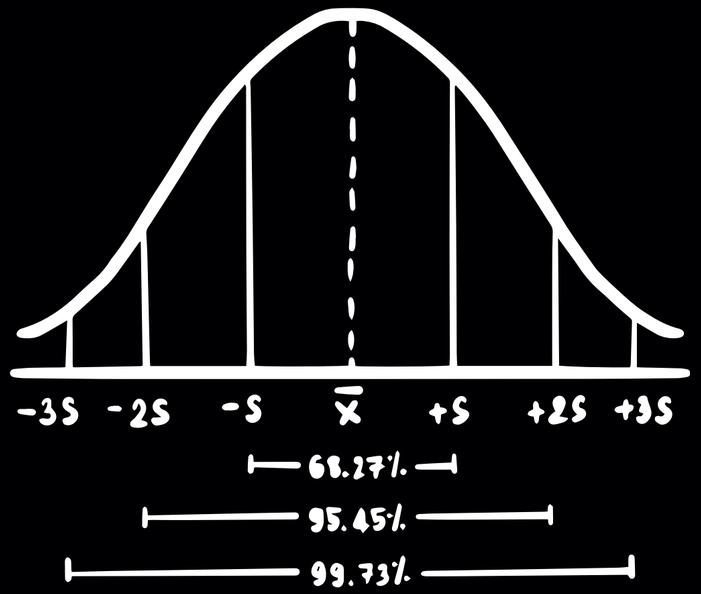
Veri Bilimi İçin İstatistik

ATIL SAMANCIOĞLU



$$\mu = \frac{\sum x_i}{n}$$

Statistics

Descriptive

Betimsel

Inferential

Çıkarımsal

Descriptive

Betimsel

Verileri düzenleme ve özetlemeden (organizing and summarizing the data sorumludur. Yani, eldeki veriyi anlamak için kullanılır

Mean (ortalama)

Median

Mode

Variance

Standart sapma

Inferential

Çıkarımsal

iyi verilerden sonuçlar çıkarma (drawing conclusions from good data) işlemidir. Bu, örneklem verilerini kullanarak daha büyük bir popülasyon hakkında genellemeler ve tahminler yapmak anlamına gelir.

Veri topla
Örneklem al
Deneyler yap
Bunlara bakarak sonuçlar çıkar

Örnek: Üniversite Sınav Sonuçları ile Açıklayıcı ve Çıkarımsal İstatistiklerin Karşılaştırılması

DESCRIPTIVE

Diyelim ki bir üniversitenin İstatistik 101 dersindeki 200 öğrencinin final sınavı notlarını topladınız. Aşağıdaki değerleri hesapladınız:

- Ortalama (Mean) not = 75
- Medyan (Ortanca) not = 78
- Standart sapma = 10
- Not dağılımını gösteren histogram

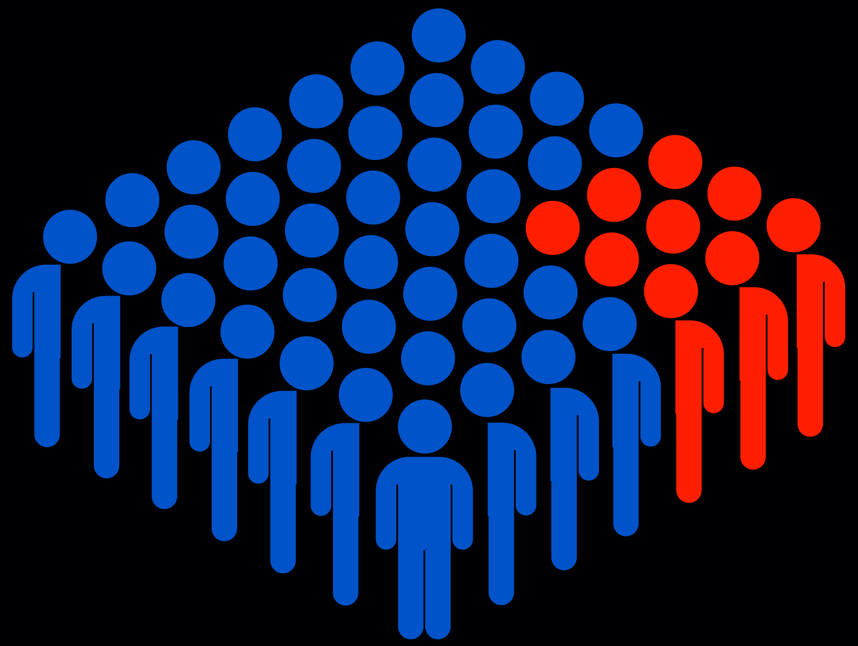
Bu veriler yalnızca elimizdeki 200 öğrenciye ait veriyi açıklar ve herhangi bir tahminde bulunmaz.

INFERENTIAL

Şimdi, bu 200 öğrencinin verilerini kullanarak tüm üniversitedeki (10.000 öğrenci) İstatistik 101 öğrencilerinin final sınav not ortalamasını tahmin etmek istiyorsunuz.

- Örneklem ortalamasını (75) kullanarak güven aralığı (confidence interval) hesaplayarak genel öğrenci ortalamasını tahmin edersiniz.
- Bu yılın sınav zorluğunun geçen yıla göre aynı olup olmadığını anlamak için bir hipotez testi yapabilirsiniz.
- Regresyon analizi kullanarak öğrencilerin çalışma saatleri ile sınav notları arasında bir ilişki olup olmadığını inceleyebilirsiniz.

