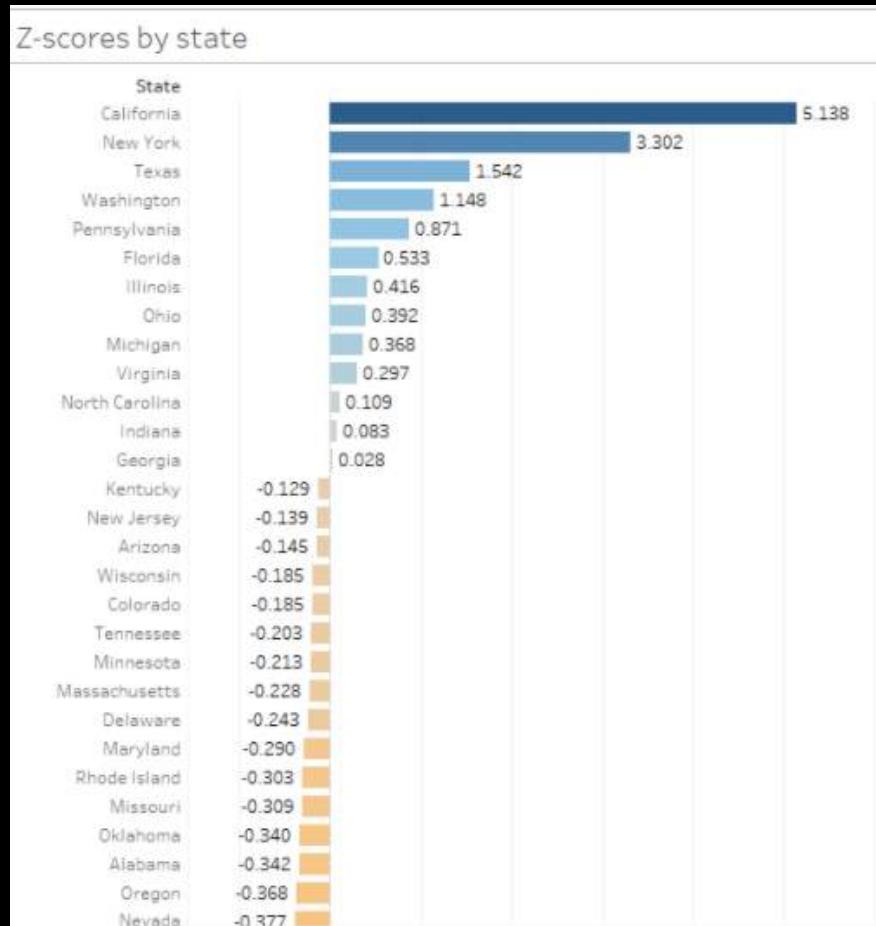


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379

z puanımız standardizasyon işlemi sonucunda -1.1 olarak bulundu. z -puan tablosuna bakıldığında toplam popülasyonun %13.57'sinin 49 ve daha altında puan aldığı tespit edilmiştir.



Genel bir kural olarak,  $-1.96$ ’dan düşük veya  $1.96$ ’dan daha yüksek olan z-skorları alışılmadık ve ilginç olarak değerlendirilir. Yani, bunlar istatistiksel açıdan belirgin aykırı değerlerdir.



California ve New York'un ikisinin de z skorları  $1,96$ 'dan büyüktür.

Buradan Kaliforniya ve New York'un

diğer eyaletlere kıyasla çok daha yüksek ortalama satışlar elde ettiğine karar verebilirsiniz.

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB

# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-5

CEMİLE YILDIZÇAKAR

15.01.2021



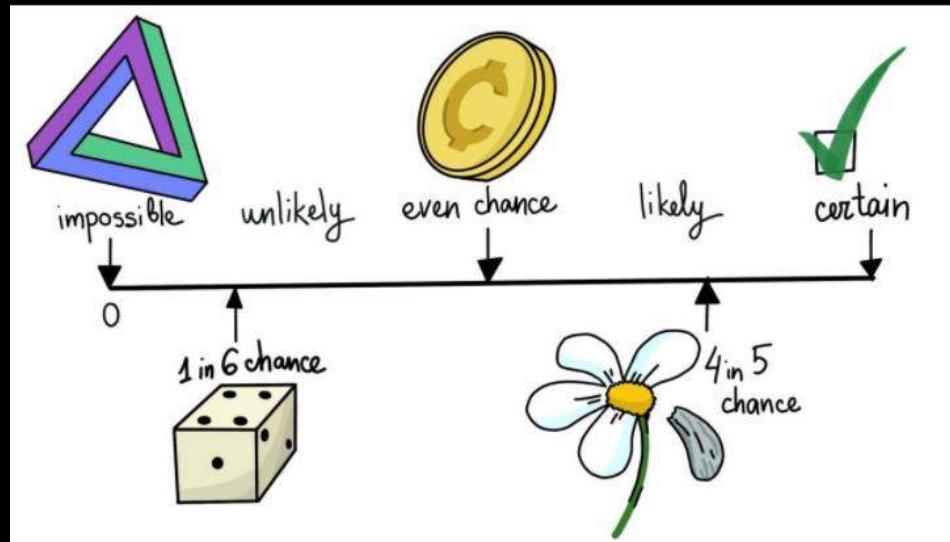


## İçindekiler

- Deney
- Örneklem Birimi
- Örneklem Uzayı
- Olay
- Olasılık Hesaplama
- İmkansız Olay, Kesin Olay
- Birleşim ve Kesişim
- Tümleyen Olaylar
- Bütüne Tamamlayan Olaylar
- Karşılıklı Ayrık Olaylar

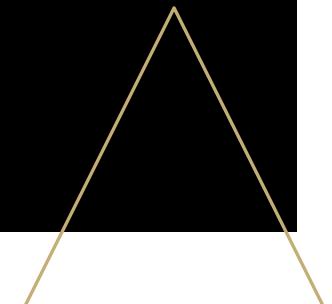
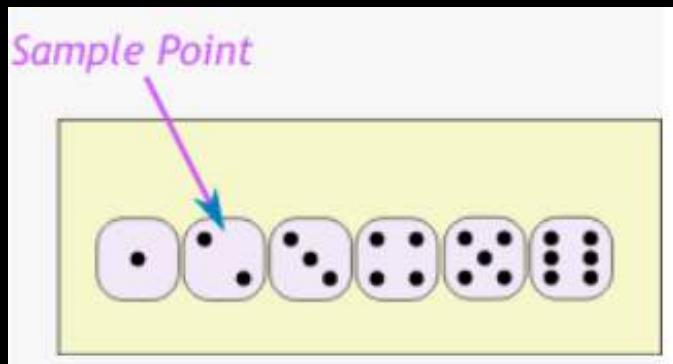
# Deney (Experiment)

- Deney, sonucu kesin olarak bilinmeyen ve tekrarlanabilen olaylara ilişkin gözlem yapma ya da veri toplama süreci olarak tanımlanabilir.
- Deney tekrarlanan denemelerden oluşur.



# Örneklem Birimi ( Sample Point)

- Bir deneyin çıktılarından herbirine örneklem birimi denir.



# Örneklem Uzayı (Sample Space)

- Bir deneyin olası tüm çıktılarına örneklem uzayı denir.
- $S = \{H,T\}$  ,  $S = \{1,2,3,4,5,6\}$

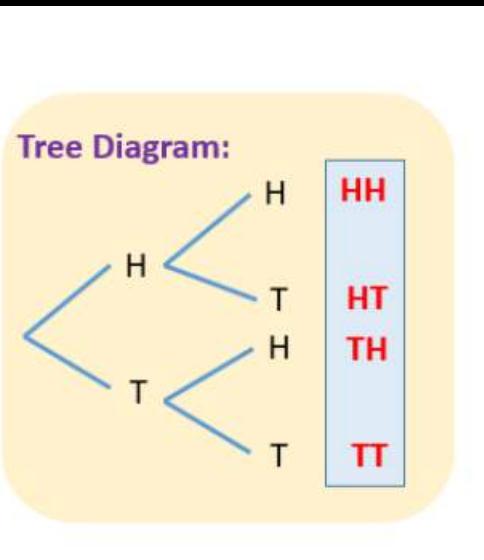
(H – Head, T – Tail)

List:

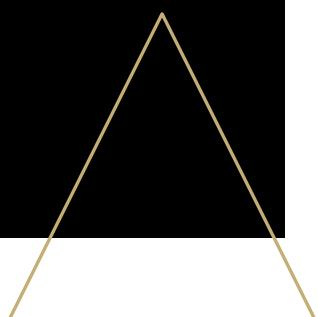
HH	HT	TH	TT
----	----	----	----

Table:

	H	T
H	HH	HT
T	TH	TT

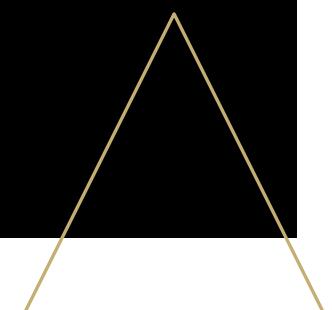
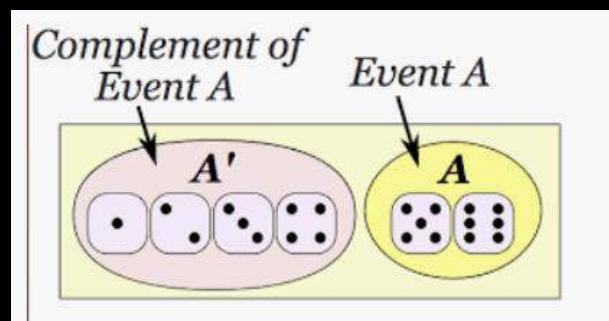


(a,b)	1	2	3	4	5	6
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

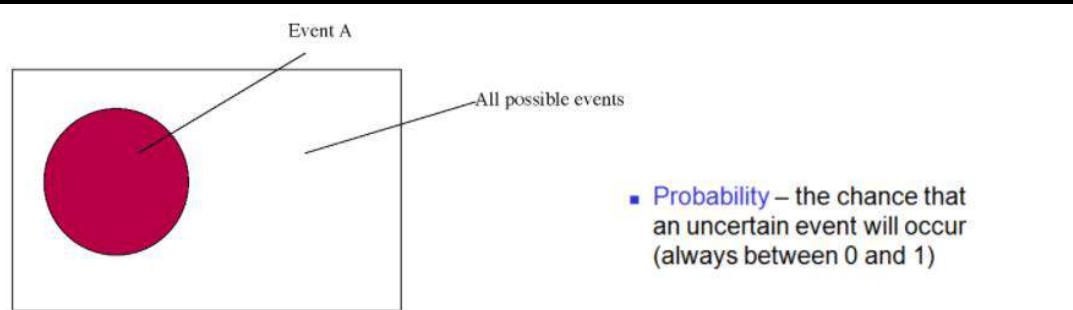


# Olay (Event)

- Bir deneyde belirli özelliğe sahip sonuçların oluşturduğu kümeye olay denir.
- Bir örneklem uzayının alt kümelerine olay denir.



# Probability of an Event



- **Probability** – the chance that an uncertain event will occur (always between 0 and 1)

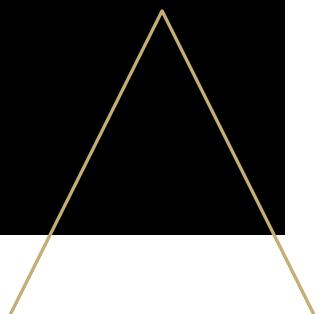
$$0 \leq P(A) \leq 1 \text{ For any event } A$$

$$P(\text{Event}) = \frac{\text{the number of ways it can happen}}{\text{the number of possible outcomes}}$$

## Probability of an Event

The probability of an event  $A$  is calculated by summing the probabilities of the sample points in the sample space for  $A$ .

**Example 6** Refer to Example 4 and 5. If  $A=\{H\}$ ,  $P(A)=p_1=0.5$ . If  $A=\{H, T\}$ ,  $P(A) = p_1 + p_2 = 1$ .



## Probability Rules for Sample Points

Let  $p_i$  represent the probability of sample point  $i$ . Then

1. All sample point probabilities *must* lie between 0 and 1 (i.e.  $0 \leq p_i \leq 1$ ).
2. The probabilities of all the sample points within a sample space *must* sum to 1 (i.e.,  
 $\sum p_i = 1$ )

# Olasılık Hesaplama Adımları

## Steps for Calculating Probabilities of Events

1. Define the experiment; that is, describe the process used to make an observation and the type of observation that will be recorded.
2. List the sample points.
3. Assign probabilities to the sample points.
4. Determine the collection of sample points contained in the event of interest.
5. Sum the sample point probabilities to get the probability of the event.



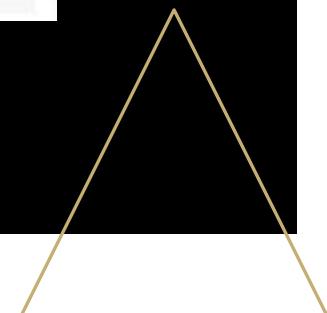
Bir zar havaya atılıyor üst yüzeye gelen sayının çift olma olasılığını inceleyiniz.

**Deneysel:** Zarın havaya atılması.

**Olay:** Zarın çift gelmesi olayıdır.

**Örnek Uzayı:** 1,2,3,4,5,6

**Cıktı:** 2,4,6



## Örnek

- A ve B oyuncuları tarafından oynanan tenis maçında A'nın kazanma şansı B'nin 2 katıdır.
- A ve B'nin 2 maç yaptığını düşünelim.
- A'nın en az 1 maç kazanması olasılığını hesaplayınız.

## Çözüm

□  $SS=\{AA, AB, BA, BB\}$  1

□  $P(A)=2/3$  2

□  $P(B)=1/3$  3

<u>Örnek uzay</u>	<u>Olasılık</u>	4
AA	4/9	
AB	2/9	
BA	2/9	
BB	1/9	

□  $P(\text{en az bir } A)=P(AA)+P(AB)+P(BA)=8/9$  5

# Impossible event , Certain event

**An impossible event**

Roll a 7 on one die



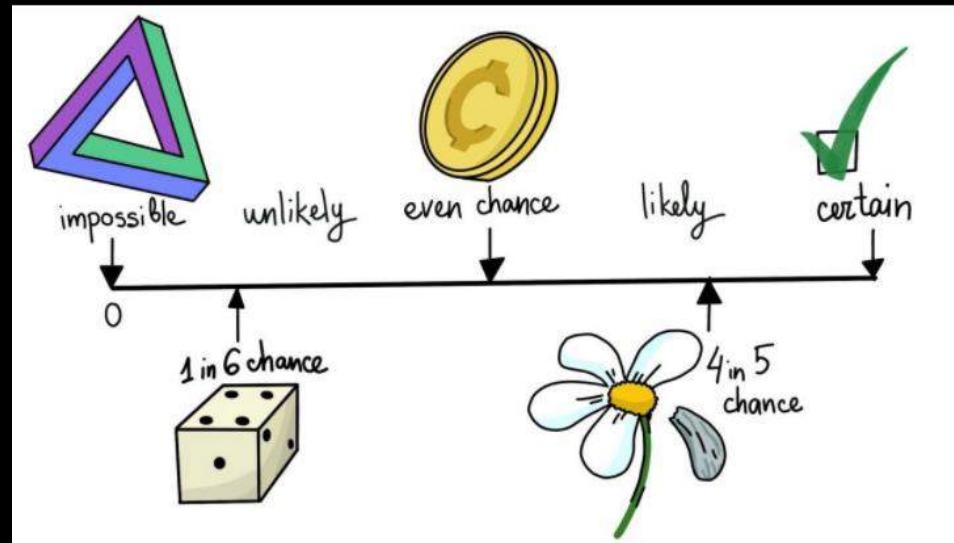
$P(E) = 0$

**An certain event**

Flip a head or tail

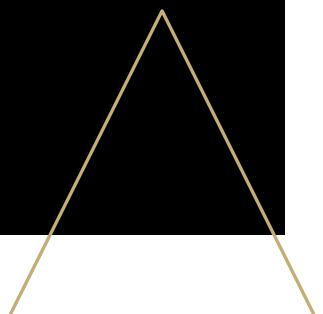


$P(E) = 1$



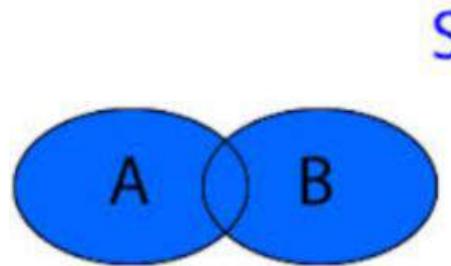
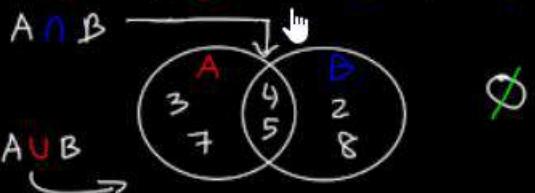
# Birleşim ve Kesişim(Union and Intersection)

- A ve B olaylarının birleşim kümesi,  
A, B ya da her ikisine ait olan elemanların oluşturduğu kümedir.
- -----
- A ve B olaylarının kesişim kümesi,  
A, B olaylarının ortak elemanlarının oluşturduğu kümedir.



## Union & Intersection

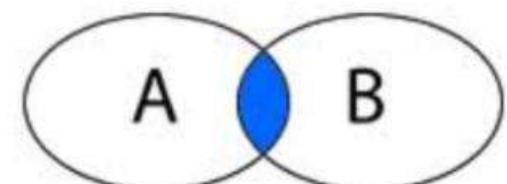
$$A = \{3, 4, 5, 7\} \quad B = \{2, 4, 5, 8\}$$



$A \cup B$

A or B

S



$A \cap B$

A and B

S

A ve B , aynı S örneklem uzayında tanımlanmış iki olay olmak üzere;  
-A ve B olaylarının birleşimi  $A \cup B$  olarak gösterilir.  $A \cup B$  olayının sonuçları ya A ya B ya da her ikisinden birinden ortaya çıkar.  
-A ve B olaylarının kesişimi  $A \cap B$  olarak gösterilir.  $A \cap B$  olayının sonuçları hem A hem de B olayından ortaya çıkar.

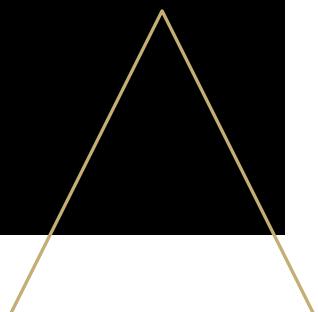
- **Ayrık Olay:** Eğer A ve B gibi iki olay aynı anda gerçekleşemiyor ise bu olaylara **ayrık**(birbirini engelleyen) olaylar denir

**Örnek:** Madeni para atılması sonucunda yazı veya tura gelmesi ayrık olaylardır.

Zarın atılması sonucu 1 ve tek sayı gelmesi olayları ayrık olaylar değildirler. Çünkü aynı anda gerçekleşebilirler.

- **Eşit Olasılıklı Olaylar:** Bir örnek uzayındaki tüm basit olayların ortaya çıkma olasılığı eşit ise bu olaylara eşit olasılıklı olaylar denir.

