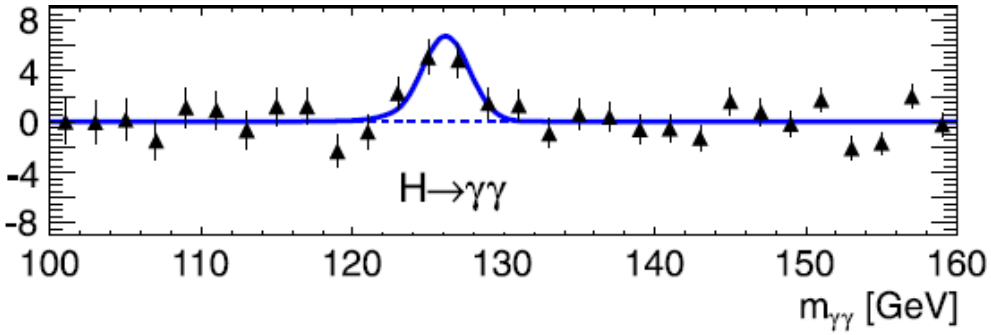
YBS514
Mühendislikte İstatistik Yöntemler

Bölüm 8

Örnekleme ve Istatistiksel Tahminleme



http://www1.gantep.edu.tr/~bingul/stat

Gaziantep Üniversitesi

Yönetim Bilişim Sistemleri

Tezsiz Yüksek Lisans Programı

Aralık 2020

İçerik

- Örnekleme
- İstatistiksel Tahminleme

Örnekleme
(Sampling)

Neden Örnekleme?

Bir istatistiksel araştırma sürecininde

- Önce ilgili anakütle tanımlarır.
- Sonra, ana kütlenin ilgilenilen parametreleri hakkında bilgi üretilmeye çalışılır.

- **A planı**: mümkünse, tanımlanan ana kütle için veri toplarken tam sayım yapılmalıdır.
  (*Zaman, bütçe, araştırmacı sayısı, araştırma şekli, vb. sorunlar yüzünden, bu her zaman mümkün olmaz!*,)
- **B planı**: Ana kütleden sınırlı sayıda örnekler alınır ve örnekleme istatistiğinden ana kütle hakkında fikir edinilir.

Anakütle & Örnek

İstatistikte anakütle ve örnek arasındaki farkı iyi anlamak gerekir.

Anakütle bir kümenin bütün elemanlardır.

Örnek bu grup içinden seçilmiş bir alt kümedir.

- Örnek(ler) anakütlenin önemli karakteristik özelliklerini (ortalama, varyans, standart sapma) ortaya koymak veya tahmin etmek için kullanılır.
- İsteğe bağlı, örneğin hacmi anakütle hacminin %1, %10 veya %50 kadar olabilir. Ama, ana kütleye hiçbir zaman eşit olamaz.

Neden Tam Sayımdan Kaçınırız?

Tam sayım, sonlu bir ana kütlenin bütün birimlerinin sayılmasıdır.

Araştırma: Yaklaşan yerel seçimlerde yurttaşların kararı.

Ana kütle: Gaziantep'de yaşayan oy kullanacak $7x10^5$ kişi

Parametre: Adayların oy oranları

Örneklem: 10 farklı mahalleden rastgele seçilen toplam 1600 kişi

Not:

Anket sonuçlarındaki hata = *sqrt(1600) = 40*

Göreli hata = *40/1600 = 0.025 = %2.5 (yanılma payı)*

Neden Tam Sayımdan Kaçınırız?

Araştırma: Bir üniversitedeki öğrencilerin kendilerine sunulan hizmetten memnun olup olmadıkları ile ilgili anket.

Ana kütle: Bütün öğrenciler (Anakütle hacmi 50 000 kişi)

Parametre: Memnuniyet (E/H)

Örneklem: 500 öğrenci

Neden Tam Sayımdan Kaçınırız?

Araştırma: Bir fabrikada üretilen cıvataların kusurlu üretilme yüzdesini hesaplamak.

Ana kütle: Cıvatalar (Anakütle hacmi tam bilinmiyor)