Burada çıkarımsal istatistikler, eldeki örneklem verilerinden yola çıkarak daha büyük bir öğrenci kitlesi hakkında tahminde bulunmamıza olanak tanır.



Population (Ana Kitle)

Diyelim ki bir ilaç firması, yeni geliştirdiği bir grip aşısının tüm Türkiye'deki yetişkinler üzerinde ne kadar etkili olduğunu öğrenmek istiyor.

Burada tüm Türkiye'deki yetişkinlerin tamamı bizim popülasyonumuzdur (ana kitle).

Ancak, 80 milyon insanın hepsini test etmek imkansızdır, bu yüzden bir örneklem seçilir.

Sample (Örneklem)

Araştırmacılar, farklı yaş gruplarından ve bölgelerden rastgele seçilmiş 5.000 kişiyi alıp bu kişilere aşı yaparak sonuçları analiz ederler.

Bu 5.000 kişi, popülasyonun küçük bir alt kümesi olduğu için bizim örneklemimizdir.
Ana Fark

- Popülasyon (Ana Kitle): Çalışmanın ilgilendiği tüm grup (Türkiye'deki tüm yetişkinler).
- Örneklem (Sample): Popülasyondan seçilmiş daha küçük bir grup (5.000 kişi).

Örneklemden elde edilen verilerle, tüm popülasyon hakkında çıkarım yapmaya çalışırız.

Measure of Central Tendency (Merkezi Yönelim)

Mean (Ortalama)

$$\bar{x} = \frac{\sum x}{n} \qquad \mu = \frac{\sum x}{n}$$

notlar = {40, 50, 60, 70, 80, 100, 120}

ortalama = (40 + 50 + 60 + 70 + 80 + 100 + 120) / 7

= 74.28

Measure of Central Tendency (Merkezi Yönelim)

Median (Medyan - Ortanca)

notlar = {40, 50, 60, 70, 80, 100, 120}

median = 40, 50, 60, 70, 80, 100, 120

= 70

Outlier data

notlar = {40, 50, 60, 70, 80, 100, 120, 1500}

ortalama = 252.5

median = (70 + 80) / 2

= 75

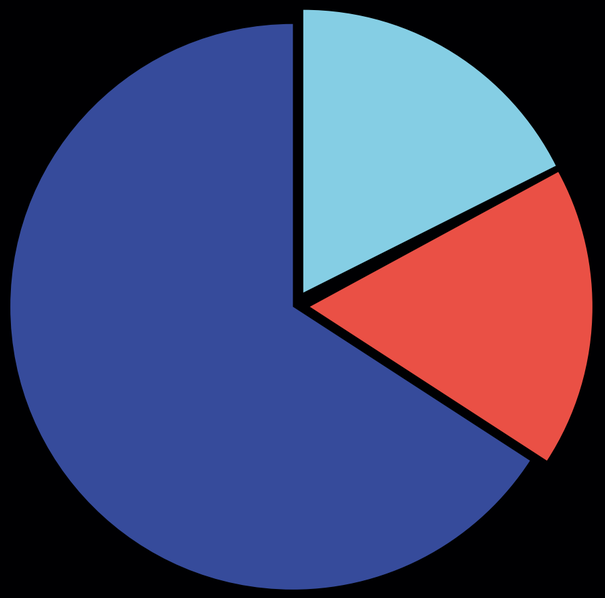
medyan ile mean arasında çok fark varsa outlier data olma ihtimali yük

Measure of Central Tendency (Merkezi Yönelim)

Mode (Mod)