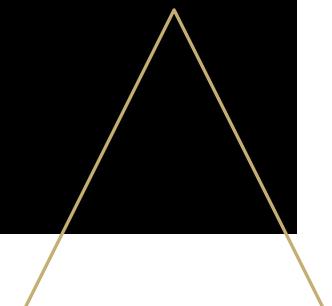
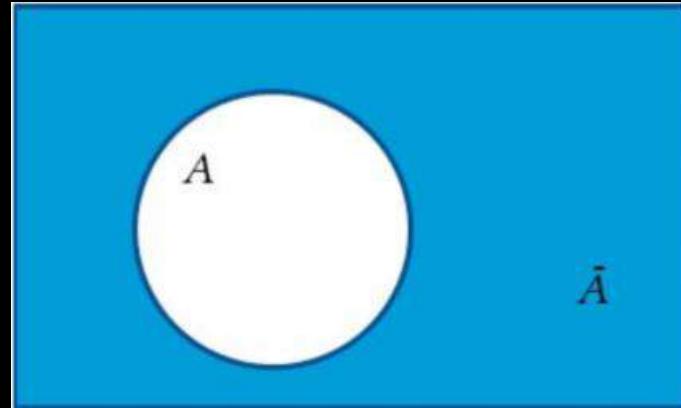
**Örnek:** Bir deste iskambil kağıdından bir adet kağıt çekilmesi.



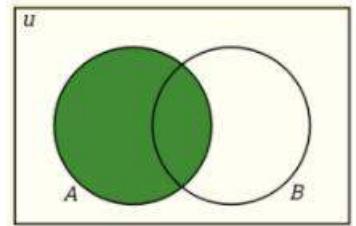
# Tümleyen Olaylar (Complementary Events)

- A kümesine ait olmayan, örnek uzaya ait olan elemanların oluşturduğu kümeye A'nın tümleyeni denir.

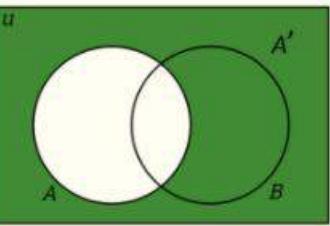
$$P(A) + P(A^c) = 1.$$



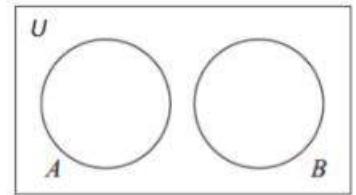
### Set Operations and Venn Diagrams



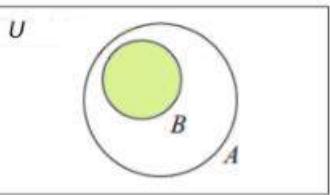
Set A



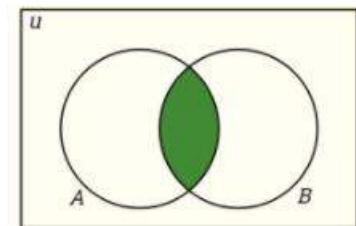
$A'$  the complement of A



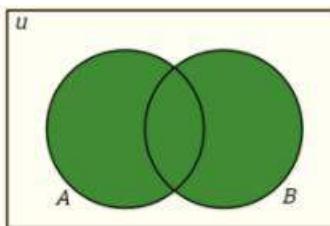
A and B are disjoint sets



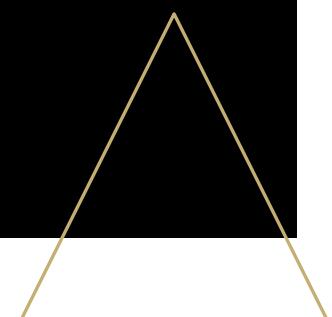
B is proper  
subset of A  
 $B \subset A$



Both A and B  
A intersect B  
 $A \cap B$



Either A or B  
A union B  
 $A \cup B$

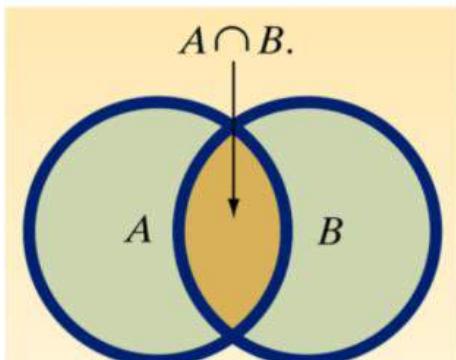


# Olasılıkta Toplama Kuralı

## Additive Rule of Probability

The probability of the union of events  $A$  and  $B$  is the sum of the probability of event  $A$  and the probability of event  $B$ , minus the probability of the intersection of events  $A$  and  $B$ ; that is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Entire shaded area is  
 $A \cup B$ .

# Bütüne Tamamlayan Olaylar (Collectively Exhaustive Events)

- E<sub>1</sub>, E<sub>2</sub>, ... Ek olaylarının kesişimleri boş küme, birleşimleri örnek uzaya eşit ise bu olaylar bütüne tamamlayan olaylardır. ( $E_1 \cup E_2 \cup \dots \cup E_k = S$ )

Mutually exclusive and exhaustive system of events : Let S be the sample space associated with a random experiment. Let A<sub>1</sub>, A<sub>2</sub> ..... A<sub>n</sub> be subsets of S such that

(i)  $A_i \cap A_j = \emptyset$  for  $i \neq j$  and

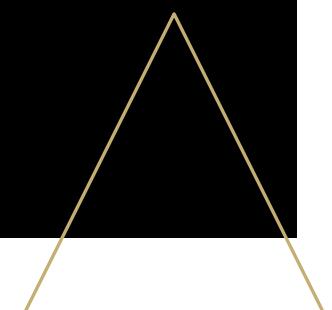
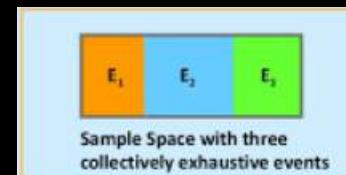
(ii)  $A_1 \cup A_2 \cup \dots \cup A_n = S$

Then the collection of events A<sub>1</sub>, A<sub>2</sub>, ..... , A<sub>n</sub> is said to form a mutually exclusive and exhaustive system of events.

In this system,

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

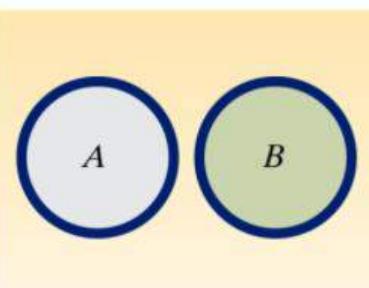
$$E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n = \bigcup_{i=1}^n E_i = S$$



# Karşılıklı Ayrık Olaylar (Mutually Exclusive Events)

Events  $A$  and  $B$  are **mutually exclusive** if  $A \cap B$  contains no sample points—that is, if  $A$  and  $B$  have no sample points in common. For mutually exclusive events,

$$P(A \cap B) = 0$$



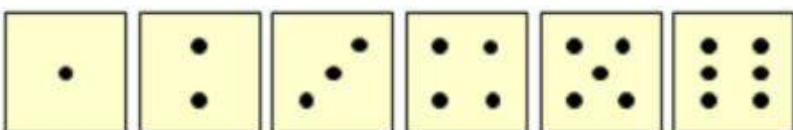
$S$

## Probability of Union of Two Mutually Exclusive Events

If two events  $A$  and  $B$  are *mutually exclusive*, the probability of the union of  $A$  and  $B$  equals the sum of the probability of  $A$  and the probability of  $B$ ; that is,  
 $P(A \cup B) = P(A) + P(B)$ .

## Örnek

Let the **Sample Space** be the collection of all possible outcomes of rolling one die:



$$S = [1, 2, 3, 4, 5, 6]$$

Let **A** be the event “Number rolled is even”

Let **B** be the event “Number rolled is at least 4”

Then

$$A = [2, 4, 6] \quad \text{and} \quad B = [4, 5, 6]$$

$S = [1, 2, 3, 4, 5, 6]$

$A = [2, 4, 6]$

$B = [4, 5, 6]$

Complements:

$$\bar{A} = [1, 3, 5]$$

$$\bar{B} = [1, 2, 3]$$

Intersections:

$$A \cap B = [4, 6]$$

$$\bar{A} \cap B = [5]$$

Unions:

$$A \cup B = [2, 4, 5, 6]$$

$$A \cup \bar{A} = [1, 2, 3, 4, 5, 6] = S$$

$$S = [1, 2, 3, 4, 5, 6]$$

$$A = [2, 4, 6]$$

$$B = [4, 5, 6]$$

- Mutually exclusive:

- A and B are **not** mutually exclusive
    - The outcomes 4 and 6 are common to both

- Collectively exhaustive:

- A and B are **not** collectively exhaustive
    - $A \cup B$  does not contain 1 or 3

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

The background of the slide features a photograph of a tropical landscape with several tall palm trees in the foreground and middle ground, and dark, silhouetted hills or mountains in the distance under a clear blue sky. A large black rectangular overlay covers the right side of the slide, containing the quote.

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB

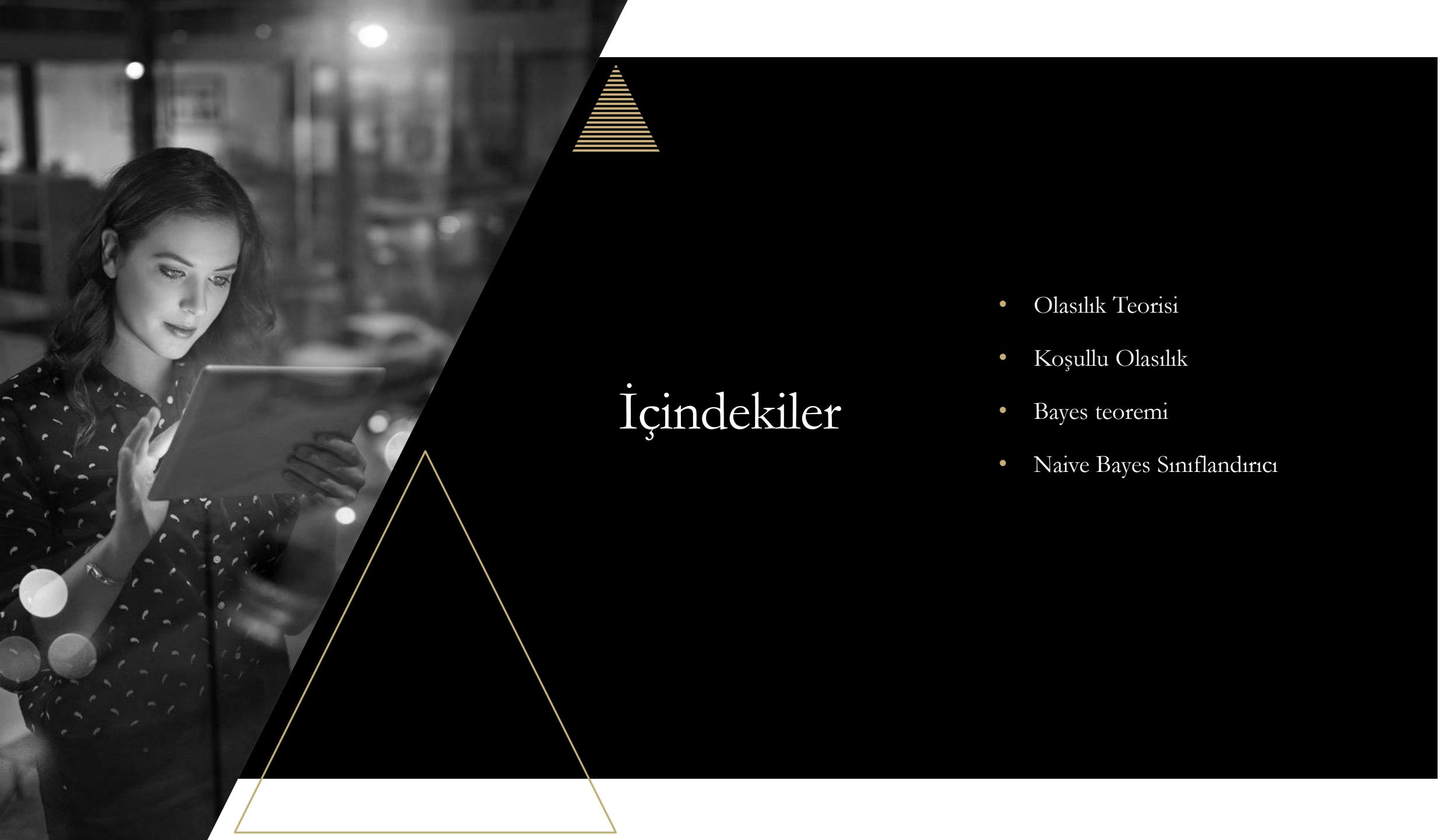
# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-6

CEMİLE YILDIZÇAKAR

19.01.2021





# İçindekiler

- Olasılık Teorisi
- Koşullu Olasılık
- Bayes teoremi
- Naive Bayes Sınıflandırıcı

# Bütüne Tamamlayan Olaylar (Collectively Exhaustive Events)

- E<sub>1</sub>, E<sub>2</sub>, ... Ek olaylarının kesişimleri boş küme, birleşimleri örnek uzaya eşit ise bu olaylar bütüne tamamlayan olaylardır. ( $E_1 \cup E_2 \cup \dots \cup E_k = S$ )

Mutually exclusive and exhaustive system of events : Let S be the sample space associated with a random experiment. Let A<sub>1</sub>, A<sub>2</sub> ..... A<sub>n</sub> be subsets of S such that

(i)  $A_i \cap A_j = \emptyset$  for  $i \neq j$  and

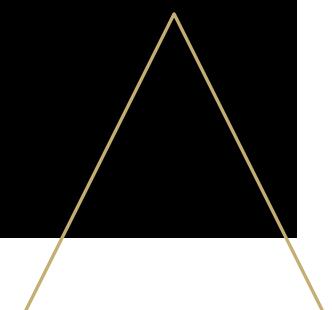
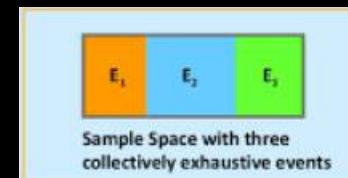
(ii)  $A_1 \cup A_2 \cup \dots \cup A_n = S$

Then the collection of events A<sub>1</sub>, A<sub>2</sub>, ..... , A<sub>n</sub> is said to form a mutually exclusive and exhaustive system of events.

In this system,

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

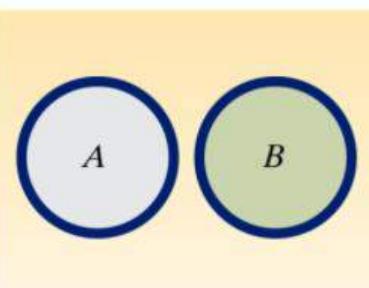
$$E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n = \bigcup_{i=1}^n E_i = S$$



# Karşılıklı Ayrık Olaylar (Mutually Exclusive Events)

Events  $A$  and  $B$  are **mutually exclusive** if  $A \cap B$  contains no sample points—that is, if  $A$  and  $B$  have no sample points in common. For mutually exclusive events,

$$P(A \cap B) = 0$$



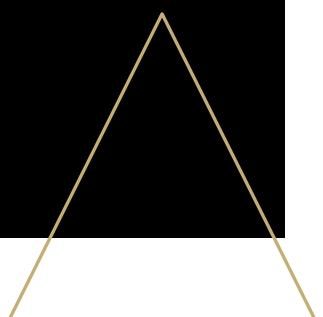
## Probability of Union of Two Mutually Exclusive Events

If two events  $A$  and  $B$  are *mutually exclusive*, the probability of the union of  $A$  and  $B$  equals the sum of the probability of  $A$  and the probability of  $B$ ; that is,  
 $P(A \cup B) = P(A) + P(B)$ .

# Olasılık Teorisi

- Klasik (A Priori) Olasılık
- Frekans (A Posteriori) Olasılığı
- Aksiyom Olasılığı.

Olasılığın tarihsel gelişim aşamalarıdır.



## KLASİK OLASILIK

- Eğer bir örnek uzayı  $n(S)$  adet ayrık ve eşit olasılıkla ortaya çıkan basit olaylardan oluşuyor ve örnek uzayındaki basit olaylardan  $n(A)$  adedi A olayının özelliğine sahip ise A'nın olasılığı:

$$P(A) = n(A) / n(S)$$

kesri ile elde edilir

## KLASİK OLASILIK NEDEN YETERSİZDİR?

- Örnek uzayının eleman sayısı sonsuz olduğu durumlarda,
- Eşit olasılıklı olay varsayıımı yapılamadığı durumlarda ,
  - *Tümdengelim çıkarımları yapılamadığında* klasik olasılık ile hesaplama yapılamayacağından dolayı yetersizdir.

## FREKANS OLASILIĞI

- Araştırılan anakütle üzerinde  $n$  adet deney uygulanır. Yapılan bu deneylerde ilgilenilen A olayı  $n(A)$  defa gözlenmiş ise A olayının göreli frekansı (yaklaşık olasılığı):

$$P(A) = n(A) / n$$

olarak bulunur.

## ÖRNEK:

Bir fabrikanın üretmiş olduğu televizyonların hatalı olma olasılığı  $p$  nedir?

Once örnek uzayı oluşturular:

$$S=\{\text{sağlam,hatalı}\}$$

Klasik olasılığa göre (eşit olasılıklı olaylar)  $p=0.5$  olup gerceği yansındığı şüphelidir.

Yapılması gereken örneklem alarak  
 $p = n(H) / n$

olasılığını hesaplamaktır.



## FREKANS OLASILIĞIN KARARLILIK ÖZELLİĞİ

- Gerçekleştirilen deney sayısı arttıkça  $P(A)$  olasılık değerindeki değişkenlik azalacak ve giderek bir sabit değere yaklaşacaktır. Bu duruma **kararlılık özelliği** adı verilir.
- Bir olayın olasılığı deneyin tekrarlama sayısı sonsuza yaklaşırken o olayın görelî frekansının alacağı limit değer olarak tanımlanır:

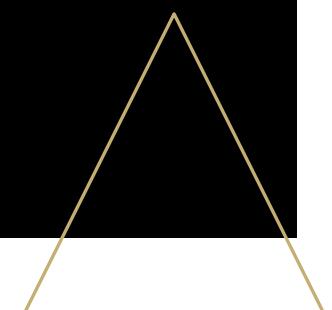
$$p = P(A) = \lim_{n \rightarrow \infty} n(A) / n$$

## FREKANS OLASILIĞI NİCİN YETERSİZDİR

- Olasılığın kararlılık değerine ulaşığı deneme sayısı kaçtır?
- Sonsuz adet deneme yapmak mümkün değildir.
- Aynı deney iki defa aynı tekrar sayısı ile gerçekleştirildiğinde elde edilen olasılıklardan hangisi olayın olasılığı olarak kabul görektir?

# Aksiyom Olasılığı Nedir?

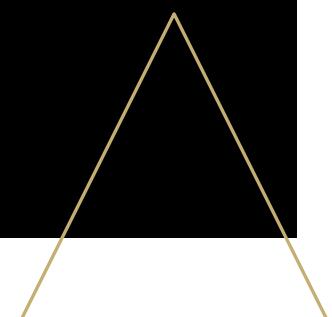
- Olasılığın matematiksel teorisini tanımlar.
- Bu teorinin oluşturduğu ideal modeller yaşadığımız dünyanın problemlerini çözmeye kullanılır.
- Klasik ve frekans Olasılık teorisi için ortak nokta; her ikisinin de, benzer koşullarda (teorikte aynı koşullarda) uygulanan deneylere gereksinim duyar. Buna karşın tekrarlı olarak uygulanamayan durumlarda olasılıkların hesaplanmasında AKSIYON OLASILIĞI yardımcı olur.





