

# BSM622 VERİ ANALİTİĞİ VE İSTATİSTİK

---

## 3. Hafta: Veri ve Örneklem Dağılımları

Doç. Dr. Onur Polat

Hacettepe Üniversitesi • Bilişim Enstitüsü

Tezsiz YL — Tüm Anabilim Dalları • Güz 2026, Ders 3 / 14

Kaynak: Bruce vd. (2026), Bölüm 2 — Data and Sampling Distributions

## Geçen Hafta ve Bu Haftanın Gündemi

### Geçen Hafta (Hafta 2)

- Konum ölçüleri: ortalama, medyan, kırpılmış
- Yayılım: std, IQR, MAD
- Dağılım keşfi: histogram, boxplot, KDE
- Korelasyon ve saçılım grafikleri
- FacetGrid koşullu görselleştirme
- AI destekli keşif (3. baskı)

### Bu Hafta (Hafta 3)

- Rassal örnekleme ve örneklem yanlılığı
- Seçim yanlılığı ve veri gözetleme
- Örnekleme dağılımı ve standart hata
- Merkezi Limit Teoremi (MLT)
- Bootstrap yeniden örnekleme
- Güven aralığı
- Normal, uzun kuyruklu, t, binom, Poisson dağılımları

## Bu Haftadan Sonra Yapabilecekleriniz

| # | Öğrenme Çıktısı                                                              |
|---|------------------------------------------------------------------------------|
| 1 | Rassal örnekleme, tabakalı örnekleme ve yanlılık türlerini ayırt etme        |
| 2 | Örnekleme dağılımı ile veri dağılımı arasındaki farkı açıklama               |
| 3 | Merkezi Limit Teoreminin ne söylediğini ve neden önemli olduğunu açıklama    |
| 4 | Bootstrap ile standart hata ve güven aralığı hesaplama                       |
| 5 | Normal, t, binom ve Poisson dağılımlarını scipy ile üretme ve görselleştirme |
| 6 | QQ-plot ile normallik değerlendirmesi yapma                                  |



# BÖLÜM I

Rassal Örneklem ve Yanlılık

## Kitle, Örneklem ve Rassal Seçim

---

- Kitle (Population): Tüm olası gözlemlerin kümesi — genellikle erişilemez.
- Örneklem (Sample): Kitleden seçilen bir alt küme — analiz buna dayanır.
- Rassal örnekleme: Her birey eşit olasılıkla seçilir → temsili örneklem.
- Yerine koymalı vs. koymadan örnekleme.
- Büyük veri çağında bile örnekleme gerekli: veri kalitesi ve hesaplama maliyeti.

*Kitap: "Büyük veri, örnekleme ihtiyacının sonu değil — aksine, değişen kalite ve ilgililik karşısında örneklemenin önemini pekiştirir." (Bruce vd., s. 55)*

## Tarihsel Örnek: Literary Digest Felaketi (1936)

---

- Literary Digest, 10 milyondan fazla kişiye anket göndererek Alf Landon'ın kazanacağını tahmin etti.
- George Gallup yalnızca 50,000 kişiyle Roosevelt'in kazanacağını doğru tahmin etti.
- Sorun: Dergi aboneleri ve telefon sahipleri kitleyi temsil etmiyordu — sistematik yanlılık.
- Gallup'un başarısı: Rassal ve temsili örneklem seçimi.
- Ders: 10 milyon yanlı gözlem, 50,000 doğru seçilmiş gözlemden daha kötüdür.

## Tabakalı Örneklem (Stratified Sampling)

---

- Kitleyi homojen alt gruplara (tabaka / stratum) böl.
- Her tabakadan ayrı ayrı rassal örneklem çek.
- Avantaj: Her alt grubun yeterince temsil edilmesini garanti eder.
- Örnek: Siyasi anketlerde demografik gruplardan orantılı örneklem.
- Bilişim uygulaması: A/B testinde kullanıcı segmentlerinden tabakalı örneklem — her segment yeterince gözlem içerir.