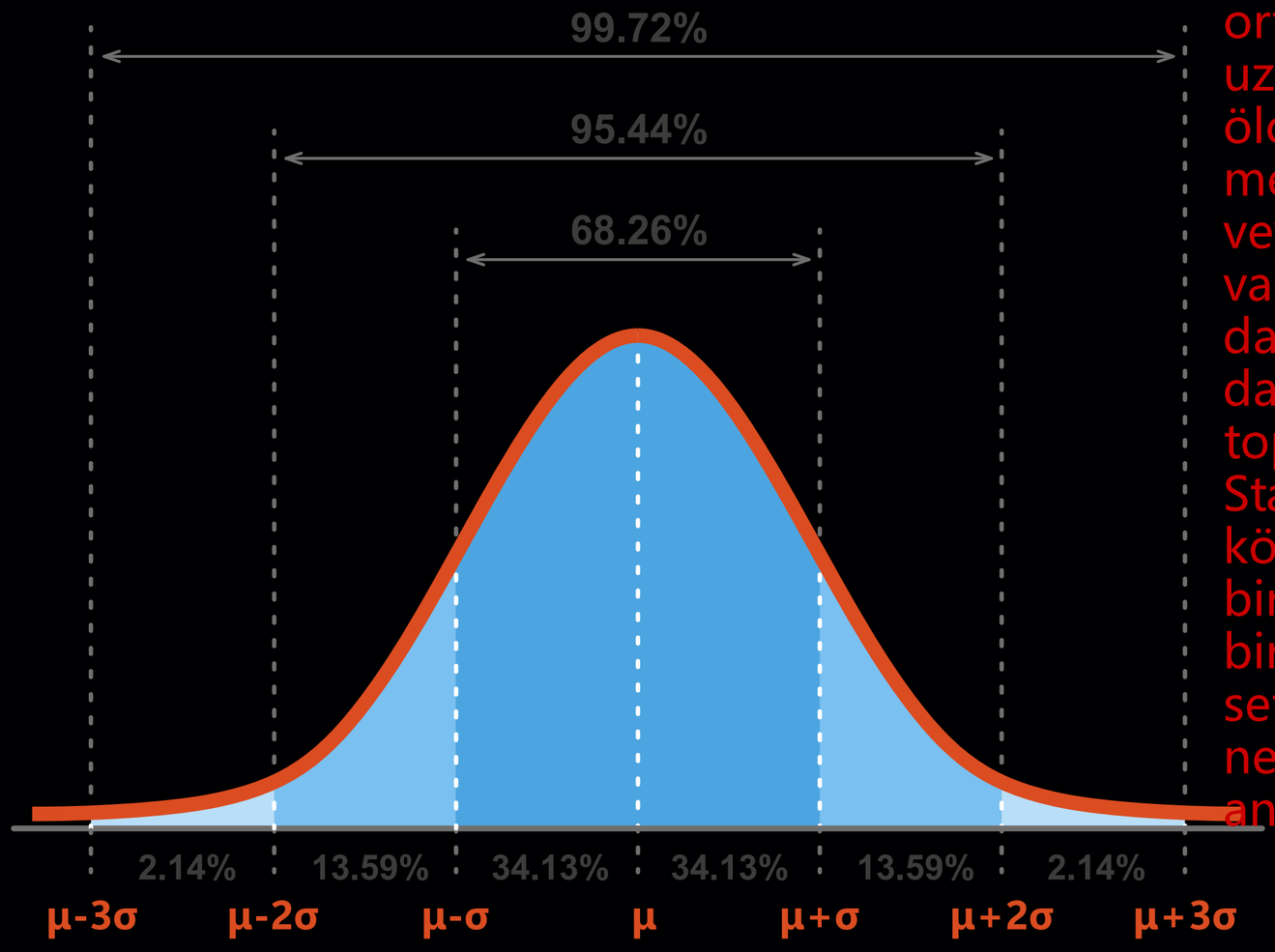
notlar = {40, 40, 40, 40, 40, 50, 50, 60, 70, 80, 200}

mod = 40



Measure of Dispersion (Yayılımın Ölçümü)

Standard Deviation (Standart Sapma) & Variance (Varyans)



Varyans: Veri noktalarının ortalamadan ne kadar uzaklaştığını karelerine alarak ölçer. Bu, veri setinin ne kadar merkezi etrafında toplandığını veya dağıldığını gösterir. Yüksek varyans, verilerin geniş bir alana dağıldığını, düşük varyans ise daha yoğun bir şekilde toplandığını gösterir.
Standart Sapma: Varyansın kare köküdür ve daha kullanıcı dostu bir ölçüdür çünkü aynı birimlerle ifade edilir. Bu da, veri setinin ortalamadan ortalama ne kadar uzak olduğunu kolayca anlamamıza yardımcı olur.

Measure of Dispersion (Yayılımın Ölçümü)

notlar = {10, 10, 50, 50}

ortalama = 30

diger_notlar = {20, 20, 40, 40}

diger_ortalama = 30

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

Variance Formülü

Measure of Dispersion (Yayılımın Ölçümü)

notlar = {10, 10, 50, 50}

ortalama = (10 + 10 + 50 + 50) / 4

= 30

diger_notlar = {20, 20, 40, 40}

ortalama = (20 + 20 + 40 + 40) / 4

= 30

$$\sum (x_i - \mu)^2$$

x_i	μ	$(x_i - \mu)^2$
10	30	400
10	30	400
50	30	400
50	30	400

1600

x_i	μ	$(x_i - \mu)^2$
20	30	100
20	30	100
40	30	100
40	30	100

400

Measure of Dispersion (Yayılımın Ölçümü)

notlar = {10, 10, 50, 50}

$$\sum (x_i - \mu)^2 \quad 1600$$

$$\sigma^2 \quad 400$$

$$\sigma \quad 20$$

diger_notlar = {20, 20, 40, 40}

$$\sum (x_i - \mu)^2 \quad 400$$

$$\sigma^2 \quad 100$$

$$\sigma \quad 10$$

Measure of Dispersion (Yayılımın Ölçümü)

Sınıf	Notlar	Ortalama	Varyans	Standart Sapma
A	70, 72, 68, 69, 71	70	2.5	1.58
B	50, 90, 40, 100, 60	68	625	25

1. Varyans, verilerin ortalamadan ne kadar saptığını ölçer ama kareli birimlerde gösterir, bu yüzden doğrudan yorumlamak zor olabilir.

- A Sınıfının varyansı (2.5) düşük, yani öğrencilerin notları birbirine çok yakın.
- B Sınıfının varyansı (625) yüksek, yani bazı öğrenciler çok yüksek, bazıları ise çok düşük not almış.

2. Standart Sapma, varyansın karekökü olduğu için verileri orijinal birimine geri döndürür ve yorumlamayı kolaylaştırır.

- A Sınıfında standart sapma 1.58, yani öğrencilerin notları çok fazla değişmiyor.
- B Sınıfında standart sapma 25, yani öğrencilerin notları çok geniş bir aralığa yayılmış.

