

# VERİ BİLİMİ İÇİN TEMEL İSTATİSTİK

hafta-8

CEMİLE YILDIZÇAKAR

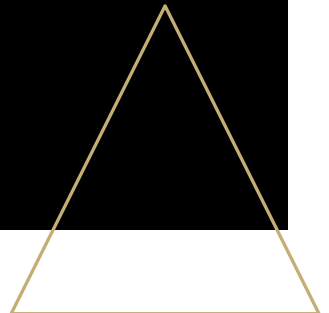
09.02.2021





# Kesikli Olasılık Dağılımları(Discrete Probability Distributions)

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

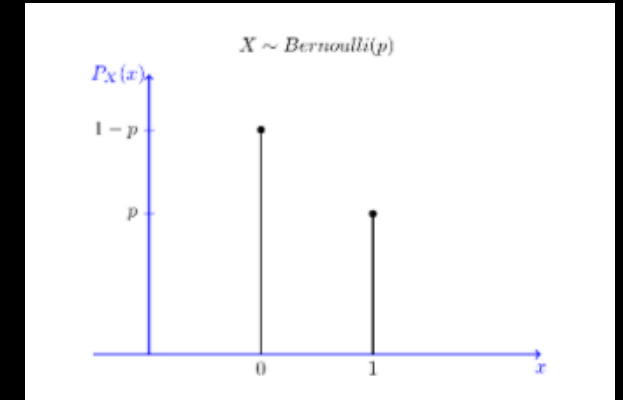


# Bernoulli Dağılımı - Bernoulli Distribution

- 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 (başarı- başarısız)
- $p$ : başarı olasılığı
- $1-p$  : başarısızlık olasılığı

Rassal Değişken  $X$ :

$x = 1$  : sonuç başarılı ise

$x = 0$  : sonuç başarısız ise

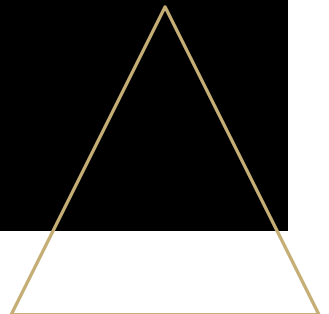
Bernoulli Olasılık Dağılımı:  $p(X = x) = p^x (1 - p)^{1-x}$   $x = 0,1$


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

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

$$E(X) = p$$

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



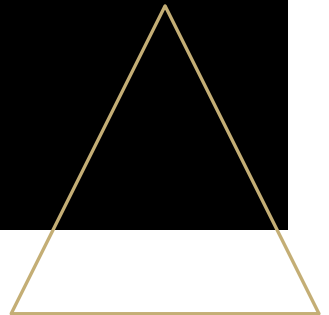


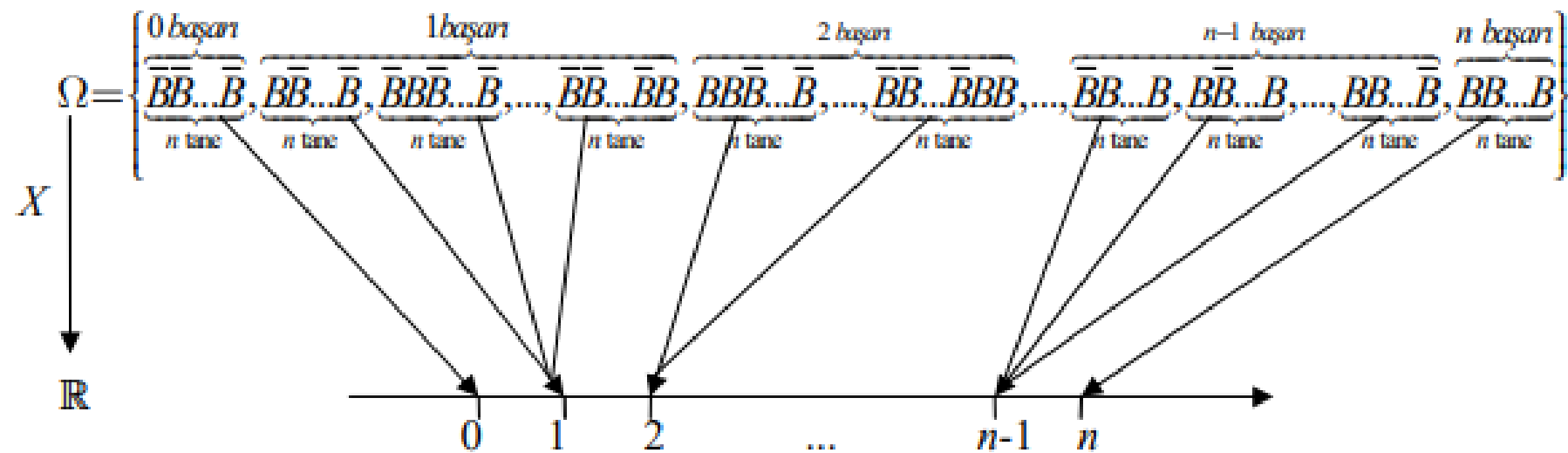
# Binom Dağılım

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.

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





$X$  rasgele değişkeninin aldığı değerlerin kümesi,

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

ve

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

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

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

...