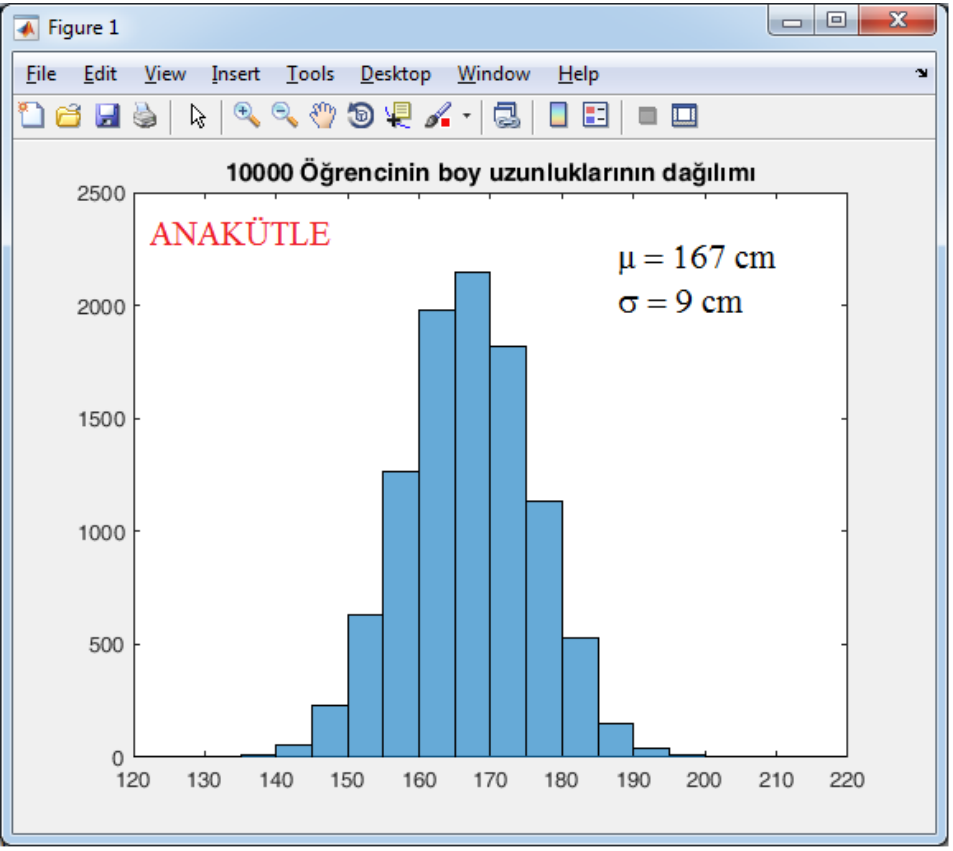
Parametre: Kusur yüzdesi

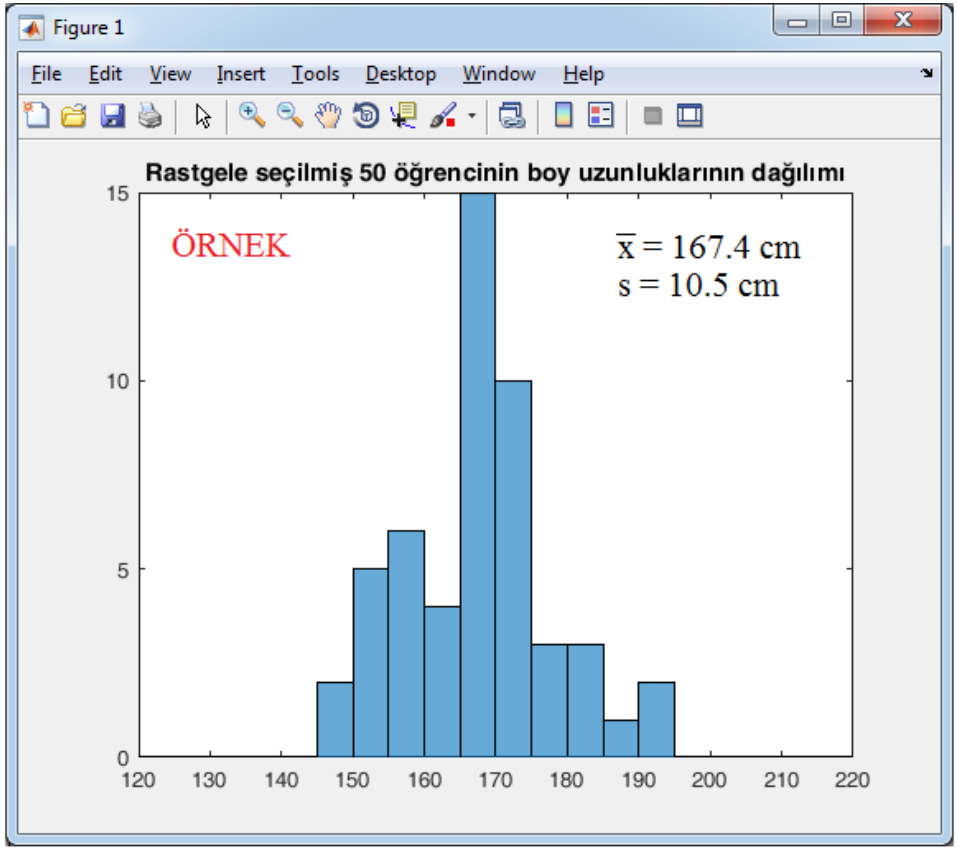
Örneklem: 6 gün boyunca her gün rastgele zamanlarda seçilen 20 cıvata. Toplam 120 cıvata.

Örnekleme Yöntemleri

- Olasılıklı olmayan örnekleme yöntemleri
- Yargısal Örnekleme
- Kota örneklemesi
- Kartopu örneklemesi
- Rassal Örnekleme

Örnekleme Dağılımı





μ = Anakütle ortalaması

σ = Anakütle std.sapması

$\overline{x}$ = Örneklem ortalaması

s = Örneklem std.sapması

Ortalama ve Standart Sapma

Anakütle ortalaması ve standart sapması:

$$\mu = \frac{anakutle\ elemanlarinin\ toplami}{anakutle\ boyutu} = \frac{\sum x_i}{N} \qquad \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Burada *N* anakütledeki eleman sayısı veya anakütle boyutudur.

Örneklem ortalaması ve standart sapması:

$$\bar{x} = \frac{örnekteki\ elemanlarin\ toplami}{örnegin\ boyutu} = \frac{\sum x_i}{n} \qquad s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Burada *n* örnek eleman sayısı veya örnek boyutudur.

Standart Hata

Bir örneğin ortalaması $\overline{X}$ olsun. $\overline{X}$ değerindeki belirsizlik (standart hata), örneklem hacmi arttıkça küçülür.

Ayrıca, *n* değeri büyüdükçe $\overline{X}$ değeri μ değerine yaklaşır.

Buna göre, standart hata aşağıdaki gibi hesaplanır.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Örnek 1

Bir sıvının yoğunluğu (g/cm^3 cinsinden) 10 farklı yöntemle ölçülmüş ve aşağıdaki veri elde edilmiştir.

$$x = \{1.10,\ 1.12,\ 1.09,\ 1.09,\ 1.07,\ 1.14,\ 1.11,\ 1.16,\ 1.07,\ 1.08\}$$

$$\overline{x} = (1.10 + 1.12 + 1.09 + 1.09 + \cdots + 1.08)/10 = 1.103\ \text{g/cm}^3$$

$$s = \sqrt{[(1.10 - 1.103)^2 + (1.12 - 1.103)^2 + \cdots + (1.08 - 1.103)^2]/9} = 0.030\ \text{g/cm}^3$$

$$s^2 = 0.0009\ (\text{g/cm}^3)^2$$

$$s_{\overline{x}} = 0.03/\sqrt{10} = 0.009\ \text{g/cm}^3$$

Merkezi Limit Teoremi

Ana kütlenin dağılım şekli ne olursa olsun, basit rassal örneklem hacmi büyüdükçe (n>20), $\bar{X}$ 'in örnekleme dağılımı normal dağılıma yaklaşır. Bu dağılmın ortalaması μ ve varyansı σ^2 / n'dir.