## BENZER KOŞULLARDA TEKRARLI OLARAK UYGULANAMAYAN DURUMLARA ÖRNEK;

- Çok hoşlandığınız bir kişi ile çıkma olasılığı nedir?
- Karşıyaka – Göztepe maçının 6-0 bitme olasılığı nedir?
- 3. Dünya savaşının çıkma olasılığı nedir?



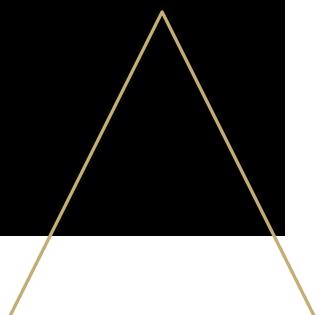
# KOLMOGOROV AKSIYOMLARI

- BİRİNCİ AKSIYOM:

Bir olayın olasılığı bir negatif-olmayan reel sayıdır ve bu sayı şöyle ifade edilir:

$$P(E) \geq 0 \quad \forall E \subseteq F$$

Burada  $F$  olay uzayıdır.



## İKİNCİ AKSIYOM

Bu **birim-ölçüsü** varsayımdır: Örnekleme uzayının tümünü kapsayan bir basit olay ortaya çıkması için olasılık 1dir. Daha belirli bir şekilde ifadeyle; Örneklem uzayını taşıan hiçbir basit olay mümkün değildir:  $P(\Omega) = 1$ .

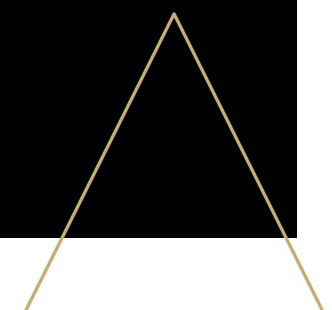
Bu aksiyom bazı hatalı olasılık hesaplamalarında çok kere temel bir hatanın ortayamasına neden olmuştur. Eğer tüm örneklem uzayı kesinlikle tanımlanamıyorsa bunun herhangi bir alt setinin tanımlanması da imkânsızdır.

## ÜÇÜNCÜ AKSIYOM

Bu  **$\sigma$ -toplanoabilirlik** varsayımdır. Herhangi bir ikişerli bağlantısız ortaya çıkan **sayılabilir** olaylar dizisi,  $E_1, E_2, \dots$  şu eşitliği tatmin eder:

$$P(E_1 \cup E_2 \cup \dots) = \sum_i P(E_i).$$

Bazı yazarlar sadece **sonsuz olmayan-toplanabilir** olasılık uzaylarını ele alırlar.

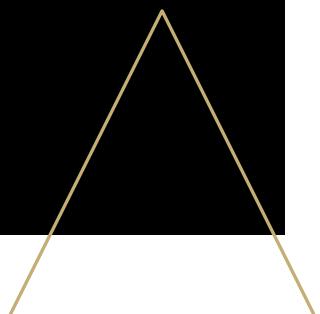


# KOŞULLU OLASILIK(CONDITIONAL PROBABILITY)

- Koşullu olasılık, bir olayın, başka bir olayın gerçekleştiği biliniyor iken, gerçekleşmesi olasılığıdır.
- A olayının gerçekleştiği bilindiğinde B olayının olması olasılığına, B olayın A olayına göre koşullu olasılığı ya da B olayın koşullu olasılığı denir.

$$P(B|A)$$

- $(B|A)$  olayı, A ile kısıtlanmış örneklem uzayında tanımlanmış bir olaydır.



Koşullu olasılık  $P(A|B)$  biçiminde gösterilir.  $P(A|B)$ 'nın anlamı,  $B$  gibi bir olayın gerçekleştiği bilindiğinde  $A$  olayının olasılığı olarak ifade edilir.

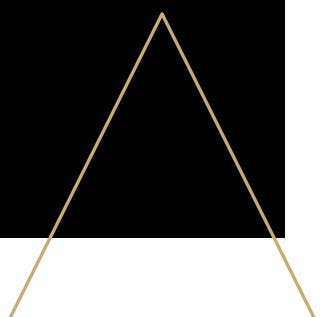
$A$  ve  $B$ , aynı örneklem uzayında tanımlanmış iki olay ve  $P(B) > 0$  olmak üzere  $B$  olayının gerçekleştiği varsayıımı altında  $A$  olayının koşullu olasılığı

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Biçiminde tanımlanır. Bu tanımdan yararlanılarak  $A$  ve  $B$  olaylarının birlikte gerçekleşme olasılığını koşullu olasılık yardımı ile

$$P(A \cap B) = P(A|B) \cdot P(B)$$

biçiminde bulabiliriz.



**Örnek:** Okulumuz öğrencilerinden %45'i istatistik, %35'i bilgisayar derslerinden ve %25'i hem istatistik hem de bilgisayar derslerinden başarısızdır. Rasgele seçilen bir öğrencinin,

- a) Bilgisayardan başarısız ise, istatistikten de başarısız olma olasılığını,
- b) İstatistikten başarısız ise, bilgisayardan da başarısız olma olasılığını,
- c) Bu iki dersten en az birinden başarısız olma olasılığını bulunuz.

$\bar{I}$ , istatistik dersinden başarısız öğrencileri; ve  $B$ , bilgisayar dersinden başarısız öğrencileri göstersin.

$$P(\bar{I})=0.45, \quad P(B)=0.35 \quad \text{ve } P(\bar{I} \cap B) = 0.25$$

$$\text{a)} \quad P(\bar{I}|B) = \frac{P(\bar{I} \cap B)}{P(B)} = \frac{0.25}{0.35} = \frac{5}{7}$$

$$\text{b)} \quad P(B|\bar{I}) = \frac{P(B \cap \bar{I})}{P(\bar{I})} = \frac{0.25}{0.45} = \frac{5}{9}$$

$$\text{c)} \quad P(\bar{I} \cup B) = P(\bar{I}) + P(B) - P(\bar{I} \cap B) = 0.45 + 0.35 - 0.25 = 0.55$$

Bulunur.

İki tabak dolusu bisküvi düşünülsün; tabak #1 içinde 10 tane çikolatalı bisküvi ve 30 tane sade bisküvi bulunduğu kabul edilsin. Tabak #2 içinde ise her iki tip bisküviden 20şer tane olduğu bilinsin. Evin küçük çocuğu bir tabağı rastgele seçip bu tabaktan rastgele bir bisküvi seçip alsın. Çocuğun bir tabağı diğerine ve bir tip bisküviyi diğerine tercih etmeyeceğine dair elimizde hiçbir gösterge bulunmamaktadır. Çocuğun seçtiği bisküvinin sade olduğu görülsün. Çocuğun bu sade bisküviyi tabak #1 den seçmiş olmasının olasılığının ne olacağı problemi burada incelenmektedir.

Sezgi ile, tabak #1de sade bisküvi sayısının çikolatalı bisküvi sayısına göre daha fazla olduğunu göz önüne alırsak incelenen olasılığın %50'den daha fazla olacağı hemen algılanır. Bu soruya cevap Bayes teoremi kullanarak kesin olarak verilebilir.

Önce soruyu değiştirdiğimizde Bayes teoremi uygulanabilecek şekilde sokmak gerekmektedir: Çocuğun bir sade bisküvi seçmiş olduğu bilinmektedir; o halde bu koşulla birlikte tabak #1den seçim yapması olasılığı ne olacaktır?

Böylece Bayes teoremi formülüne uymak için  $A$  olayı çocuğun tabak #1den seçim yapması;  $B$  olayı ise çocuğun bir sade bisküvi seçmesi olsun. İstenilen olasılık böylece  $\Pr(A|B)$  olacaktır ve bunu hesaplamak için şu olasılıkların bulunması gereklidir:

- $\Pr(A)$  veya hiçbir diğer bilgi olmadan çocuğun tabak #1'den seçim yapması olasılığı;

İki tabak arasında tercih olmayıp seçimin eşit olasılığı olduğu kabul edilmektedir.

- $\Pr(B)$  veya hiçbir diğer bilgi olmadan çocuğun bir sade bisküvi seçmesi olasılığı: Diğer bir ifade ile, bu çocuğun her bir tabaktan bir sade bisküvi seçme olasılığıdır. Bu olasılık, önce her iki tabaktan ayrı ayrı olarak seçilen bir tabaktan bir sade bisküvi seçme olasılığı ile bu tabağı seçme olasılığının birbirine çarpılması ve sonra bu iki çarpımın toplanması suretiyle elde edilir. Tabaklarda olan sade bisküvinin sayısının toplama orantısından bilinmektedir ki tabak #1'den bir sade bisküvi seçme olasılığı ( $30/40=$ ) 0,75; tabak #2'den sade bisküvi seçme olasılığı ( $20/40=$ ) 0,5 olur. Her iki tabaktan seçme olasılığı ise her tabak aynı şekilde uygulama gördüğü için 0,50 olur. Böylece bu problemin tümü için bir sade bisküvi seçme olasılığı  $0,75 \times 0,5 + 0,5 \times 0,5 = 0,625$  olarak bulunur.

- $\Pr(B|A)$ , veya çocuğun tabak #1'den seçim yaptığı bilirken bir sade bisküvi seçmesi.: Bu 0,75 olarak bilinmektedir çünkü tabak #1'deki toplam 40 bisküviden 30'u sade bisküvidir.

Şimdi bu açıklanan tüm olasılık değerleri Bayes teoremi formülüne konulabilir:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.75 \times 0.5}{0.625} = 0.6$$

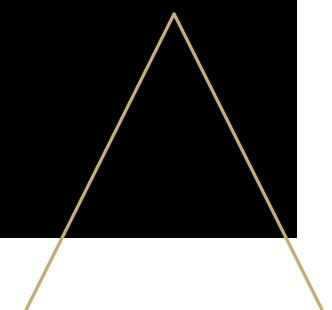
Böylece çocuğun sade bisküvi seçimi bilindiğine göre tabak #1'den alma olasılığı %60'tır ve sezgimize göre seçtiğimiz %50'den daha büyütür.

# Çarpım Kuralı (Multiplication Rule)

- A ve B gibi birlikte ortaya çıkan olayların olasılığı  $P(A \cap B)$  şeklinde ifade edilir. İki olayın kesişiminin olasılığı, bir olayın olasılığı ile ikinci olayın koşullu olasılığının çarpımına eşittir ve bu kurala çarpma kuralı denir.

$$P(A \cap B) = P(A) * P(B | A)$$

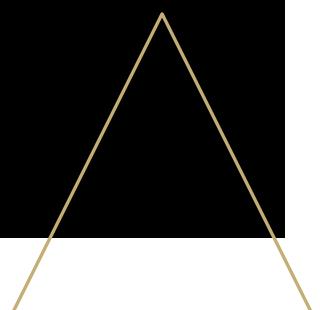
Koşullu olasılık  $P(B | A)$  biçiminde gösterilir.  $P(B | A)$ : A olayının gerçekleştiği bilindiğinde B olayının olma olasılığını ifade eder.



# Koşullu Olasılık Ve Bağımsız Olaylar

Olasılık konusunda iki olayın bağımsız olduğunu söylediğimizde, bir olayın olmasının diğer olayın olasılığını değiştirmedigini söylüyoruz.

Örneğin, hilesiz bir paranın yazı tura atışında "tura" gelme olasılığı  $1/2$ 'dir. Peki ya yazı tura atılan günün Salı günü olduğunu bilseydik? Bu, "tura" gelmesi olasılığı değiştirir miydi? Tabii ki değiştirmezdi. Salı olduğu bilindiğinde, "tura" gelme olasılığı hala  $1/2$ 'dir. O zaman, yazı tura atışının sonucuya günün Salı olması bağımsız olaylardır; Salı olduğunu bilmek, "tura" gelme olasılığını değiştirmez.

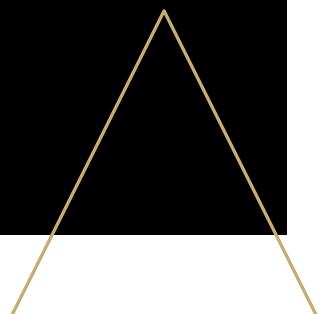




Her durum bu kadar açık değildir. Ya cinsiyet veya hangi elin kullanıldığı (solak ve sağlam)? Bir kişinin cinsiyetiyle solak veya sağlam olması tamamen bağımsız olaylar gibi görülmektedir. Yine de olasılıklara baktığımızda, tüm kişilerin %10'unun solak, ama erkeklerin yaklaşık %12'sinin solak olduğunu görürüz. O zaman, bu olaylar bağımsız değildir, çünkü rastgele bir kişinin erkek olduğunu bilmek onun solak olma olasılığını artırır.

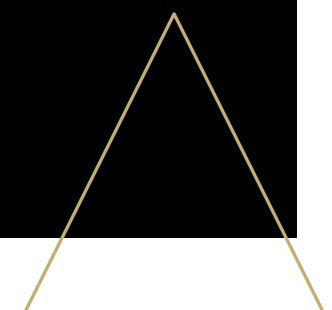
Buradaki temel fikir şu: Olasılıkta olayların bağımsız olup olmadığına bakarız.

Eğer  $P(A | B) = P(A)$  ve  $P(B | A) = P(B)$  ise, iki olay bağımsızdır.

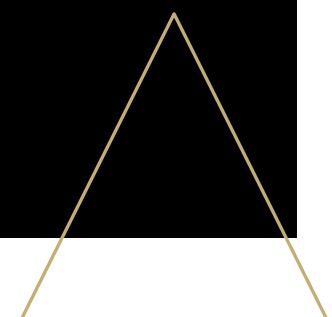


- ◻ A ve B olaylarından birinin elde edilmesi, diğerinin elde edilmesini etkilemiyorsa A ve B bağımsız olaylardır.
- ◻ Eğer aşağıdaki eşitlik sağlanıyorsa, A ve B olayları istatistiksel olarak bağımsızdır.

$$P(A \cap B) = P(A)P(B)$$



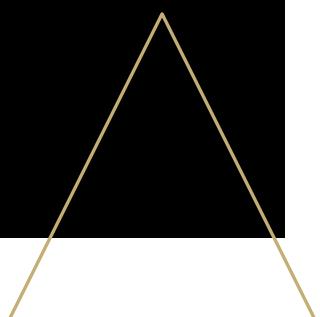
- 
- Gerçek hayattan veri setlerinde olayların bağımsızlığına baktığımızda, mükemmel eşitlikte olasılıklar elde etmek çok nadir görülür. Şans oyunları içermeyen neredeyse tüm gerçek olaylar bir derece bağımlıdır.
  - Pratikte, genelde olayların bağımsız olduğunu varsayarız ve örneklem verisi üzerinde bu varsayıımı test ederiz.
  - Son olarak, veriler iyi tasarlanmış bir deneyden gelmediği sürece, neden sonuç ilişkisi hakkında bir sonuca varmaktan kaçının.<sup>st</sup> ederiz. Olasılıklar çok farklıysa, olayların bağımsız olmadığı sonucuna varırız.



# BAYES TEOREMİ

Bayes ağları, bir rastlantı değişkenleri kümесinin çok değişkenli olasılık dağılımlarını etkili bir biçimde göstermeye ve modellemeye yarayan bir kavramdır. Bir olayın meydana gelmesinde birden fazla etkenin olması koşulunda, olayın hangi etkenin etkinliği ile ortaya çıktığını gösteren bu teoremde, rassal bir süreçle bağlı olarak ortaya çıkan rasgele bir  $X$  olayı ile diğer bir rasgele  $Y$  olayı için koşullu olasılıklar ve marjinal olasılıklar arasındaki ilişki tanımlanır. Bu ilişkiyi ilk kez Thomas Bayes ortaya atmış ve aşağıdaki eşitliği önermiştir .

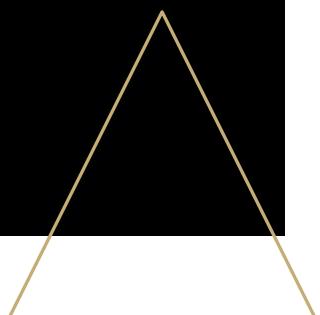
$$P(Y|X) = \frac{P(Y \cap X)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$



# ► Naive Bayes Sınıflandırıcı

## Sınıflandırıcı nedir?

Sınıflandırıcı, belirli özellikleri temel alan farklı nesneleri ayırt etmek için kullanılan bir makine öğrenmesi modelidir.

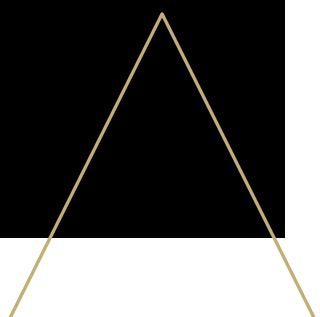


# Naive Bayes Sınıflandırma Modeli

Bir sınıflandırma problemi bir çok özellikten ve bir sonuç (hedef) değişkeninden oluşur.

$$p(C|F_1, \dots, F_n) = \frac{P(C)p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}$$

C verilen hedef ve F özelliklerimiz temsil eder. Naive bayes sınıflandırıcı basitçe bütün koşullu olasılıkların çarpımıdır.



Hava durumu üzerinden futbol oynamayıp oynamayacağımıza karar vermeye çalışalım.

Özellikler	Hedef
Hava Durumu	Futbol Oyna
Yağmurlu	Hayır
Yağmurlu	Hayır
Bulutlu	Evet
Güneşli	Evet
Güneşli	Evet
Güneşli	Hayır
Bulutlu	Evet
Yağmurlu	Hayır
Yağmurlu	Evet
Güneşli	Evet
Yağmurlu	Evet
Bulutlu	Evet
Bulutlu	Evet
Güneşli	Hayır

		Futbol Oyna	
		Evet	Hayır
Hava Durumu	Güneşli	3	2
	Bulutlu	4	0
	Yağmurlu	2	3



		Futbol Oyna	
		Evet	Hayır
Hava Durumu	Güneşli	3/9	2/5
	Bulutlu	4/9	0/5
	Yağmurlu	2/9	3/5
		9/14	5/14

Hava güneşli olduğunda futbol oynama olasılığını tahmin edelim. Bunun için yukarıdaki tablolardan hesaplamalar yapacağız.

Beklenti 1: Güneşliyken Futbol Oyna Evet =  $P(\text{Evet} \mid \text{Güneşli}) = P(\text{Güneşli} \mid \text{Evet}) * P(\text{Evet}) / P(\text{Güneşli})$

$$P(\text{Güneşli} \mid \text{Evet}) = 3/9 = 0.333, P(\text{Güneşli}) = 5/14 = 0.357, P(\text{Evet}) = 9/14 = 0.643$$

$$P(\text{Evet} \mid \text{Güneşli}) = 0.333 * 0.643 / 0.357 = 0.600$$

Beklenti 2: Güneşliyken Futbol Oyna Hayır =  $P(\text{Hayır} \mid \text{Güneşli}) = P(\text{Güneşli} \mid \text{Hayır}) * P(\text{Hayır}) / P(\text{Güneşli})$

$$P(\text{Güneşli} \mid \text{Hayır}) = 2/9 = 0.222, P(\text{Güneşli}) = 5/14 = 0.357, P(\text{Hayır}) = 5/14 = 0.357$$

$$P(\text{Hayır} \mid \text{Güneşli}) = 0.222 * 0.357 / 0.357 = 0.222$$

Son aşamada beklen 1 ile beklen 2 kıyaslanır. Beklen 1 daha büyük değere sahip olduğu seçili. Naive Bayesian Classifier havayı güneşli gördüğünde futbol oynaya izin verir.