## Yanlılık Türleri

| Yanlılık Türü       | Tanım                                              | Örnek                                     |
|---------------------|----------------------------------------------------|-------------------------------------------|
| Örneklem yanlılığı  | Örneklem kitleyi temsil etmiyor                    | Literary Digest anketi                    |
| Kendi kendine seçim | Katılımcılar kendileri seçiliyor                   | Yelp / Google yorumları                   |
| Veri gözetleme      | Veride 'ilginç' şey bulana kadar arama             | p-hacking, data dredging                  |
| Geniş arama etkisi  | Çok sayıda model/soru → sahte sonuç                | 100 özellikten en iyisini seçme           |
| Ortalamaya dönüş    | Uç değerler bir sonraki ölçümde daha merkeze kayar | Parlak bir yıl sonrası sıradan performans |

## Seçim Yanlılığı ve Korunma Yolları

---

- Hipotez belirle → deneme tasarla → veri topla → test et: bu sıra yanlılığı önler.
- Tersî: veriyi tara → ilginç şey bul → sonradan hipotez oluştur → p-hacking.
- Tutma seti (holdout set): Modeli eğitmediğiniz veriyle doğrulayın.
- Hedef karıştırma (target shuffling): Yanıt değişkenini karıştırıp modeli tekrar çalıştırın.
- Önceden kayıt (pre-registration): Analizinizi önceden belirleyin.

*"Ne aradığınızı bilmiyorsanız, yeterince sıkı bakarsanız bulursunuz." — Yogi Berra'dan uyarlama (Bruce vd.)*



# BÖLÜM II

Örnekleme Dağılımı ve Merkezi Limit Teoremi

# Örnekleme Dağılımı vs. Veri Dağılımı

## Veri Dağılımı

- Bireysel gözlemlerin dağılımı
- Çarpık, çok modlu, ağır kuyruklu olabilir
- Örneklem büyüklüğüyle değişmez
- EDA ile doğrudan gözlemlenir

## Örnekleme Dağılımı

- Bir istatistiğin (örn.  $\bar{x}$ ) birçok örneklemden hesaplanan dağılımı
- Düzenli ve çan biçimli olma eğiliminde
- $n$  arttıkça daralır ve normalleşir
- MLT bunun teorik temelidir

## Merkezi Limit Teoremi (CLT) — İstatistiğin En Önemli Teoremi

**Kaynak kitlenin dağılımı ne olursa olsun, yeterince büyük örneklemelerden hesaplanan ortalamaların dağılımı Normal dağılıma yakınsar.**

- "Yeterince büyük" genellikle  $n \geq 30$  olarak kabul edilir (kaba kural).
- Kaynak veri ne kadar çarpıksa, normalleşme için o kadar büyük  $n$  gerekir.
- MLT, t-testleri ve ANOVA gibi klasik yöntemlerin teorik temelidir.
- Ortalama dışında (medyan, oran, std) istatistikler için de geçerlidir ama yakınsama hızı farklıdır.
- MLT sayesinde parametrik testleri çoğu durumda güvenle kullanabiliriz.

## MLT'yi Gözlemleme — Python Simülasyonu

# Gelir verisinden MLT simülasyonu (Kitap s. 67-68)

loans\_income = pd.read\_csv(f'{DATA\_URL}/loans\_income.csv').squeeze()

```
results = []
# Ham veri dağılımı
results.append(pd.DataFrame({'income': np.random.choice(loans_income, 1000),
                           'type': 'Ham Veri'}))

# 5'lik ortalamaların dağılımı
results.append(pd.DataFrame({
    'income': [np.random.choice(loans_income, 5).mean() for _ in range(1000)],
    'type': '5 Ortalaması'}))

# 20'lik ortalamaların dağılımı
results.append(pd.DataFrame({
    'income': [np.random.choice(loans_income, 20).mean() for _ in range(1000)],
    'type': '20 Ortalaması'}))

all_data = pd.concat(results)
g = sns.FacetGrid(all_data, col='type', col_wrap=1, height=2, aspect=3)
g.map(plt.hist, 'income', range=[0, 200_000], bins=40, edgecolor='white')
plt.show()
```

## Standart Hata (SE) — Örneklem Belirsizliğinin Ölçüsü

---

$$SE(\bar{x}) = s / \sqrt{n}$$

- Standart hata, bir istatistiğin örneklemden örneklem ne kadar değişeceğini ölçer.
- Standart sapma  $\neq$  standart hata: sapma veriyi, hata tahmini ölçer.
- $n$  arttıkça SE azalır:  $\sqrt{n}$  ile ters orantılı.
- SE'yi 4'e bölmek için örneklem büyüklüğünü 16'ya çıkarmanız gerekir.
- Bootstrap, SE'yi hesaplamamanın modern ve genel yöntemidir.



# BÖLÜM III

Bootstrap Yeniden Örneklem

## Bootstrap Fikri

---

- Eldeki örneklemden yerine koyarak tekrar tekrar çekiş yapın.
- Her çekışten ilgilendiğiniz istatistiğı (ort, med, std) hesaplayın.
- Bu istatistiklerin dağılımı → örnekleme dağılımının tahmini.
- Dağılım varsayımı gerektirmez — tüm bilgi veriden gelir.
- Bradley Efron (Stanford, 1979) tarafından popülerleştirildi.