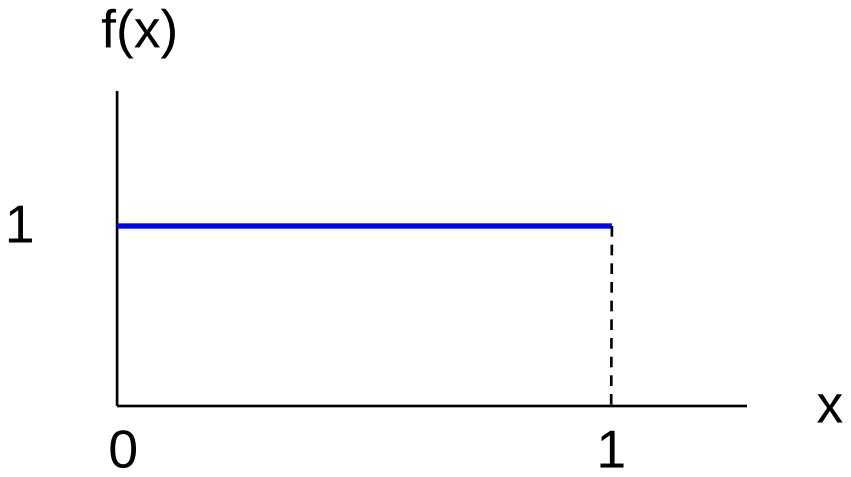
Yani, herhangi bir dağılımdan seçilen örneklere ait ortalamaların ($\bar{X}$ değerlerinin) dağılımı, ortalaması μ ve varyansı σ^2 / n olan normal bir dağılımdır.

Ayrıca, aşağıdaki dönüşümle, söz konusu dağılım standart normal dağılıma dönüşür.

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Örnek 2

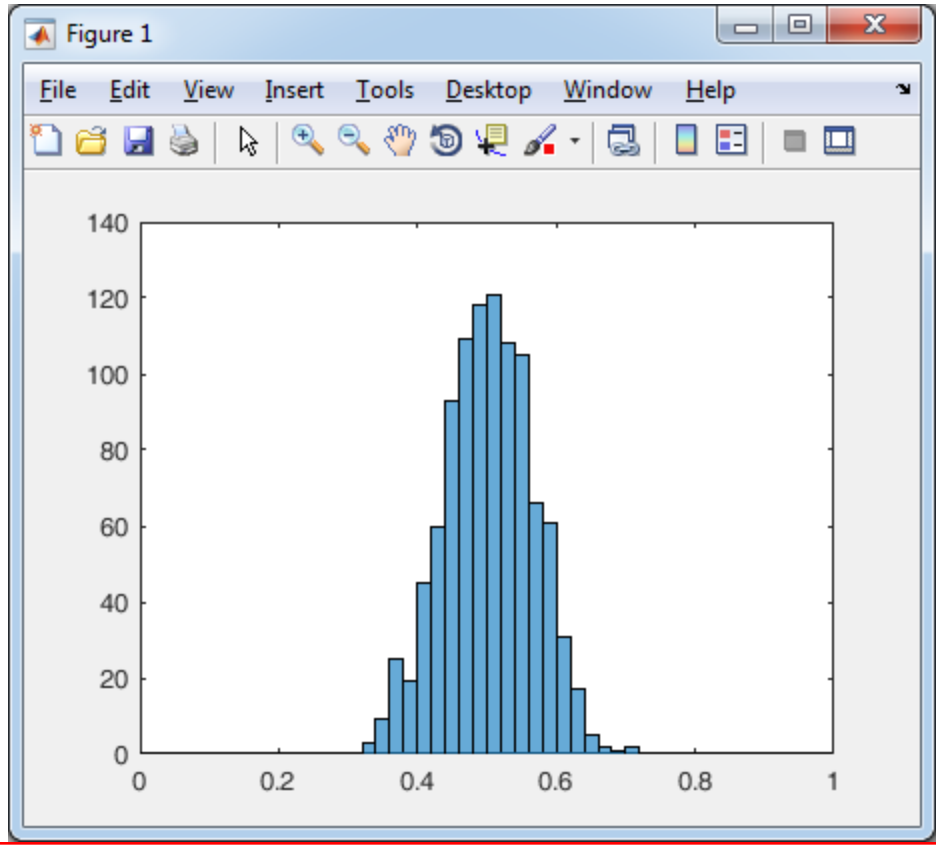
(0, 1) arasında düzgün dağılmış rastgele sayıları düşünelim. Bu dağılımdan, herbirinin hacmi n = 20 olan rastgele m = 1000 adet örnek alalım.




```
% merkezi limit teoremi
m = 1000; % birbirinden farklı örnek sayısı
n = 20;   % her örneğin hacmi
x = rand(n,m);
Ortalama = mean(x);     % ortamaların saklandığı dizi (1000 elemanlı)
Sapma = std(Ortalama); % ortalamaların standart sapması
histogram(Ortalama,'BinEdges',0:0.01:1)
fprintf('Hesaplanan sapma = %f\n',Sapma);
fprintf('Beklenen   sapma = %f\n',(1/sqrt(12))/sqrt(n));
```