# Örnek Bir tez;

- “E-Ticaret Sistemlerinde Reklam Ürünlerinin Bayes Teoremine Göre Yerleştirilmesi” - Mehmet Akif BÜLBÜL
- Bu tez çalışmasında e-ticaret ile alışveriş yapan üye ziyaretçilerin ürünler üzerindeki her bir hareketi kayıt altına alınmış ve bu kayıtlar analiz edilerek kullanıcının bir sonraki ürün satın alma veya ürün inceleme ihtimali yüksek olan ürün grubu Bayes Teoremi ile hesaplanmıştır.
- Sunulan çalışmada, reklam ürünlerinin dinamik yerleştirilmesi ve öngörü sistemi ile web tabanlı kişiselleştirilmiş prototip yapısı Bayes Teoremi kullanılarak oluşturulmaktadır. Ayrıca, önceki araştırmalarda kullanıcıların tercihleri/girişleri manuel olarak gerçekleştirılmıştır. Bu çalışma sayesinde kullanıcıların site içi anlık hareketleri ile ürün inceleme veya ürün satın alma arasındaki ilişki otomatik olarak Bayes Teoremi kullanarak hesaplanmakta ve kişiselleştirilmiş dinamik ürün yerleştirme ile özgün bir model ortaya konulmaktadır.

1. Başla
2. Eğer kullanıcı sisteme giriş yaptı ise 4. Adıma git
3. Rastgele reklam ürünü yerleştir 10. Adıma git
4. Kullanıcı hareketlerini veritabanından süz
5. Reklam ürün gruplarını Bayes Teoremi ile hesapla
6. En yüksek oranlı ürün grubunu belirle
7. Web arayüzündeki reklam alanlarını belirlenen ürün grubu ürünler ile güncelle
8. Eğer ziyaretçi kategori değiştirdi ise 4. Adıma geri dön
9. Eğer ziyaretçi ürün aldı veya ürün baktı ise 4. Adıma geri dön
10. Bitir

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB

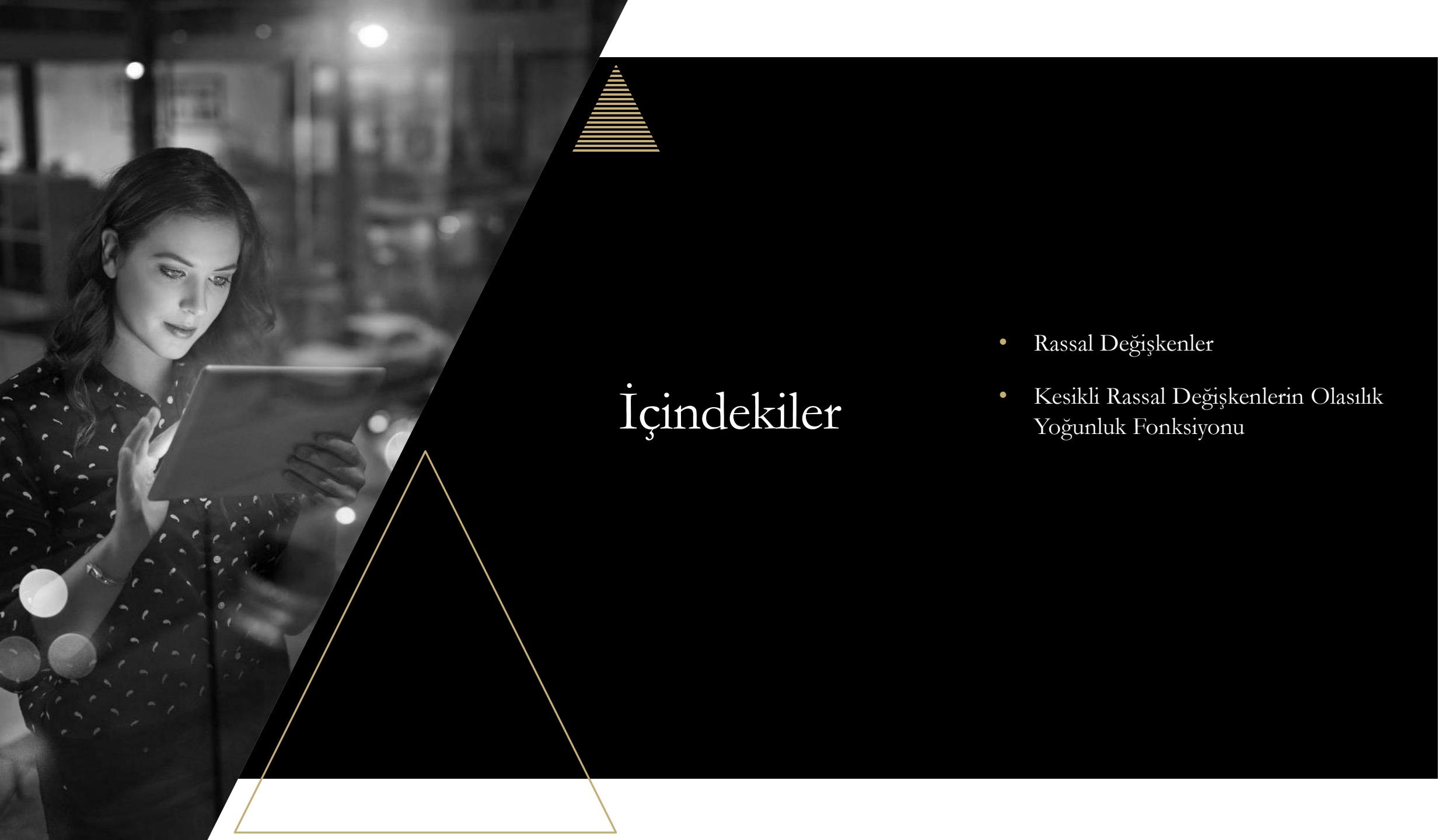
# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-7

CEMİLE YILDIZÇAKAR

26.01.2021



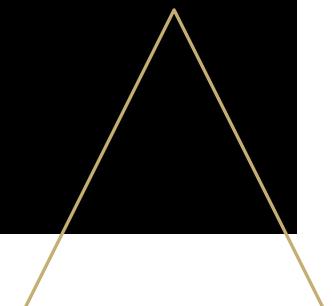


# İçindekiler

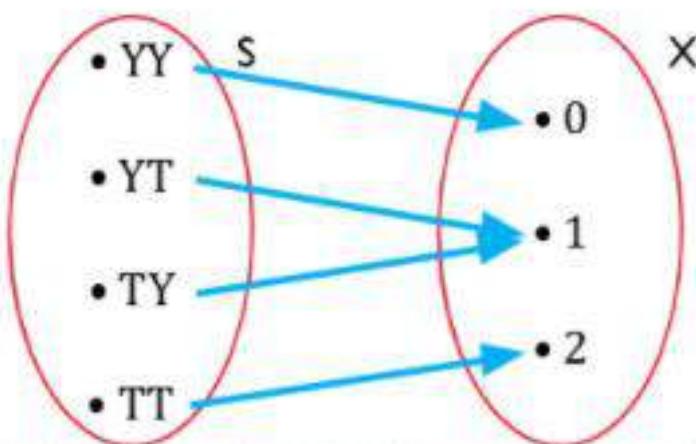
- Rassal Değişkenler
- Kesikli Rassal Değişkenlerin Olasılık Yoğunluk Fonksiyonu

## Rassal Değişkenler (Random Variable)

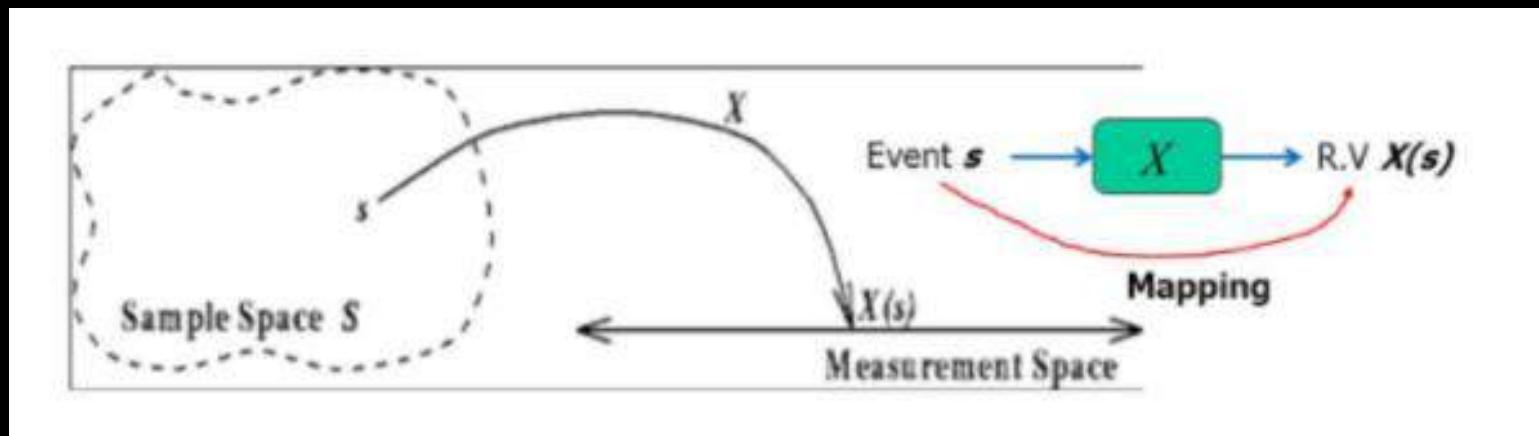
- Rassal değişkenin temelinde rastgele gerçekleşen olaylar yer alır.
- Bir deney ya da gözlemin şansa bağlı sonucu bir değişkenin aldığı değer olarak düşünülür ise, olasılık ve istatistikte böyle bir değişkene rassal değişken adı verilir.



- Deneylerin olası sonuçlarını sayılar ile temsil etmek istediğimizde, bunu rassal değişkenler aracılığıyla yaparız.
- Örneğin;  
“iki para atışında gelen tura sayısı” bir **rassal değişkendir**.



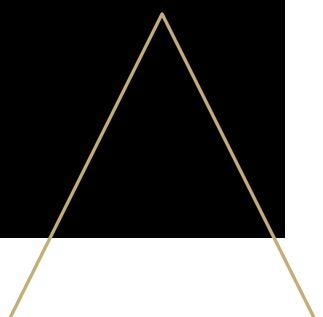
- Örnek uzayının her bir elemanını **gerçel sayılar kümesine taşıyan** fonksiyona rassal değişken denir.
- Deney tekrarlandıkça, rassal değişkenin aldığı değer değişir.

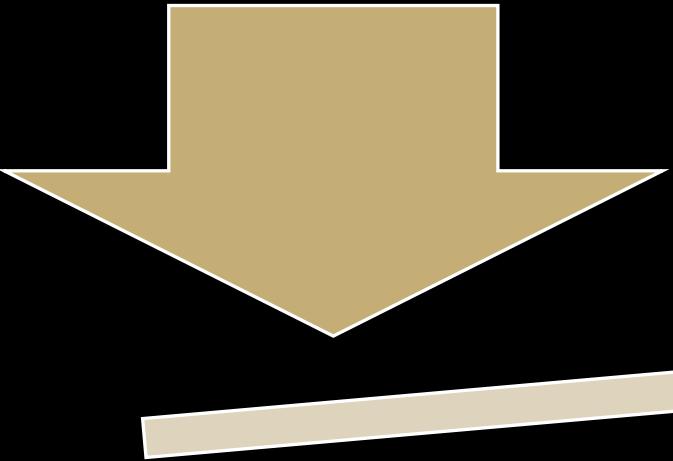




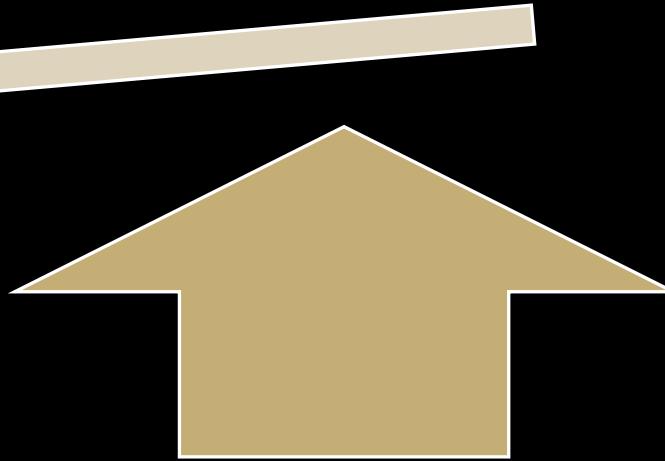
## Örnekler

- Bir futbol takımının herhangi bir maçta atacağı gol sayısı
- Bir fabrikada günde üretilen şeker miktarı
- Bir otobüsün Bornova'dan kampüsüne geliş süresi
- Herhangi bir günde polikliniğe gelen hasta sayısı

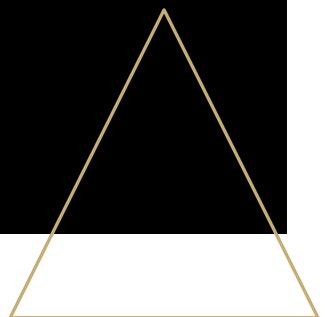




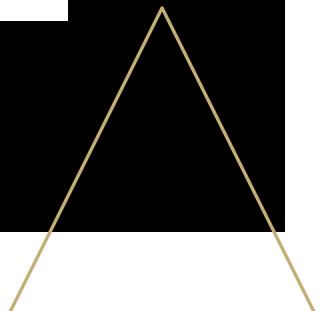
Kesikli Rassal  
Değişken



Sürekli Rassal  
Değişken

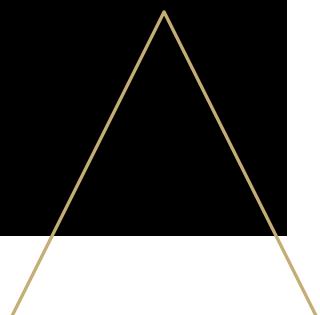


- 
- **Kesikli rassal değişkenler** sonlu sayıda ya da sayılabilir sonlu sayıda değerler alabilirler.
    - ▶ Örnek: Marketteki müşteri sayısı  $x = 0, 1, 2, \dots$
  - ▶ **Sürekli rassal değişkenler** bir aralıktaki tüm değerleri alabilirler.
    - ▶ Örnek: Bebeklerin doğum ağırlığı  $2000 < x < 5500$



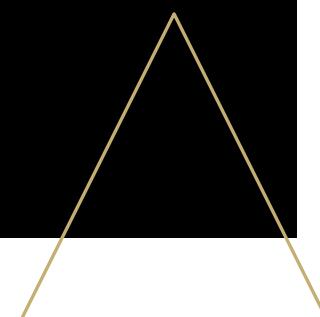
## Kesikli Rasgele Değişkenler İle İlgili Örnekler

- Bir süpermarkete 5 dakikalık süre içerisinde gelen müşteri sayısı,
- Bir madeni paranın üç kez atılması sonucunda yazı gelme sayısı,
- Bir bayanın sahip olduğu ayakkabı sayısı,
- Anaokuluna giden çocukların ağızındaki çürük diş sayısı,
- Bir aşçının günlük kullandığı yumurta sayısı.



## Sürekli Rasgele Değişkenler İle İlgili Örnekler

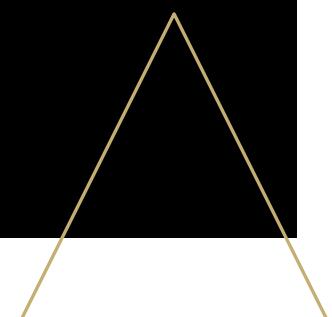
- Bir kişinin ağırlığı,
- Sınavda bir sorunun çözülme süresi,
- Bir arsanın fiyatı,
- Bir çağrı merkezine gelen telefonların arasındaki geçen süre,
- Bir mandıranın günlük sattığı süt miktarı.





Aşağıdaki rassal değişkenlerin türünü belirleyiniz.  
Alabilecekleri değerleri düşününüz.

- Bir restaurantta kullanılan günlük sıvı yağ miktarı.
- Bahçedeki bir ağacın yapraklarının sayısı.
- Bir televizyonun ömrü (yaşam süresi).
- İki arkadaşın telefonda konuşma süresi.



Rassal değişkenleri ifade etmek için büyük harfler kullanılır. (**X, Y, Z, ...**)

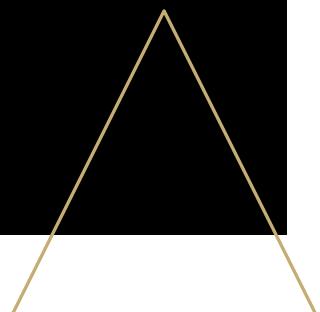
Rassal değişkenlerin alabildiği değerleri ifade etmek için küçük harfler kullanılır. (**x, y, z, ...**)

Örnek:

Tek bir hilesiz zar atıldığında;

$X$  : zarın üst yüzündeki noktaların sayısı

$x = 1, 2, 3, 4, 5, 6$



# KESİKLİ RASSAL DEĞİŞKENLERİN OLASILIK DAĞILIMLARI

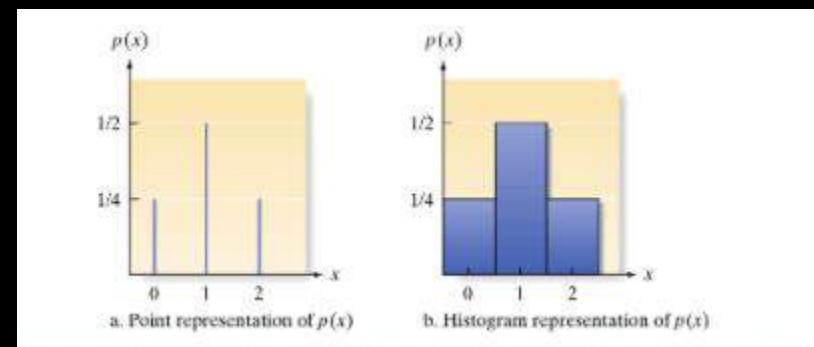
- Olasılık dağılımı (Probability Distribution) :

Bir kesikli rassal değişkenin, alabildiği tüm değerlere karşılık gelen olasılıkları veren fonksiyonudur.

$P(X)$  veya  $f(x)$  şeklinde gösterilir.

