Olasılık ve İstatistik

Fırat İsmailoğlu, PhD

Merkezi Limit Teoremi

Daha önce istatistiği iki başlık altında toplamıştık: Tanımlayıcı İstatistik ve Çıkarımsal İstatistik:

İstatistik

Tanımlayıcı İstatistik (Descriptive Statistics)	Çıkarımsal İstatistik (Inferential Statistics)
Elimizdeki veriyi/numuneyi (sample'ı) tanımlar. Merkezi eğilimi(mod, ortalama) ve dağınıklığı ölçer.	Elimizdeki verinin toplandığı populasyon ile ilgili (büyük resim ile ilgili) çıkarım yapmamızı sağlar.

Bu haftaya kadar tanımlayıcı istatistik ile ilgilendik. Elimizdeki veriyi anlamaya çalıştık. Şimdi ise çıkarımsal istatistik ile ilgileneceğiz. Verinin toplandığı populasyon hakkında bir çıkarım yapmaya çalışacağız.

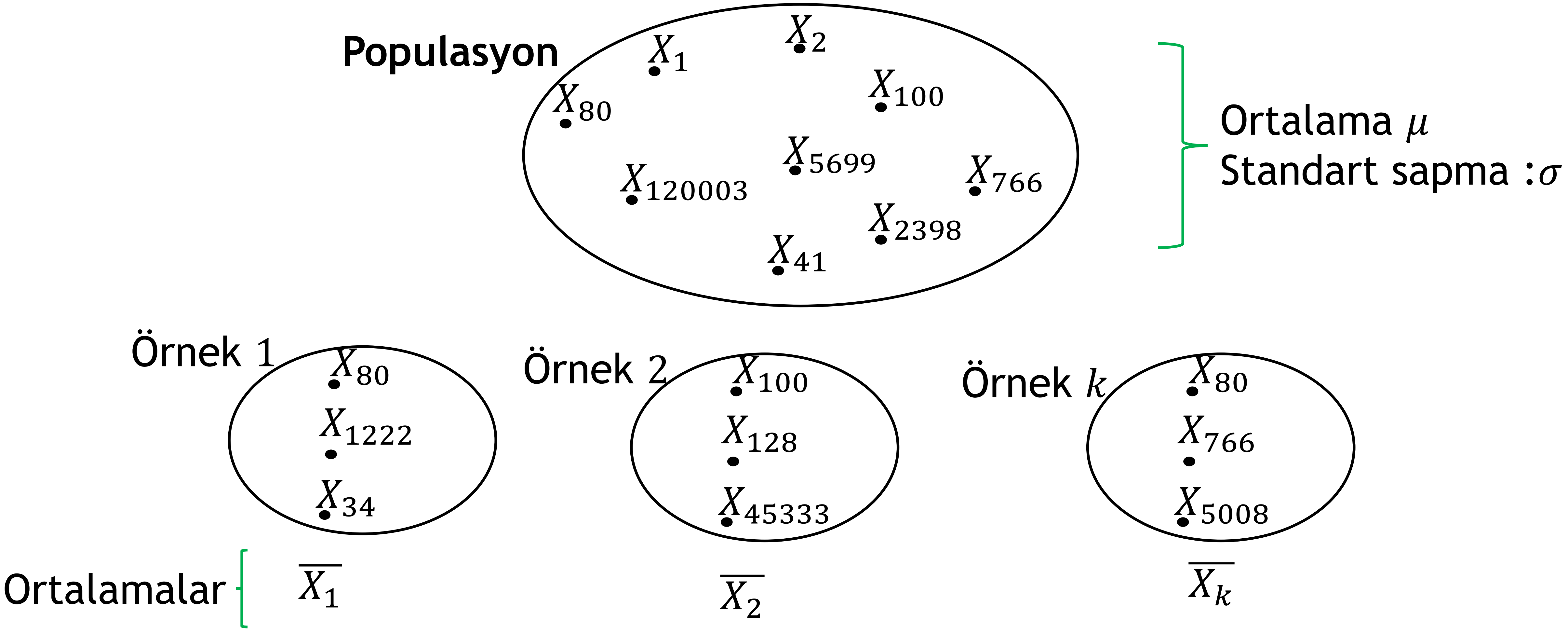
Merkezi Limit Teoremi (Central Limit Theorem)

Populasyonun her bir bireyine bir sayı eşleyelim. Bu sayı örneğin bir kisinin boyu, kilosu, vucut sicakliği; yada bir fabrikada üretilen bir malin raf ömrü olabilir. Bu sayıları X_i rastgele değişkeni ile gösterelim (i. bireyin sayısal değeri). Bu rastgele değişkenlerin ortalaması μ ve standart sapması σ olsun. (Bunlar populasyonun parametreleri).