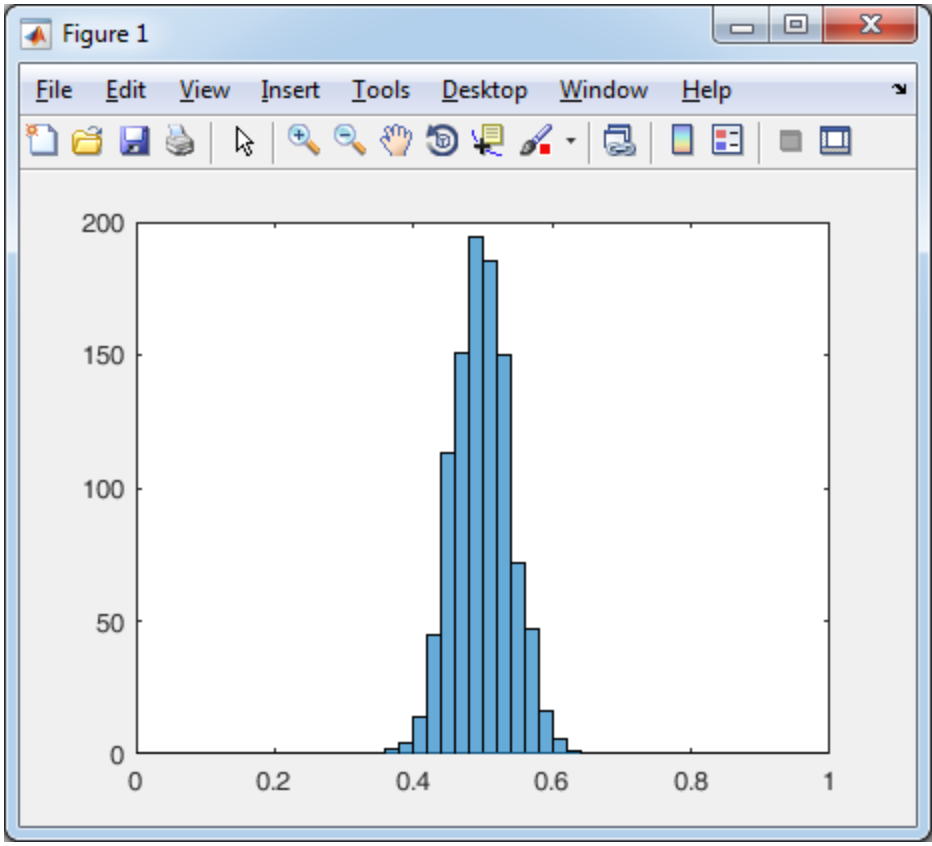
Örnek2 - devam

```
m = 1000;
n = 20;
Hesaplanan sapma = 0.063602
Beklenen   sapma = 0.064550
```



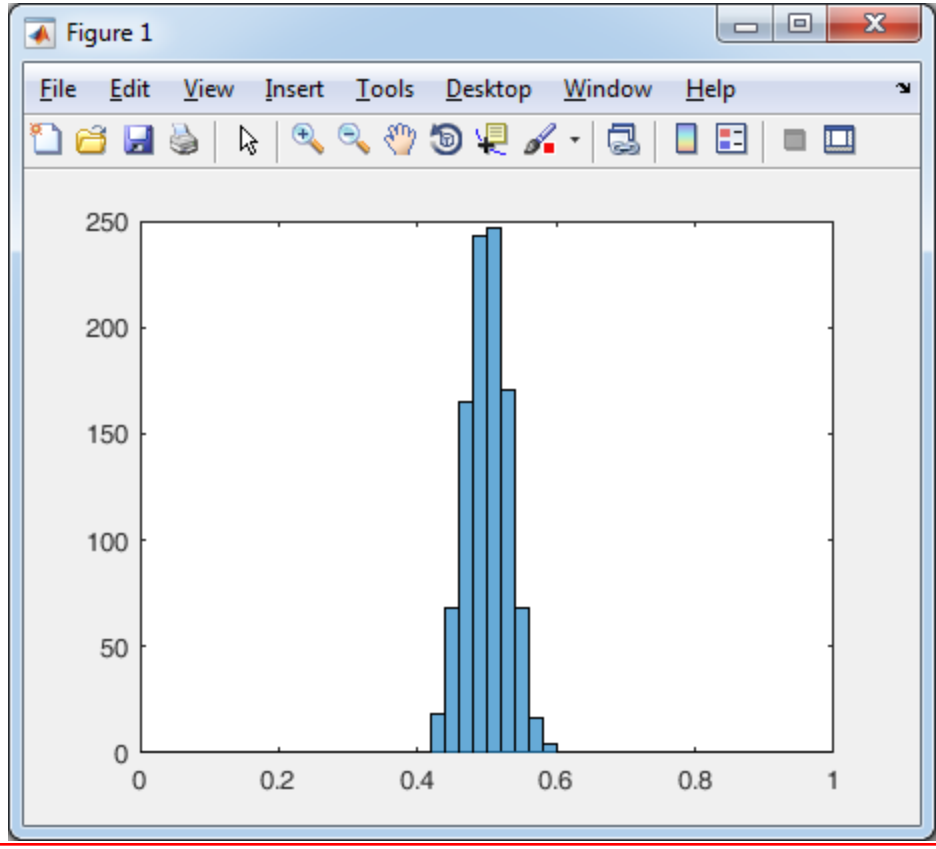
Örnek2 - devam

```
m = 1000;
n = 50;
Hesaplanan sapma = 0.040042
Beklenen   sapma = 0.040825
```



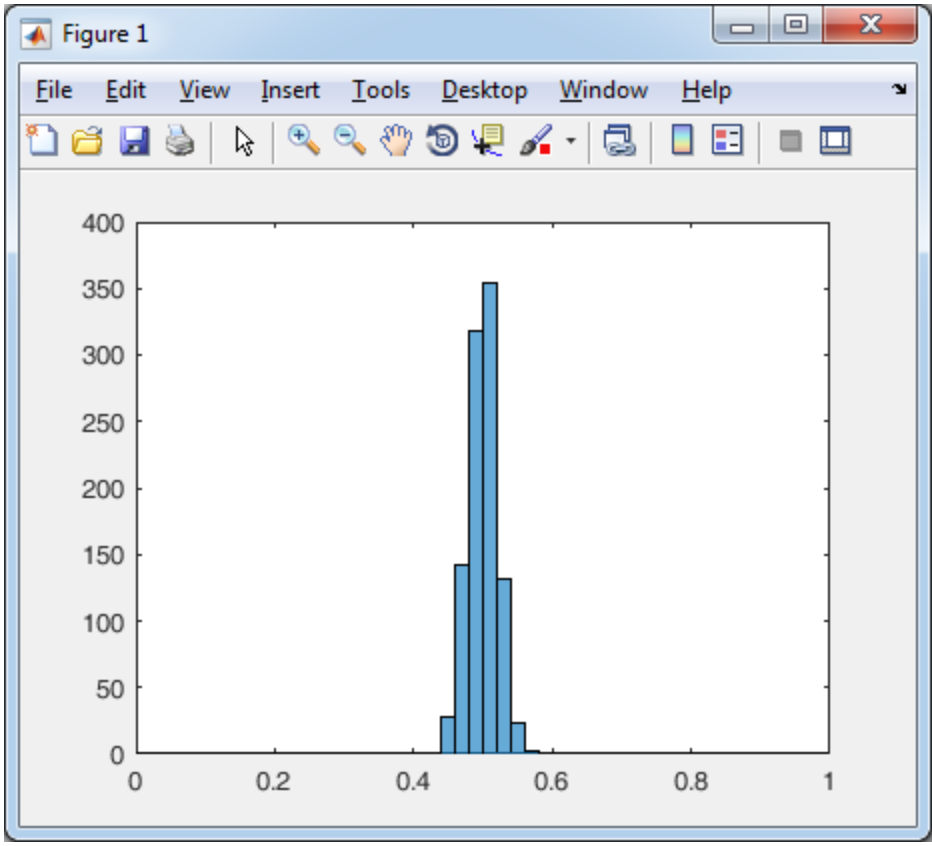
Örnek2 - devam

```
m = 1000;
n = 100;
Hesaplanan sapma = 0.028826
Beklenen   sapma = 0.028868
```



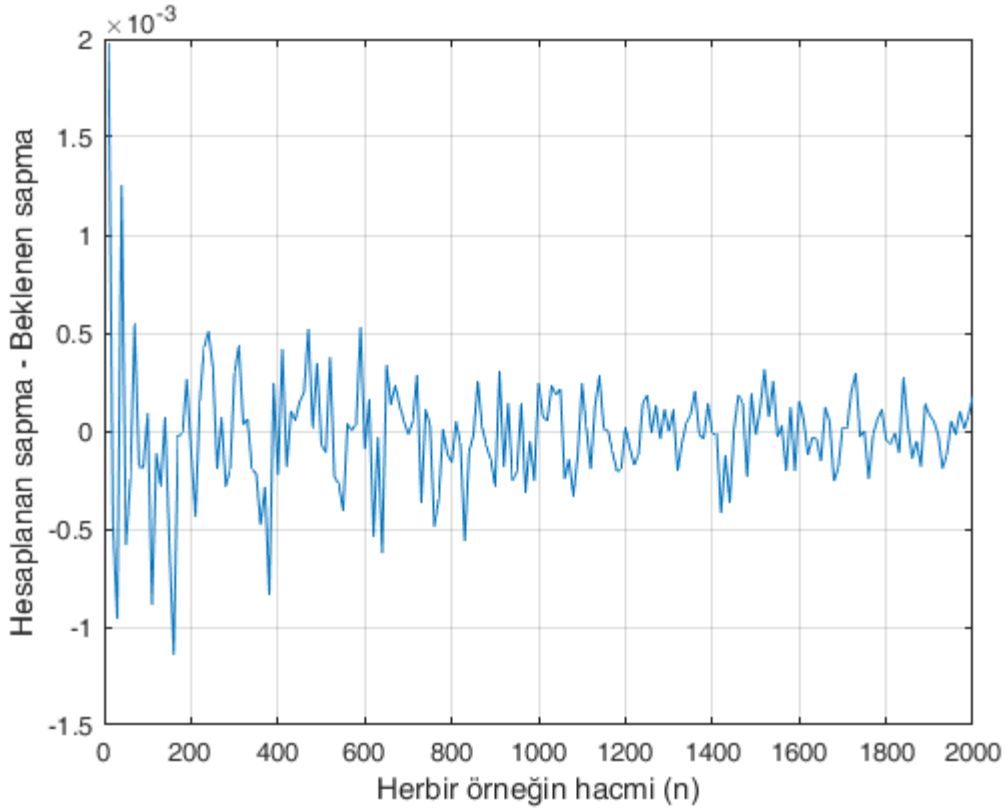
Örnek2 - devam

```
m = 1000;
n = 200;
Hesaplanan sapma = 0.020422
Beklenen   sapma = 0.020412
```



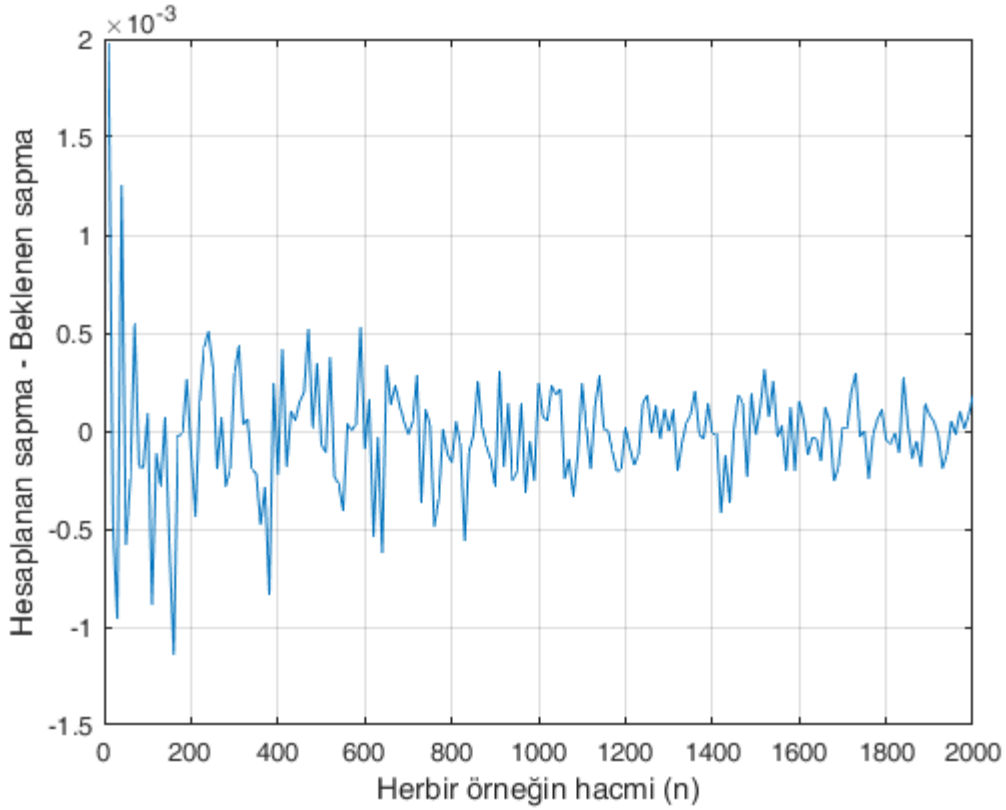
Örnek 2 - devam

n değerine karşılık sapmaların farkları



Alıştırma 1

Aşağıdaki grafiği çıkaran bir program yazın



Alıştırma 2

Örnek 2'deki analizi m = 1000 ve n = 50 için bir üçgen dağılım için tekrarlayın.

MATLAB'da üçgen dağılım elde etmek için:

```
>> x = sqrt(rand(250,1));
>> histogram(x,20)
```

Istatistiksel Tahminleme
(Statistical Estimation)

Amaç

Örneklemeler ana kütle parametreleri hakkında bilgi üretmek amacıyla seçilir.

Örneklem istatistiği kullanılarak ana kütle parametreleri hakkında genelleme yapma süreci, istatistiksel yorumlamadır.

Tahminleme, tanımlanan ana kütleden seçilen rastgele örneklemden hesaplanan istatistikler yardımıyla, ana kütlenin uyduğu dağılım parametre değerlerini araştırmak demektir.

Mesela, $\bar{x}$ ve s verilirse, μ ve σ değerlerini tahmin etmek.

Tahminleme Türleri

Nokta Tahminlemesi

Bir rassal örneklemden hesaplanan bir parametre a,
ana kütle parametresi A değerine eşit kabul eden tahminleme yöntemidir.

Örnek 3:

Bir şirketin son üç aylık ortalama elektrik gideri $\bar{x}$ = 425 TL ve standart sapması *s* = 50 TL iolarak hesaplanmıştır. Aylık ödeme turarını tahmin ediniz (μ = ?).

Yanıt:

Nokta tahminlemesine göre cevap μ = 425 TL.

Tahminleme Türleri

Aralık Tahminlemesi

Nokta tahminlemesi tahminin güvenilirliği hakkında bilgi vermez. Güvenilirliği somut bir şekilde ortaya koymak için aralık kavramı geliştirilmiştir.

Güvenlik Aralığı (=Confidence Interval)

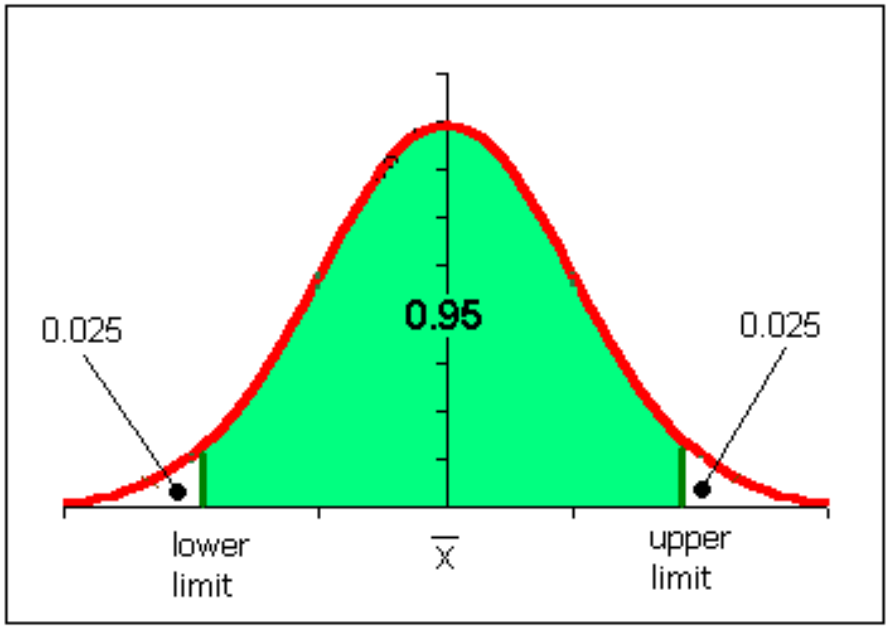
İstatistikte güvenlik aralığı (GA), anakütle parametrelerini tahmin etmede kullanılan bir aralıktır.

Güvenlik Seviyesi (GS) (veya Güvenlik Düzeyi), anakütle parametresini içine alma derecesini gösteren bir sayıdır.

http://en.wikipedia.org/wiki/Confidence_level

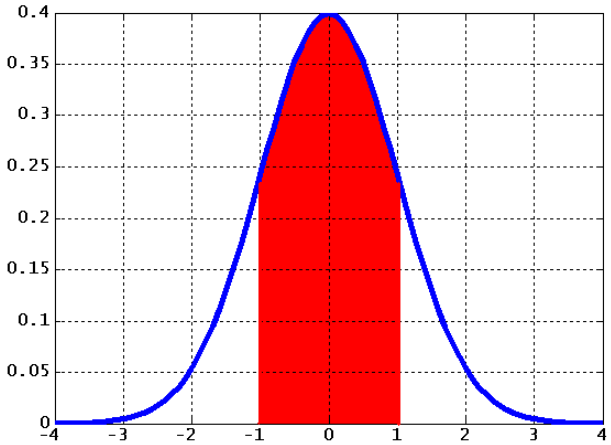
Güvenlik seviyeleri şöyle tanımlanır:

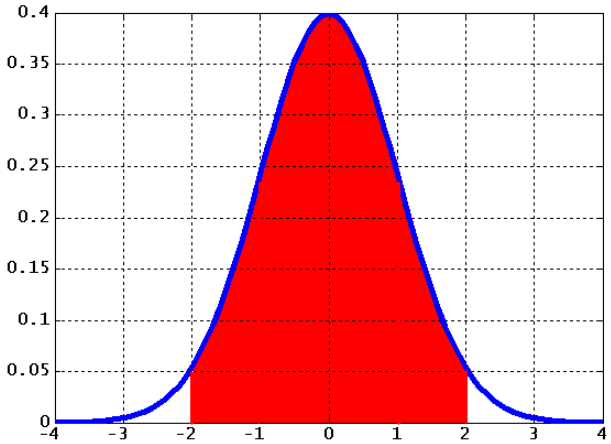
GS	+- sigma
0.800	1.28σ
0.900	1.65σ
0.950	1.96σ
0.990	2.58σ
0.999	3.29σ

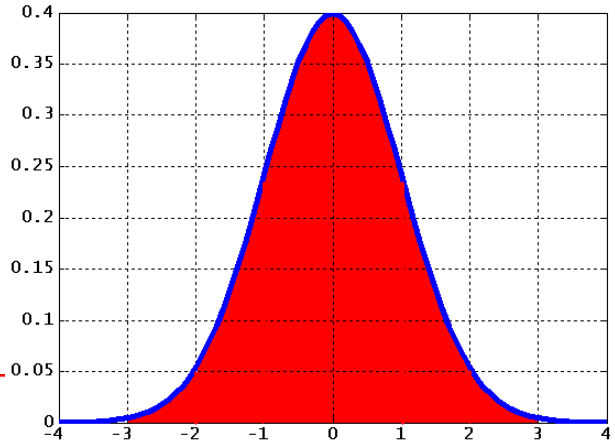


Uygulamada, genellikle
yaklaşık +- 2σ bölgesine karşılık gelen
%95 Güvenlik Seviyesi kullanılır.

CS	Eğri altında kalan alan
%68	$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.6827$
%95	$\int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.9545$
%99.7	$\int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.9973$







Güvenlik seviyeleri şöyle tanımlanır:

α	G.S. $1-\alpha$	$z_{\alpha/2}$
0.01	0.99	2.576
0.05	0.95	1.960
0.10	0.90	1.645
0.20	0.80	1.282
0.40	0.60	0.842
0.50	0.50	0.674

Örnek ortalama: $\bar{x} = \frac{\sum x_i}{n}$

Örnek standart sapma: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Ortalama değerindeki standart hata aşağıdaki gibi tanımlanır:

$$s_x = \frac{s}{\sqrt{n}}$$

Sonuçlar:

+- 1σ hata ile = %68 güvenlik seviyesi ile: $\bar{x} \pm s_x$

+- 2σ hata ile = %95 güvenlik seviyesi ile: $\bar{x} \pm 2s_x$

Örnek 4

Otomobil lastiği üreticisi bir fabrikanın yönetim organı üretilen lastiklerin ortalama ömrünü km olarak tahminlemek istiyor. Bu amaçla rassal olarak 32 lastik seçilmiş lastiklerin ortalama ömrünün 30000 km ve standart sapmasının da 1500 km olduğu tespit edilmiştir. %99 güven düzeyi için istenilen tahminlemeyi yapınız.

Yanıt:

n = 32 lastik,

<x> = 30000 km,

s = 1500 km,

G.D. = %99,

α = 0.01

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{1500}{\sqrt{32}} = 265.2\ \text{km}$$

$$\bar{x} - Z_{\alpha/2} < \mu < \bar{x} + Z_{\alpha/2}$$

$$\bar{x} - (2.576)s_{\bar{x}} < \mu < \bar{x} + (2.576)s_{\bar{x}}$$

$$30000 - (2.576)(265.2) < \mu < 30000 + (2.576)(265.2)$$

$$29317\ \text{km} < \mu < 30683\ \text{km}$$

Örnek 5

Bir bilgisayar programı 36 kez çalıştırılmış ve saniye cinsinden çalışma süreleri (real run-time) aşağıda verilmiştir.

```
371 373 374 362 352 375 384 352 354 358 397 366
396 386 367 350 391 378 378 375 368 367 366 355
366 372 361 365 344 366 381 379 399 396 362 367
```

$\langle x \rangle$ = 371 s, s = 14 s, s_x = 2.3 s

Buna göre ana kütle ortalamasının bulunduğu güven aralıkları:

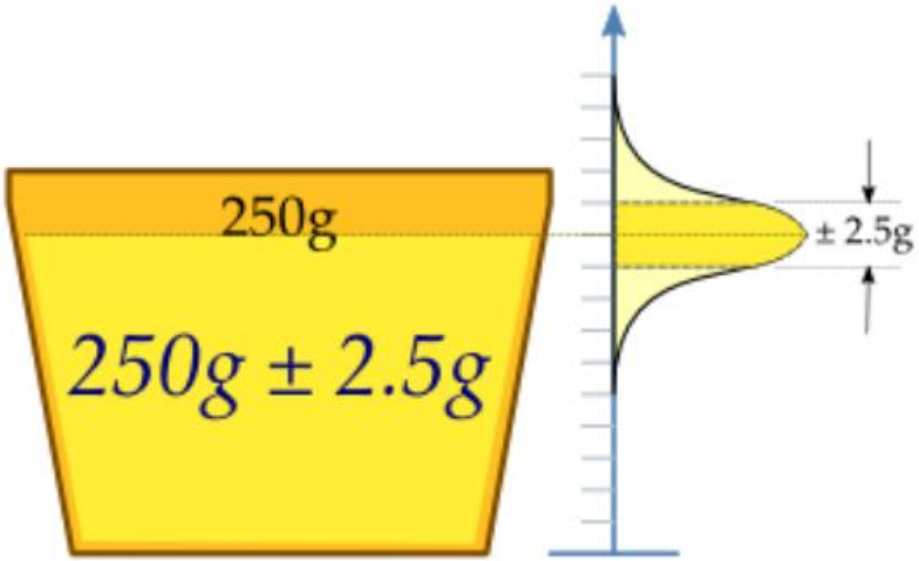
α	$1-\alpha$	$z_{\alpha/2}$	μ Güvenlik aralığı
0.01	0.99	2.576	[364.9894, 377.0106]
0.05	0.95	1.960	[366.4267, 375.5733]
0.10	0.90	1.645	[367.1617, 374.8383]
0.20	0.80	1.282	[368.0087, 373.9913]

Örnek 6:

Bir çay makinasının (üretici firmaya göre) bardak başına verdiği çay miktarı 250 gram olması gerekmektedir. Bunun test etmek için makinadan 20 örnek (numune) alınmış ve aşağıdaki veri toplanmıştır.

X = {247.1, 250.0, 250.1, 249.8, 246.7,
254.4, 249.2, 249.4, 247.0, 247.0,
245.0, 253.3, 251.2, 250.7, 250.6,
247.3, 248.5, 248.0, 243.6, 250.2}

Buna göre, %95 güvenlik seviyesi ile makinanın doğru kalibre edilmiş olup olmadığını sınayın.



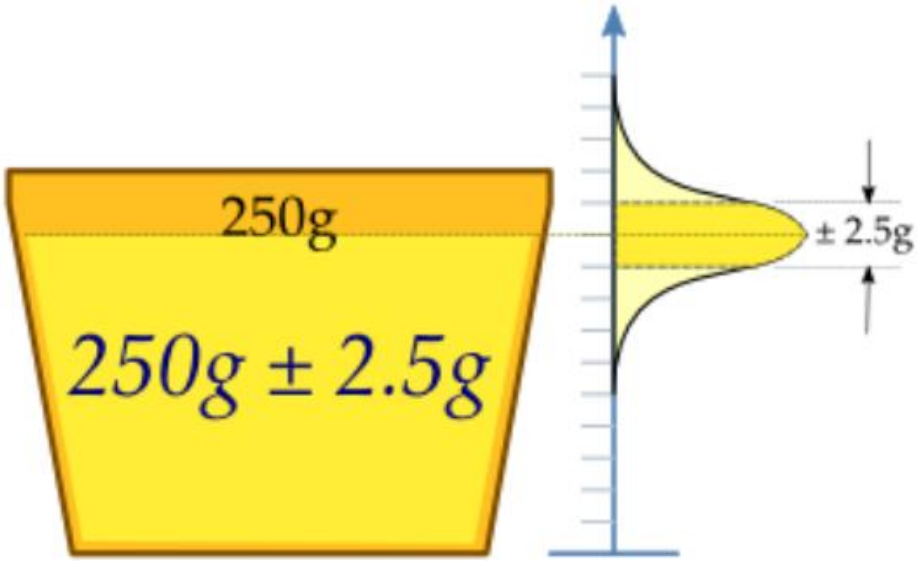
Örnek 6 -devam:

X={247.1, 250.0, 250.1, 249.8, 246.7, 254.4, 249.2, 249.4, 247.0, 247.0, 245.0, 253.3, 251.2, 250.7, 250.6, 247.3, 248.5, 248.0, 243.6, 250.2}

$$\bar{x} = \frac{\sum x_i}{20} = 248.955\,\text{g}$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{20-1}} = 2.6015\,\text{g}$$

$$s_x = \frac{s}{\sqrt{n}} = \frac{2.6015}{\sqrt{20}} = 0.5817\,\text{g}$$



Örnek 6 -devam:

Ana kütle ortalaması %95 güvenlik seviyesi ile şu aralıktadır:

$$\bar{x} - 2s_x \le \mu \le \bar{x} + 2s_x$$

$$248.955 - 2(0.5817) \le \mu \le 248.955 + 2(0.5817)$$

$$247.7916 \le \mu \le 250.1184$$

Yani, anakütle ortalaması (doğru ortalama) 95% olasılıkla [247.7916, 250.1184] aralığındadır.

- Firma bu değeri μ = 250 g olarak öngörmüştü.
- *Buna göre, bizim örneğimizde elde ettiğimiz aralık, ana kütle değerini içine almaktadır. Yani, makine %95 güvenlik seviyesinde doğru kalibre edilmiştir.*

Alıştırma 3

Dersin web sayfasında bulunan aşağıdaki belgeyi kullanarak,

http://www1.gantep.edu.tr/~bingul/stat/yenidogan.xls

Erkek bebeklerin ağırlıkların ortalamasını %95 güvenlik düzeyinde tahmin edin.