Tablo formül veya grafik şeklinde oluşturulabilir.

$$P(X_i) = p_i \quad i = 1, 2, 3 \dots, n \quad ya da \quad i = 1, 2, 3, \dots$$



Kesikli olasılık dağılımının sağlanması gereken koşullar (Requirements for discrete probability distributions)

Her  $X_i$  için  $0 \leq P(X_i) \leq 1$

$$\sum_{i=1}^n P(X_i) = 1$$

$X_i$	$P(X_i)$
1	$P_1$
2	$P_2$
3	$P_3$
.	.
.	.
.	.
n	$P_n$

**Requirements for the Probability Distribution of a Discrete Random Variable x**

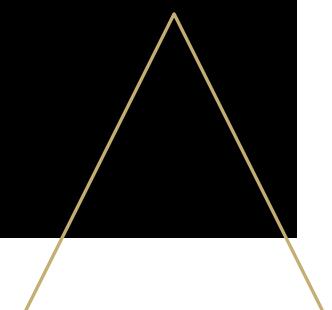
1.  $p(x) \geq 0$  for all values of  $x$ .

2.  $\sum p(x) = 1$

where the summation of  $p(x)$  is over all possible values of  $x$ .\*

## Örnek

- Hilesiz bir bozuk para 4 kez atılıyor.
  - Gelen tura sayısı  $Y$  rassal değişkeni olarak tanımlansın.
  - $Y$  rassal değişkeninin olasılık dağılımını oluşturunuz.



Bu deneyde Y rastgele değişkeni turların  
sayısı olarak tanımlanırsa,  
Y'nin alabileceği değerler YY için 0,  
TY ve YT için 1 ve TT için 2 olacaktır.  
Böylece örnek uzayındaki noktalar reel sayılar  
ile ifade edilebilmektedir. Bu örnek için  
olasılık fonksiyonunu yazabilmemiz için ilgili  
durumların olasılıkları hesaplanmalıdır.  
YY için  $P(X=0)=1/4$ ,  
YT ve TY için  $P(X=1)=(1/4)+(1/4)=2/4$   
iken  
TT için  $P(X=2)=1/4$  olacaktır.  
Bu durumda X rastgele değişkeninin olasılık  
fonksiyonu aşağıdaki gibidir:

X	P(X)
0	1/4
1	2/4
2	1/4

- 
- Soru: Bir torbada bulunan 3 kırmızı ve 4 beyaz bilye arasından **2 bilye ard arda seçiliyor**. X rastgele değişkeni kırmızı bilyelerin **sayısı olmak üzere**
    - a) X'in olasılık fonksiyonunu bulunuz.
    - b)  $P(X=1)$  ve  $P(X<2)$  olasılıklarını hesaplayınız.

Çözüm:

Deneyin örnek uzayı :

$$S = \{\text{KK, KB, BK, BB}\}$$

X rastgele değişkeni kırmızı bilyelerin sayısı olduğuna göre

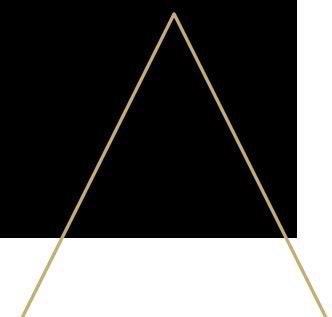
BB için  $X=0$ ,

KB ve BK durumları için  $X=1$  iken

KK durumu için  $X=2$  olacaktır.

O halde X rastgele değişkeninin alabileceği değerler 0, 1 ve 2 olmaktadır.

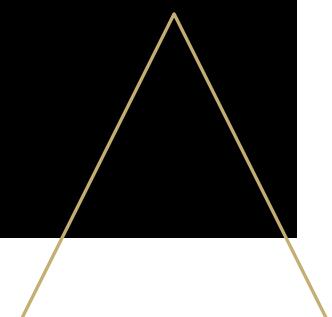
X'in bu değerleri alma olasılıkları ise aşağıdaki gibi hesaplanabilir.




$$P(X = 0) = P(BB) = \frac{4}{7} \times \frac{3}{6} = \frac{2}{7}$$

$$P(X = 1) = P(KB) + P(BK) = \left(\frac{3}{7} \times \frac{4}{6}\right) + \left(\frac{4}{7} \times \frac{3}{6}\right) = \frac{4}{7}$$

$$P(X = 2) = P(KK) = \frac{3}{7} \times \frac{2}{6} = \frac{1}{7}$$



X	P(X)
0	2/7
1	4/7
2	1/7

$$P(X = 1) = \frac{4}{7} ; \quad P(X < 2) = P(X = 0) + P(X = 1) = \frac{2}{7} + \frac{4}{7} = \frac{6}{7}$$

Soru: X rastgele değişkeni için olasılık fonksiyonu aşağıdaki gibi verilmiştir. X'in olasılık fonksiyonu olabilmesi için k sabiti hangi değeri almalıdır?

$$P(X) = \begin{cases} k(X+1) & X = 1, 2, 3 \\ 0 & \text{diğer durumlar için} \end{cases}$$

ÇÖZÜM:

$$\sum P(X_i) = 1$$

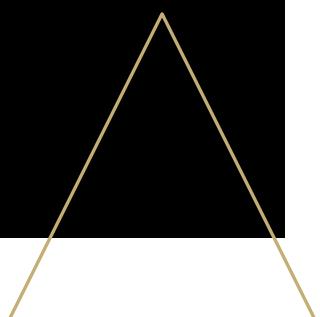
$$\sum_{i=1}^3 k(X+1) = k[2+3+4] = 9k = 1 \Rightarrow k = \frac{1}{9}$$

## Örnek

- X kesikli rassal değişkenine ait olasılık fonksiyonu aşağıdaki gibi verilmiştir.

$$P(X = x) = \begin{cases} kx & x = 2, 4, 6 \\ k(x - 2) & x = 8 \\ 0 & \text{otherwise} \end{cases}$$

k sabitinin değerini bulunuz.



## Örnek

- Aşağıdaki fonksiyonun olasılık dağılımı olup olamayacağını inceleyiniz.

$$f(x) = \frac{x+2}{25}, x=1, 2, 3, 4, 5$$

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB



# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-8

CEMİLE YILDIZÇAKAR

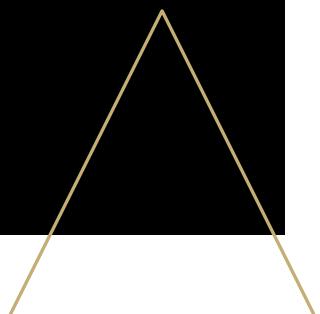
05.02.2021



## Birikimli Dağılım Fonksiyonu (Cumulative Distribution Function cdf)

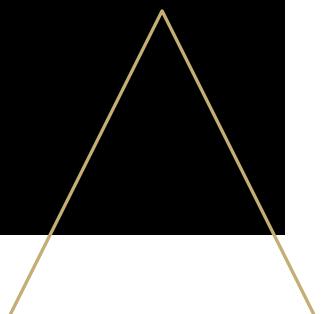
- $X$  rassal değişkeninin  $x$  değerine eşit ya da daha küçük bir değer alma olasılığını veren fonksiyondur.
- Tüm rassal değişkenler için elde edilebilir.
- $F(X)$  şeklinde gösterilir.

$$F(x) = P(X \leq x) \text{ for } -\infty < x < \infty$$



## Birikimli dağılım fonksiyonunun sağlanması gereken koşullar Requirements for Cumulative Distribution Function

- $0 \leq F(x) \leq 1$  for all  $x$
- $\lim_{x \rightarrow -\infty} F(x) = 0$
- $\lim_{x \rightarrow \infty} F(x) = 1$
- $F(x)$  is a *nondecreasing* function of  $x$

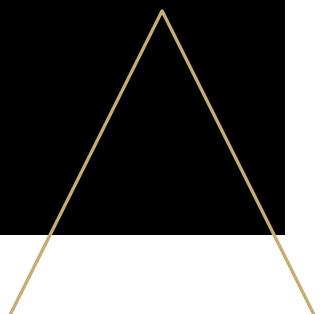


## Teorem:

- X rassal değişkeni için  $x_1 < x_2 < x_3 < \dots < x_n$  ise;

$$f(x_1) = F(x_1)$$

$$f(x_i) = F(x_i) - F(x_{i-1}) \text{ for } i = 2, 3, \dots, n$$



## Örnek

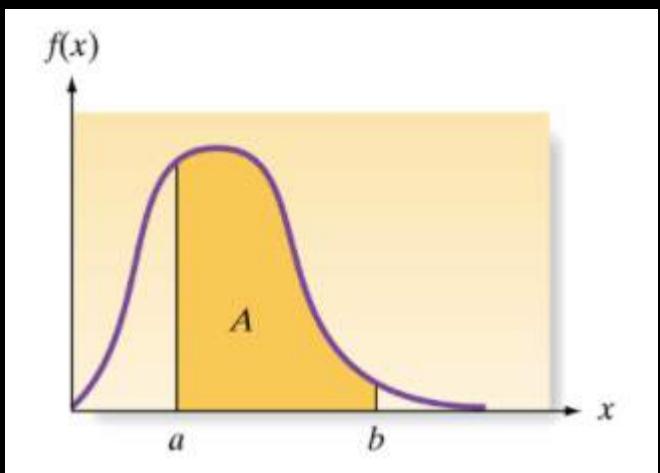
- Bir futbol takımının yaptığı maçlarda attığı gol sayısının fonksiyonu aşağıdaki gibi elde edilmiştir.

X	0	1	2	3	4	5
P(X)	0.1	0.2	0.4	0.15	0.1	0.05

- Bu fonksiyonun olasılık dağılımı olup olamayacağını inceleyiniz.
- Atılan gol sayısının birikimli dağılım fonksiyonunu oluşturunuz.

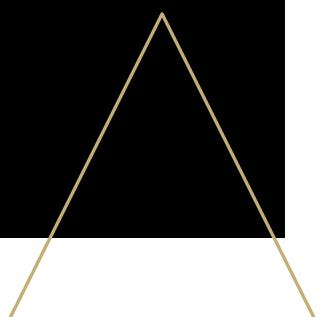
# Sürekli Rassal Değişkenlerin Olasılık Dağılımları

- Sürekli rassal değişkenlerin olasılık dağılımları bir eğri oluştururlar.  $X$ 'in bir fonksiyonu olan bu eğriye olasılık yoğunluk fonksiyonu (probability density function -pdf) denir.



- $f(x)$  fonksiyonu hiçbir değerin olasılığını göstermez. Bu fonksiyonun integrali alınırsa olasılık elde edilebilir.
- Herhangi bir  $f(x)$  fonksiyonunun grafisinin altında  $a$  ve  $b$  noktaları arasında kalan alan;  $P(a < X < b)$  olasılığını verir.

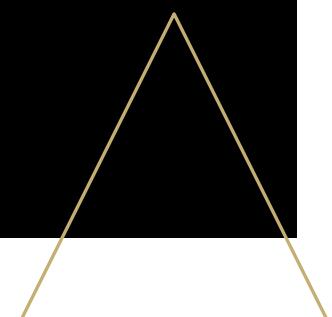
$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x) dx$$





## Sürekli olasılık dağılımının sağlanması gereken koşullar (Requirements for continuous probability distributions)

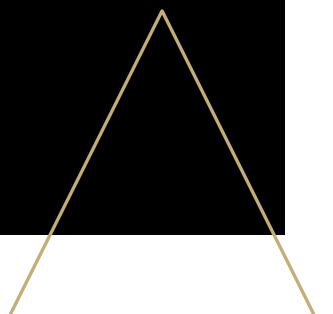
- $f(x) > 0$  for all possible intervals of  $x$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$



# Birikimli Dağılım Fonksiyonu

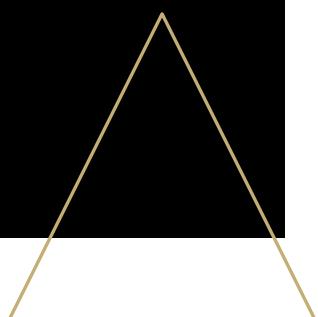
## Cumulative Distribution Function (cdf)

$$F(y_0) = P(Y \leq y_0) = \int_{-\infty}^{y_0} f(y) dy$$



# Birikimli Dağılım Fonksiyonunun Özellikleri

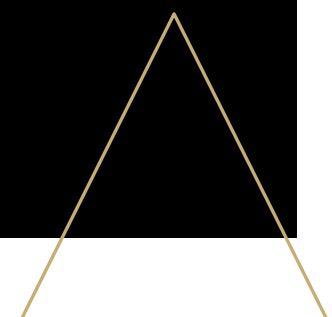
1. The CDF is non-negative:  $F(x) \geq 0$ . Probabilities are never negative.
2. The CDF goes to zero on the far left:  $\lim_{x \rightarrow -\infty} F(x) = 0$ .  $X$  is never less than  $-\infty$ .
3. The CDF goes to one on the far right:  $\lim_{x \rightarrow \infty} F(x) = 1$ .  $X$  is never more than  $\infty$ .
4. The CDF is non-decreasing:  $F(b) \geq F(a)$  if  $b \geq a$ . If  $b \geq a$ , then the event  $X \leq a$  is a sub-set of the event  $X \leq b$ , and sub-sets never have higher probabilities. (This was a problem in HW2.)



## Teorem

- $X$  sürekli bir rassal değişken,  $a$  ve  $b$  reel sayılar ise ( $a \leq b$ );

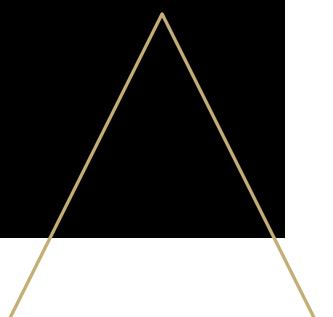
$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$$



## Teorem

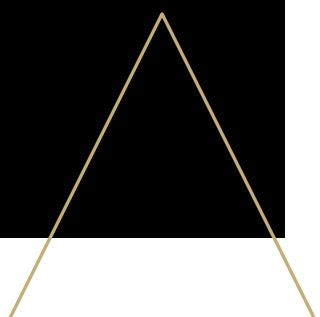
■  $X$  sürekli bir rassal değişken,  $a$  ve  $b$  reel sayılar ise ( $a \leq b$ ):

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$$



# Beklenen Değer (Expected Value)

- Bir rassal değişkenin beklenen değeri veya ortalaması; ilgili deney çok fazla sayıda tekrarlandığında gözlenmesi beklenen değerdir.
- $E(X)$  şeklinde ifade edilir.



$X_i$  'nin alabileceği tüm mümkün değerleri  $X_1, X_2, \dots, X_k$  ve olasılıkları sırasıyla  $p_1, p_2, \dots, p_k$  olan bir tesadüfi değişken olsun. Bu durumda  $X_i$ 'nin beklenen değeri:

$$E(X_i) = \sum_i^k X_i p_i \text{ şeklinde tanımlanır.}$$

$n$  örnek hacmi olmak üzere  $\bar{x} = \frac{1}{n} \sum_i^k x_i$  şeklinde ifade edilen örnek ortalamasının beklenen değerini

bulmak istediğimizde  $\bar{x}$ 'nın tüm mümkün değerlerinin bulunması gereklidir.  $\bar{x}$ 'nın mümkün bütün değerlerini bulabilmek için  $N$  hacimli bir populasyondan seçilebilecek  $n$  hacimli tüm mümkün örneklerin bulunması gereklidir. Bu durumda  $\bar{x}$ 'nın beklenen değeri:

$$E(\bar{x}) = \sum_i^k \bar{x}_i p_i \text{ şeklinde tanımlanır.}$$

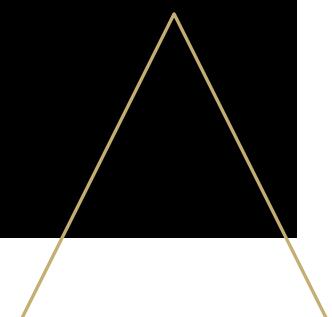
Sürekli rassal değişkenler için :

$$E(X) = \int x f(x) dx$$

all x

# Beklenen değerin özellikleri

- $E(c) = c$
- $E(cX)=cE(X)$
- $E(cX)=cE(X)$
- $E(X+Y)= E(X) + E(Y)$



## Örnek

- Hilesiz bir zar atılsın.
  - Tek sayı gelirse gelen sayı kadar para kazanılıyor.
  - Çift sayı gelirse, 4 lira kaybediliyor.
- Bu oyunu oynarsanız ne kadar kazanc beklersiniz?

Let  $X$  = your earnings

$$X=1 \quad P(X=1) = P(\{1\}) = 1/6$$

$$X=3 \quad P(X=1) = P(\{3\}) = 1/6$$

$$X=5 \quad P(X=1) = P(\{5\}) = 1/6$$

$$X=-4 \quad P(X=1) = P(\{2,4,6\}) = 3/6$$

$$E(X) = 1*1/6 + 3*1/6 + 5*1/6 + (-4)*1/2 = 1/6 + 3/6 + 5/6 - 2 = -1/2$$

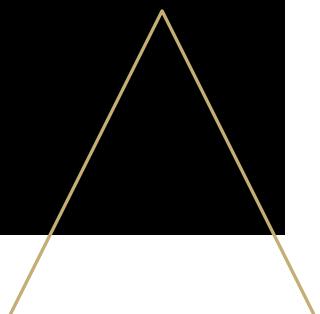
## Örnek

For the variable  $X$  with p.d.f.

$$f(x) = \begin{cases} \frac{1}{2}x, & 0 \leq x \leq 2 \\ 0, & \text{elsewhere} \end{cases}$$

find  $E(X)$

$$E(X) = \int_0^2 \frac{1}{2}x \cdot x \, dx = \left[ \frac{1}{6}x^3 \right]_0^2 = \frac{8}{6} = \frac{4}{3}$$



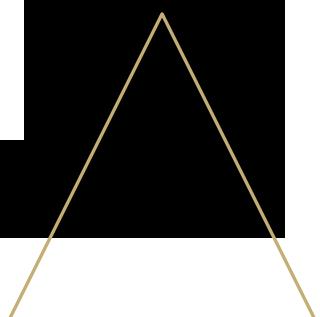
## Örnek

- Aşağıda pdf'i verilen  $X$  rassal değişkeninin ortalamasını hesaplayınız.

$$f(x) = \begin{cases} 2(1-x) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

## Çözüm

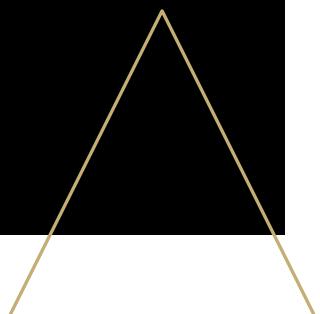
$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_0^1 x[2(1-x)]dx \\ &= 2 \int_0^1 (x - x^2)dx \\ &= 2\left(\frac{x^2}{2} - \frac{x^3}{3}\right)\Big|_0^1 \\ &= 1/3\end{aligned}$$



# Varyans (Variance)

$$\sigma_x^2 = E[(X - \mu_x)^2]$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

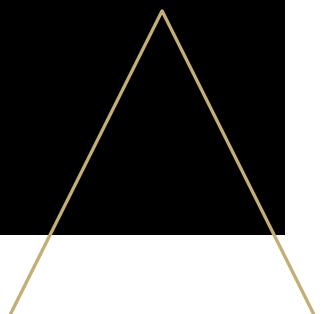


Kesikli rassal değişkenler için:

$$Var(X) = \sigma^2 = \sum_{\text{all } x} (x_i - \mu)^2 P(x_i)$$

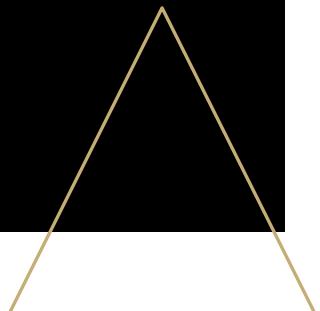
Sürekli rassal değişkenler için :

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x_i - \mu)^2 f(x_i) dx$$



$X$  ve  $Y$  herhangi rassal değişkenler,  $c$  sabit bir sayı ise;

- $\text{Var}(c) = 0$
- $\text{Var}(c+X) = \text{Var}(X)$
- $\text{Var}(cX) = c^2 \text{Var}(X)$
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) \rightarrow X \text{ ve } Y \text{ bağımsız ise!!!!}$



## *Varyansın Özellikleri*

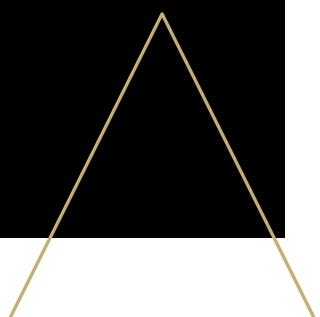
$\forall a, b, c \in \mathbb{R}$  için,

1)  $Var(a) = 0$

2)  $Var(bX + c) = b^2Var(X)$

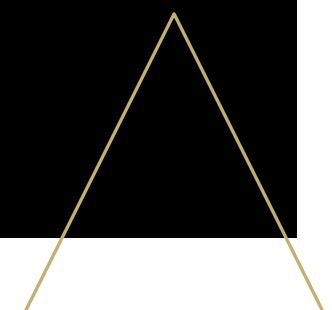
3)  $X$  ve  $Y$  tesadüfi değişkenler ise,

$$Var(aX \pm bY) = a^2Var(X) \pm 2abCov(X, Y) + b^2Var(Y)$$



# KOVARYANS

- Kovaryans iki değişken arasındaki doğrusal ilişkinin değişkenliğini ölçen bir kavramdır. Betimsel istatistiktir. Yani var olan bir şeyi bize söyler. Ortada tahmin yoktur. Sonucun pozitif olması artan bir doğrusal ilişkiyi, negatif olması azalan bir doğrusal ilişkiyi ve sıfır civarında olması ilişkinin olmadığını gösterir.



# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

The background of the slide features a photograph of a tropical landscape with several tall palm trees in the foreground and middle ground, and dark, silhouetted hills or mountains in the distance under a clear blue sky. A large black rectangular overlay covers the right side of the slide, containing the quote.

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB

# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-8

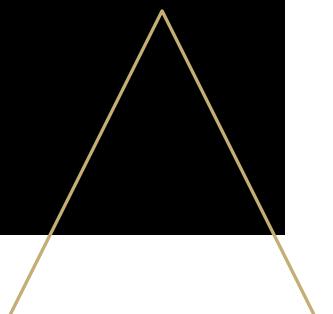
CEMİLE YILDIZÇAKAR

09.02.2021



# Kesikli Olasılık Dağılımları(Discrete Probability Distributions)

- Bernoulli Dağılımı (Bernoulli Distribution)
- Binom Dağılımı (Binomial Distribution)
- Hipergeometrik Dağılım(Hypergeometric Distribution)
- Geometrik Dağılım (Geometric Distribution)
- Negatif Binom Dağılım  $F(x) = P(X \leq x) \text{ for } -\infty < x < \infty$
- Poisson Dağılımı (Poisson Distribution)

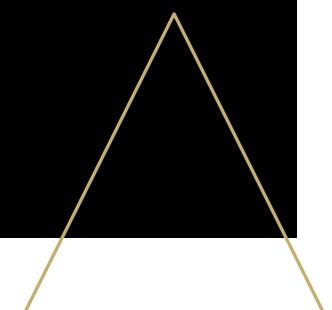
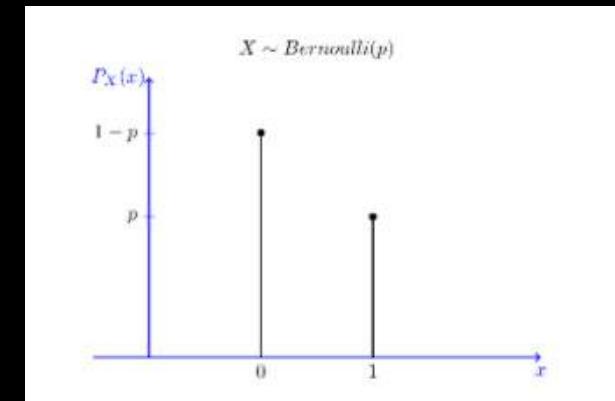


# Bernoulli Dağılımı - Bernoulli Distribution

- Rasgele bir deneme yapıldığında iki olası sonuç, başarı(success) ya da başarısızlık (failure) elde ediliyorsa, bu denemeye Bernoulli denemesi denir.

Örnekler:

- Bir anket çalışmasında bir soruya evet ya da hayır sonucunun verilmesi.
- Doğacak çocuğun cinsiyeti (kız-erkek)
- Madeni para atışında gelen sonuç (yazı-tura)
- Yapılan bir çalışmanın bir önceki çalışmaya göre elde edilen sonuçların kategorisi ( iyi – kötü)



# DİKKAT EDİLMESİ GEREKENLER

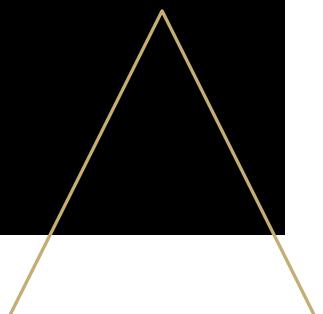
- Deneyin iki çıktısı olacak (başarılı- başarısız)
- $p$ : başarı olasılığı
- $1-p$  : başarısızlık olasılığı

Rassal Değişken X:

$x = 1$  : sonuç başarılı ise

$x = 0$  : sonuç başarısız ise

Bernoulli Olasılık Dağılımı:  $p(X = x) = p^x (1 - p)^{1-x}$   $x = 0, 1$



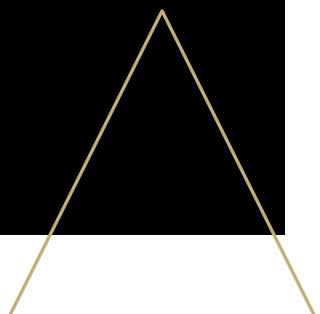


**$X \sim \text{Bernoulli}(p)$**

$$P(X = x) = \begin{cases} p^x (1-p)^{1-x} & x = 0, 1 \\ 0 & \text{d.d.} \end{cases}$$

**$E(X) = p$**

**$V(X) = p(1-p)$**

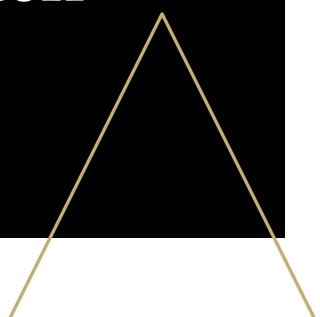


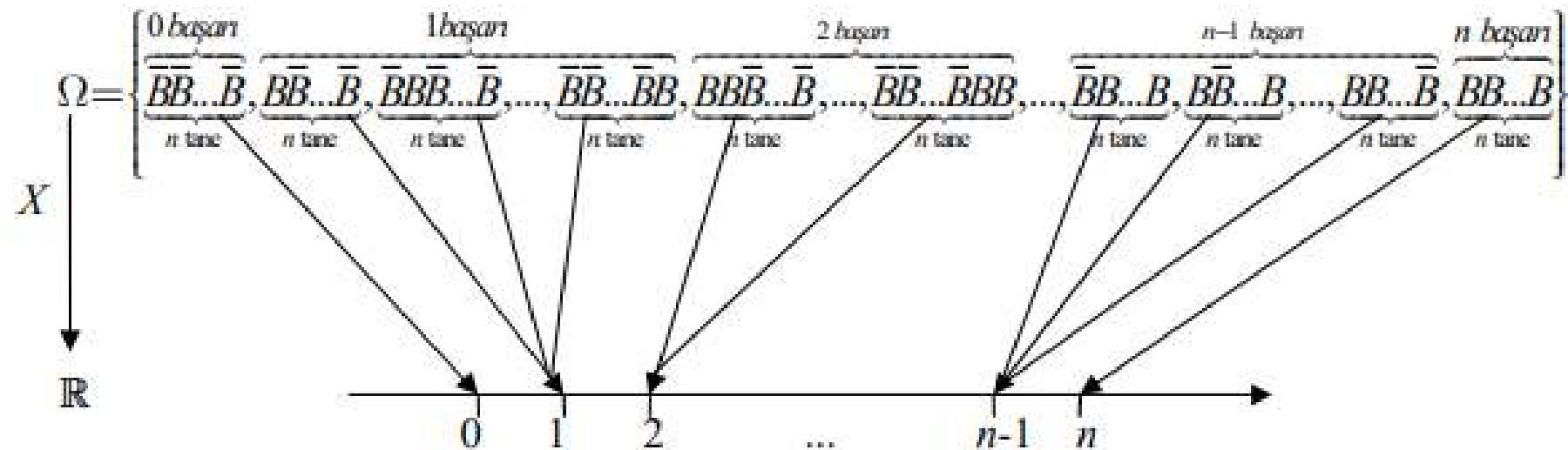
# Binom Dağılım

Başarı olasılığı olan bir Bernoulli denemesinin aynı şartlar altında (bağımsız olarak) n kez tekrarlanması ile oluşan deneye binom deneyi denir.

Binom rassal değişkeni X, n denemedeki başarı sayısını ifade etmektedir.

n deneme de en az 0, en fazla n adet başarı gözlenebileceğinden  
 $S = \{ x / 0,1,2,\dots,\dots,n \}$  olur.





$X$  rasgele değişkeninin aldığı değerlerin kümlesi,

$$D_X = \{0, 1, 2, \dots, n-1, n\}$$

ve

$$P(X = 0) = P(\underbrace{\overline{B} \overline{B} \dots \overline{B}}_{n \text{ tane}}) = \underbrace{q q \dots q}_{n \text{ tane}} = q^n$$

$$P(X = 1) = P(\underbrace{B \overline{B} \dots \overline{B}}_{n \text{ tane}} \text{ veya } \overbrace{\overline{B} B \overline{B} \dots \overline{B}}_{n \text{ tane}} \text{ veya } \dots \text{ veya } \underbrace{\overline{B} \overline{B} \dots \overline{B} B}_{n \text{ tane}}) = nq^{n-1} p = \binom{n}{1} p^1 q^{n-1}$$

$$P(X = 2) = P(\underbrace{B B \overline{B} \dots \overline{B}}_{n \text{ tane}} \text{ veya } \dots \text{ veya } \underbrace{\overline{B} \overline{B} \dots \overline{B} B B}_{n \text{ tane}}) = \binom{n}{2} p^2 q^{n-2}$$

...

$$P(X = n) = P(\underbrace{B B \dots B}_{n \text{ tane}}) = p^n$$

olup,  $X$  in olasılık fonksiyonu,

$$f(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \dots, n$$

dir. Moment çıkarılan fonksiyon,

$X$  tesadüfi değişkeni  $n$  tane bağımsız Bernoulli denemesinin başarılı olanlarının toplam sayısı olsun. Yani  $i = 1, 2, \dots, n$  için  $X_i \sim \text{Bernoulli}(p)$  ve  $X = X_1 + X_2 + \dots + X_n$  olmak üzere  $X$  tesadüfi değişkenine binom tesadüfi değişkeni denir ve olasılık fonksiyonu

$$P(X = x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0, 1, 2, \dots, n \\ 0 & \text{d.d.} \end{cases}$$

birimindedir.  $X \sim \text{Binom}(n, p)$  ile gösterilir. Burada  $n$  ve  $p$  dağılıma ait parametrelerdir.

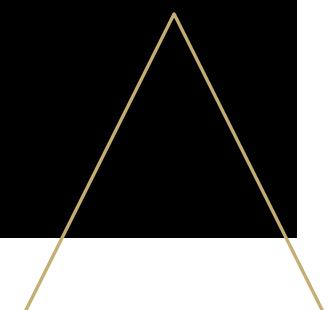
$X$  :  $n$  bağımsız Bernoulli denemesindeki başarıların sayısı.

$$X \sim \text{Binomial}(n, p)$$



## Örnekler;

- Bir fabrikanın deposundan seçilen 10 üründen 2'sinin hatalı olması ,
- Bir madeni para 5 kez atıldığında hiç tura a gelmemesi, üst yüze yazı veya tura gelmesi,
- Hilesiz bir zar 4 kez atıldığında zarın en çok 1 kez çift gelmesi,

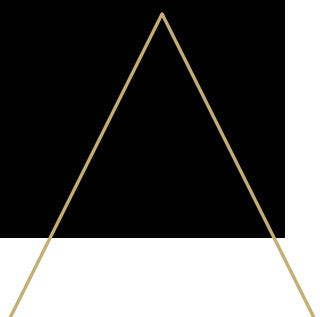


## *Beklenen Değer ve Varyans*

$$\begin{aligned}E(X) &= \sum_{D_X} xP(X=x) \\&= np(p + q)^{n-1} = np\end{aligned}$$

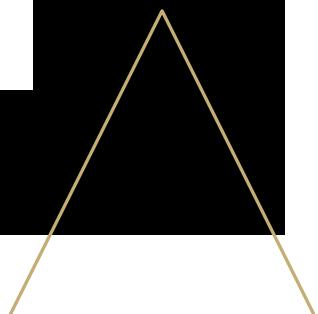
$$Var(X) = E(X^2) - (E(X))^2$$

$$Var(X) = p^2 n(n - 1) + (np)^2 = npq$$

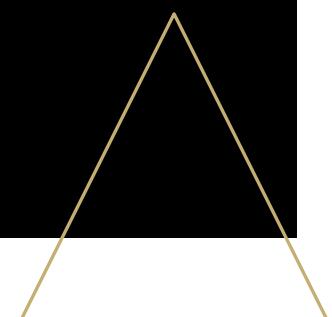


## Dikkat

- Bir deneyde seçimler yerine koymadan (**without replacement**) tekrarlanırsa başarı sayısının dağılımı binom dağılımına uymaz.
- Sonlu bir kitleden, örneklem yerine koymadan (**without replacement**) çekilirse, denemeler birbirine bağımlı hale gelir. Ve başarı sayısı farklı bir dağılım gösterir.

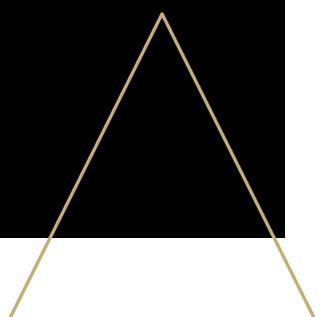


- İki tür obje içeren bir kitleye sahip olduğumuzu düşünelim;
- Bir torbada yeşil ve kırmızı top bulunması
- Bir kutuda hatalı ve hatasız ürünlerin bulunması
- Erkek ve kadınlar oluşan bir kitle



# Hipergeometrik Dağılım

- $n$  deneme benzer koşullarda tekrarlanır.
- • Her denemenin 2 mümkün sonucu vardır.
- • Sonlu kitleden iadesiz (without replacement) örneklem seçilir.
- • Örneklem iadesiz olduğundan, denemeler bağımlı hale gelir.
- Ve başarı olasılığı ( $p$ ) deneyden deneye değişir.



## Hipergeometrik Dağılımın Olasılık Dağılımı

N : kitle büyüklüğü

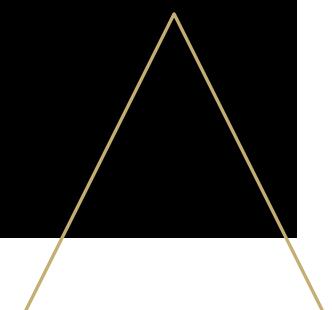
K : kitlede başarı sayısı olasılığı bulunmak istenen obje sayısı

N - K : Kitlede bulunan ikinci tür obje sayısı

n : örneklem büyüklüğü

X: Seçilen i. tür obje sayısı, i=1,2

$$h(x) = h(x; N, n, K) = \frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}} \quad x = 0, 1, \dots,$$



- Mean or Expected Value of  $X$

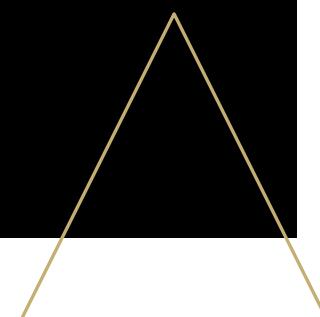
$$\mu = n \left( \frac{K}{N} \right)$$

- Standard Deviation of  $X$

$$\sigma = \left[ \frac{N-n}{N-1} \cdot n \cdot \frac{K}{N} \left( 1 - \frac{K}{N} \right) \right]^{1/2}$$

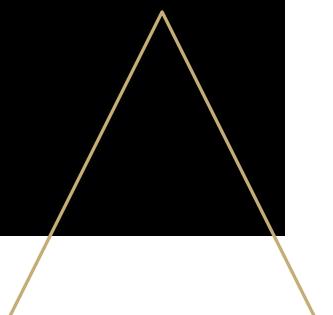
# Geometrik Dağılım (Geometric Distribution)

Binom olasılık dağılımında olduğu gibi bir Bernoulli sürecinden türetilen rassal deneylerin çıktıları ile ilgilenilsin. Bu süreçte çıktı ayrık iki olayı (başarı ve başarısızlık) tanımlar ve başarı olasılığı  $p$  deneyden deneye değişmez. Şans değişkeni  $x$  ilk başarı elde edilinceye kadar gerçekleştirilen deney sayı olarak tanımlandığında, şans değişkeninin dağılımı geometrik olasılık dağılımına uygundur. İlk başarının elde edilmesi için gerekli denemelerin sayısı  $X$ , geometrik rasgele değişkenidir.

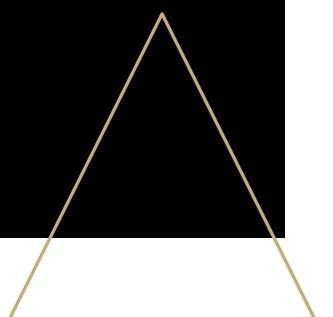


!!!!!!

- Hipergeometrik dağılımın binom dağılımına yaklaşımı:  
Anakütle eleman sayısı  $N$  çok büyük ise  $n$  ve  $p$  sabit  
kaldıkça hipergeometrik dağılım binom dağılımına  
yaklaşır.



- Denemeler başarı elde edilene kadar tekrarlanır.
- Denemeler birbirinden bağımsızdır.
- Başarı olasılığı  $p$ , her deneme için sabittir.
- Rassal değişken  $X$ , ilk başarı gerçekleşene kadar yapılan deneme sayısıdır.



