İstatistik: Daha kolay bir yolu var mı?

Meetup: Data İstanbul

5 Nisan, 2017

H. Sait Ölmez

Olasılık ve İstatistik

Olasılık

Olasılık perspektifinden: Ana kütlede bilinen oranlar üzerinden sarışın bir kişinin seçilme olasılığı

Çıkarımsal istatistik

İstatistik perspektifinden: Ana kütleden örneklemeyle sarışınların oranı üzerine çıkarım

Image source: mesmes.deviantart.com

Çıkarımsal İstatistik

ANA KÜTLE

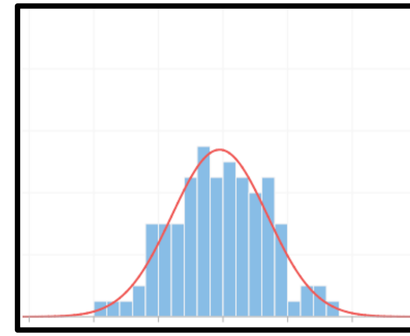
Örnek (n)

Örnek (n)

Örnek (n)

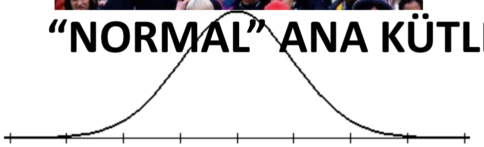
. . . .

. . . .

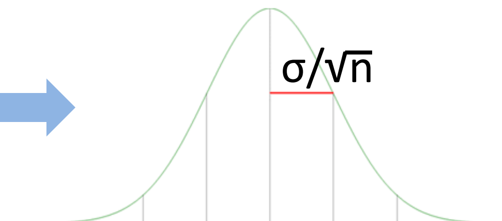


Örneklem dağılımı

"NORMAL" ANA KÜTLE

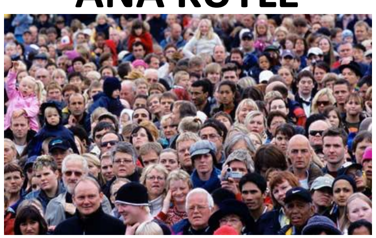


TEORİ

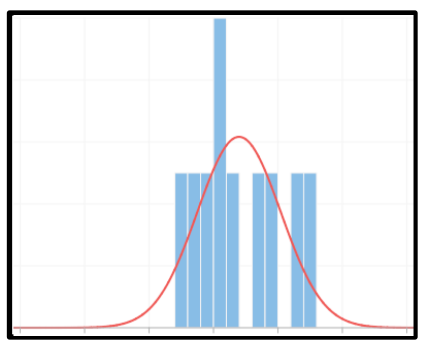


Örneklem dağılımı

ANA KÜTLE



Yalnızca 1 Örnek (n)



Yeniden örnekleme (n)

$\bar{x}$

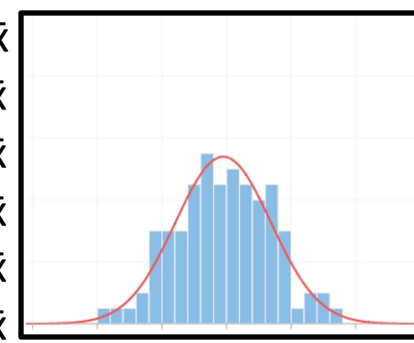
$\bar{x}$

$\bar{x}$

$\bar{x}$

$\bar{x}$

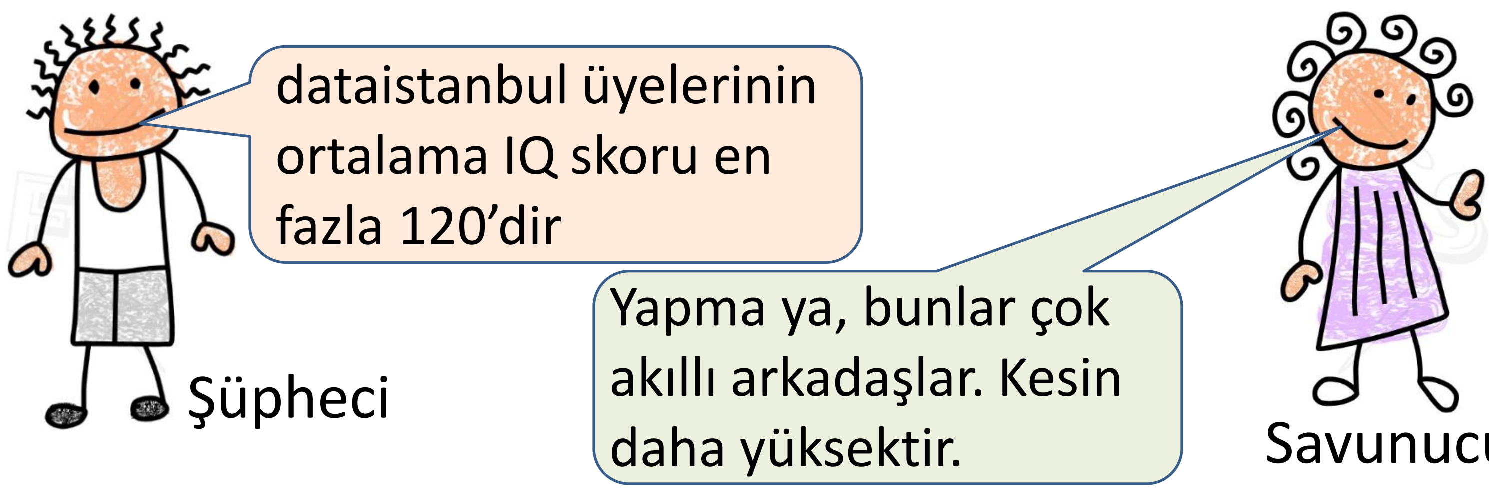
$\bar{x}$



Bootstrap dağılımı

- **Hipotezlerin sınanması**
- **Adım 1**: Problemin tanımlanması ve hipotezlerin oluşturulması

Problem: **dataistanbul**'lular zeki midir?



Hipotezler:

Sıfır Hipotezi (H$_0$) : $\mu_{IQ} \leq 120$ (doğru varsayılan)

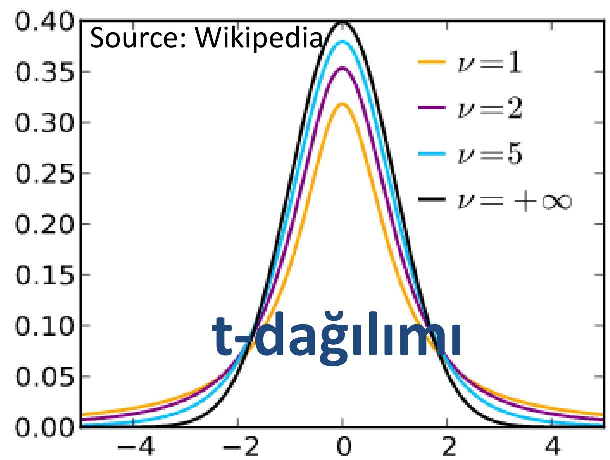
Alternatif Hipotez (H$_A$) : $\mu_{IQ} > 120$

- **Adım 2**: Güven/anlamlılık seviyesi: **α**
- **α**, Sıfır hipotezini doğruyken reddetme olasılığıdır. Konvansiyonel değeri: %5 (%95 güven seviyesi)
- H_0 hipotezini doğru olduğunda yanlışlıkla reddetmekle yapmayı göze aldığımız hata değeridir.
- **Adım 3**: Veriyi topla

Örnek = { 129,125,124,120,117,134,122,
123,122,118,123,122,120,124,
119,123,120,121,119,129 }

μ_{IQ} = 122.7

- **Adım 4**: Örnekleme dağılımını seç ve test istatistiğini belirle
  - Ana kütle dağılımı : ?
  - Ana kütle std sapması : ?
  - Örnek büyüklüğü : 20

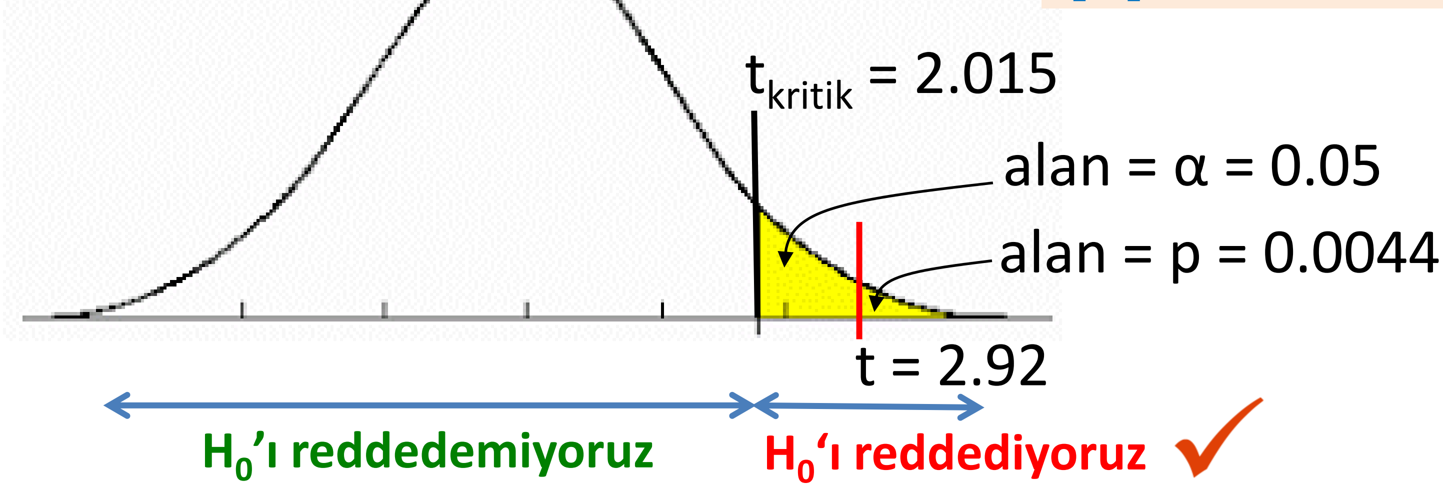


- **Adım 5**: Test istatistiğini ve kritik değerleri hesapla

t-istatistiği: $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{122.7 - 120}{4.131/\sqrt{20}} = 2.92$$

burada $\bar{x}$ örnek ortalaması, ***s*** örnek standart sapması ve ***n*** örnek büyüklüğüdür.

```
> 1 - pt(2.92, 19)
[1] 0.004391871
```

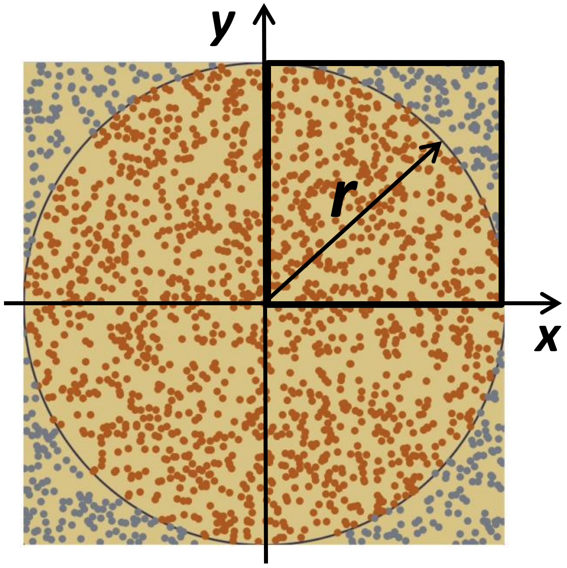


İstatistik ve Simülasyon

π sayısının hesabı

- N adet noktayı 0 ve 1 arasında rastgele örnekle
- Çeyrek daire içine isabet eden noktaları say
- Bu sayının N’ye oranı **π** sayısının çeyreğini verecektir:
- **π** sayısının ne hassasiyette tahmin edildiğini görebilmek için farklı N değerleri deneyelim:

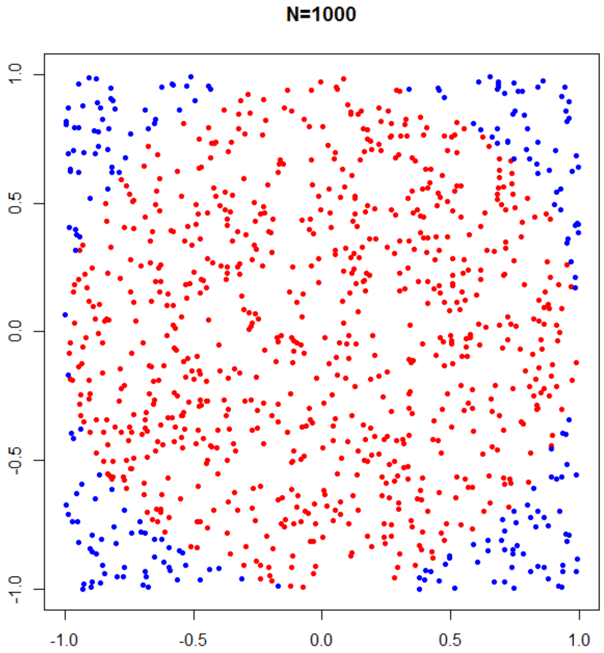
```
piR <- function(N) {
  x <- runif(N,0,1)
  y <- runif(N,0,1)
  d <- sqrt(x^2 + y^2)
  return(4*sum(d < 1.0)/N)
}
set.seed(7)
cat(piR(1000),piR(10000),
    piR(100000),piR(1000000))

3.192 3.1312 3.14284 3.141764
```

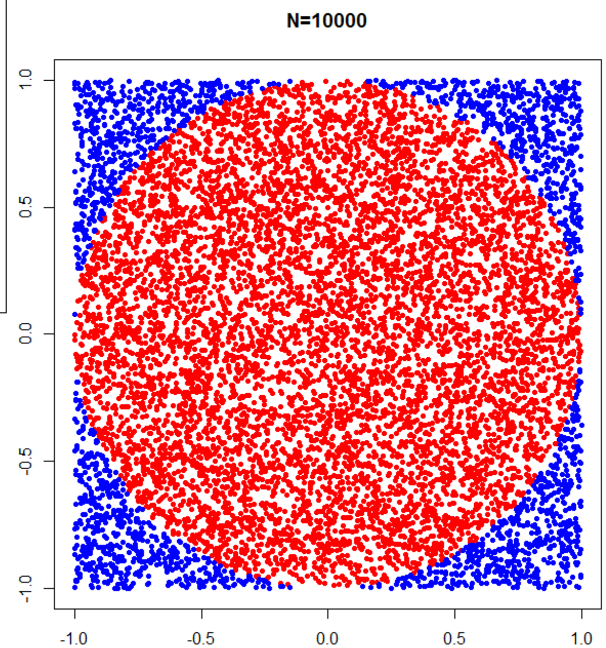


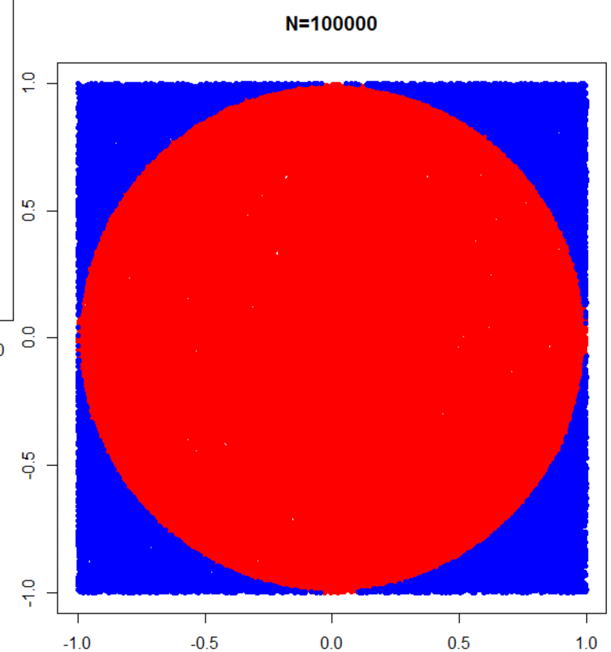
$$Oran = \frac{Alan_{1/4daire}}{Alan_{1/4kare}}$$

$$Oran = \frac{\pi r^2 / 4}{r^2} = \frac{\pi}{4}$$

π sayısının hesabı





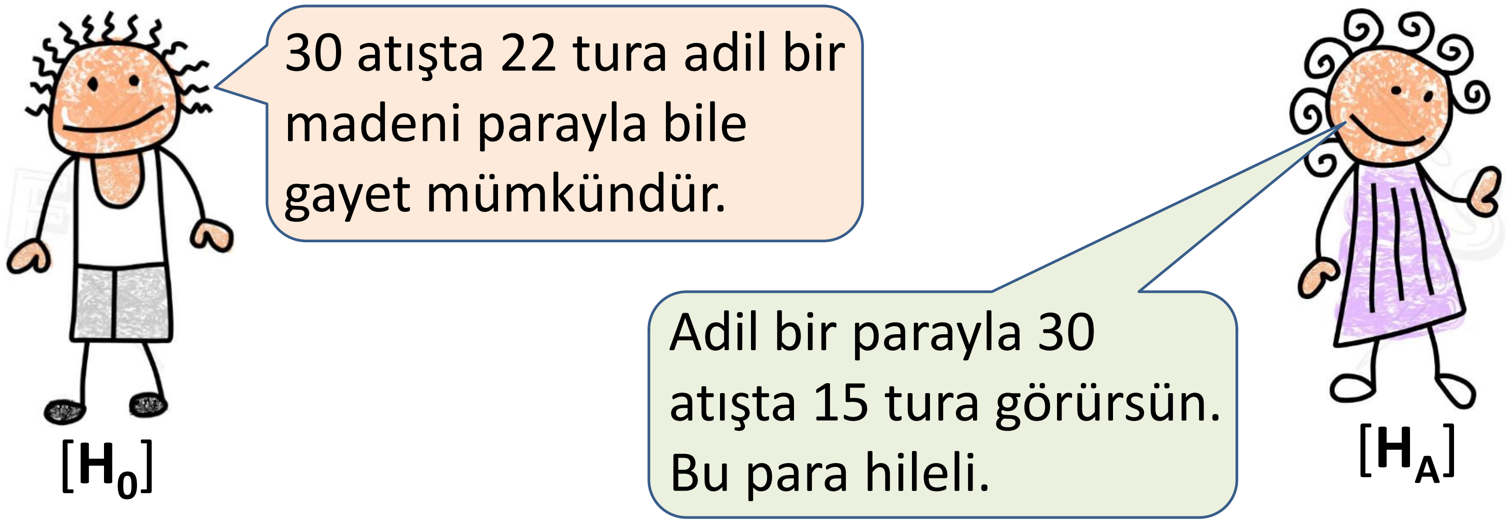


İstatistikte Simülasyon Yöntemleri

- Hesaplama kapasitesi ve yöntemlerinin gelişmesiyle istatistikte birçok problem bilgisayar simülasyonlarıyla çözülebilir hale geldi.
- Örnekleme dağılımı hesabı : Zor
  Örnekleme dağılımı simülasyonu : Kolay
- Kolay istatistik için yöntemler
  - Direkt simülasyon
  - Shuffling/Random permutations (karılma)
  - Random sampling (basit rastgele örnekleme)
  - Bootstrapping (bootstrap örneklemesi)

DİREKT SİMÜLASYON

- **Problem**: Madeni parayı 30 kez atarak 22 tura sayıyorsunuz. Bu adil bir para mıdır?



- Klasik yöntem: Sıfır hipotezinin (H_0) doğruluğunu kabul et ve hipotezi sına
- Adil bir madeni para ile 30 atışta 22 tura gelmesi olasılığı nedir?

Example taken from: “Statistics for Hackers”, Jake Vanderplas, PyCon 2016

Direkt simülasyon

- Bu problemin çözümü için teorik bir model mevcut (binom dağılımı) => 30 atışta 22 tura olasılığı:

$$p(k,n) = \binom{n}{k} \underbrace{p^k}_{k \text{ tura olasılığı}} \underbrace{(1-p)^{n-k}}_{(n-k) \text{ yazı olasılığı}} \qquad \text{ve} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Muhtemel yerleştirmelerin sayısı (binom katsayıları)

$$p(k \geq n_T, n) = \sum_{k=n_H}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

n atışta ***n_T*** ya da daha fazla sayıda tura gelme olasılığı

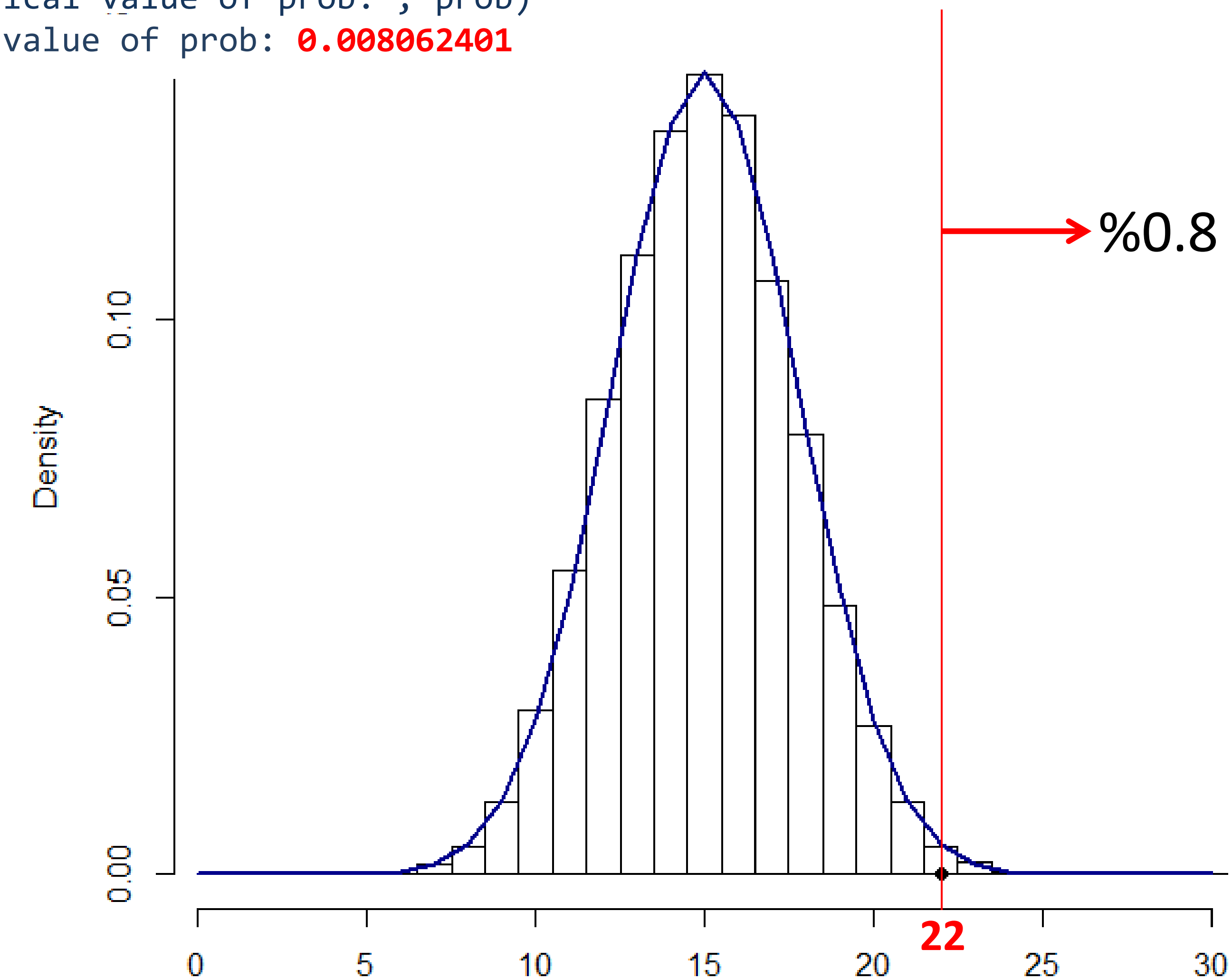
$$p(k \geq 22,30) = \sum_{k=22}^{30} \binom{30}{k} (0.5)^k (0.5)^{30-k} \approx 0.008 = \%0.8$$

- Sıfır hipotezinin doğru olduğu varsayımıyla (adil para) tesadüfen bu veriye ulaşma olasılığı %0.8'dir. Bu da 0.05 p değerinden daha küçük olduğu için H_0'ı reddediyor ve bu paranın hileli bir para olduğu sonucuna varıyoruz.

Direkt simülasyon

```
print(hist(rbinom(10000,30,.5),freq=FALSE, breaks=seq(0.5,30.5,1), ylim=c(0,0.15)))
print(lines(seq(0,30,1),dbinom(seq(0,30,1),30,0.5), col="darkblue", lwd="2"))
points(x=22, y=0, pch=16) ; abline(v=22, col="red")
prob <- print(1-pbinom(21, 30, .5))
cat("Theoretical value of prob:", prob)
Theoretical value of prob: 0.008062401
```

Binom dağılımı (analitik)



Direkt simülasyon

- Daha kolay bir yol var mı?
- Simülasyon?

```
N = 10000 ; M = 0
set.seed(10)
for (i in 1:N) {
  x1 <- sample(0:1, 30, replace=T)
  if (sum(x1) >= 22) {
    M = M + 1
  }
}
cat("Trials with more than 22 heads : ", M, "\n")
cat("Ratio of M to N                : ", M/N, "\n")

Trials with more than 22 heads :  81
Ratio of M to N                :  0.0081
```

p-değeri=0.05'den daha küçük
H_0'ı reddet (hileli para!)

KARILMA
(Rastgele Permütasyonlar)

RESEARCH ARTICLE — OPEN ACCESS

Beer Consumption ? Human Attractiveness to Malaria Mosquitoes

Article | Metrics | Related Content | Comments: 0

Thierry Lefèvre[1*], Louis-Clément Gouagna[2,3], Kounbobr Roch Dabiré[3,4], Eric Elguero[1], Didier Fontenille[2], François Renaud[1], Carlo Costantini[2,5], Frédéric Thomas[1,6]

To add a note, highlight some text. Hide notes

Make a general comment

- Araştırma konusu: Bira içmek sizi sivrisineklere karşı daha kolay yem haline getirir mi?
- Elimizde bira içen 25 denek ve sadece su içen 18 denekle bir deney yürütüyoruz.
- Her bir grupta kaç adet sivrisineğin denekleri hedef aldığını kaydediyoruz. İşte sonuçlar:

Example taken from: “Statistics without the agonizing pain”, John Rauser, Strata Conf. 2014

- Bira içen denekleri ısıran sivrisinek sayısı yalnızca su içenleri ısıranlardan ortalama 4.4 daha fazla...
- Bu sonuç istatistiki olarak anlamlı mıdır? Yoksa tamamen rastlantı mıdır?

BİRA

27	19	20	20	23
17	21	24	31	26
28	20	27	19	25
31	24	28	24	29
21	21	18	27	20

$\text{Ortalama}_B : \mu_B = 23.6$

SU

21	19	13
22	15	22
15	22	20
12	24	24
21	19	18
16	23	20

$\text{Ortalama}_S : \mu_S = 19.2$

- Ortalamalardaki fark: $\delta = \mu_B - \mu_S = 4.4$
- δ istatistiki olarak anlamlı mı (statistical significance)?

- **Analitik Çözüm**
- Şüpheci => Sıfır Hipotezi H_0: $\mu_B = \mu_S$
- Savunucu => Alternatif Hip. H_A: $\mu_B \neq \mu_S$
- 2 ana kütleye ilişkin bir hipotez problemi: Varyanslar bilinmiyor ve birbirlerinden farklı
  Uygun dağılım : **t-dağılımı**
  İstatistiki test : **t-test** (Student's t-test)
- t-skoru için gerekli ortalama ve varyans hesapları:
  - Ortalamalar: $\mu_S = 23.6$ ve $\mu_B = 19.2$
  - Varyanslar: $S_B^2 = 17.08$, $S_S^2 = 13.48$; $N_B = 25$ ve $N_S = 18$

$$S_B^{\;2} = \frac{\sum_{i=1}^{N_B}(X_i - \mu_B)^2}{N_B - 1} = 17.08 \qquad S_S^{\;2} = \frac{\sum_{i=1}^{N_S}(X_i - \mu_S)^2}{N_S - 1} = 13.48$$

- t-skoru:

$$t = \frac{(\mu_B - \mu_S)}{\sqrt{(S_B{}^2 / N_B) + (S_S{}^2 / N_S)}} \approx 3.68$$

- Eğer şüpheci haklıysa (şayet sıfır hipotezi doğruysa), bu durumda t aşağıdaki formüle göre dağılıyor demektir (t için olasılık yoğunluk fonksiyonu):

$$p(t,\upsilon) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\,\Gamma(\upsilon/2)}\left(1+\frac{t^2}{\upsilon}\right)^{-\frac{\upsilon+1}{2}}$$

Burada **Γ** Gamma fonksiyonu, ***ν*** ise aşağıdaki formüle göre hesaplanan serbestlik derecesidir (df)

$$df \approx \frac{[(S_B{}^2 / N_B) + (S_S{}^2 / N_S)]^2}{\frac{(S_B{}^2 / N_B)^2}{N_B - 1} + \frac{(S_S{}^2 / N_S)^2}{N_S - 1}} = \frac{(0.683 + 0.749)^2}{\frac{0.683^2}{25-1} + \frac{0.749^2}{18-1}} = \frac{2.051}{0.0194 + 0.033} = 39.14$$

Karılma

- Verilen dağılımdan t-skorunun 3.68'den daha büyük olma olasılığını bulmamız gerekiyor.
- Alternatif olarak, df=39.1 and α=0.025 (anlamlılık seviyesi) değerleri için t-dağılımı tablosundan kritik t-skorunu okuyabiliriz => $\mathbf{t_{kritik}=2.02}$

t distribution critical values

	Upper-tail probability *p*						
df	.25	.20	.15	.10	.05	.025	.02
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147
39.1 40	0.681	0.851	1.050	1.303	1.684	2.021	2.123

- Tablodan okuyamıyorsanız, aşağıdaki R komutunu kullanabilirsiniz:

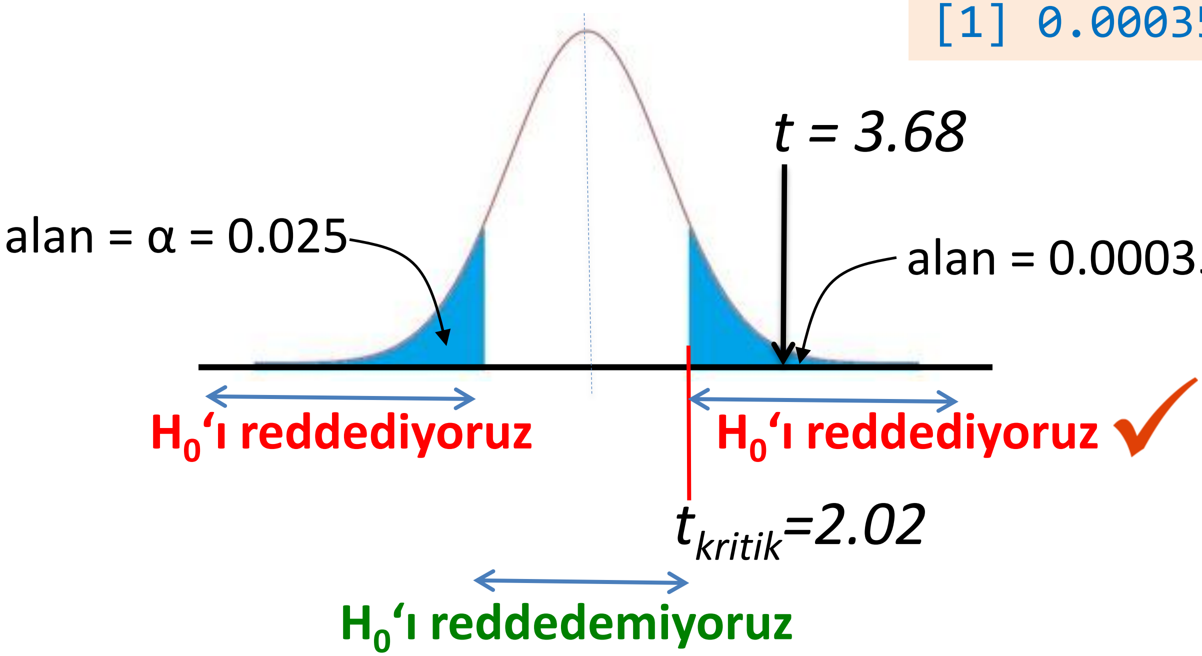
```
> qt(p=0.975, df=39.14)
[1] 2.02246
```

bu da size α=0.025 değeri için kritik t değerini verecektir.

Karılma

- t-skoru = 3.68 > 2.02 olduğu için Sıfır Hipotezini reddediyoruz. Ortalama 4.4 fark istatistiki olarak anlamlı (p-değeri = 0.00035 < 0.025)
- **Sonuç**: Bira insanları sivrisinek ısırıklarına karşı daha cazip hale getiriyor.

```
> 1 - pt(3.68, 39.14)
[1] 0.0003506143
```



Karılma

- Daha kolay bir yolu var mı diye yine sorabilirsiniz...
- Bir önceki problemdeki gibi teorik bir modelimiz (binom) yok. Elimizde yalnızca kaç adet sivrisinek tarafından cazip bulunduğunu bildiğimiz bir denek listesi var.

BİRA				
27	19	20	20	23
17	21	24	31	26
28	20	27	19	25
31	24	28	24	29
21	21	18	27	20

SU		
21	19	13
22	15	22
15	22	20
12	24	24
21	19	18
16	23	20

- **Sıfır Hipotezi**: Sivrisineklere cazip gelme konusunda bira ve su içenler arasında bir fark yok! Şayet bu doğruysa, ölçümlerde kullandığımız etiketlerin (su veya bira) de bir önemi olmamalı.
- Kısacası, bu etiketleri dilediğimiz gibi karıştırabilir, istediğimiz etiketi istediğimiz deneğe atayabiliriz.

- **Fikir**
  - Etiketleri sürekli karıştırarak bir dağılım simülasyonu yaratacağız.
- **Yöntem**
  - Bira ve su gruplarından rastgele kayıtlar seçerek her iki grup için ortalamaları hesaplayacak ve bunu birçok kez tekrarlayacağız.
- **Sonuç**
  - Sıfır hipotezinde iddia edildiği gibi iki grup arasında bir fark yoksa, hangi verinin neyle etiketlendiği (su veya bira) sonucu değiştirmeyecektir.

- **Prosedür (1)**

Etiketleri karıştır:

$N_B = 25$

BİRA				
27	19	20	20	23
17	21	24	31	26
28	20	27	19	25
31	24	28	24	29
21	21	18	27	20

$N_S = 18$

SU		
21	19	13
22	15	22
15	22	20
12	24	24
21	19	18
16	23	20

- **Prosedür (2)**

  Düzenle: Tüm bira rengi etiketleri birleştir ve yeni bir "bira" grubu oluştur. Geri kalanları su grubunda topla:

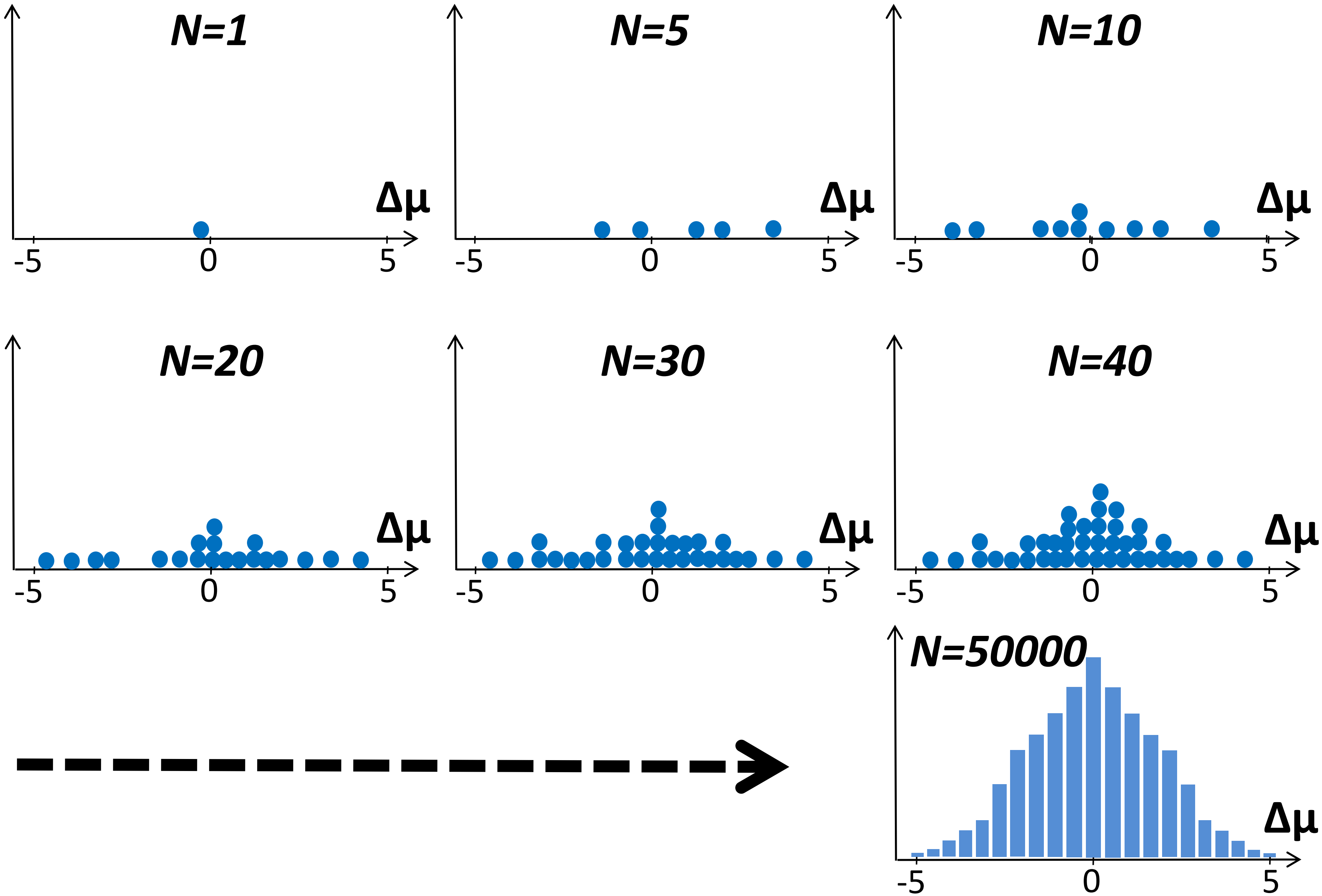
BİRA				
27	20	23	21	24
26	28	19	25	24
28	21	21	18	20
21	19	22	22	12
24	21	19	16	20

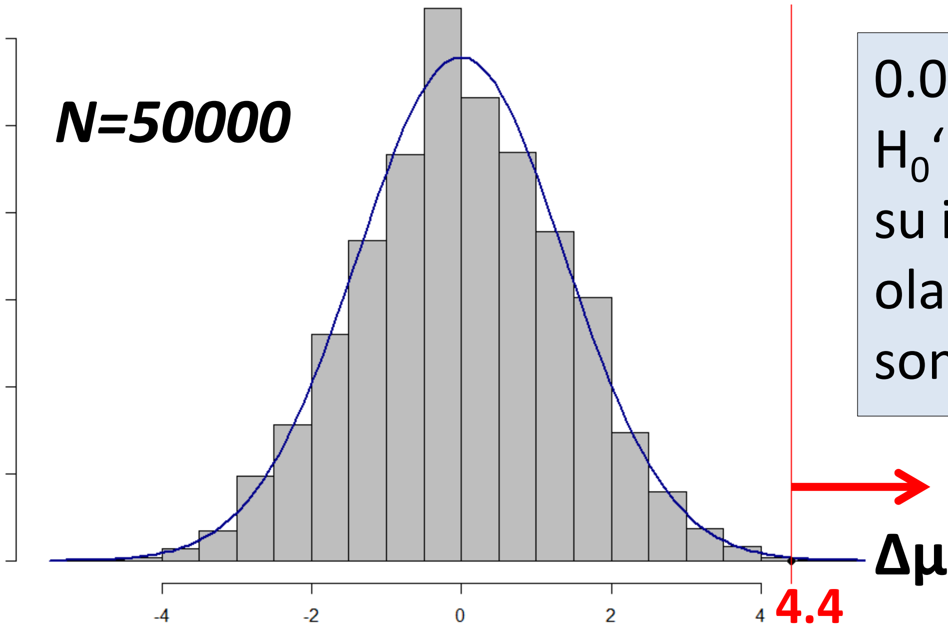
SU		
19	20	17
31	20	27
31	24	29
27	13	22
15	15	20
24	18	23

Ortalamalar farkı: $\Delta\mu = 21.64 - 21.94 = -0.3$

- **Prosedür (3)**
  - Tekrar karıştır ve düzenle... Bu işlemi (1) ve (2) üzerinde bir döngü ile N defa tekrarla
  - i=1,...,N için $\Delta\mu_i$ değerlerinden oluşan **sıklık dağılım**ını oluştur

Karılma





0.025’den küçük. Böylece H_0‘ı reddediyoruz ve bira ile su içenler arasında istatistiki olarak anlamlı bir fark olduğu sonucuna varıyoruz.

- Burada gerçek dağılımı temsilen bir örnek vekil oluşturduk. Farkın 4.4’ten büyük olduğu ($\Delta\mu > 4.4$) örneklerin toplam iterasyon sayısına oranı:

$$\frac{N_{>4.4}}{N_{tot}} = \frac{18}{50000} = 0.00036$$

- Bu değer, H_0’ın doğru olduğu varsayımıyla en az elimizdeki örneklemde gördüğümüz kadar ekstrem bir etki görme olasılığıdır [pr(veri|H_0)].
  Görünen etki rastgele bir değişkenliğin sonucu değildir.

- Karar:

RESEARCH ARTICLE — OPEN ACCESS

Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes

Article | Metrics | Related Content | Comments: 0

Thierry Lefèvre[1*], Louis-Clément Gouagna[2,3], Kounbobr Roch Dabiré[3,4], Eric Elguero[1], Didier Fontenille[2], François Renaud[1], Carlo Costantini[2,5], Frédéric Thomas[1,6]

To **add a note**, highlight some text. Hide notes

Make a general comment

Rastgele Örnekleme (Random Sampling)

İstatistikte nokta tahminleri

- Elimizde tespit edilmesi istenen bir ana kütle parametresi (ortalama gibi) olduğunu varsayalım:
  - Örnek: Türkiye'de kadınların ortalama boyu
- Bu parametreyi tahmin etmek (nokta tahmini) üzere ana kütleden "***n***" büyüklüğünde bir örnek alalım.
- Nokta tahmini: Örnekteki tüm gözlemlerin bir tahmin edicide (ortalama hesap formülü) yerine konmasıyla bulunan "en iyi" tahmin (tek bir değer).
- <u>Bazı nokta tahminleri:</u>
  - Ana kütle ortalamasını tahmin etmek üzere örnek ortalaması
  - **Ana kütle varyansı**nı tahmin etmek üzere **örnek varyansı**

İstatistikte nokta tahminleri

- İyi bir tahmin edici (estimator) nedir?
- Kritik soru: Örnek varyansı ana kütle varyansından sistematik bir şekilde farklı mı? Herhangi bir yanlılık?
- ***N*** büyüklüğünde bir ana kütle ve ***n*** gözlemden oluşan bir örnek için varyans hesapları:

	Ana kütle (parametre)	Örnek (istatistik)
Varyans	$\sigma^2 = \dfrac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$	$s^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$

yanlı

$$s^2_{unbiased} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

yansız

- Örnek varyansında, ana kütle varyansını düşük tahmin etme eğilimi mevcuttur. Bu nedenle (n-1) ayarı yapılmış örnek varyansı yanlılıktan arınmış bir tahmindir. İspat:

$$E[\sigma^2 - S^2{}_{biased}] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}\Big((x^2{}_i - 2x_i\mu + \mu^2) - (x_i{}^2 - 2x_i\bar{x} + \bar{x}^2)\Big)\right]$$

$$= E\left[\mu^2 - \bar{x}^2 + \frac{1}{n}\sum_{i=1}^{n}\big(2x_i(\bar{x}-\mu)\big)\right] = E[\mu^2 - 2\bar{x}\mu + \bar{x}^2]$$

$$= E[(\bar{x}-\mu)^2] = Var(\bar{x}) = \frac{\sigma^2}{n} > 0$$

$$S^2{}_{biased} = \sum_{i=1}^{n}(X_i - \bar{X})^2 / n$$

Varyansı düşük tahmin eder!

- Tahmin edicilerin beklenenen değerleri:

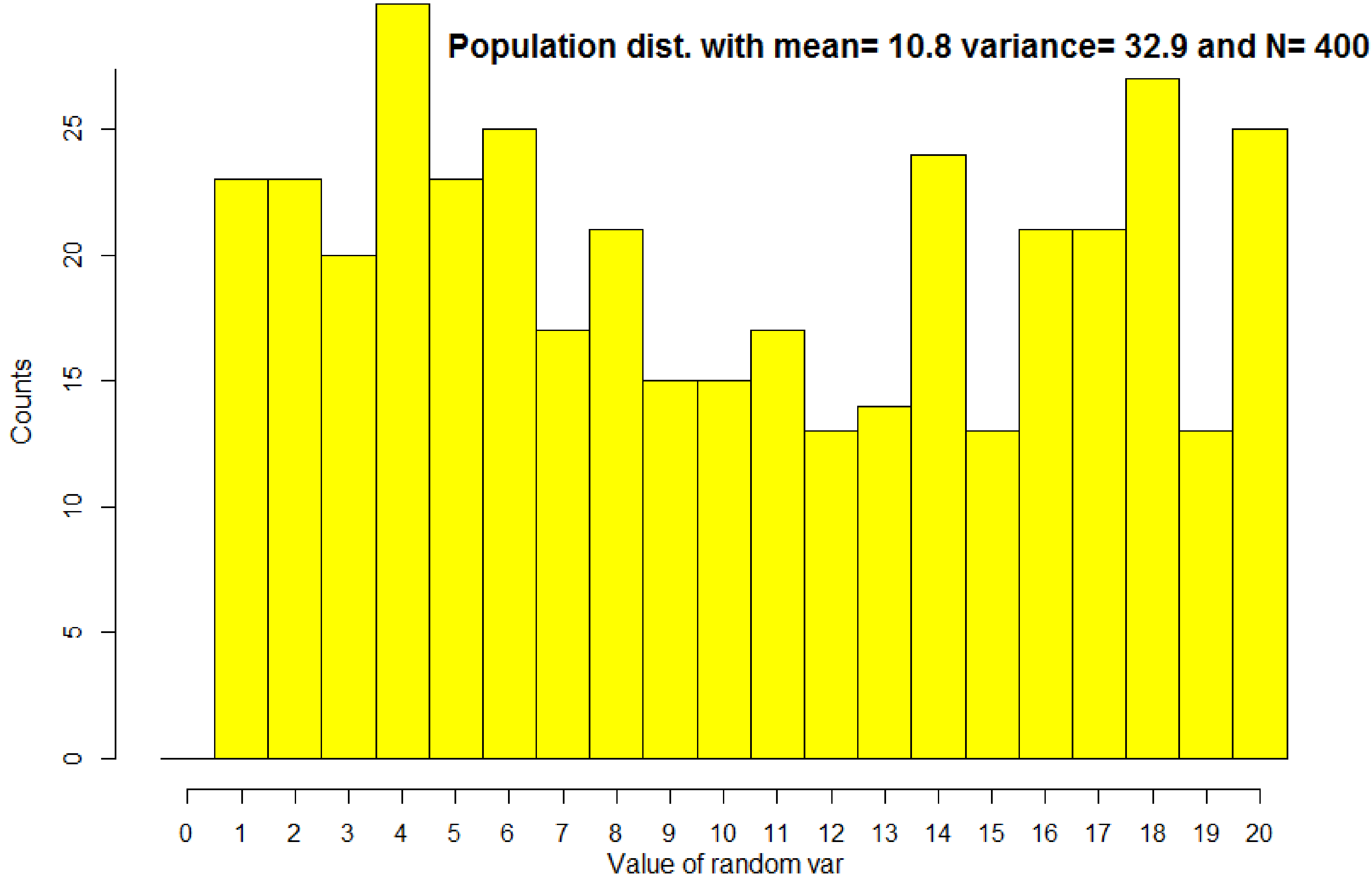
$$E[S^2{}_{biased}] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

$$S^2{}_{unbiased} = \frac{n}{n-1}S^2{}_{biased}$$

n-1 düzeltmeli S^2 bu nedenle yansız bir tahmin edici

İstatistikte nokta tahminleri

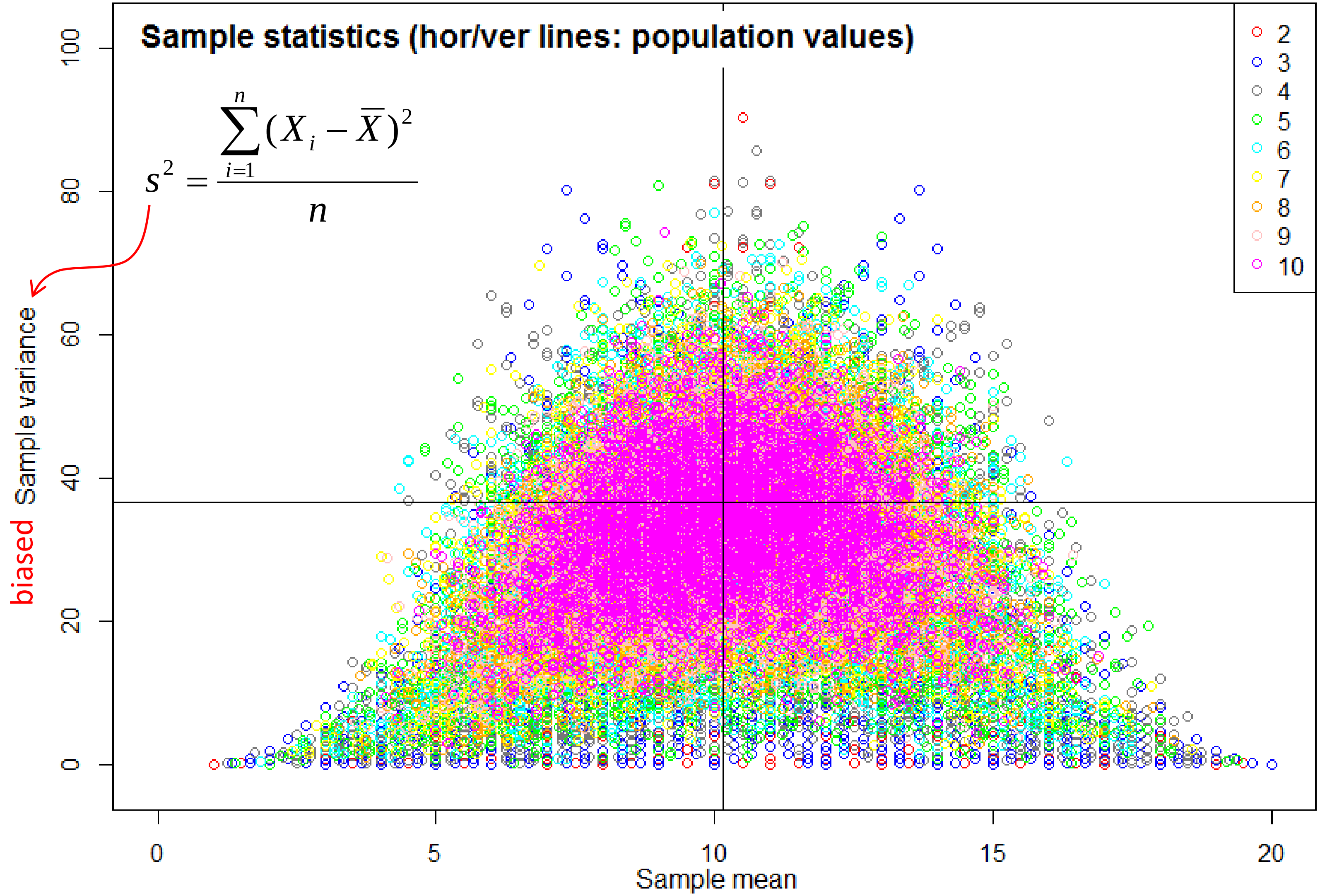
- Değerleri 1 ile 20 arasında değişen ve 400 gözlemden oluşan bir veri kümesi (ana kütle) yaratalım:



Ref: Simulation showing bias in sample variance | Probability and Statistics | Khan Academy

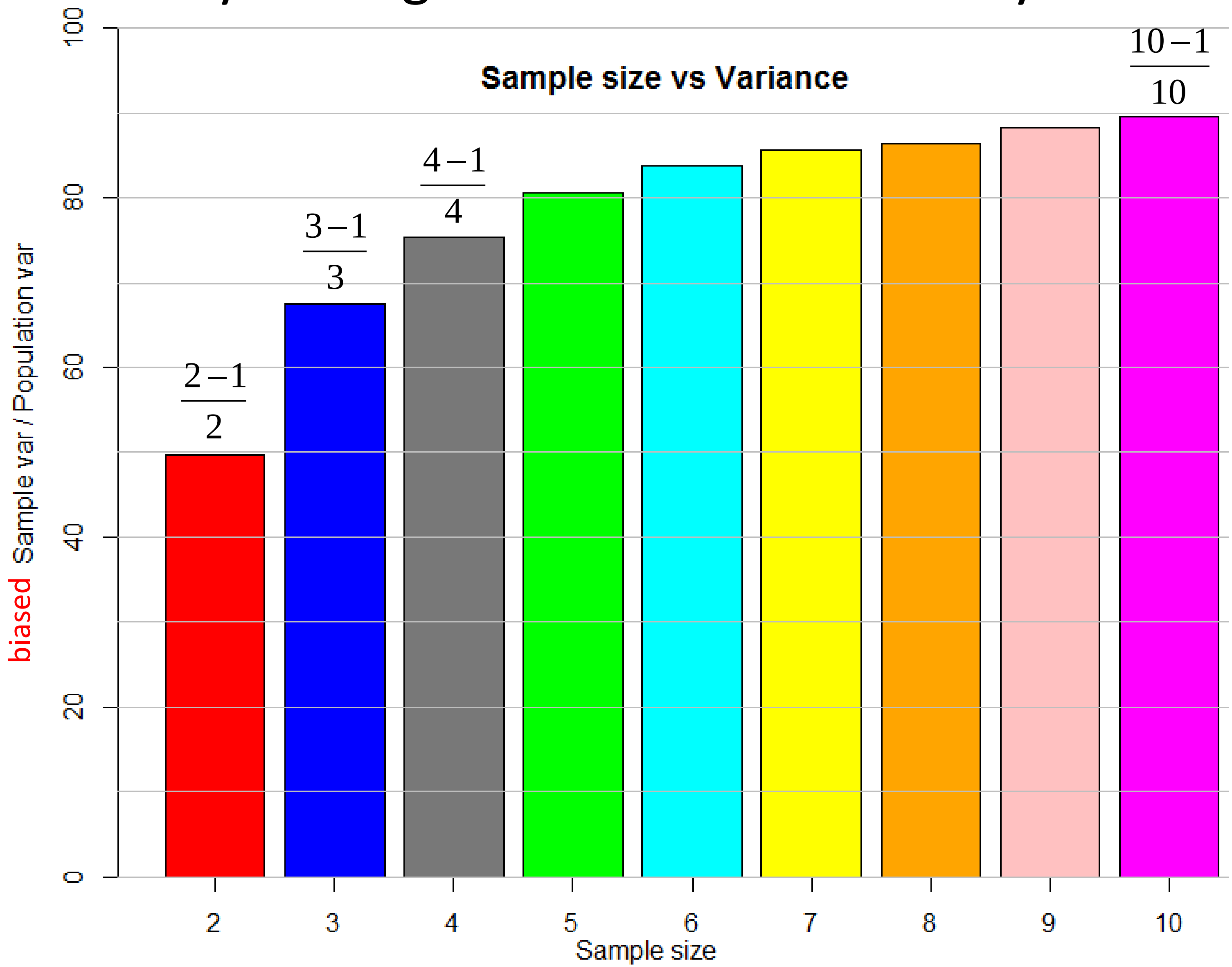
İstatistikte nokta tahminleri

- Ana kütleden eleman sayısı 2 ile 10 arasında değişen örnekler alalım (her birinden 5000 kez):



İstatistikte nokta tahminleri

- Örnek boyutuna göre örnek-ana kütle varyans oranı:



İstatistikte nokta tahminleri

- Örnek varyansını hesaplamak üzere aşağıdaki formülü kullandığımız zaman:

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

- Ana kütle varyansına yaklaşamıyoruz ama yanlı tahmin olan **(n-1)/n** çarpanlı ana kütle varyansına yaklaşıyoruz:

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} \Rightarrow \frac{n-1}{n}\sigma^2$$

- Bunu nasıl yansız hale getireceğiz? Gerçek ana kütle varyansı için en iyi tahmini elde etmek üzere her iki tarafı **n/(n-1)** ile çarparak yansız tahmini buluyoruz:

$$\frac{\cancel{n}}{n-1}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\cancel{n}} \Rightarrow \frac{\cancel{n}}{\cancel{n-1}}\frac{\cancel{n-1}}{\cancel{n}}\sigma^2$$

$$s_{unbiased}^{2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

BOOTSTRAPPING (bootstrap örneklemesi)

- **Bootstrapping nedir?**
  - Orijinal örnek üzerinde yerine koyarak yapılan örnekleme işlemi (re-sampling)
  - Örnek için ana kütle neyse, bootstrap örnekleri için de orijinal örnek aynı şeydir.
- **Nerelerde kullanılır?**
  - Örnekleme dağılımlarının tespitinde
  - Geçerli ortalama, standart hata, güven aralıkları bulunmasında
  - Bir tahmin edici (estimator) veya bir öğrenme yöntemine ilişkin belirsizliğin tespit ve değerlendirmesinde
  - Teorik birikimin yetersiz olduğu durumlarda bir yöntemin performans değerlendirmesinde

  Tüm bunları dağılımla ilgili herhangi bir varsayımda bulunmadan yapar (normallik, simetri, aykırı değerler vb)
- Geçerli teorik temellere dayanır…

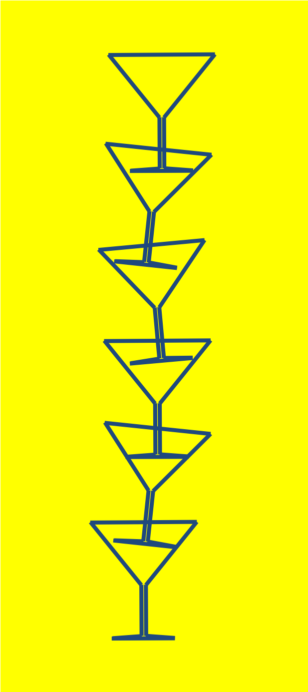
Bootstrapping

- **Ne büyüklükte örneklerle kaç kez tekrarlanır?**
  - Orijinal büyüklükteki örneği (yerine koyarak) oluştu-racak şekilde alt-örnekleme binlerce kez tekrarlanır
- **Ne zaman iyi sonuç vermez?**
- Seçilen örnek ana kütleyi temsil etmekten uzaksa
  - Seçimde yanlılık (selection bias)
- Gözlemler arasında bağımlılık mevcutsa
- Bootstrapping işlemini maksimum değeri bulmak için kullanıyorsak
  - Ana kütledeki maksimumu her zaman daha düşük hesaplar (sıralama tipi işlemlerde iyi sonuç vermez)
- Çok küçük örnek büyüklükleriyle çalışıyorsak
  - N > 20 kabaca sınır kabul edilebilir

Bootstrapping

- Kadehleri üst üste koyarak ne yükseklikte bir bardak kulesi oluşturabilirsiniz?
- 20 deneme sonucunda her bir kule için kullanılan bardak sayısı (yükseklik):

  48 24 32 61 51 12 32 18 19 24
  21 41 29 21 25 23 42 18 23 13



- Kulelerin ortalama yüksekliği?
- Tahmin üzerindeki belirsizlik?
- Kule yüksekliklerini yeteri kadar uzun gözleyecek olursak bu yükseklik değerlerindeki yayılma (spread) ne olurdu? Bardak kulelerinin yükseklik dağılımını nasıl karakterize edebiliriz?

Example taken from: “Statistics for Hackers”, Jake Vanderplas, PyCon 2016

- **Klasik yöntem:**
  - Örnek ortalaması ve ortalamadaki standart sapma:

$$\overline{X} = 28.85$$

$$\sigma_{\overline{X}} = \frac{1}{\sqrt{N}}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2} = 2.97$$

(ortalamadaki standart hata: örnek ortalamalarındaki std sapma)

Belirsizlik: $X = \overline{X} \mp t_{0.025, df=19}\, \sigma_{\overline{x}}$

```
> qt(0.975, 19)
[1] 2.093024
```

$$X = 28.85 \mp 6.216$$

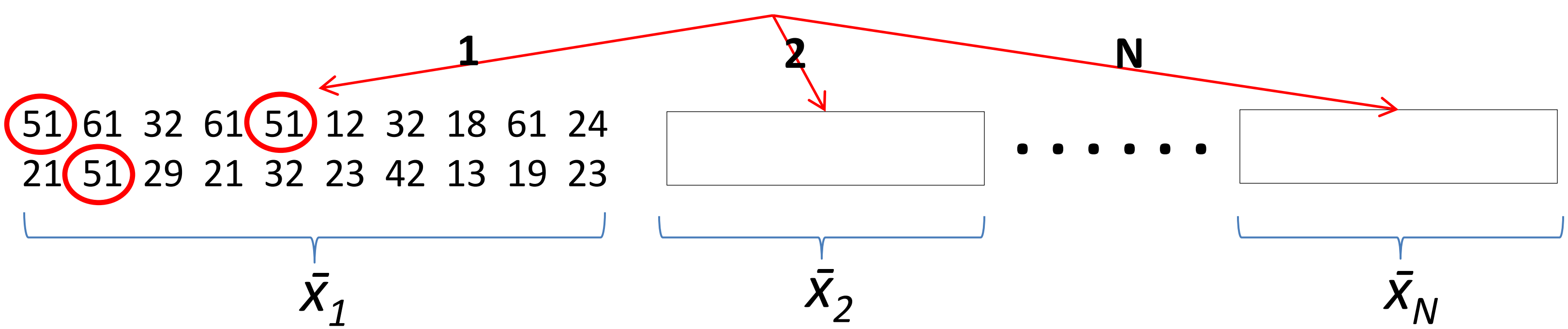
$CI\,[\,22.633\ ,\ 35.067\,]$ (%95 güven aralığı)

- Bu formüllere ne tür varsayımların girdiğini bilmiyor olabilirsiniz. 1. örnekteki gibi parametrik bir modelimiz yok. Elde birbirleriyle karşılaştırılabilecek iki ayrı grup da bulunmuyor. Bu nedenle "karılma" yöntemi de işlevsiz.

Bootstrapping

- **Çözüm:** Bootstrap örneklemi
- **Yöntem**: Yerine koyarak örnekleme ile dağılım simülasyonu
  - Orijinal veri kümesinden sürekli “yerine koyarak” aynı büyüklükte örneklem oluştur (veri tekrarı olabilir)
  - Her iterasyon sonunda örneklem ortalamalarını hesapla
  - İşlemi binlerce kez tekrarlayarak ortalama dağılımını bul

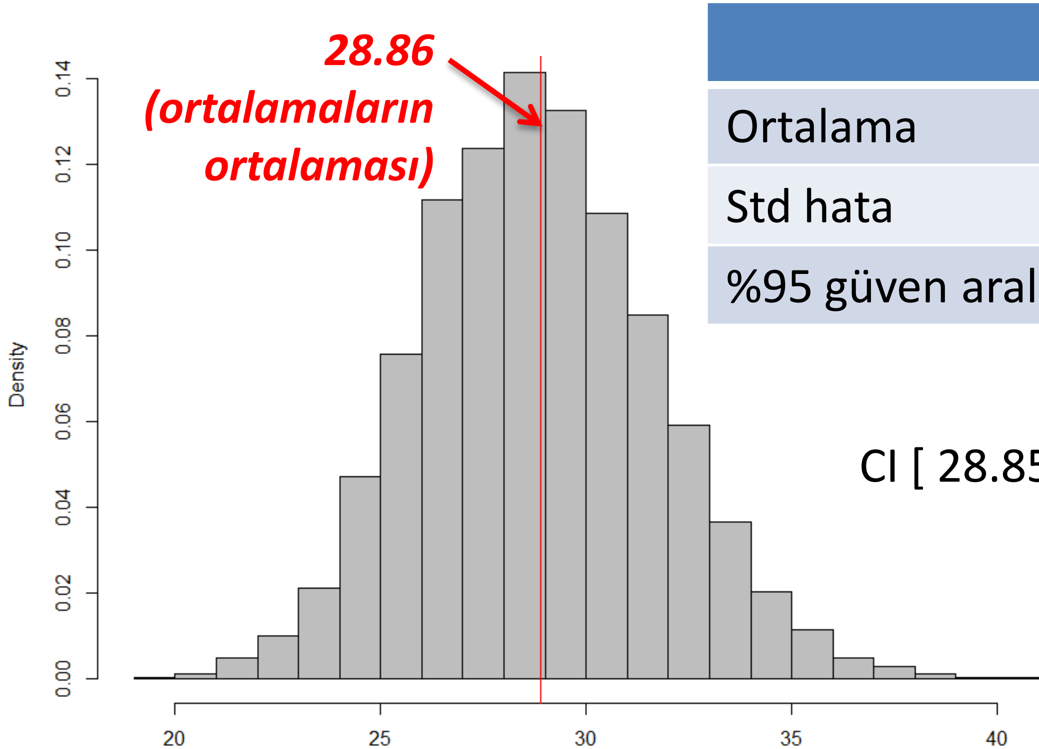
48 24 32 61 51 12 32 18 19 24
21 41 29 21 25 23 42 18 23 13



Bootstrapping

- Gözlemlerden N=10000 defa rastgele örnek al
- Her örnek için ortalamayı hesapla

```
x<-c(48,24,32,61,51,12,32,18,19,24,21,41,29,21,25,23,42,18,23,13)
randx <- replicate(10000, mean(sample(x, length(x), replace=T)))
bs_mean <- mean(randx) ; bs_sd <- sd(randx)
cat("Mean_bs: ", bs_mean, "  Std.dev_bs: ", bs_sd, "\n")
CI <- quantile(randx, c(0.025,0.975))
cat("CI for bootstrapped samples:", CI)
```



	ANALİTİK	BS
Ortalama	28.85	28.86
Std hata	2.97	2.87
%95 güven aralığı	22.63—35.07	23.6—34.8

CI [28.85 ± 6.22]

CI [28.86 -5.26 , 28.86 + 5.94]

Bootstrapping ve Regresyon modelleri

- Bootstrapping yöntemi daha karmaşık problemlere de uygulanabilir...
- Doğrusal regresyon için Bootstrapping:
  - Bardak kulesi yüksekliği ile rüzgar hızı arasındaki ilişki?
- Veri: Yükseklik – Rüzgar hızı

```
> summary(windsp)
 Min. 1st Qu.  Median    Mean  3rd Qu.   Max.
8.100   9.050   9.600  9.832   10.550 12.600

> summary(height)
 Min. 1st Qu.  Median    Mean  3rd Qu.   Max.
8.00    12.00   15.00  14.58    17.00  21.00
```

Windspeed	Height
8.1	21
8.4	19
8.7	16
8.8	18
9	15
9.1	17
9.2	17
9.3	17
9.4	19
9.6	14
9.9	14
10	15
10	11
10.5	12
10.6	12
10.6	13
11.2	10
11.9	8
12.6	9

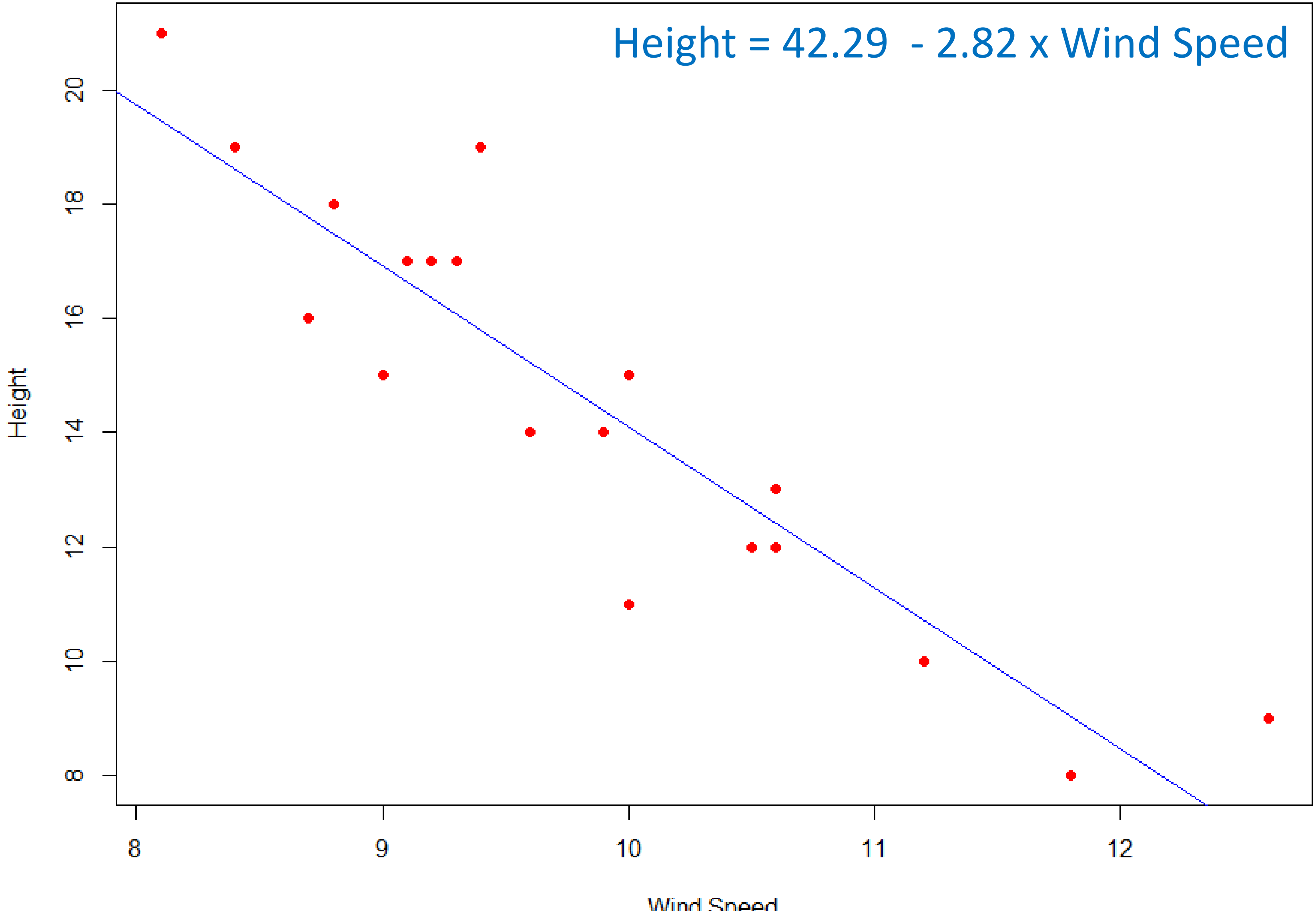
R kodu ve çıktısı

```
...
fit0 <- lm(height ~ windsp)
print(summary(fit0))
```

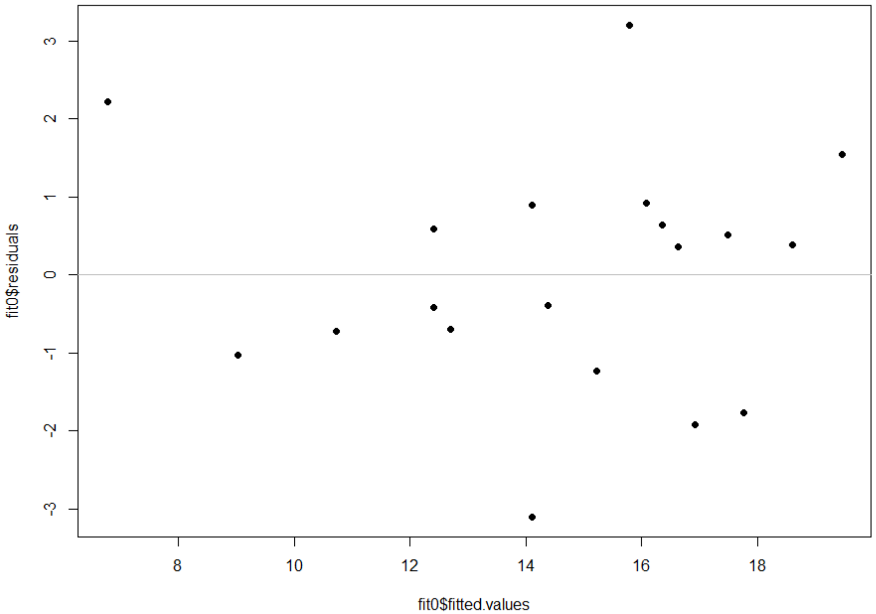
```
Call:
lm(formula = height ~ windsp)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1043 -0.8767  0.3592 0.7684 3.2047

Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  42.2879      3.0928    13.673  1.33e-10 ***
windsp       -2.8184      0.3125    -9.019  6.87e-08 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.551 on 17 degrees of freedom
Multiple R-squared: 0.8271, Adjusted R-squared: 0.817
F-statistic: 81.35 on 1 and 17 DF, p-value: 6.875e-08
```

Basit doğrusal regresyon

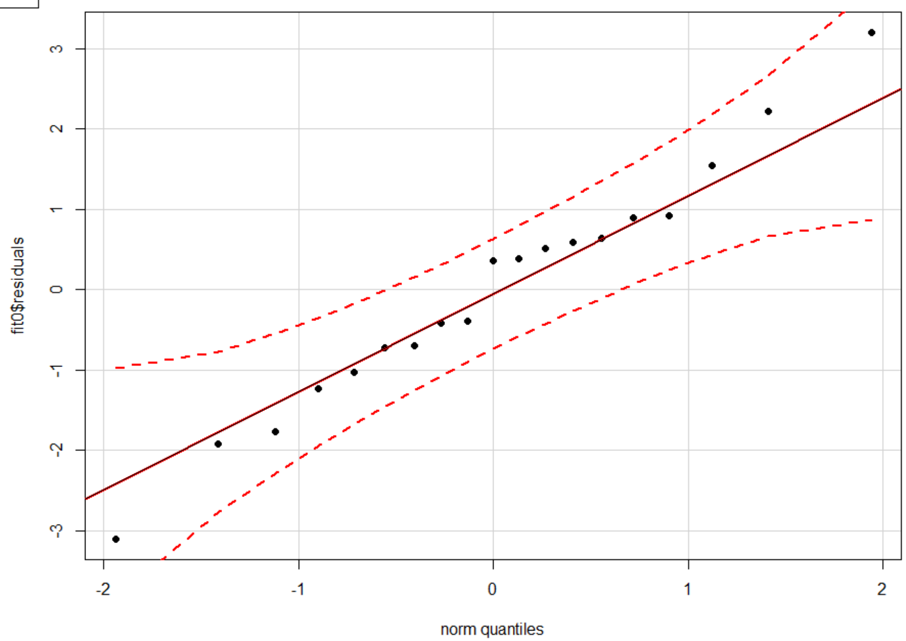
- Orijinal örnek ile parametrik sonuçlar:



Artıkların dağılımı

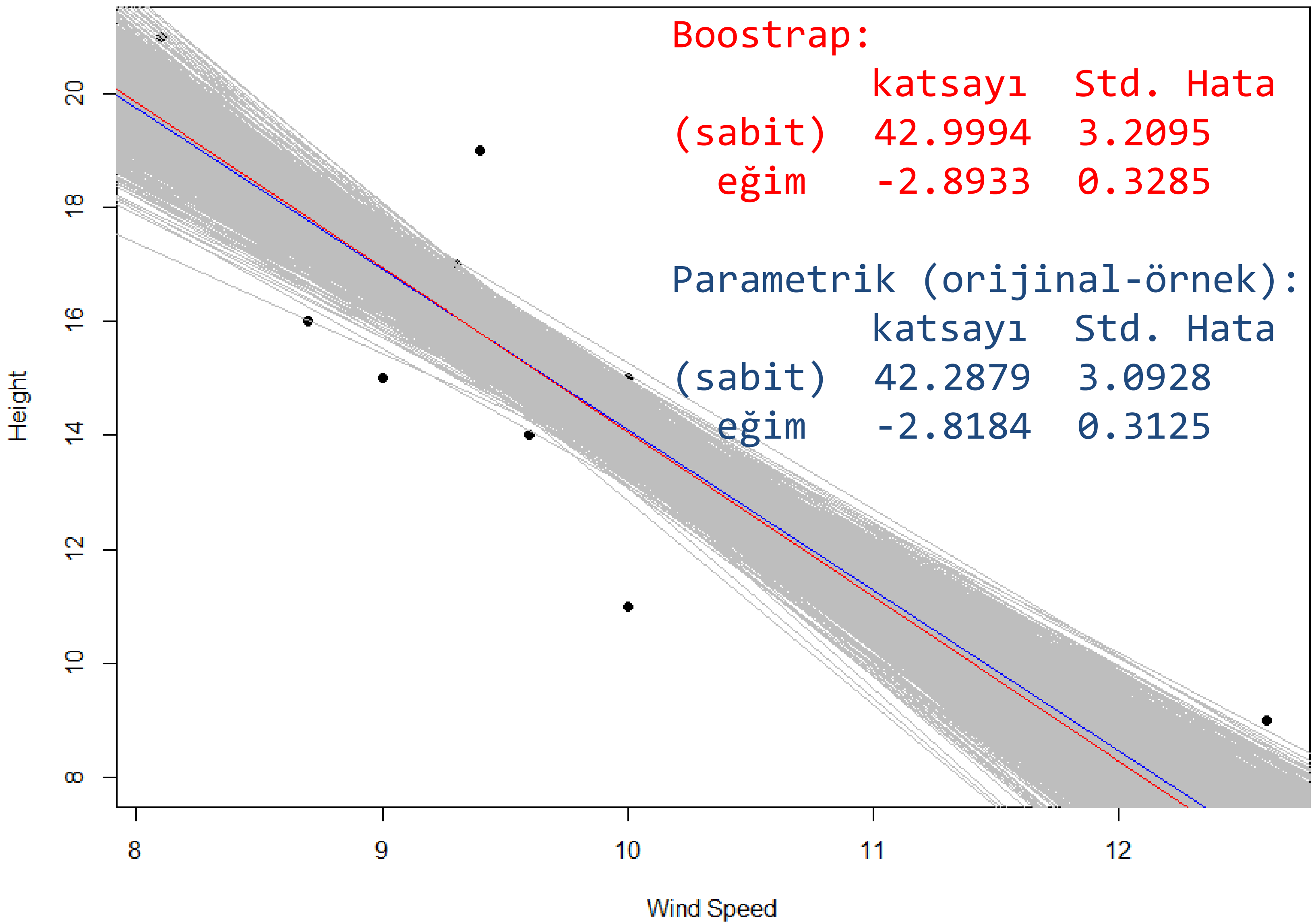


Artıklar için Normallik kontrolü



Bootstrap uyumlamaları

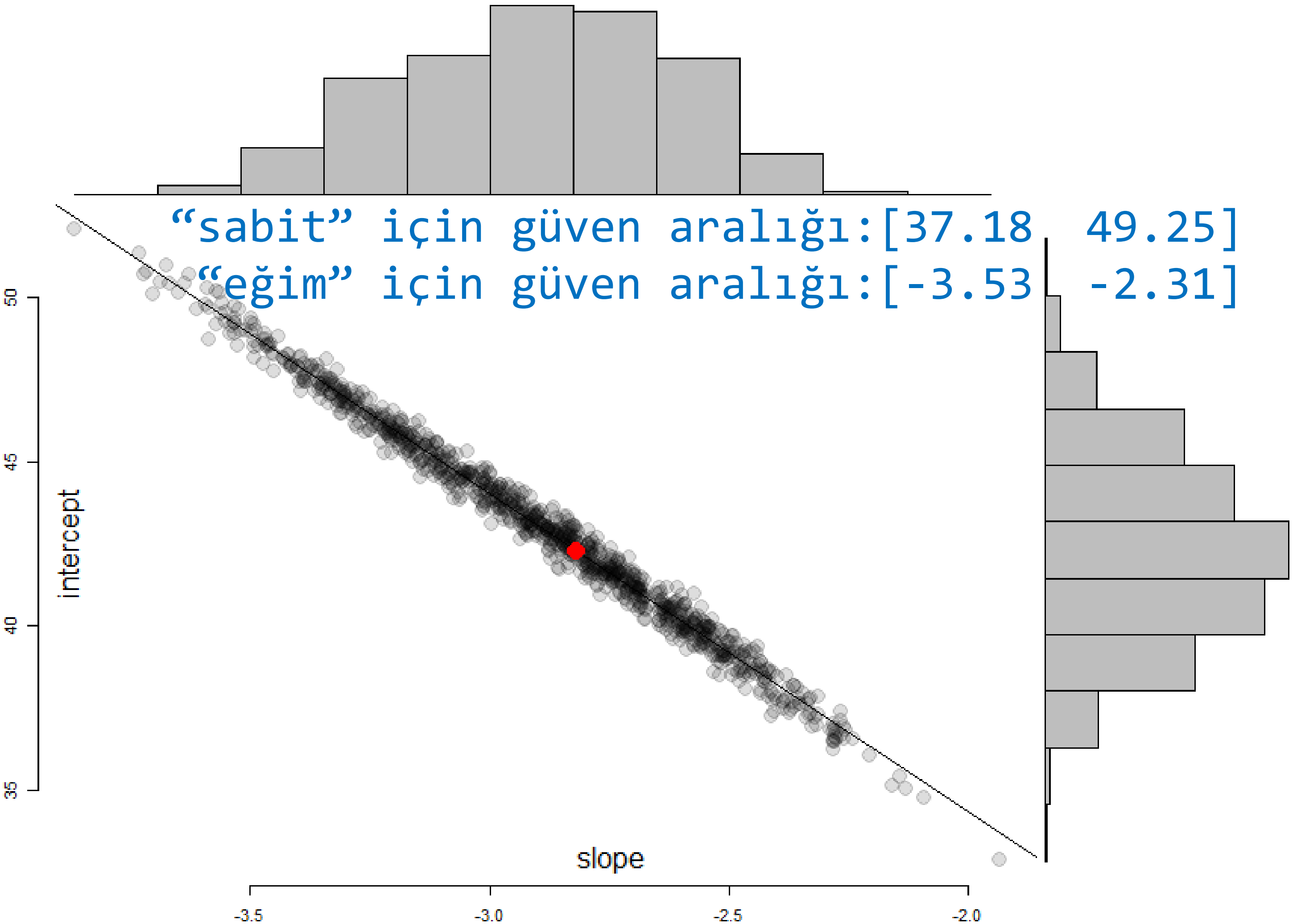
- En iyi uyumlamaların grafiği (*i=1,1000*)



Regresyon modelleri ile Bootstrapping

- Bu bileşik örneklem dağılımı bize ne aralıkta “sabit” ve “eğim” değerleri beklendiği konusunda fikir veriyor.

“sabit” için güven aralığı:[37.18 49.25]
“eğim” için güven aralığı:[-3.53 -2.31]



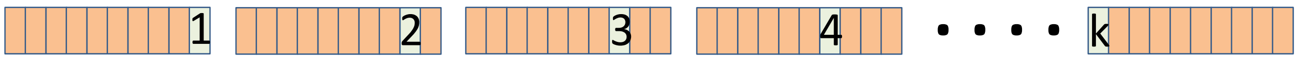
Orijinal veri kümesinden elde edilen “sabit” ve “eğim” değerleri: 42.29 ve -2.82

CROSS VALIDATION
(çapraz geçerlilik)

Eğitim ve test verisinde "bootstrapping"

- Model seçimi ve model performans değerlendirmesi için veri kümesinin bölünmesi: eğitim/test
- Yöntemler:
  - "Holdout" değerlendirme

  - k-katlı çapraz geçerlilik (k-fold cross validation)



k ↗ => varyans ↗

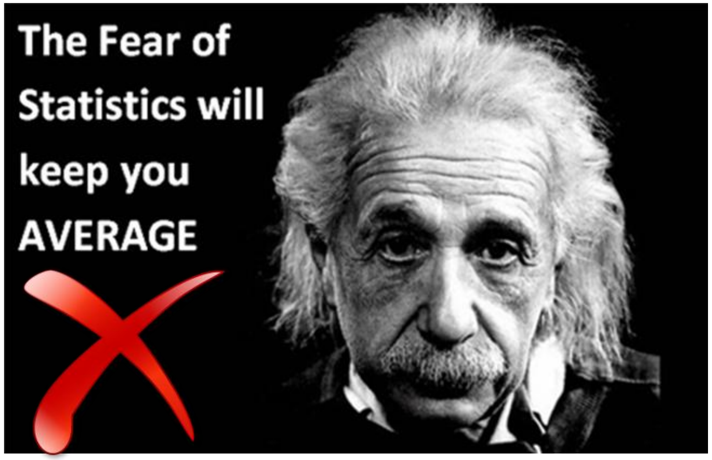
  - **Bootstrapping**

Tüm veri kümesi: X_1 X_2 X_3 X_4 X_5

	Eğitim kümesi	Test kümesi
iterasyon 1	X_3 X_1 X_3 X_3 X_5	X_2 X_4
İterasyon 2	X_5 X_5 X_3 X_1 X_2	X_4
İterasyon 3	X_5 X_5 X_1 X_2 X_1	X_3 X_4
⋮		⋮
iterasyon N	X_4 X_4 X_3 X_4 X_4	X_1 X_2 X_5

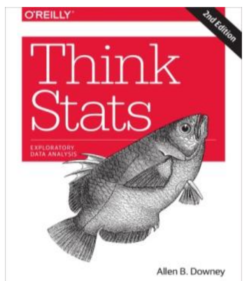
Özetle...

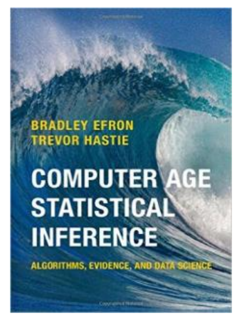
- Buradaki metodolojiyi takip edebildiyseniz...
- Rastgele sayılar üretebiliyorsanız...
- Basit bir döngü yazabiliyorsanız (Python, R vb)...



Kaynaklar

- Think Stats: Probability and Statistics for Programmers, Allen Downey
- Computer Age Statistical Inference, Bradley Efron, Trevor Hastie
- Resampling: The new statistics, Julian L. Simon





- Statistics for Hackers, Jake Vanderplas, Pycon 2016
- Statistics without the agonizing pain, John Rauser, Strata+Hadoop World, 2014
- Sunum ve R programları: github.com/solmez

H. Sait Ölmez, PhD

Sabancı Üniversitesi

Mühendislik ve Doğa Bil. Fakültesi

olmez@sabanciuniv.edu

@saitolmez

solmez

solmez

Veri Analitiği Araştırma ve Uygulama Merkezi
Center of Excellence in Data Analytics (CEDA)

Sabancı Üniversitesi MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

VERİ ANALİTİĞİ

Sabancı Üniversitesi MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

BİLİŞİM TEKNOLOJİLERİ