Bu populasyondan aynı büyüklükte k tane rastgele örnekler (sample'lar) alalım. Örneğin bu büyüklük n olsun. Bu örneklerin ortalamalarını $\bar{X}_i$ ile gösterelim ($\bar{X}_i, i$. örneğin ortalaması).

- Örneklerin ortalamaları $\overline{X_1}$, $\overline{X_2}$,…, $\overline{X_k}$ 'da bir rastgele değişkedir ve normal dağılıma sahip olur.
- Ortalamaların beklenen değeri populasyonun ortalaması μ'ye eşit olur: $E[\bar{X}] = \mu$.
- Ortalamaların standart sapması populasyonun standart sapmasının $\sqrt{n}$'e bölünmesiyle elde edilir: $\frac{\sigma}{\sqrt{n}}$

Merkezi Limit Teoremi (Central Limit Theorem)



$$\frac{\overline{X_1} + \overline{X_2} + \cdots + \overline{X_k}}{k} \cong \mu$$

ör.30 yaşındaki erkeklerin büyük tansiyonlarının ortalaması 120, standart sapması 10 olsun. Bu durumda rastgele seçilen 9 kişilik bir grubun büyük tansiyonlarının ortalamasının 130'dan fazla olma olasılığı nedir?

Çözüm.

30 yaşındaki tüm erkeklerin kümesi populasyonumuz olur. Bu populasyondaki kişilerin büyük tansiyonlarının ortalaması $\mu = 120$ ve standart sapması $\sigma = 10$ olarak verilmiş.

$\bar{X}$ rastgele değişkeni, 30 yaşındaki erkeklerden oluşan 9 kişilik herhangi bir grubun buyuk tansiyonlarının ortalaması olsun. Biz $P(\bar{X} > 130)$ olasılığı ile ilgileniyoruz.

Merkezi limit teoremi gereği $\bar{X}$ rastgele değişkenleri normal dağılıma sahiptir.

$$P(\bar{X} > 130) = P\left(Z > \frac{130 - 120}{\frac{10}{\sqrt{36}}}\right) = P(Z > 3) = 1 - P(\leq 3) = 1 - 0.9987 = 0.0013$$

ör.Bir fabrikada çalışan işçilerin kolestrol seviylerinin ortalaması 202, standart sapması 14 olsun.

i) Işçilerden 36 kişilik bir grup oluşturulduğunda, bu grubun kolestrol seviyelerinin 198 ile 206 arasında olma olasılığı nedir?

ii) Işçilerden 64 kişilik bir grup oluşturulduğunda bu grubun kolestrol seviyelerinin yine 198 ile 206 arasında olma olasılığı nedir?

Çözüm.

i) 36 kişilik grupların kolestrol seviylerinin ortalamasını gösteren $\bar{X}$ rastgele değişkeni merkezi limit teoremi gereği normal dağılıma sahip olur. O halde

$$P(198 \leq \bar{X} \leq 206) = P\left(\frac{198-202}{14/\sqrt{36}} \leq Z \leq \frac{206-202}{14/\sqrt{36}}\right) = P(-1.714 \leq Z \leq 1.714)$$

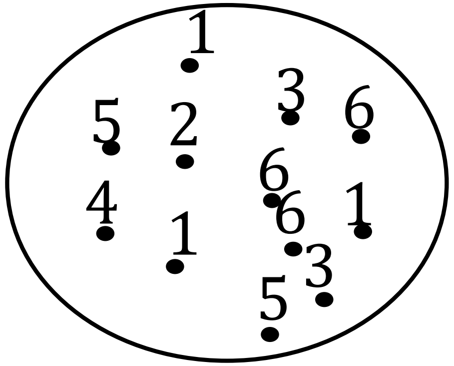
$$= P(Z \leq 1.714) - P(Z \leq -1.714) = 0.9564 - 0.0436 = 0.9128$$

ii) $P(198 \leq \bar{X} \leq 206) = P\left(\frac{198-202}{14/\sqrt{64}} \leq Z \leq \frac{206-202}{14/\sqrt{64}}\right) = P(-2.286 \leq Z \leq 2.286) = 0.995$

Sonuç olarak grubun büyüklüğü artırıldığında grubun ortalaması populasyon ortalamasına yaklaşır.