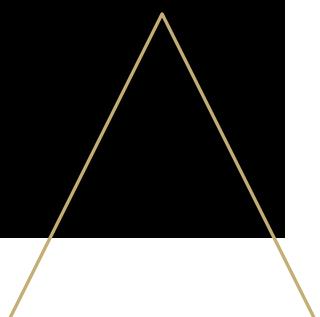
**Teorem:**  $X$ , bir tek denemede başarısızlık olasılığı  $q = 1 - p$  ve başarı olasılığı  $p$  olan geometrik rasgele değişken ise,  $X$  in olasılık fonksiyonu;

$$f(x) = P(X = x) = q^{x-1}p \quad x = 1, 2, 3, \dots$$

**İspat:** İlk başarının elde edilmesi için gereken denemelerin sayısı  $X, 1, 2, 3, \dots$  değerlerinden biri olabilir.  $x-1$  ilk başaridan önceki başarısızlıkların sayısı olsun. Bu durum aşağıdaki gibi gösterilebilir.

$$\underbrace{FF\dots F}_{x-1} S$$

O halde  $x-1$  başarısızlığı, başarının takip ettiği dizinin olasılığı  $q^{x-1}p$  dir. Bu nedenle  $X$  rasgele değişkeninin olasılık fonksiyonu;

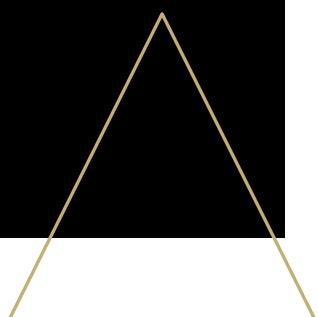


# Negatif Binom Dağılımı

Geometrik dağılımin genel şeklidir. Bir deney birbirinden bağımsız Bernoulli denemelerinden oluşmaktadır. Deneye  $K$  başarı elde edilinceye kadar devam edersek  $K$  başarının elde edilmesi için gerekli denemelerin sayısı negatif binom rasgele değişkenidir.

Negatif binom dağılımında, denemelerin sayısı bir rasgele değişkendir ve başarıların sayısı sabittir; binom dağılımda başarının sayısı rasgele değişkendir ve denemelerin sayısı sabittir.

Yani deneme sayısı sabit değil rassal değişkendir.



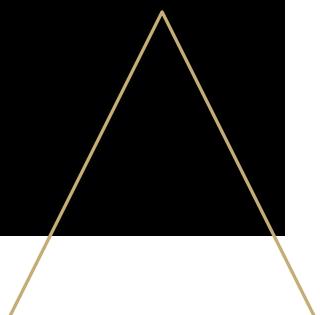
- $X$ : r. başarı gerçekleşene kadar yapılan deneme sayısı
- $p$ : başarı olasılığı

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad x = r, r+1, r+2, \dots \quad r = 1, 2, 3, \dots, x$$

$X \sim N\text{bin}(r,p)$

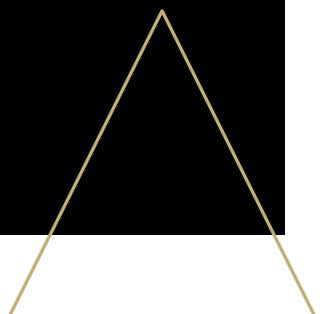
- Beklenen değer ve Varianc:

$$E(X) = \frac{r}{p} \quad \text{Var}(X) = \frac{r(1-p)}{p^2}$$



# Örnekler;

- Bir parayı 5 kez tura gelinceye kadar attığımızda 5. turayı elde ettiğimiz deneme sayısı,
- Bir basketbolcunun 3 sayılık atışlarda 10. isabeti sağlaması için gerekli olan atış sayısı.



# Poisson dağılımı (Poisson Distribution)

Bir çok deney sürekli bir zaman aralığında yapılır. Böyle bir ortamda gözlenen sonuçlar kesikli olabilir. Birim zaman aralıkları (dakika, saat, gün, ay, yıl gibi) veya birim uzunluk (alan veya hacim gibi) sürekli ortamlardır. Örneğin, bir mağazaya belli bir saat dilimi içinde gelen müşterilerin sayısı, böyle bir deneye örnektir. Bu tür deneylere Poisson deneyleri denir.  $X$  sürekli ortamdaki kesikli sonuçların sayısını göstermek üzere,  $X$  in (Poisson dağılımının) olasılık fonksiyonu,  $\lambda > 0$  için,

$$P(X = x) = e^{-\lambda} \lambda^x / x! , x = 0, 1, 2, 3, \dots$$

şeklindedir.  $X$  rasgele değişkeni bu olasılık fonksiyonuna sahipse,  $X$  Poisson dağılımına sahiptir denir ve  $X \sim Poisson(\lambda)$  ile gösterilir.  $e^\lambda$  fonksiyonunun sıfır noktası komşuluğundaki Taylor serisi açılımı

$$\sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

olup verilen fonksiyon bir olasılık fonksiyonudur. :

### Probability Distribution, Mean, and Variance for a Poisson Random Variable

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (x = 0, 1, 2, \dots) \quad \mu = \lambda \quad \sigma^2 = \lambda$$

where

$\lambda$  = Mean number of events during a given unit of time, area, volume, etc.

$e = 2.71828\dots$

$\lambda$  (lambda) is the expected number of events per unit.

X	$\lambda$								
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494
4	0.0000	0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Example: Find  $P(X = 2)$  if  $\lambda = .50$

$$P(X = 2) = \frac{e^{-\lambda} \lambda^X}{X!} = \frac{e^{-0.50} (0.50)^2}{2!} = .0758$$

$$\mu = E(X) = \lambda$$

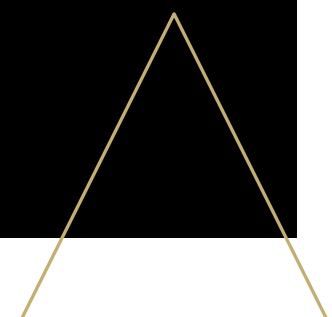
$$\sigma^2 = E[(X - \mu)^2] = \lambda$$

**Teorem 5.1.** Başarı olasılığı  $p$ , 0'a ya da 1'e yaklaştığında ve  $n \rightarrow \infty$  iken  $X \sim Binom(n, p) \approx Poisson(\lambda)$  olur. Yani,

$$p_X(x) = \binom{n}{x} p^x q^{n-x} \approx \frac{e^{-\lambda} \lambda^x}{x!}$$

 **Örnek 5.9:** Bir bölgede bir hastalığa yakalanma oranının 0.001 olduğu biliniyor. Tesadüfi olarak seçilen 2000 kişilik bir örneklemle çalışıldığında,

- a) En az iki kişinin bu hastalığa yakalanma olasılığı nedir?
- b) En çok dört kişinin bu hastalığa yakalanma olasılığı nedir?
- c) Hiç kimsenin bu hastalığa yakalanmama olasılığı nedir?



*Cözüm.*  $p = 0,001$  ve  $n = 2000$  olduğundan  $E(X) = np = 2$  olur

$$X \sim Binom(2000; 0,001) \approx Poisson(2)$$

elde edilir.

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) \\ &= 1 - \sum_{x=0}^1 \frac{e^{-2} 2^x}{x!} \\ &= 1 - 3e^{-2} = 0,594 \end{aligned}$$

b)

$$P(X \leq 4) = \sum_{x=0}^4 \frac{e^{-2} 2^x}{x!} = 7e^{-2} = 0,947$$

c)

$$P(X = 0) = \frac{e^{-2} 2^0}{0!} = e^{-2} = 0,1353$$

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB

# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

## hafta-10

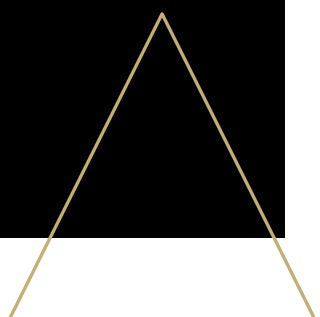
CEMİLE YILDIZÇAKAR

26.02.2021



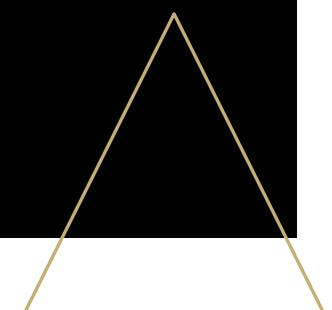
## Sürekli Olasılık Dağılımları(Continuous Probability Distributions)

- GAMMA DAĞILIMI
- ÜSTEL DAĞILIM (Exponential Distribution)
- NORMAL DAĞILIM



# GAMMA DAĞILIMI

- Rasgele değişken bir poisson işlem içinde  $k$  kadar değer oluncaya kadar ki mesafe aralığı gamma dağılımına sahiptir.
- Gamma dağılımı iki parametreli bir olasılık dağılımıdır.
- Parametrelerden biri ölçek ( $k$ ),  
diğeri ise şekil parametresidir.



- Gamma fonksiyonu:

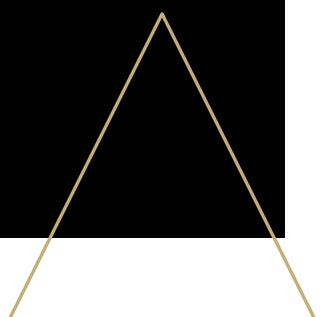
$$\Gamma(k) = \int_0^{\infty} x^{k-1} e^{-x} dx \quad \text{for } k > 0$$

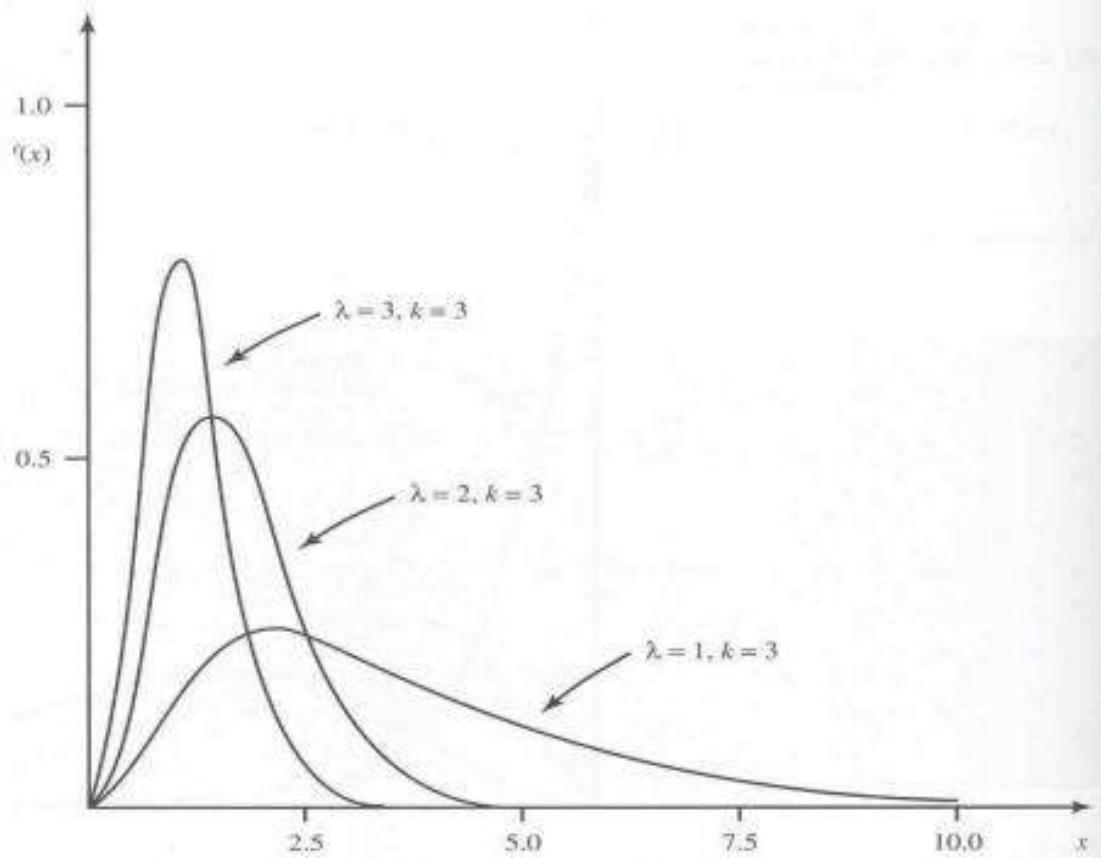
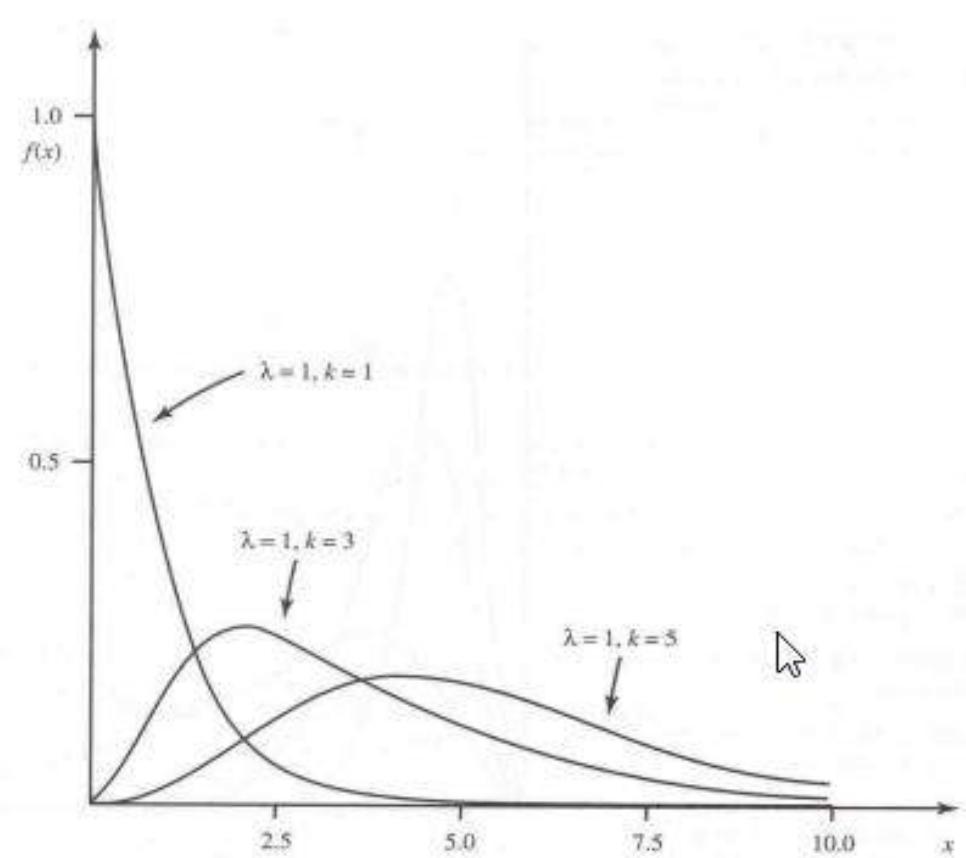
- $k > 0$  ve  $\lambda > 0$  için Gamma olasılık yoğunluk fonksiyonu:

$$f(x) = \frac{\lambda(\lambda x)^{k-1} e^{-\lambda x}}{\Gamma(k)} \quad \text{for } x \geq 0$$

- Ortalama ve varyans:

$$E(X) = k / \lambda \quad \text{and} \quad V(X) = k / \lambda^2$$





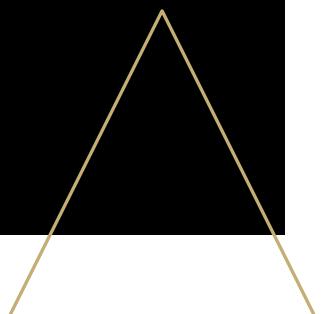
- Gamma rasgele değişkeninin özellikleri:

Eğer  $X_i, i = 1, \dots, n$ ,  $(k_i, \lambda)$ , parametreleri ile bağımsız gamma rasgele değişkenleri ise o zaman  $\sum_{i=1}^n X_i$ ,  $(\sum_{i=1}^n k_i, \lambda)$ . parametresi ile gamma dağılır.

- Parametresi  $(1, \lambda)$  olan gamma rasgele değişkeni, parametresi  $\lambda$ . olan üstel rasgele değişkene eşdeğerdir.

# ÜSTEL DAĞILIM

- Özellikle sanayi ürünlerinin dayanma sürelerinin incelenmesinde yaygın olarak kullanılan sürekli olasılık dağılımıdır.
- Bağımsız olaylar arasındaki zaman aralığını modelleştirdirken bir üstel dağılım doğal olarak ortaya çıkar.
  - ✓ Müşterilerin gelişleri arasında geçen zaman
  - ✓ Bozulmalararası geçen zaman
  - ✓ Belli bir bölgedeki iki deprem arasında geçen zaman



Bir üstel dağılım için olasılık yoğunluk fonksiyonu şu şekli alır:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

Burada  $\lambda > 0$  dağılım için tek parametredir ve çok zaman *oran parametresi* olarak anılır. Dağılım için destek  $[0, \infty)$  aralığında verilir. Eğer  $X$  rassal değişkeni bu üstel dağılım gösteriyorsa bu şöyle yazılır:

$$X \sim \text{Üstel}(\lambda).$$

Ancak bir diğer şekilde değişik parametreleme ile ise üstel dağılım için olasılık yoğunluk fonksiyonu şöyle ifade edilir:

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & , x \geq 0, \\ 0 & , x < 0. \end{cases}$$

Burada  $\beta > 0$  bir *ölçek parametresidir* ve yukarıda tanımlanan *oran parametresi* olan  $\lambda$ 'nın bir üstü değeri *çarpım tersi*, yani  $\beta=1/\lambda$ ; dır. Bu çeşit tanımlamada  $\beta$  *kalım parametresi* çünkü eğer bir rassal değişken  $X$  bir biyolojik veya mekanik sistem  $M$  için ömür geçirme zaman uzunluğu ise ve  $X \sim \text{Üstel}(\beta)$  ise

- Örnek: Bir aracın aküsü tükeninceye kadar alınan yol uzunluğu, ortalaması 10.000 mil olan üstel dağılıma sahip olduğunu varsayıyalım. Bir kişi 5000 mil yolculuk yapmak istiyorsa, aküsünü değiştirmeden yolculuğunu tamamlayabilme olasılığı nedir?
- $X$  bataryanın kalan ömrünü (bin mil olarak) içeren rasgele bir değişken olsun. O zaman,

$$E[X] = 1/\lambda = 10 \Rightarrow \lambda = 1/10$$

$$P\{X > 5\} = 1 - F(5) = e^{-5\lambda} = e^{-1/2} \approx 0.604$$

$X$ , üstel rasgele değişken değilse ne olur?

$$P\{X > t + 5 | X > t\} = \frac{1 - F(t + 5)}{1 - F(t)}$$

Yani,  $t$  hakkında ek bir bilgi bilinmelidir

## Hafızasızlık özelliği (Lack of memory)

$$\square P(x > 10 | x > 3) = P(x > 7+3 | x > 3) = P(x > 7)$$

$$P(X > s+t | X > s) = P(X > t)$$

# Üstel dağılım ile Poisson dağılımı arasındaki ilişki

- Poisson olayları arasında geçen süre genellikle üstel dağılım ile açıklanır.
- Kasaya birim zamanda gelen müşteri sayısı: Poisson
- Kasaya gelişler arasında geçen süre: Üstel
- Servis için başvuran müşteri sayısı: Poisson
- Servis için başvuran müşterilerin gelişleri arasında geçen süre: Üstel

## Poisson

- number of events in a time interval
- Discrete  
 $X = 0, 1, 2, \dots$

number of  
↓ lines



## Exponential

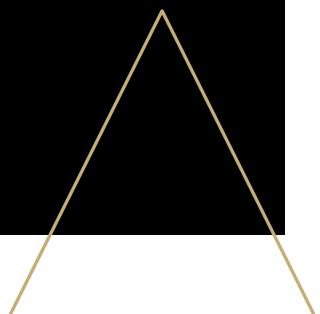
- time between two events
- Continuous  
on an interval



# NORMAL DAĞILIM



- Normal dağılımın ilk uygulamaları doğada gerçekleşen olaylara karşı başarılı bir biçimde uyum göstermiştir. Dağılımın göstermiş olduğu bu uygunluk adının Normal Dağılım olması sonucunu doğurmuştur.
- İstatistiksel yorumlamanın temelini oluşturan Normal Dağılım, bir çok rassal süreçlerin dağılımı olarak karşımıza çıkmaktadır.
- Normal dağılış kullanımının en önemli nedenlerinden biride bazı varsayımların gerçekleşmesi halinde kesikli ve sürekli bir çok şans değişkeninin dağılımının normal dağılışa yaklaşım göstermesidir.