**Kavramsal: Örneklemınızı binlerce kez kopyalayarak hayali bir kitle oluşturun, sonra bu kitleden örneklem çekin.**

## Bootstrap Algoritması

| Adım | Eylem                                                                               |
|------|-------------------------------------------------------------------------------------|
| 1    | Orijinal örneklemden (n gözlem) yerine koyarak n gözlem çekin → bootstrap örnekleme |
| 2    | İstatistiği hesaplayın (ort, med, std, ...)                                         |
| 3    | 1-2'yi R kez tekrarlayın (R = 1,000 veya 10,000)                                    |
| 4    | R adet istatistiğin standart sapmasını hesaplayın → SE tahmini                      |
| 5    | Gerekirse persantil güven aralığı oluşturun                                         |

## Python'da Bootstrap — Medyan İçin SE ve GA

```
# Bootstrap: Gelir medyanı için SE ve %95 GA (Kitap s. 70-72)
R = 5_000
medians = np.array([
    np.median(np.random.choice(loans_income, len(loans_income), replace=True))
    for _ in range(R)
])

se_boot = medians.std()
ci_low, ci_high = np.percentile(medians, [2.5, 97.5])

print(f'Original medyan: {np.median(loans_income):.0f}')
print(f'Bootstrap SE: {se_boot:.0f}')
print(f'%95 GA: [{ci_low:.0f}, {ci_high:.0f}'])
```

Bootstrap vs. Permütasyon: Bootstrap yerine koyarak çeker (örnekleme dağılımı tahmini); permütasyon koymadan karıştırır (sıfır hipotezi testi). Farklı sorulara cevap verirler.



# BÖLÜM IV

Güven Aralığı

## Güven Aralığı — Nokta Tahmininden Aralık Tahminine

- Nokta tahmini: tek sayı ( $\bar{x} = 72,000$  TL). Aldatıcı hassasiyet hissi.
- Güven aralığı: makul değerler aralığı ( $[68,000; 76,000]$  TL). Belirsizliği gösterir.
- %95 GA: Bu prosedürle oluşturulan aralıkların %95'i gerçek parametreyi kapsar.
- GA'nın genişliği: örneklem büyüklüğü ve değişkenlikle belirlenir.
- Dar GA  $\rightarrow$  hassas tahmin; geniş GA  $\rightarrow$  belirsiz tahmin.

*İnce nokta: "%95 olasılıkla gerçek değer bu aralıktadır" demek teknik olarak yanlış — doğru ifade aralıkların %95'inin kapsayacağıdır. Pratikte çoğu kişi ilk yorumu yapar.*

## Bootstrap ile Güven Aralığı — İki Yöntem

### Persantil Yöntemi

GA =  $[P_{2.5}, P_{97.5}]$  bootstrap istatistikleri

- Basit ve sezgisel
- Dağılım varsayımı gerektirmez
- Küçük örnekleme sapmalı olabilir
- Bu derste varsayılan yöntemimiz

### t-Dağılımı Yaklaşımı

GA =  $\bar{x} \pm t_{\{n-1\}}(0.025) \times s/\sqrt{n}$

- Klasik formül
- Normal dağılım varsayar (veya büyük n ile MLT'ye güvenir)
- Hızlı hesaplama
- Hafta 4'te detaylı işlenecek



# BÖLÜM V

İstatistiksel Dağılımlar

## Normal (Gauss) Dağılımı

---

- Çan biçimli, simetrik,  $\mu$  ve  $\sigma$  ile tam olarak tanımlanır.
- %68 veri  $\mu \pm 1\sigma$ , %95 veri  $\mu \pm 2\sigma$ , %99.7 veri  $\mu \pm 3\sigma$  içinde.
- Ham veri genellikle Normal DEĞİLDİR — ancak örnekleme dağılımları çoğunlukla öyledir (MLT).
- z-skoru:  $(x - \mu) / \sigma \rightarrow$  standart Normal'e dönüştürme, karşılaştırma amaçlı.
- z-skoruna dönüştürmek veriyi Normal YAPMAZ — yalnızca ölçeği standartlaştırır.