Örneklerin büyüklükleri arttıkça (n arttıkça) örneklerin ortalamaları, populasyon ortalaması olan μ'ye daha çok yaklaşır; ve örnek ortalamalarının standart sapması azalır, yani örneklerin ortalamarı birbirine daha çok benzer ve μ etrafında toplanır.

Bunun bir uygulaması olarak, herbir örnek (sample) bir zarın 50 defa atılmasının sonuçlarını içersin. Böylece her örnekte 50 eleman olur ve bunlar $1-6$ arası değerlerdir. Bu şekilde 100 tane örnek oluşturalım. Yani 100 tane 50 büyüküğünde örnek var olsun.



100 adet içinde 50 tane $1-6$ arası sayı olan bu şekilde örnek var.

Bu örnekleri R programlama dilinde yada başka bir dilde kolayca oluşturabiliriz. R'de herbir örnek için:

```
>>sample(1:6,50, replace=True)
```

Bu örneklerin ortalamasını basitçe **mean** komutuyla alabiliriz:

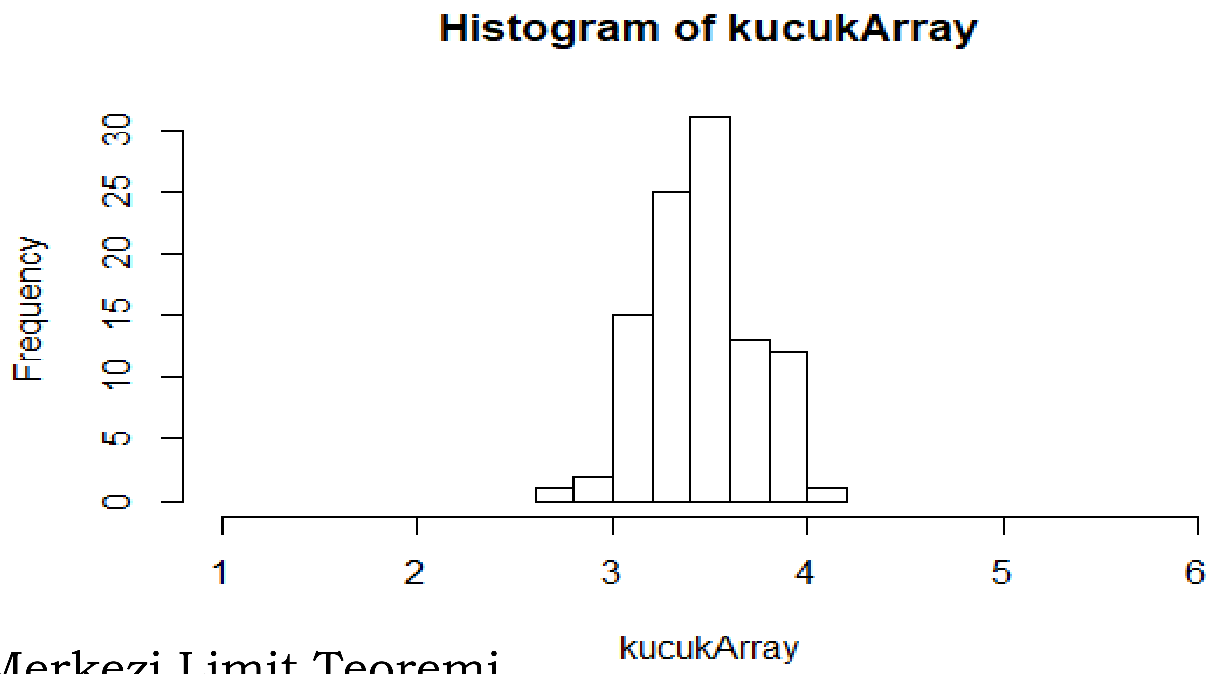
```
>>mean(sample(1:6,50, replace=True))
```

Üretilen her örneğin ortalamasını, bir arrayde tutalım:

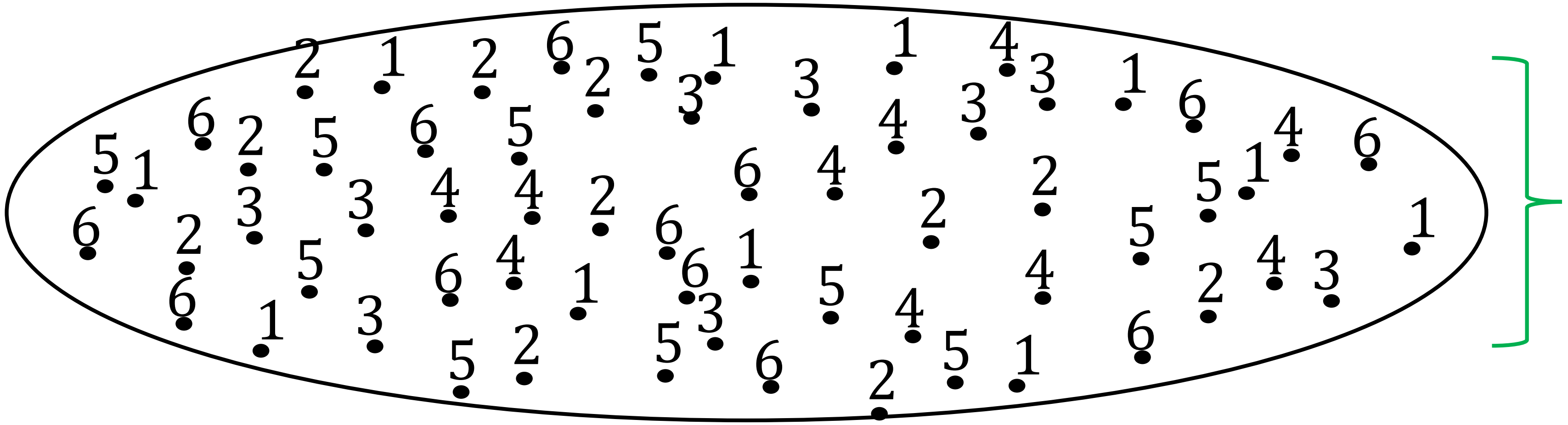
```
>>kucukArray=c(1:100) # once arrayi tanimlayalim.
>> for (i in 1:100){
kucukArray[i]=mean(sample(1:6,50, replace=T))}
```

Bu array'in histogramını çizerek 50'lik örneklerde ortalamanın daha çok hangi değerler aldığını görelim

```
>>hist(kucukArray,xlim=c(1,6))
```



Şimdi de herbir örnek (sample) bir zarın 1000 defa atılmasının sonuçlarını içersin. Böylece her örnekte 1000 eleman olur ve bunlar $1-6$ arası değerlerdir. Bu şekilde 1000 tane örnek oluşturalım.



100 tane içinde 1000 tane $1-6$ arası sayı olan bu şekilde örnek var.

Bu örnekleri R'de

```
>>sample(1:6,1000, replace=True)
```

kodu ile oluşturabiliriz.

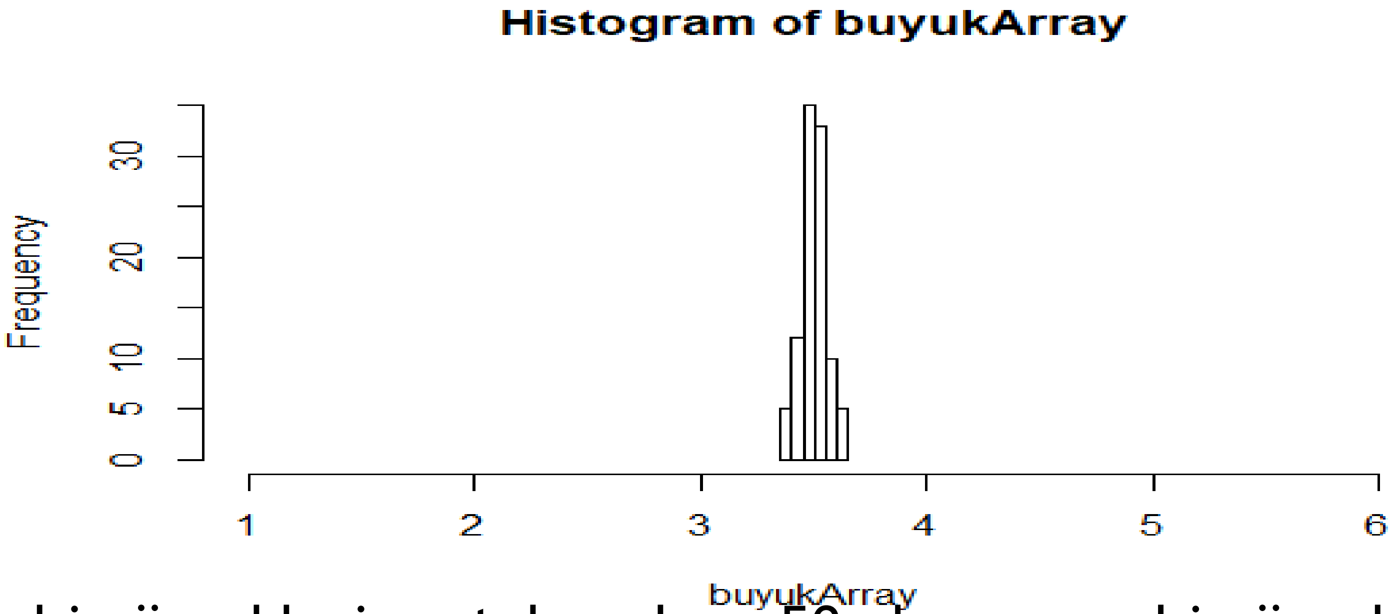
Oluşan örneklerin ortalamasını bir arrayde tutatlım:

```
>> buyukArray=c(1:100)
```

```
>> for (i in 1:100){
buyukArray[i]=mean(sample(1:6,1000, replace=T))}
```

Bu array'in histogramını çizerek 1000'lik örneklerde ortalamanın daha çok hangi değerler aldığını görelim:

```
>>hist(buyukArray,xlim=c(1,6))
```



1000 elemana sahip örneklerin ortalamaları, 50 elemana sahip örneklerin ortalamalarına göre gerçek populasyon ortalaması olan 3.5 değerine çok daha yakın ve bu ortalamarın birbirinden uzaklığı (sapmaları) daha az.

```
>> for (i in 1:100){
buyukArray[i]=mean(sample(1:6,1000, replace=T))}
```

Bu array'in histogramını çizerek 1000'lik örneklerde ortalamanın daha çok hangi değerler aldığını görelim:

```
>>hist(buyukArray,xlim=c(1,6))
```

ör. Pil üreten bir fabrikanın ürettiği pillerin ömrünün ortalaması μ, standart sapması 100 olsun. Rastgele 10 pil alındığında bu pillerin ömürlerinin ortalamalarının μ'den 20 fazla olma olasılığı ne olur?

Çözüm.

$\bar{X}$ rastgele değişkeni 10 büyüklüğündeki pil gruplarının ortalama ömrü olsun.Merkezi limit teoremi gereği $\bar{X}$ normal dağılıma sahiptir.

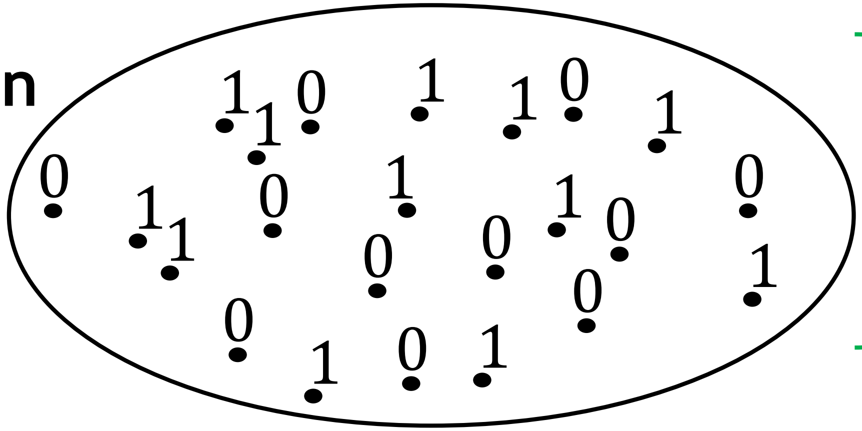
$$P(\bar{X} \geq \mu + 20) = P\left(Z \geq \frac{\mu + 20 - \mu}{\frac{100}{\sqrt{10}}}\right) = P(Z \geq 0.63) = 1 - P(Z < 0.63)$$

$$= 1 - 0.7357$$
$$= 0.2643$$

Merkezi Limit Teoremi Yardımıyla Normal Dağılım Kulllanarak Binomial Dağılıma Yaklaşmak

Bernoulli rastgele değişkenlerinden oluşan bir populasyonumuz olsun. (Yani populasyonda yalnızca 0 (başarısızlık) ve 1 (başarı) değerleri var. Buradan n elemana sahip bir örnek alalım.

Populasyon



Populasyon ortalaması: p
Populasyon standart sapması: $\sqrt{p(1-p)}$

Örnek

$X_1 = 0$
$X_2 = 1$
$X_n = 1$

Bu örneğin ortalaması:

Toplam başarı sayısı!

$$\bar{X} = \frac{\boxed{X_1 + X_2 + \cdots + X_n}}{n} = \frac{X}{n}$$

Merkezi limit teoremi gereği $\bar{X} = \frac{X}{n}$ normal dağılıma sahip olur. Bunu standart

normal dağılımdaki karşılığı: $Z = \dfrac{\frac{X}{n} - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = \dfrac{\frac{X-np}{n}}{\sqrt{\frac{p(1-p)}{n}}} = \dfrac{X-np}{\sqrt{np(1-p)}}$

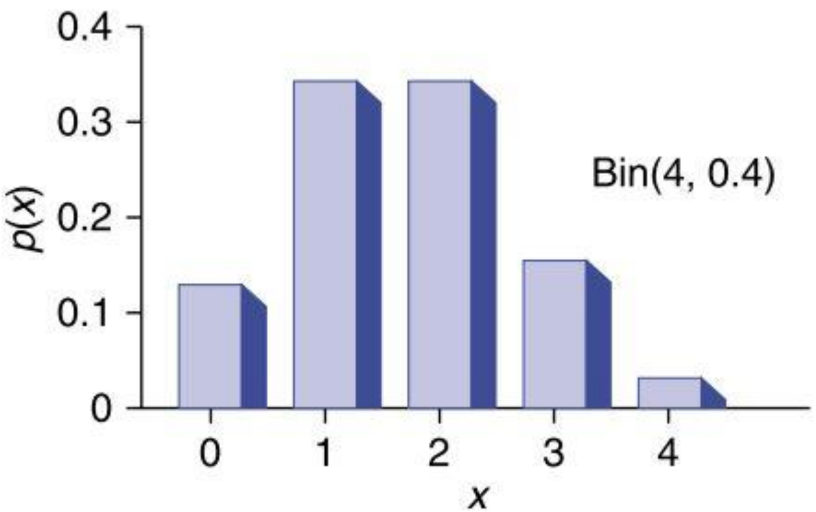
$\dfrac{X-np}{\sqrt{np(1-p)}}$ yaklaşık olarak normal dağılıma sahip olur. Burada X, toplam başarı sayısını gösterdiğinden binomial rastgele değişken olur.

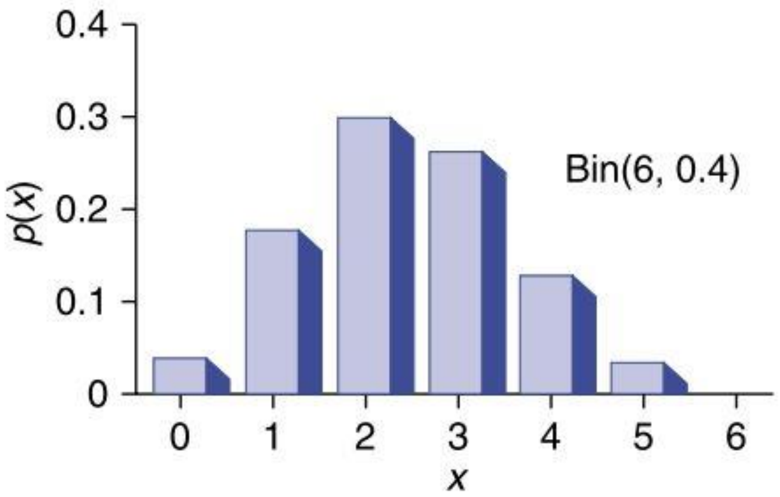
ör. Varsayalımki bir populasyonun $\%46$'sı bir A adayını desteklesin. Bu populasyondan rastgele seçeceğimiz 200 kişinin en az 100'ünün A adayını destekleme olasılığı nedir?

Çözüm.

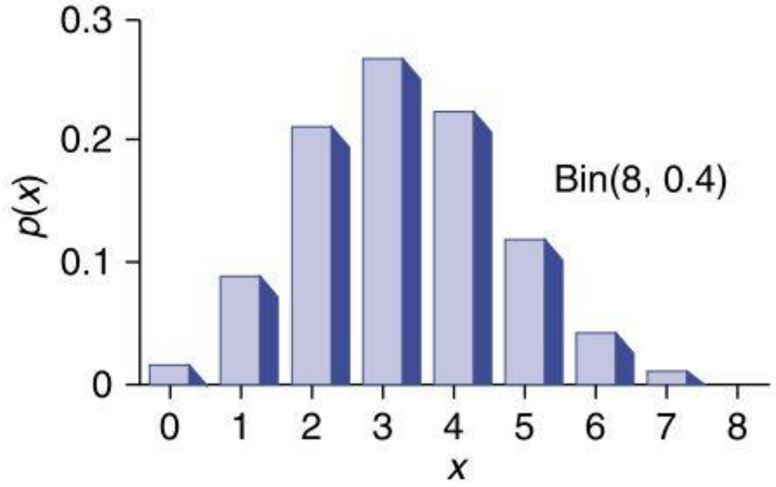
X rastgele değişkeni seçilen 200 kişiden A'yı destekleyen toplam kişi sayısını göstersin. $Biz\ P(X \geq 100)$ ile ilgileniyoruz.

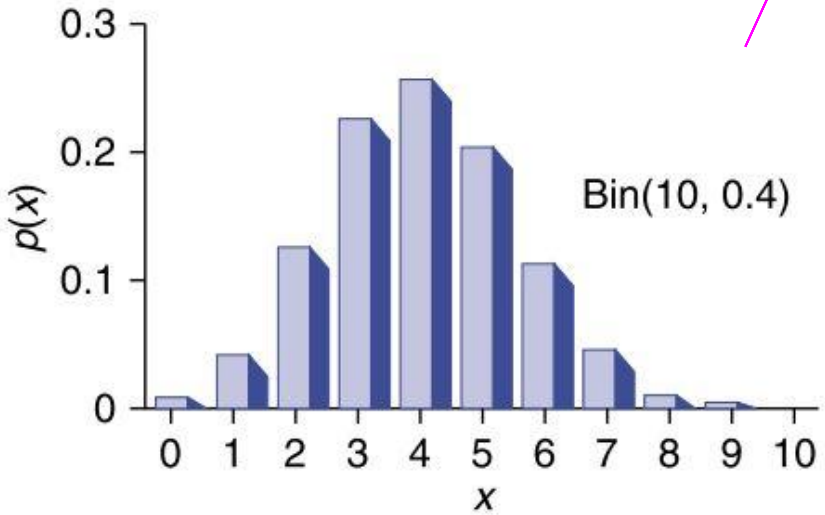
$$P(X \geq 100) = P\left(Z \geq \frac{100 - 200 \times 0.46}{\sqrt{200 \times 0.46(1 - 0.46)}}\right) = P(Z \geq 1.135) = 1 - P(Z < 1.135)$$
$$= 0.0885$$





n (yani deney tekrar sayısı) arttıkça binomial dağılım , normal dağılıma benzer.





ör. Bir fabrikanın ürettiği malların %10'u bozuk olsun. Bu fabrikanın ürettiği mallardan rastgele seçilen 8 malın,

i) hiçbirinin

ii) %15'inden fazlasının

bozuk olma olasılığı nedir?

Çözüm.

X rastgele değişkeni bu fabrikanın ürettiği mallardan rastgele seçilen 8 maldaki toplma bozuk sayısı olsun.

$p = 0.1$ başarı oranı

$i)\quad P(X = 0) = (1 - 0.1)^9 = 0.43$

$$ii)\quad P\left(X > 8 \times \frac{15}{100}\right) = P(X > 1.2) = P\left(Z > \frac{1.2 - 8 \times 0.1}{\sqrt{8 \times 0.1 \times (0.9)}}\right) = P(Z > 0.471) =$$

$$= 1 - P(Z \leq 0.471) = 1 - 0.68 = 0.32$$