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

olup,  $X$  in olasılık fonksiyonu,

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

dır. Moment çıkaran 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

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

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

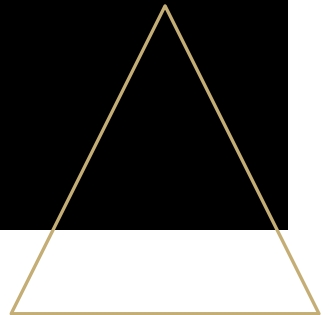
$X$  :  $n$  bağımsız Bernoulli denemesindeki başarıların sayısı.

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



## 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*


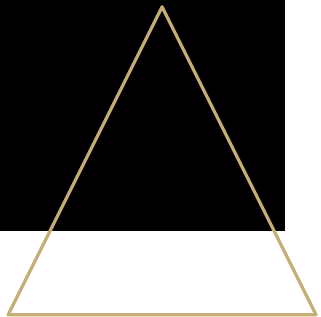
$$\begin{aligned} E(X) &= \sum_{D_X} xP(X = x) \\ &= np(p + q)^{n-1} = np \end{aligned}$$

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

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

# Dikkat

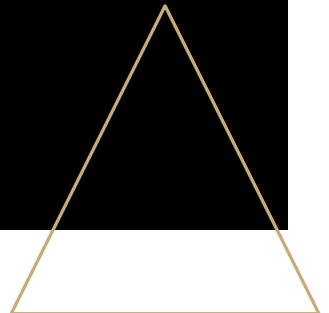
- Bir deneyde seçimler yerine koymadan (**without replacement**) tekrarlanırsa başarı sayısının dağılımı binom dağılımına uymaz.
- Sonlu bir kitleden, örneklem yerine koymadan (**without replacement**) çekilirse, denemler birbirine bağımlı hale gelir. Ve başarı sayısı farklı bir dağılım gösterir.

- 
- 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  $X$

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



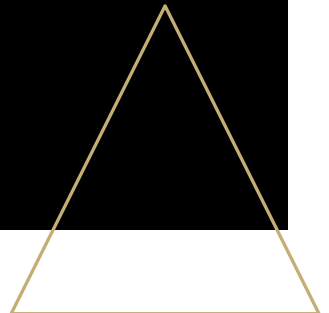
# Geometrik Dağılım (Geometric Distribution)


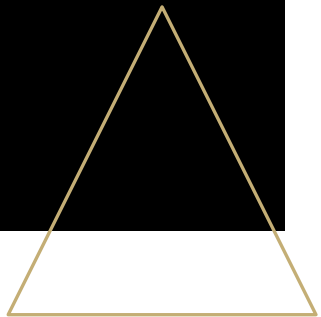
Binom olasılık dağılımında olduğu gibi bir Bernoulli sürecinden türetilen rassal deneylerin çıktısı ile ilgilenilsin. Bu süreçte çıktı ayrık iki olayı (başarı ve başarısızlık) tanımlar ve başarı olasılığı  $p$  deneyden deneye değişmez. Şans değişkeni  $x$  ilk başarı elde edilinceye kadar gerçekleştirilen 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 yaklaşıır.



- 
- 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 başarıdan 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 dağılımın 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 dağılımında 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

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

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

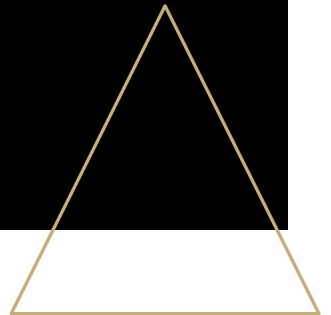
- Beklenen değer ve Varianans:

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



# 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,

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

şeklindedir.  $X$  rasgele değişkeni bu olasılık fonksiyonuna sahipse,  $X$  Poisson dağılımına sahiptir denir ve  $X \sim \text{Poisson}(\lambda)$  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  $\lambda = .50$

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

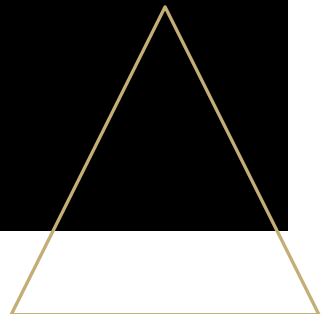
$$\mu = E(X) = \lambda$$


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



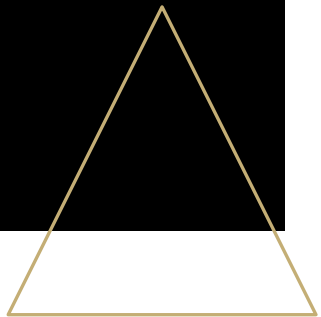
**Teorem 5.1.** Başarı olasılığı  $p$ , 0' a ya da 1'e yaklaştığında ve  $n \rightarrow \infty$  iken  $X \sim \text{Binom}(n, p) \approx \text{Poisson}(\lambda)$  olur. Yani,

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





**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?
- 

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

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

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](mailto:yildizcakar.cemile@gmail.com)

Cemile YILDIZÇAKAR

A life without love  
is like a year  
without summer.

A SWEDISH PROVERB