## Normal Dağılım — Python Uygulaması

```
# Normal Dağılımın Ortalama ve Standart Sapma ile Z-Skorları  
from scipy import stats
```

```
x = stats.norm.rvs(loc=100, scale=15, size=1000, random_state=42)
```

```
fig, axes = plt.subplots(1, 2, figsize=(12, 4))  
axes[0].hist(x, bins=30, edgecolor='white', density=True)  
xr = np.linspace(50, 150, 200)  
axes[0].plot(xr, stats.norm.pdf(xr, 100, 15), 'r-', lw=2)  
axes[0].set_title('Normal Dağılım:  $\mu=100$ ,  $\sigma=15$ ')
```

```
# z-skorları  
z = (x - x.mean()) / x.std()  
stats.probplot(z, plot=axes[1])  
axes[1].set_title('QQ-Plot')  
plt.tight_layout(); plt.show()
```

## QQ-Plot — Normallik Değerlendirmesi

---

- QQ-plot, örneklem kantillerini teorik Normal kantilleriyle karşılaştırır.
- Noktalar doğru üzerindeyse → veri yaklaşık Normal.
- S-biçimli sapma → ağır kuyruklar (leptokurtik).
- J-biçimli sapma → çarpıklık.
- Finansal getiriler tipik olarak S-biçimli sapma gösterir — ağır kuyruklar!

```
# Netflix getirilerinin QQ-plot'u (Kitap Şekil 2-12)
nflx = sp500_px['NFLX']
nflx_ret = np.diff(np.log(nflx[nflx > 0].values))

fig, ax = plt.subplots(figsize=(4, 4))
stats.probplot(nflx_ret, plot=ax)
ax.set_title('NFLX Getirileri QQ-Plot')
plt.show()
```

## Uzun Kuyruklu Dağılımlar — "Çoğu Veri Normal Değildir"

---

- Uzun kuyruklu dağılımlarda uç değerler beklenenden çok daha sık görülür.
- Finansal getiriler, gelir dağılımı, internet trafiği → tipik uzun kuyruklu veriler.
- Normal dağılım bu verileri modellersek → uç olayların olasılığını ciddi ölçüde hafife alırız.
- Dağılımı veriye uydurma (fitting) bilim kadar sanattır — alan bilgisi gerektirir.
- QQ-plot, uzun kuyrukları tespit etmenin en hızlı görsel aracıdır.

## Student t-Dağılımı — Küçük Örneklem İçin

---

- W. S. Gosset (Guinness Bira Fabrikası, 1908) "Student" takma adıyla yayımladı.
- Normalin kalın kuyruklu versiyonu — serbestlik derecesi (sd) ile kontrol edilir.
- sd artışı  $\rightarrow$  t dağılımı Normal'e yaklaşır (sd > 30'da neredeyse aynı).
- Güven aralığı:  $\bar{x} \pm t_{n-1}(\alpha/2) \times s/\sqrt{n}$
- scipy: `stats.t.rvs(df=10, size=1000)`

## Binomial Dağılım — Evet/Hayır Sonuçları

---

- $n$  denemede, her birinde  $p$  olasılıkla başarının kaç kez gerçekleşeceğinin dağılımı.
- Bernoulli denemesi: iki sonuçlu tek deneme (yazı/tura, tıklama/tıklamama).
- Ortalama =  $np$ , Varyans =  $np(1-p)$ .
- Yeterince büyük  $n$ 'de Normal'e yakınsar.
- Bilişim uygulaması: "Tıklama olasılığı 0.02 ise 200 ziyaretten 0 dönüşüm olasılığı nedir?"

```
# 5 denemede tam 2 başarı olasılığı (p = 0.10)
print(f'P(X=2) = {stats.binom.pmf(2, n=5, p=0.1):.4f}')
```

```
print(f'P(X≤2) = {stats.binom.cdf(2, n=5, p=0.1):.4f}')
```

## Poisson ve Üstel Dağılım — Olaylar ve Bekleme Süreleri

- Birim zamanda (veya alanda) sabit oranda gerçekleşen olayların sayısı → Poisson.
- Olaylar arasındaki süre → Üstel (Exponential) dağılım.
- Parametre:  $\lambda$  (birim zamandaki ortalama olay sayısı).
- Örnekler: dakikadaki çağrı sayısı, m<sup>2</sup> başına hata, sayfadaki yazım hatası.
- Zamanla değişen olay oranı → Weibull dağılımı (cihaz ömrü modellemesi).

```
# Poisson:  $\lambda=2$  ile 100 dakika simülasyonu
```

```
olaylar = stats.poisson.rvs(mu=2, size=100, random_state=42)
print(f'Simüle edilen olay sayıları: ort={olaylar.mean():.2f}, std={olaylar.std():.2f}')
```

```
# Üstel: olaylar arası bekleme süresi
```

```
bekleme = stats.expon.rvs(scale=1/2, size=100, random_state=42)
print(f'Bekleme süreleri: ort={bekleme.mean():.2f} dakika')
```

## Temel Dağılımlar — Özet Tablo

| Dağılım  | Ne Modeller?                  | Parametre                         | scipy             |
|----------|-------------------------------|-----------------------------------|-------------------|
| Normal   | Simetrik, çan biçimli sürekli | $\mu, \sigma$                     | stats.norm        |
| t        | Küçük n, kalın kuyruk         | sd (df)                           | stats.t           |
| Binomial | n denemede başarı sayısı      | n, p                              | stats.binom       |
| Poisson  | Birim zamanda olay sayısı     | $\lambda$                         | stats.poisson     |
| Üstel    | Olaylar arası bekleme         | $\lambda$ (veya $1/\lambda$ )     | stats.expon       |
| Weibull  | Değişen oranlı ömür süresi    | k, $\lambda$                      | stats.weibull_min |
| Ki-Kare  | Sapma testi                   | sd (df)                           | stats.chi2        |
| F        | Varyans oranı (ANOVA)         | df <sub>1</sub> , df <sub>2</sub> | stats.f           |



# BÖLÜM VI

AI Destekli Keşif (3. Baskı) ve Uygulama

## Kitabın 3. Baskı AI Alıştırmaları (Bölüm 2)

---

Kitabın AI Exercises bölümünden (s. 94):

1. AI aracına LendingClub gelir verisini gönderin, bootstrap örneklemeleri üretmesini isteyin.
2. Bootstrap'ten güven aralığı hesaplatın.
3. t-istatistiği bazlı güven aralığı türetmesini isteyin.
4. "home\_ownership" durumuna göre tabakalı örneklem oluşturmasını isteyin.

Dikkat: AI yanıtlarını daima doğrulayın — AI hallüsinasyon yapabilir.

## Hafta 3 — Temel Çıkarımlar

### Kavramsal

- ✓ Rassal örnekleme yanlılığı önler
- ✓ Büyük  $n$ , yanlı örnekleme düzeltmez
- ✓ Veri dağılımı  $\neq$  örnekleme dağılımı
- ✓ MLT: ortalamalar normalleşir
- ✓ Bootstrap dağılım varsayımı gerektirmez
- ✓ GA, nokta tahmininin belirsizliğini gösterir
- ✓ Çoğu ham veri Normal değildir

### Uygulamalı

- ✓ Gelir verisinden MLT simülasyonu
- ✓ Bootstrap SE ve GA hesaplama
- ✓ QQ-plot ile normallik değerlendirmesi
- ✓ scipy ile Normal, t, Binom, Poisson üretme ve görselleştirme
- ✓ Tabakalı örnekleme uygulama
- ✓ Uzun kuyukları QQ-plot ile tespit

## Ödev 3 — Örneklem Dağılımları ve Bootstrap

- G1. Kitap deposundan loans\_income verisini yükleyin.
- G2. Ham veri dağılımının histogramını çizin. Normal mi? QQ-plot ile doğrulayın.
- G3. MLT simülasyonu:  $n=5, 20, 50$  için 1000 ortalama çekin, histogramlarını FacetGrid ile çizin.
- G4. Medyan için bootstrap ( $R=5,000$ ): SE hesaplayın, persantil GA oluşturun.
- G5. Bootstrap GA ile t-dağılımı GA'yı karşılaştırın — hangisi daha geniş, neden?
- G6. state veri setinden Murder.Rate için bootstrap GA hesaplayın.
- G7. tips veri setinde smoker/non-smoker grupları için tip\_ratio ortalamasının bootstrap GA'sını hesaplayın.  
İki GA örtüşüyor mu? Bu ne anlama geliyor?

Teslim: 4. Hafta ders saatinden önce • Moodle → BSM622 → 3. Hafta → Ödev

## Hafta 4 — İstatistiksel Deneyler ve Anlamlılık Testi

---

### Hazırlık

- Bruce vd., Bölüm 3: Hipotez testi ve permütasyon bölümlerini okuyun
- Ödev 3'ü teslim edin
- Gelecek hafta:  $H_0/H_1$ , p-değeri, permütasyon testi, t-testi, Cohen's d, güven aralığı

Akademik İletişim: [onurpolat@hacettepe.edu.tr](mailto:onurpolat@hacettepe.edu.tr)