A Sınıfında öğrencilerin notları birbirine yakın (düşük değişkenlik).
B Sınıfında bazı öğrenciler çok iyi, bazıları çok kötü olduğu için notlar arasında büyük farklılıklar var (yüksek değişkenlik).

Measure of Dispersion (Yayılımın Ölçümü)

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

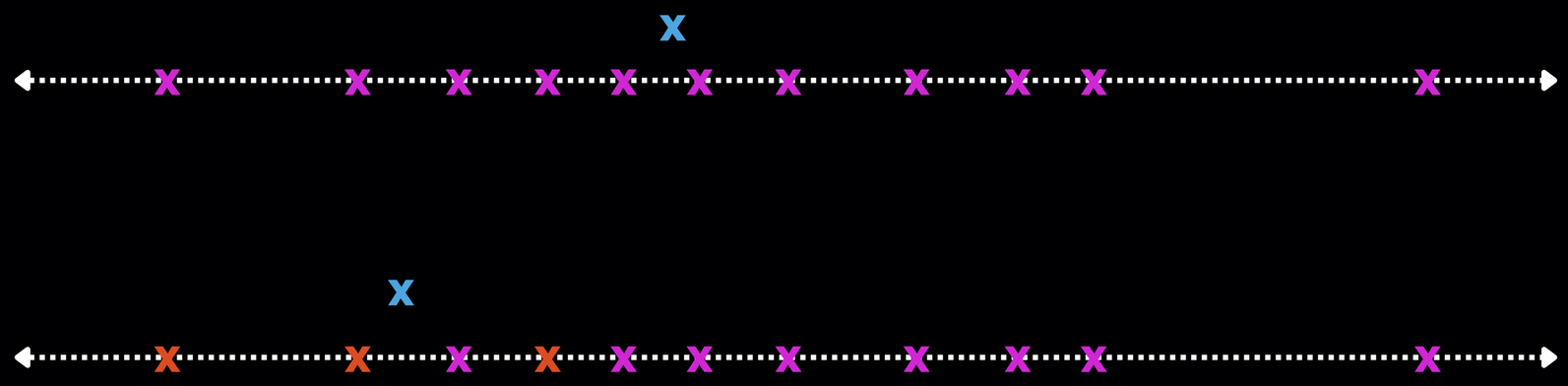
Population Variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Sample Variance

Bessel’s Correction (n-1)

Measure of Dispersion (Yayılımın Ölçümü)



Temel sebep, bir örneklem kullanarak tüm popülasyonu tahmin ettiğimizde oluşan yanlılığı düzeltme gerekliliğidir. Örneklem varyansı formülü, popülasyon varyansını daha doğru tahmin etmek için düzeltilmiştir.

Yukarıdaki örnekteki örneklem seçildiği takdirde varyans olması gerektiğinden daha küçük çıkacaktır. n-1 düzeltmesiyle bu bias (eğilim, yanlılık) düzeltilebilir.

Measure of Dispersion (Yayılımın Ölçümü)

Örnek: Bir Okuldaki Öğrencilerin Boy Uzunlukları

Bir okulda 500 öğrenci (popülasyon) olduğunu düşünelim. Tüm öğrencilerin ortalama boyu 160 cm ve gerçek popülasyon varyansı 18 cm^2 olsun.

Durum 1: Popülasyon Varyansını Kullanma

Eğer 500 öğrencinin tamamının boylarını ölçersek, varyansı doğrudan N (500) ile bölerek hesaplarız. Çünkü tüm veriye sahibiz ve herhangi bir tahminde bulunmamıza gerek yok.

Durum 2: Örneklem Varyansını Kullanma

Şimdi, popülasyondan rastgele 5 öğrenci seçtiğimizi düşünelim:

- Boy uzunlukları: 155, 162, 159, 165, 157
- Örneklem ortalaması ($\bar{x}$) = 159.6 cm
- **Örneklem varyansı (n) ile yapılırsa = 12.64**
- **Örneklem varyansı (n-1) ile yapılırsa = 15.75**

Eğer varyansı n (5) ile bölersek, gerçek popülasyon varyansı 18 cm^2'den daha küçük bir değer elde ederiz, yani tahminimiz yanlış olur.

Bunun yerine n - 1 (4) ile böldüğümüzde, örneklem varyansını daha doğru bir tahmine dönüştürmüş oluruz.

Variable (Değişken)

İstatistikte değişken (variable), incelenen birimlere ait ölçülebilen veya sınıflandırılabilen özelliklere denir. Değişkenler farklı değerler alabilir ve iki ana gruba ayrılır:

1. Nicel (Sayısal) Değişkenler

Bu değişkenler sayılarla ifade edilir ve matematiksel işlemler yapılabilir.

a) Sürekli Değişkenler (Sonsuz değer alabilir, Continuous Variable)

- Öğrencinin sınav puanı → (85.5, 90.3, 76.8)
- Öğrencinin boyu → (165.2 cm, 178.9 cm)
- Öğretmenin maaşı → (12.500 TL, 15.750 TL)

b) Kesikli Değişkenler (Sayılabilir, Discrete Variable)

- Bir sınıftaki öğrenci sayısı → (25, 30, 35)
- Bir öğrencinin aldığı ders sayısı → (4, 5, 6)
- Bir öğretmenin verdiği sınav sayısı → (2, 3, 5)

2. Nitel (Kategorik) Değişkenler

Bu değişkenler sayısal değer taşımaz, gruplandırma veya sınıflandırma için kullanılır.

a) Adlandırılmış (Nominal) Değişkenler (Sıralama yok)

- Öğrencinin okuduğu bölüm → (Matematik, Türk Dili, Tarih, Fizik)
- Öğrencinin cinsiyeti → (Kadın, Erkek)
- Eğitim türü → (Uzaktan, Örgün, Hibrit)

b) Sıralı (Ordinal) Değişkenler (Sıralama var ama aralık belli değil)

- Öğrencinin mezuniyet seviyesi → (İlkokul < Ortaokul < Lise < Üniversite < Yüksek Lisans)
- Bir öğretmenin kıdem seviyesi → (Çaylak, Deneyimli, Uzman)
- Öğrencinin ders notu → (Çok kötü, Kötü, Orta, İyi, Çok iyi)

Rastgele Değişken (Random Variable)

bir deney veya olay sonucunda önceden kesin olarak bilinmeyen ancak olasılıkla belirlenebilen değiskenlerdir

İstatistikte rastgele değişken (random variable), bir deneyin olası sonuçlarını sayısal olarak temsil eden bir değişkendir. Bu değişkenin aldığı değerler rastgele belirlenir ve her bir değerin belirli bir olasılığı vardır.

Bir rastgele değişken genellikle X, Y, Z gibi büyük harflerle gösterilir ve belirli bir değeri x, y, z gibi küçük harflerle ifade edilir.

sınıftan rastgele seçilen öğrencinin boyu
bir zar attığımızda gelecek sayı

Rastgele Değişken Türleri

1. Ayrık Rastgele Değişken (Discrete Random Variable)

- Sayılabilir ve belirli değerler alabilen değişkenlerdir.

- Genellikle tam sayılar ile ifade edilir.

Örnekler:
- Bir sınavda doğru yapılan soru sayısı ($X = \{0, 1, 2, ..., 50\}$)
- Bir sınıfta bulunan öğrenci sayısı ($Y = \{20, 21, ..., 35\}$)
- Bir öğrencinin bir yıl içinde aldığı ders sayısı ($Z = \{1, 2, 3, ..., 10\}$)

2. Sürekli Rastgele Değişken (Continuous Random Variable)

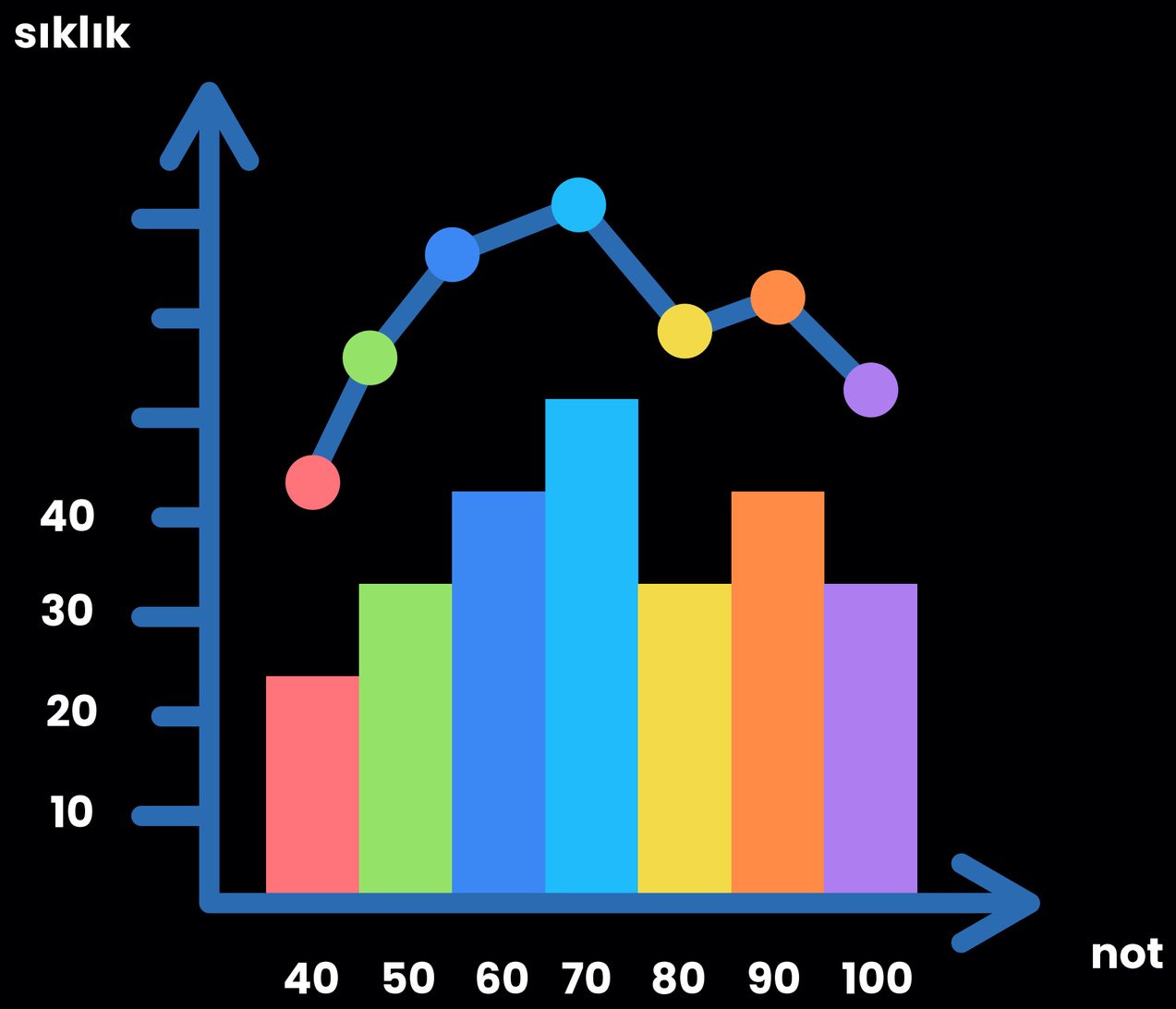
• Bir aralıktaki herhangi bir reel sayıyı alabilen değişkenlerdir.

Örnekler:

• Bir öğrencinin sınavdan aldığı puan (X = [0, 100])
• Bir öğretmenin maaşı (Y = [10.000 TL, 20.000 TL])
• Bir öğrencinin boyu (Z = [140 cm, 200 cm])

Deney	Rastgele Değişken (X)	Türü
Bir öğrencinin sınavdan kaç doğru yaptığı	X: Doğru cevap sayısı (0, 1, ..., 50)	Ayrık
Bir öğrencinin sınavdan aldığı puan	X: Puan (0-100 arasında herhangi bir değer)	Sürekli
Bir öğrencinin derse katılma süresi	X: Katılım süresi (0-120 dakika)	Sürekli
Bir sınıfta devamsızlık yapan öğrenci sayısı	X: Devamsız öğrenci sayısı (0, 1, ..., 30)	Ayrık

Histogram



Yüzde (Percentage)

Yüzde, bir sayının 100'ün bir kesri olarak ifade edilmesidir. Bir bütünün parçalarını karşılaştırmak veya oranları ifade etmek için kullanılır.

Diyelim ki bir sınavda 50 üzerinden 45 puan aldınız. Yüzdeyi hesaplamak için:

(45 / 50) * 100 = %90

Bu, sınavda %90 aldığınız anlamına gelir.

Yüzdelik (Percentile)

Yüzdelik, bir veri setinde belirli bir yüzdenin altında kalan değeri ifade eder. Veriyi 100 eşit parçaya böler.

Önemli Noktalar:
50. yüzdelik, veri setinin medyanıdır (ortanca değer).

Yüzdelikler, bireysel performansı daha büyük bir grupla karşılaştırmak için kullanılır (örneğin, sınav puanları, büyüme eğrileri).

Yüzdelik (Percentile)

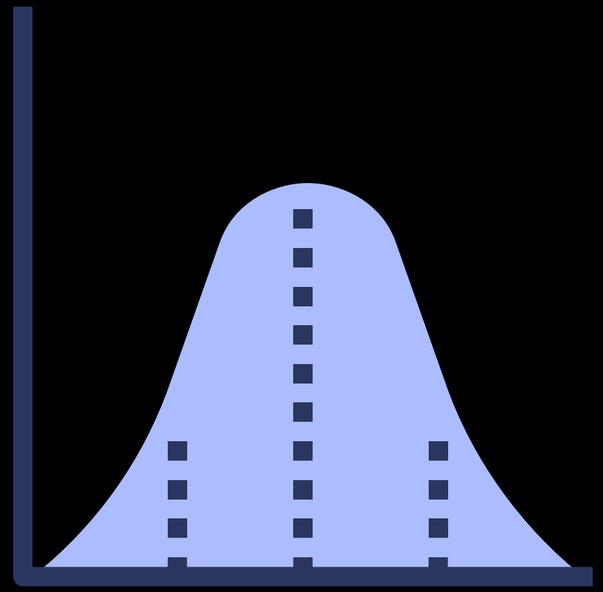
$$\text{Percentile} = \left(\frac{k}{100}\right) * n$$

numaralar = (2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22)

Bu listeyi kullanarak yüzdelikleri hesaplayalım:

25. Yüzdelik:
Pozisyon = (25 / 100)×11 = 2.75
2.75 tam sayı olmadığı için 3'e yuvarlıyoruz.
25. yüzdelik’e denk gelen 3. sıradaki numara **= 6**

50. Yüzdelik:
Pozisyon = (50/100)×11 = 5.5
5.5 tam sayı olmadığı için 6'ya yuvarlıyoruz.
50. yüzdelik’e denk gelen 6. sıradaki 12



Çeyreklik (Quartile)

Çeyreklik, bir veri setini dört eşit parçaya böler. Üç çeyreklik vardır: Q1, Q2, ve Q3.

Önemli Noktalar:

- Q1 (Birinci Çeyreklik): Verinin alt yarısının medyanı (25. yüzdelik).
- Q2 (İkinci Çeyreklik): Tüm veri setinin medyanı (50. yüzdelik).
- Q3 (Üçüncü Çeyreklik): Verinin üst yarısının medyanı (75. yüzdelik).

numaralar = (2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22)

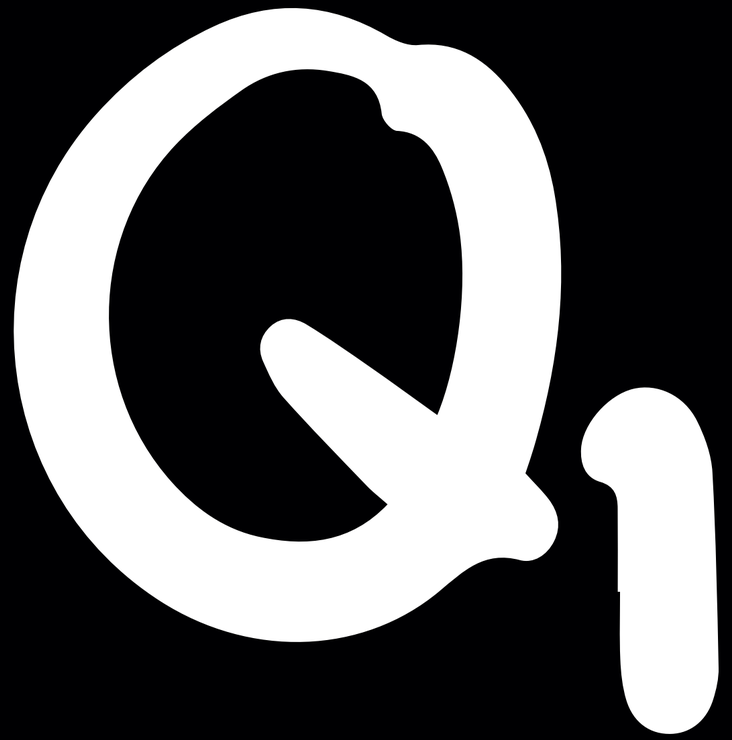
75. Yüzdelik:
Pozisyon = (75 / 100)×11 = 8.25
8.25 tam sayı olmadığı için 9'a yuvarlıyoruz.
75. yüzdelik'e denk gelen 9. sıradaki numara **= 18**

5 Numara Özetleri

1) Minimum
2) 1. Çeyrek
3) Medyan
4) 3. Çeyrek
5) Maximum

numaralar = (2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22)

1) Minimum = 2
2) 1. Çeyrek = 6
3) Medyan = 12
4) 3. Çeyrek = 18
5) Maximum = 22



Covariance

Covariance, iki değişkenin birlikte nasıl değiştiğini gösteren bir ölçümdür. Pozitif veya negatif olabilir:

- **Pozitif Kovaryans** → Bir değişken artarken diğeri de artıyorsa (veya ikisi de azalıyorsa).
- **Negatif Kovaryans** → Bir değişken artarken diğeri azalıyorsa (ters yönlü hareket).

$$Covariance_{sample}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

$$Covariance_{pop}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{n}$$

Covariance

- Hava sıcaklığı arttıkça dondurma satışı da artar.
- Bu durumda kovaryans pozitif olur.

Gün	Sıcaklık (°C)	Dondurma Satışı
1	20	40
2	22	45
3	24	50
4	26	55
5	28	60

xi	xi − $\bar{X}$	yi	yi − Y	(xi − $\bar{X}$) (yi − Y)
20	−4	40	−10	40
22	−2	45	−5	10
24	0	50	0	0
26	2	55	5	10
28	4	60	10	40
Σ xi = 120		**Σ yi = 250**		**Σ (xi − $\bar{X}$) (yi − Y) = 100**

Bu örnekte toplam 100 / 5 = 20 kovaryans değerine ulaşırız

Kovaryansın birimi değişkenlerin birimlerine bağlıdır (°C × Satış), bu yüzden korelasyon daha sık kullanılır çünkü ölçekten bağımsızdır.

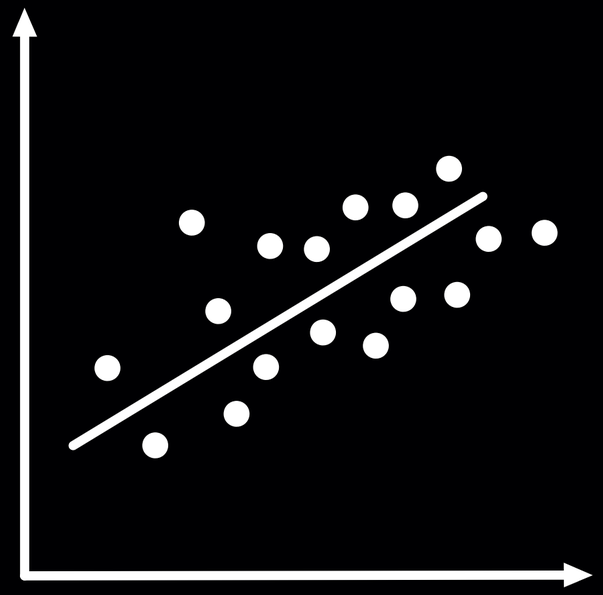
Correlation (Korelasyon)

Korelasyon, iki değişken arasındaki ilişkinin gücünü ve yönünü ölçen bir değerdir. Kovaryanstan farklı olarak, korelasyon her zaman -1 ile 1 arasında bir değerdir, bu da onu daha anlaşılır ve karşılaştırılabilir yapar.

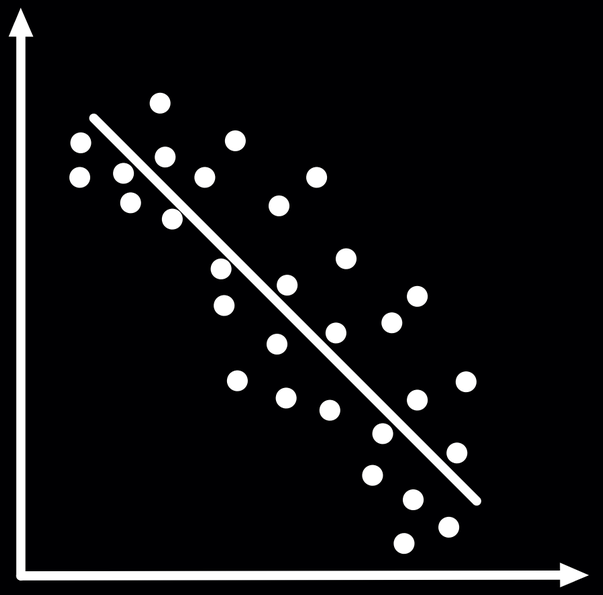
- **Pozitif Korelasyon (+1'e yakın)** → Bir değişken artarken diğeri de artıyorsa.
- **Negatif Korelasyon (-1'e yakın)** → Bir değişken artarken diğeri azalıyorsa.
- **0'a yakın Korelasyon** → Değişkenler arasında belirgin bir ilişki yoksa.

$$\rho_{XY} = \frac{Covariance(X, Y)}{\sigma_X \sigma_Y}$$

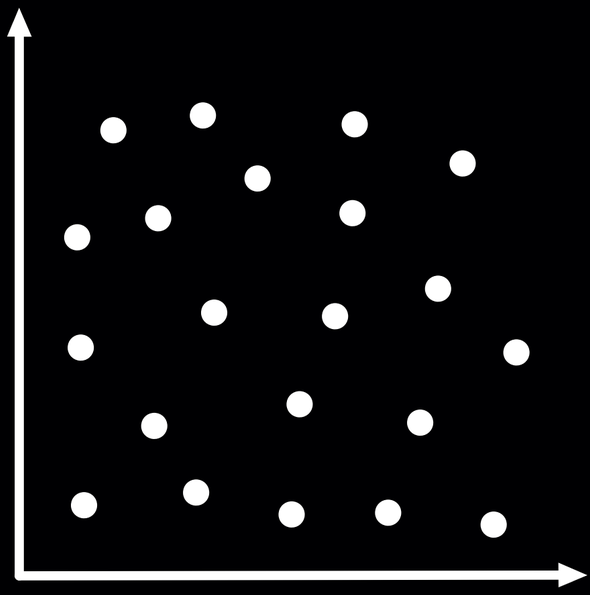
Correlation (Korelasyon)



Korelasyon Pozitif



Korelasyon Negatif



Korelasyon Nötr

Correlation (Korelasyon)

xi	xi – X̄	yi	yi – Y	(xi – X̄) (yi – Y)
20	-4	40	-10	40
22	-2	45	-5	10
24	0	50	0	0
26	2	55	5	10
28	4	60	10	40
	Σ (xi- X̄)^2 = 4 0		**Σ (xi- Y)^2 = 250**	**Σ (xi – X̄) (yi – Y) = 100**

$$\rho_{XY} = \frac{Covariance(X,Y)}{\sigma_X \sigma_Y}$$

Kovaryans = 20

SigmaX = sqrt(40 / 5)
SigmaY = sqrt(250 / 5)
SigmaX * SigmaY = sqrt (400)
SigmaX*SigmaY = 20

Korelasyon = 20 / 20
= 1

Correlation (Korelasyon)

Pearsons vs Spearmans vs ...

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

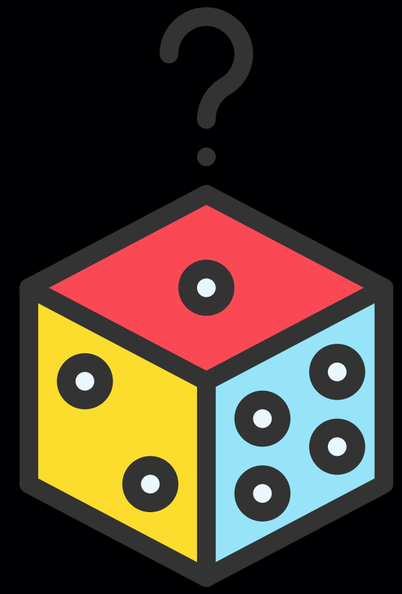
n = number of observations

Probability (Olasılık)

Olasılık, bir olayın gerçekleşme ihtimalini ölçen matematiksel bir kavramdır. 0 ile 1 arasında bir değerdir:

- **0** → Olayın gerçekleşmesi imkansızdır. (Örn: Zar atıp 7 gelmesi)
- **1** → Olayın kesinlikle gerçekleşeceğini gösterir. (Örn: Güneşin doğması)
- **0.5** → Olayın gerçekleşme ihtimali %50'dir. (Örn: Dürüst bir para atışında yazı veya tura gelmesi)

$$P(A) = \frac{\text{istenilen durum sayısı}}{\text{tüm olası durum sayısı}}$$



Probability (Olasılık)

Madeni para havaya atıldığında tura gelme ihtimali nedir?

$$P(T) = \frac{1}{2} \qquad P(T) = P(Y) = \%50$$

$$= 0.5$$

$$= \mathbf{50\%}$$

Bir zar atıldığında 4 gelme ihtimali nedir?

$$P(4) = \frac{\mathbf{1}}{\mathbf{6}} \qquad P(1) = P(2) = P(3) = P(4) = P(5) = P(6)$$