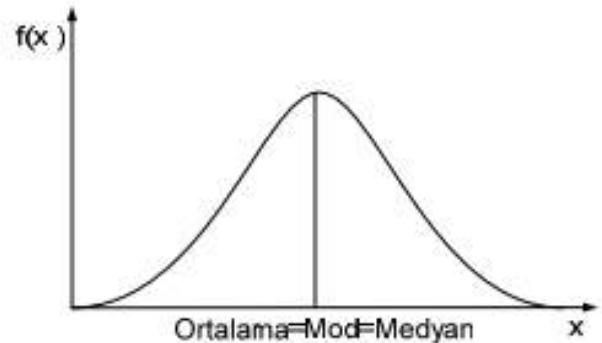
## Örnek;

- Yetişkinlerin boy uzunlukları, kütleleri vb.
- Örneğin 1000 yetişkinin zekâ düzeylerinin dağılımı frekans poligonu üzerinden incelense, gözlemlerin ortalama değeri olan 100 civarında kümelendiği, daha yüksek ve daha düşük IQ'lu birey sayısının daha az olduğu görülecektir.

## Normal Dağılımın Özellikleri

- Çan eğrisi şeklindedir.
- Simetrik bir dağılıstır.
- Normal Dağılımın parametreleri,

$$E(x) = \mu \quad \text{Var}(x) = \sigma^2$$



## Normal Dağılımın Olasılık Yoğunluk fonksiyonu

$$f(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} & , -\infty < x < \infty \\ 0 & , \text{diger yerlerde} \end{cases}$$

$\pi = 3,14159\dots$

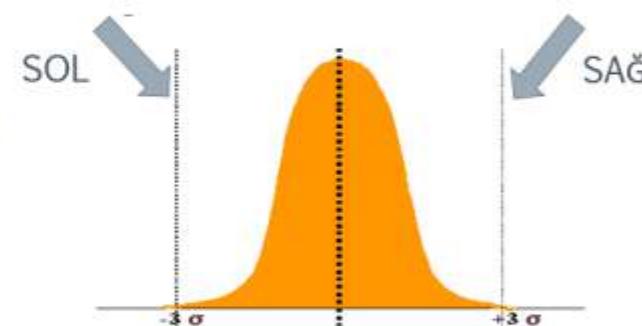
$e = 2,71828$

$\sigma$  = populasyon standart sapması

$\mu$  = populasyon ortalaması

## Normal Dağılım Eğrisinin Özellikleri

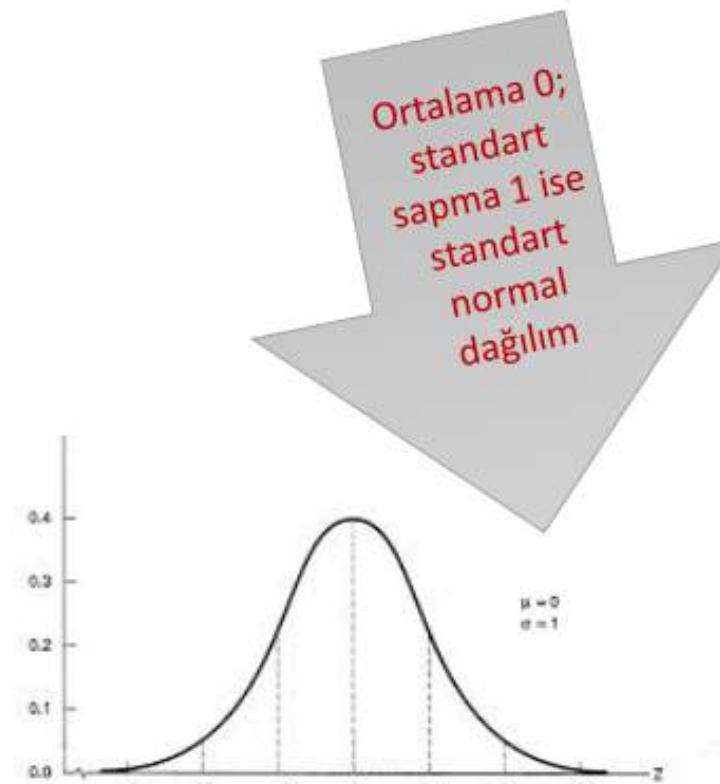
- Eğri, dikey eksene göre **simetiktir**. Puanların yarısı eksenin sağ, diğer yarısı da sol tarafındadır.
- Puanlar merkez etrafında kümelenme eğilimi gösterir.
- Mod, ortanca ve ortalama birbirine eşittir.
- Dağılımin her iki ucu giderek yatay eksene yaklaşır, ancak hiçbir zaman bu eksene demez (asimptomatik). Normal dağılım eğrisi atındaki alan sınırsızdır.



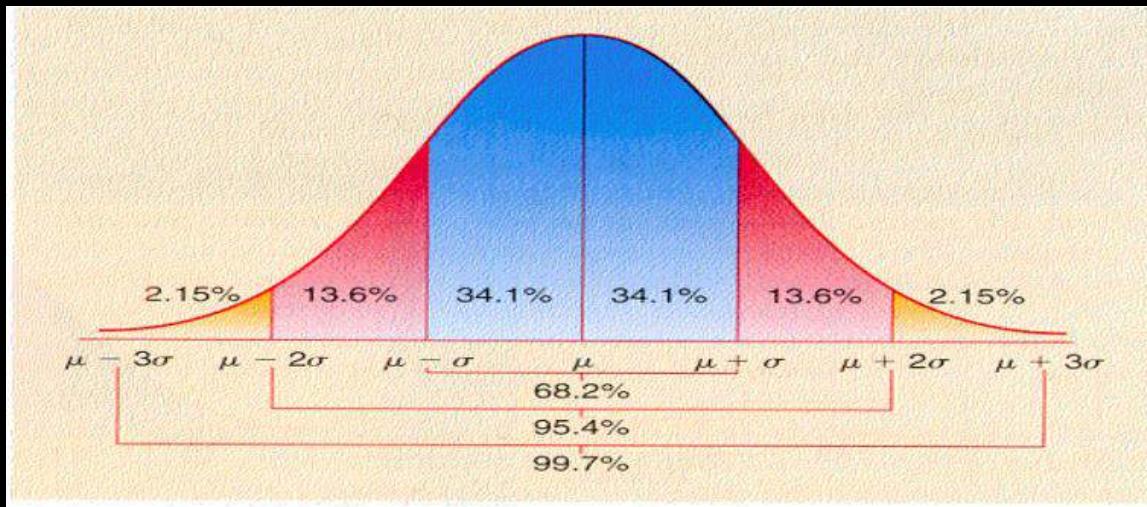
(Ferguson ve Takane, 1989; Ravid, 1994)

## Standart Normal Dağılım

- Standart normal dağılımda, ortalama 0, standart sapma 1'dir.
- Ortalamanın sol tarafındaki (altındaki) birimler negatif, sağındakiler pozitiftir.
- İki standart sapma arasındaki uzaklıklar birbirine eşittir.
- İki standart sapma arasında kalan alanlardan merkeze yakın olanlar, uzak olanlara göre daha fazla puan kapsar (Ravid, 1994).

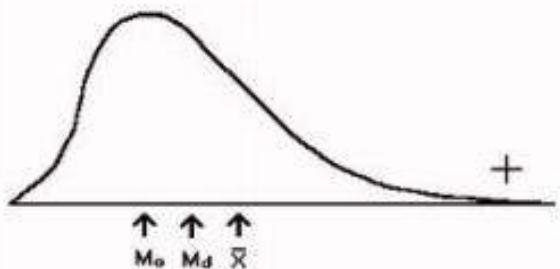


- Ortalama: 25 ve Standart Sapma: 5 olsun;
  - Puanların % 68.3' 20 ile 30 puan arasındadır.
  - Puanların % 95.4'ü 15 ile 35 puan arasındadır.
  - Puanların % 99.7'si 10 ile 40 puan arasındadır.
  - Puanların % 47.7'si 25 ile 35 puan arasındadır. (\*2'den %95.4'ü 15 ile 35 arası)
  - Puanların % 49.8'i 10 ile 25 arasındadır

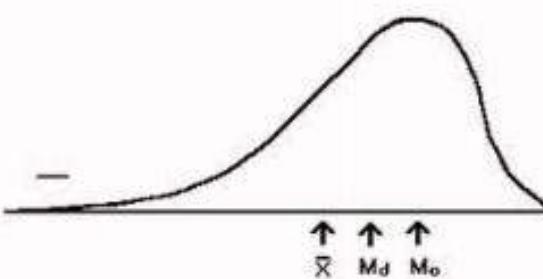


## ÇARPIK VE BASIK DAĞILIMLAR

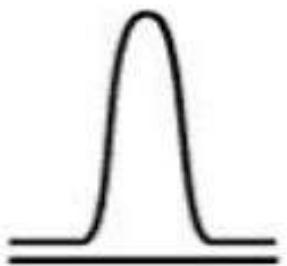
Aşağıda normal dağılımdan farklılaşan dağılımlar, dağılımın şekilleri ile gösterilmiştir.



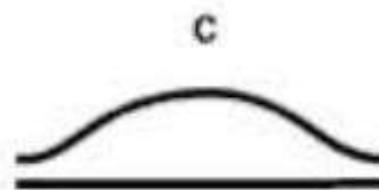
Şekil 1. Sağa Çarpık Dağılım



Şekil 2. Sola Çarpık Dağılım



Şekil 3. Sivri Dağılım



Şekil 4. Basık Dağılım

## DAĞILIM NORMALLİĞİNİN İNCELENMESİ

- a) Verilerin normal dağılım gösterip göstermediğini belirlemenin yollarından biri dağılımin grafiğini çizmek ve bu grafiği yorumlamaktır.
- b) Verilerin dağılımının normal dağılım gösterip göstermediğini belirlemenin bir diğer yolu ortalama, mod ve medyan değerlerine bakmaktır. Normal dağılımda bu değerler çakışmaktadır. Bu istatistikler birbirine yaklaştığı ölçüde dağılım normal dağılıma yaklaşır. Birbirinden uzaklaşlığı ölçüde dağılım çarpıklaşır. Fakat bu yakınlığın düzeyi ile ilgili belirli bir standart yoktur. Bu nedenle burada verilen diğer yöntemlerle birlikte değerlendirilmesi önerilir.
- c) Normal dağılımı test etmenin bir diğer yolu da basıklık ve çarpıklık katsayılarına bakmaktır. Çarpıklık (skewness) katsayısı normal dağılımda 0'dır. Negatif çarpıklık katsayısı sağa çarpık dağılıma, pozitif çarpıklık katsayısı sola çarpık dağılıma işaret eder. Basıklık (kurtosis) katsayısı da normal dağılımda 0'dır. Pozitif basıklık katsayısı sivri dağılıma, negatif basıklık katsayısı ise basık bir dağılıma işaret eder. Dağılımin normal dağılımdan manidar düzeyde farklılaşmamak için bu değerlerin (-1, +1) aralığında kalması beklenir.

## Normal Dağılımın önemi

- Birçok istatistiksel test Normal dağılım varsayıımına dayanmaktadır. Bunun anlamı, ancak normal dağılım varsayıımı sağlandığı takdirde, bu testler en iyi sonucu vermektedir.
- Sadece güçlü 'robust' olarak tanımlanan testler, normalden sapmaları tolere etmektedir.

## Normallik Sınaması (Assessing Normality )

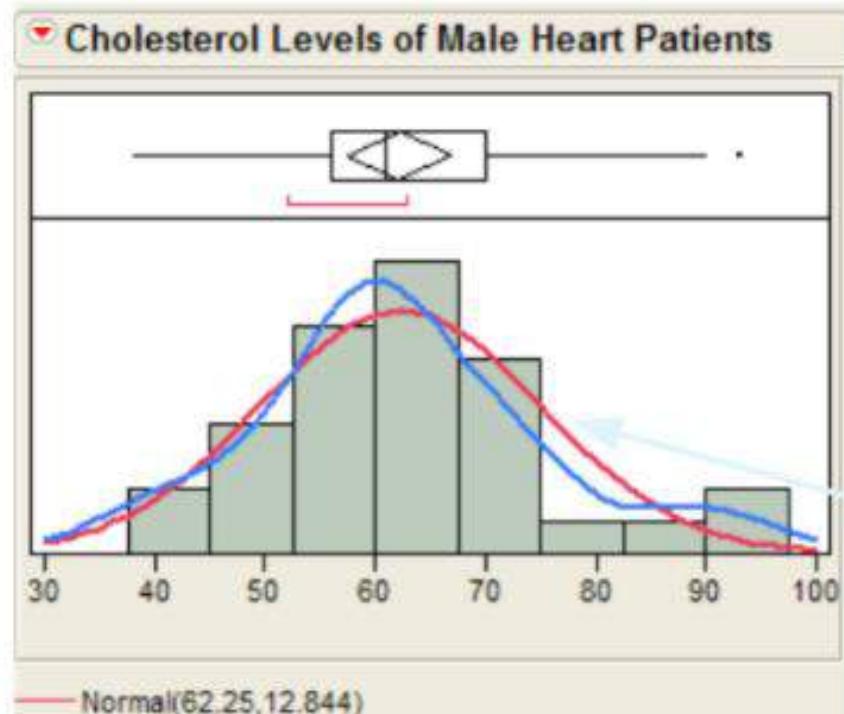
- Normallik sınavması, verilerin normal dağılıp dağılmadığına karar verme sürecidir.
- Normal dağılan bir kitleden çekilen bir örneklem her zaman normal dağılmayabilir.
- Örneklemeler her seferinde değiştiğinden , her örneklemen dağılımı da değişir.
- İstatistiksel testler, küçük veri setleri için ( $n < 30$ ) kullanıldığında, kitlenin normal ya da normale yakın dağılım gösterdiği varsayımları yapılır. Küçük veri setlerinin histogramları kitlenin dağılımını her zaman yansıtmayabilir. Bu sebeple, kitlenin normalliğini sağlamak için farklı yöntemlere ihtiyaç duyulmaktadır.
- Bununla birlikte, örneklem normal dağılan bir kitleden geliyor ve yeterli büyüklüğe sahip ise, dağılımı normale yakın olabilir.

## Normallik Sınaması için Kullanılan Yöntemler

---

- Histogram
- Boxplot
- Normal Quantile Plot (Normal Olasılık Grafiği)  
(Normal probability Plot))
- Uyum İyiliği Testleri (Goodness of Fit Tests)

## Histogram ve Boxplot

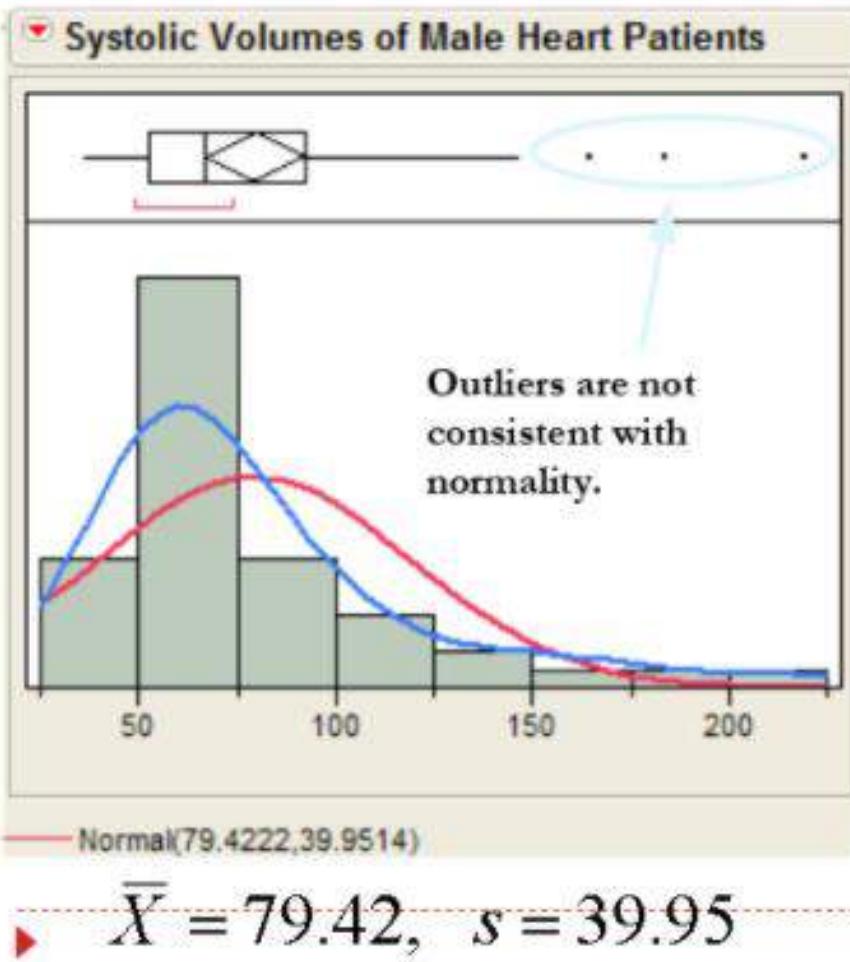


$$\bar{X} = 62.25, \ s = 12.84$$

Sağdan çarpıklık için ortalamanın ortancadan büyük olması gibi bazı kanıtlar olmasına rağmen hastaların kolesterol seviyesi yaklaşık olarak normal dağılmaktadır.

Kırmızı eğri normal dağılımın bu dağılıma fit edilmiş halidir. Mavi eğri ise bu verilerin olasılık yoğunluk fonksiyonu kestirimidir. Bu veri normal dağılıyor olsaydı, iki eğri üst üste gelecekti.

## Histogram and Boxplot



Historama göre; Erkek kalp hastalarının sistolik değerleri bu örneklem sağıdan çarpık bir dağılan bir kitleden çekildiğini göstermektedir.

## Uyum İyiliği Testleri (Goodness of Fit Tests)

### □ Ampirik Dağılım Fonksiyonuna Dayalı Testler

- Kolmogorov-Smirnov Test
- Kuiper's Test
- Lilliefors Test
- Cramér-Von Mises Test
- Anderson-Darling Test
- Watson Test

### □ Regresyon ve Korelasyona Dayalı Testler:

- Shapiro-Wilk Test

# Teşekkür Ederim



## LinkedIn

<https://www.linkedin.com/in/cemile-yildizcakar-34782248/>



## Email

yildizcakar.cemile@gmail.com

Cemile YILDIZÇAKAR

The background of the slide features a photograph of a tropical landscape with several tall palm trees in the foreground and middle ground, and dark, silhouetted hills or mountains in the distance under a clear blue sky. A large black rectangular overlay covers the right side of the slide, containing the quote.

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB