

2025

İstatistik Ders Notları

BEYZA KÜÇÜK

İçindekiler Tablosu

1. Bölüm : Temel İstatistik Kavramları.....6

- **İstatistiğin Anlamı**
- Eisenhower Matrisi
- Karakterizasyon
 - Veri Toplama
 - Analiz
 - Görselleştirme
 - Çıkarım
 - Sunum
 - İstatistiğin Nedenleri
 - İstatistiğin Önemi
- Veri Bilimi vs İstatistik
 - Ne Kadar İstatistik Bilgisi Gerekiyor?
 - İstatistik Türleri
- Betimsel İstatistik (Descriptive)
- Çıkarımsal İstatistik (Inferential)
 - Veri Türleri
 - Parametreler ve İstatistikler
 - Olasılık vs İstatistik
 - Ölçüm Düzeyleri
- Nominal
- Ordinal
- Aralıklı (Interval)
- Oranlı (Ratio)
- İstatistiğin Gerçek Hayatta Kullanım Alanları
- Veri Bilimi vs İstatistik

2. Bölüm : Veri Görselleştirme ve Temel Analiz.....12

- **Veri Görselleştirme - Grafiksel Temsil**
 - **Desenler**
- Merkez (Center)
- Yayılmış (Spread)
- Şekil (Shape)
- Simetri (Symmetric)
- Tepe Sayısı (Number of Peaks)
- Çarpıklık (Skewness)
- Düzgün Dağılım (Uniform)
 - **Olağandışı Özellikler**
- Boşluklar (Gaps)
- Aykırı Değerler (Outliers)

- **Frekans Tablosu**
- Göreli Frekans (Relative Frequency)
- Kümülatif Frekans (Cumulative Frequency)
 - **Çubuk Grafik (Bar Chart)**
 - **Pasta Grafik (Pie Chart)**

- Histogram
- Popülasyonlar ve Örneklemeler
- Parametreler ve İstatistikler
 - Merkezi Eğilim (Measure of Centre)
- Ortalama (Mean)
- Medyan (Median)
- Mod (Mode)
 - Yayılım (Measure of Spread)
- Aralık (Range)
- Çeyrekler Arası Aralık (IQR)
- Standart Sapma (Standard Deviation)
- Ampirik Kural (Empirical Rule)
 - Varyasyon (Variation)

3. Bölüm : İlişki Analizi.....20

- Saçılım Grafiği (Scatter Plot)
- Line Of Best Fit
- Doğrusallık (Linearity)
- Eğim (Slope)
- Güç (Strength)
 - Olağanüstü Özellikler
- Kümeler (Clusters)
- Boşluklar (Gaps)
- Aykırı Değerler (Outliers)
 - Kutu Grafiği (Box Plot)
- Min ve Max Değerler
- $1.5 \times \text{IQR}$ (John Tukey Kuralı)
 - Kovaryans (Covariance)
 - Korelasyon
- Pearson Korelasyon Katsayısı
- Korelasyon - Doğrusal İlişki

4. Bölüm : Regresyon Analizi.....32

- Lineer Regresyon
- Bağımlı ve Bağımsız Değişkenler
- Regresyon Denklemi
- Pearson's r Hesaplaması
- Artık Terim (Residual Term - e)
 - Determinasyon Katsayısı – R^2

5. Bölüm : Olasılık.....41

- Olasılık
- Büyük Sayılar Yasası (Law of Large Estimates)
- Örneklem Uzayı – Olay (Sample Space – Event)
- Bağımsız – Bağımlı Olaylar (Independent – Dependent Event)
- İki Bağımsız Olayın Olasılığı
- Kesişim, Birleşim, Tümleyen (Intersection, Union, Complement)
- Permütasyon (Permutation)

- Kombinasyon (Combination)
- Koşullu Olasılık (Conditional Probability)
 - Bağımsızlık Kontrolü (Independence Check)
 - **Bayes Teoremi**

6. Bölüm : Rastgele Değişkenler ve Dağılımlar.....48

- **Rastgele Değişkenler (Random Variables)**
- **Olasılık Dağılımları**
- **Kesikli Olasılık Dağılımları**
 - Olasılık Kütle Fonksiyonu (PMF)
 - Kümülatif Dağılım Fonksiyonu (CDF)
 - Binom Dağılımı (Binomial Distribution)
 - Bernoulli Dağılımı (Bernoulli Distribution)
 - Poisson Dağılımı (Poisson Distribution)
 - Geometrik, Hipergeometrik, Negatif Binom
- **Sürekli Olasılık Dağılımları**
 - Düzgün Dağılım (Uniform Distribution)
 - Normal Dağılım (Normal Distribution)
 - Z Tablosu (Z Table)
 - Standart Dağılım (Standard Distribution)
 - T Dağılımı (Student's T-Distribution)
 - Üstel, Gamma, Ki-Kare, F Dağılımları

7. Bölüm : Örneklem Dağılımları ve Güven Aralıkları.....59

- **Örneklem Dağılımı (Sample Distribution)**
- Basit Rastgele Örnekleme (Simple Random Sampling - SRS)
- Ortalama için Standart Hata (Standard Error of the Mean)
- Merkezi Limit Teoremi (Central Limit Theorem)
 - Normal Dağılımin Avantajları
- Güven Aralığı (Confidence Interval)

8. Bölüm : Hipotez Testleri.....66

- **Hipotez (Anlamlılık) Testi**
- Hipotez Testi Adımları
 - Varsayımlar (Assumptions)
 - Hipotezler (Hypotheses)
 - Null Hipotez (Null Hypotheses)
 - Alternatif Hipotez (Alternative Hypotheses)
 - Test İstatistiği (Test Statistic)
 - P-Değeri (P-Value)
 - Sonuçlar (Conclusions)
- Anlamlılık Düzeyi (α – Alpha)
- Tip I – II Hatalar (Type I – II Error)
- Tek – İki Kuyruk Testleri (One – Two Tail Tests)
 - Sol Kuyruk Testi (Left Tail Test)

- Sağ Kuyruk Testi (Right Tail Test)
- İki Taraflı Test (Two Side Test)
- T Testi
- Z Testi

9. Bölüm : İleri Hipotez Testleri.....72

- **Bağımsız Örneklemler T Testi (Independent Samples T Test)**
- **Bağımlı T Testi – Eşleştirilmiş T Testi (Paired T Test)**
- **Tek Yönlü ANOVA (One Way ANOVA)**
- Test İstatistikleri – ANOVA Tablosu
 - Regresyon Kareler Toplamı (SSR)
 - Hata Kareler Toplamı (SSE)
 - Toplam Kareler Toplamı (SST)
 - Regresyon Ortalama Kareler (MSR)
 - Hata Ortalama Kareler (MSE)
 - F İstatistiği
 - **Kategorik Veri Analizi**
 - **Ki-Kare Testi (Chi-Square Test)**

1. Bölüm : Temel İstatistik Kavramları

İstatistikin Anlamı

İstatistik, verileri toplama, analiz etme, yorumlama ve sunma bilimidir. Verilerden anlamlı sonuçlar çıkarmak için kullanılır.

Metod Olarak İstatistik

İstatistiğe konu olabilen olaylara ait nicel verilerin derlenmesi, işlenmesi, analizi ve yorumlanmasıında kullanılan teknikler bütündür.

Eisenhower Matrisi

Önceliklendirme yapmak için kullanılan bir araçtır. İşleri aciliyet ve önem durumuna göre sınıflandırır.



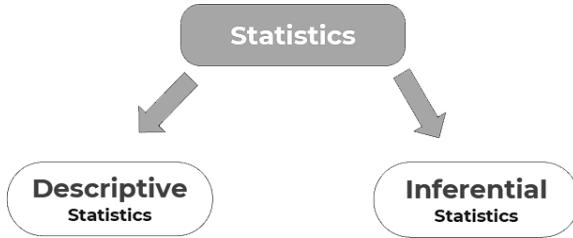
Karakterizasyon

- Verilerin özelliklerini belirleme ve tanımlama sürecidir. Bu süreç şu adımları içerir:
 - Veri Toplama:** Verilerin sistematik bir şekilde toplanması.
 - Analiz:** Verilerin istatistiksel yöntemlerle incelenmesi.
 - Görselleştirme:** Verilerin grafikler veya tablolar halinde sunulması.
 - Çıkarım:** Verilerden anlamlı sonuçlar çıkarma.
 - Sunum:** Bulguların anlaşılır bir şekilde paylaşılması.
 - İstatistikin Nedenleri:** Verilerin doğru yorumlanması ve karar verme süreçlerine rehberlik etmesi.
 - İstatistikin Önemi:** Bilimsel araştırmalarda, iş dünyasında ve günlük生活中 doğrudan kararlar almak için kritiktir.

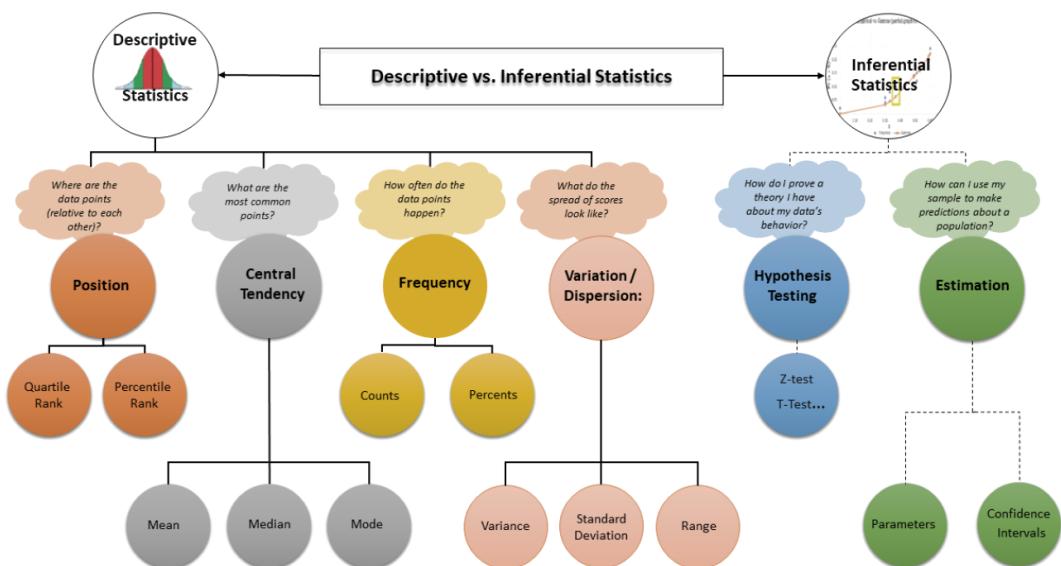
Ne Kadar İstatistik Bilgisi Gerekiyor?

Veri bilimciler, istatistiksel yöntemleri kullanarak verilerden anlamlı bilgiler çıkarır. Temel istatistik bilgisi, veri bilimi için gereklidir.

Istatistik Türleri:

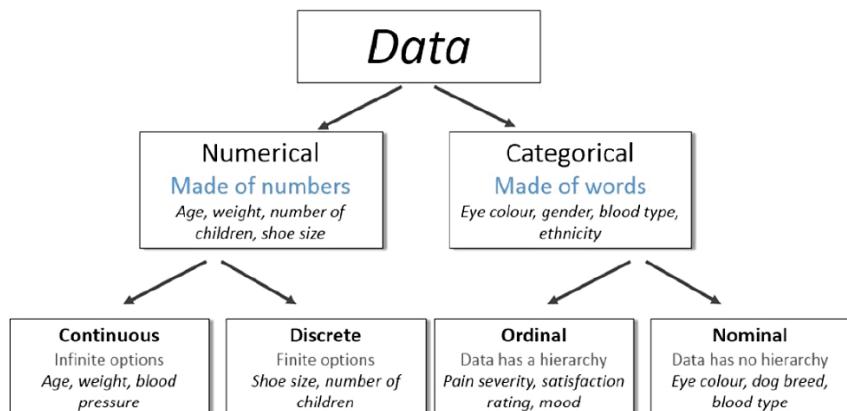


- **Betimsel İstatistik (Descriptive)**
 - Verilerin özetlenmesi ve tanımlanması için kullanılır. Ortalama, medyan, mod, standart sapma gibi ölçütler kullanılır.
- **Çıkarımsal İstatistik (Inferential)**
 - Verilerden genel sonuçlar çıkarmak için kullanılır. Hipotez testleri ve güven aralıkları bu kapsamdadır.



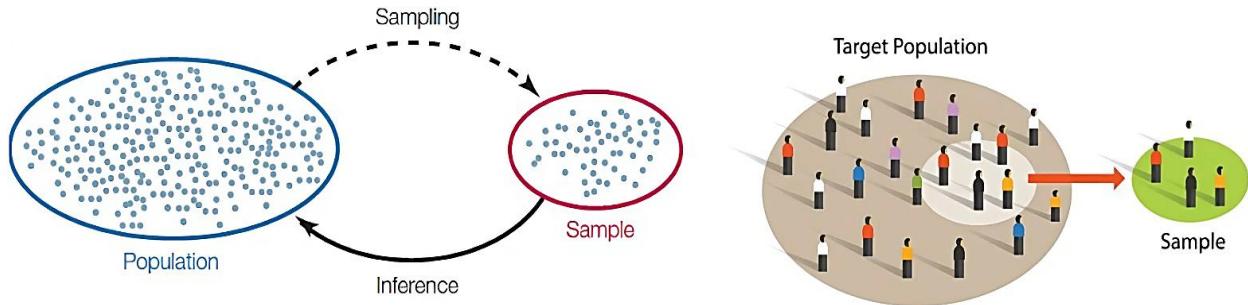
Veri Türleri:

- Nicel (sayısal) ve nitel (kategorik) veriler.



Parametreler ve İstatistikler:

- Parametre: Popülasyonun özelliklerini tanımlar.
- İstatistik: Örneklemenin özelliklerini tanımlar.



Olasılık vs İstatistik:

- **Olasılık:** Olayların gerçekleşme ihtimalini inceler.
- **İstatistik:** Sonuç bilinir ve süreç hakkında bir çıkarımda bulunmak (Inferention) için kullanılır.

Ölçüm Düzeyleri:

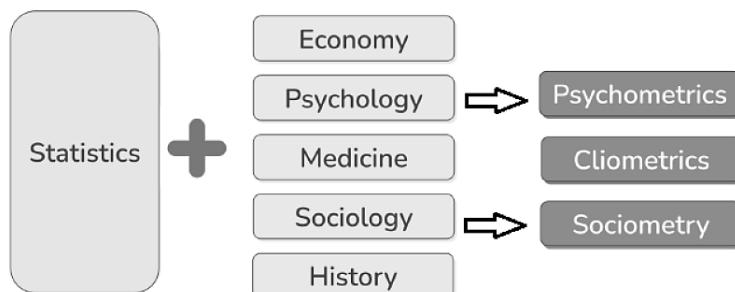
Verilerin ölçüldüğü farklı düzeyler:

- **Nominal:** Kategorik veriler (örneğin, cinsiyet).
- **Ordinal:** Sıralı kategorik veriler (örneğin, eğitim düzeyi).
- **Aralıklı (Interval):** Sayısal veriler, sıfır noktası keyfi (örneğin, sıcaklık).
- **Oranlı (Ratio):** Sayısal veriler, mutlak sıfır noktası (örneğin, ağırlık).

Levels of Measurement

Incremental Progress	Measure Property	Mathematical Operators	Advanced Operations	Central Tendency
Nominal	Classification, Membership	=, !=	Grouping	Mode
Ordinal	Comparison, Level	>, <	Sorting	Median
Interval	Difference, Affinity	+, -	Yardstick	Mean, Deviation
Ratio	Magnitude, Amount	*, /	Ratio	Geometric Mean, Coeff. of Variation

İstatistikin Gerçek Hayatta Kullanım Alanları



- **Tıbbi Çalışmalar (Medical Study)**

İstatistik, tıbbi araştırmalarda ilaçların etkinliğini test etmek, hastalıkların yaygınlığını ölçmek ve tedavi yöntemlerini değerlendirmek için kullanılır.

Örnek: COVID-19 aşlarının etkinliği istatistiksel yöntemlerle test edilmiştir.

- **Hava Durumu Tahmini (Weather Forecasts)**

Meteorologlar, istatistiksel modeller kullanarak hava durumu tahminleri yapar.

Örnek: Yağmur olasılığı veya sıcaklık tahminleri istatistiksel verilere dayanır.

- **Kalite Kontrol (Quality Testing)**

Şirketler, ürünlerinin kalitesini test etmek ve en iyi kaliteyi sunmak için istatistiksel yöntemler kullanır.

Örnek: Bir otomobil üreticisi, her bir aracın güvenilirliğini test eder.

- **Borsa (Stock Market)**

Yatırımcılar, hisse senedi fiyatlarını tahmin etmek ve riskleri değerlendirmek için istatistiksel analizler kullanır.

Örnek: Hisse senedi trendleri ve volatilite analizleri.

- **Tüketici Ürünleri (Consumer Goods)**

Şirketler, tüketici tercihlerini analiz ederek ürünlerini geliştirir.

Örnek: Yeni bir ürünün piyasaya sürülmeden önce test edilmesi.

- **Devlet (Government)**

Devletler, nüfus sayımı, ekonomik planlama ve politika oluşturma süreçlerinde istatistik kullanır.

Örnek: Vergi politikalarının etkisinin değerlendirilmesi.

- **Acil Durum Hazırlığı (Emergency Preparedness)**

İstatistik, doğal afetlerin etkilerini tahmin etmek ve acil durum planları oluşturmak için kullanılır.

Örnek: Deprem risk analizleri.

- **Siyasi Kampanyalar (Political Campaigns)**

Politikacılar, seçmen tercihlerini analiz ederek kampanya stratejileri belirler.

Örnek: Anketler ve seçim tahminleri.

- **Spor (Sports)**
Sporcular ve takımlar, performanslarını artırmak için istatistiksel verileri kullanır.
Örnek: Bir basketbol oyuncusunun şut yüzdesi.
- **Araştırma (Research)**
Bilimsel araştırmalarda, verilerin analizi ve sonuçların yorumlanması için istatistik kullanılır.
Örnek: Yeni bir ilaçın klinik deneyleri.
- **Eğitim (Education)**
Eğitim kurumları, öğrenci performansını değerlendirmek ve eğitim politikalarını geliştirmek için istatistik kullanır.
Örnek: Sınav sonuçlarının analizi.
- **Tahmin (Prediction)**
İstatistik, gelecekteki olayları tahmin etmek için kullanılır.
Örnek: Satış tahminleri veya ekonomik büyümeye projeksiyonları.
- **Hastalık Tahmini (Predicting Disease)**
Epidemiyologlar, hastalıkların yayılma hızını ve etkilerini tahmin etmek için istatistik kullanır.
Örnek: COVID-19'un yayılma hızının modellenmesi.
- **Sigorta (Insurance)**
Sigorta şirketleri, riskleri değerlendirmek ve primleri belirlemek için istatistik kullanır.
Örnek: Araba sigortası primlerinin hesaplanması.
- **Finansal Piyasalar (Financial Market)**
Finansal analistler, piyasa trendlerini ve riskleri değerlendirmek için istatistik kullanır.
Örnek: Portföy yönetimi ve risk analizi.
- **İş İstatistikleri (Business Statistics)**
Şirketler, karar verme süreçlerinde verileri analiz etmek için istatistik kullanır.
Örnek: Satış verilerinin analizi ve pazarlama stratejileri.

İstatistiğin Günlük Hayattaki Önemi

- **Karar Verme:** İstatistik, doğru ve bilinçli kararlar almak için verileri analiz etmemizi sağlar.
- **Tahminler:** Gelecekteki olayları tahmin etmek ve planlama yapmak için kullanılır.
- **Kalite ve Verimlilik:** Ürün ve hizmetlerin kalitesini artırmak için istatistiksel yöntemler kullanılır.
- **Risk Yönetimi:** Riskleri değerlendirmek ve azaltmak için istatistiksel analizler yapılır.

Özet Tablo

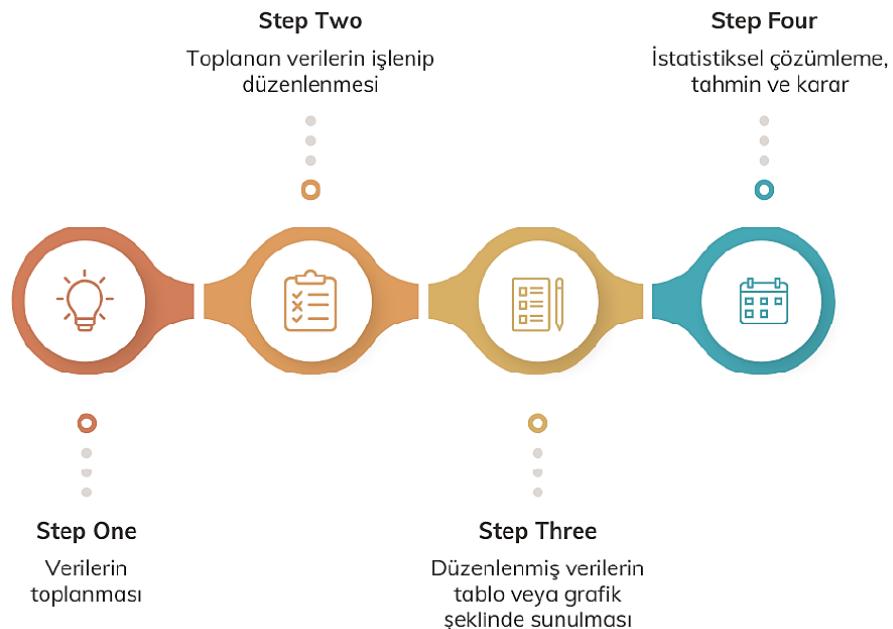
Kavram	Açıklama
İstatistiğin Anlamı	Verileri toplama, analiz etme, yorumlama ve sunma bilimi.
Eisenhower Matrisi	Önceliklendirme için kullanılan bir araç.
Karakterizasyon	Verilerin özelliklerini belirleme ve tanımlama süreci.
Veri Bilimi vs İstatistik	Veri bilimi, istatistiksel yöntemleri kullanarak verilerden bilgi çıkarır.
Betimsel İstatistik	Verilerin özetlenmesi ve tanımlanması.
Çıkarımsal İstatistik	Verilerden genel sonuçlar çıkarma.
Veri Türleri	Nicel ve nitel veriler.
Ölçüm Düzeyleri	Nominal, Ordinal, Aralıklı, Oranlı.

Veri Bilimi vs İstatistik

Kriter	Veri Bilimi (Data Science)	İstatistik (Statistics)
Matematik Anlayışı	Matematiksel modeller ve algoritmalar kullanır.	Matematiksel teorilere ve istatistiksel yöntemlere odaklanır.
Problem İnceleme	Büyük veri setlerindeki sorunları araştırır.	Verilerin analizi ve yorumlanması üzerine odaklanır.
Keşifsel Veri Analizi (EDA)	Verileri keşfetmek ve ön analiz yapmak için kullanılır.	Verilerin dağılımını ve ilişkilerini anlamak için kullanılır.
Trend Analizi	Verilerdeki trendleri ve desenleri analiz eder.	Verilerdeki trendleri istatistiksel yöntemlerle inceler.
Tahminler Oluşturma	Makine öğrenmesi modelleri kullanarak tahminler yapar.	İstatistiksel modeller kullanarak tahminler yapar.
Görselleştirme	Verileri görselleştirmek için araçlar ve kütüphaneler kullanır.	Verilerin grafiksel sunumu için istatistiksel yöntemler kullanır.
Teknik Olmayan Kullanıcılarla Raporlama	Bulguları teknik olmayan kullanıcılarla anlaşılır şekilde sunar.	Bulguları raporlar ve teknik olmayan kullanıcılarla açıklar.

- Veri Bilimi:** Büyük veri setlerini analiz etmek, makine öğrenmesi modelleri oluşturmak ve otomatik sistemler geliştirmek için kullanılır.
- İstatistik:** Verilerin analizi, hipotez testleri ve istatistiksel modeller oluşturmak için kullanılır.

Istatistikte İzlenen Sıra

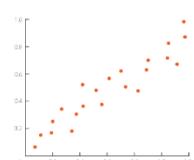
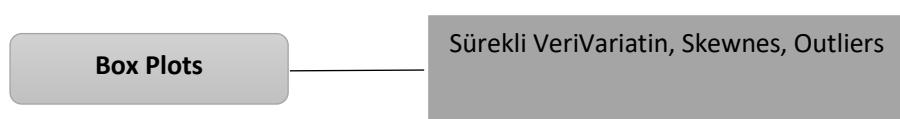
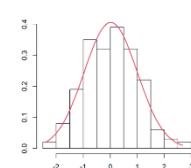
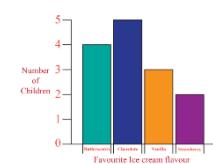


2.Bölüm : Veri Görselleştirme ve Temel Analiz

Veriler için Grafiksel Özetleme



Colour	Tally marks	Frequency
Black		1
Blue		5
Pink		2
White		4
		Total = 12

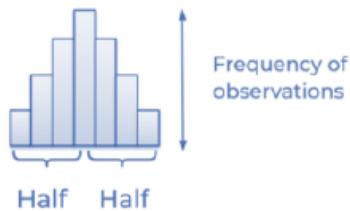


Veri Görselleştirme(Data Visualization) - Grafiksel Temsil

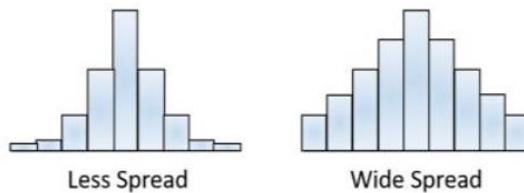
- **Tanım:** Verilerin grafikler veya şemalar kullanılarak görsel olarak temsil edilmesidir.
- **Örnek:** Çubuk grafikler, pasta grafikleri, histogramlar.

Desenler(Data Patterns)

- **Tanım:** Verilerin dağılımında gözlemlenen örüntüler veya eğilimlerdir.
- **Türleri:**
 - **Merkez(Center):** Dağılımin merkezi grafiksel olarak dağılımin medyanında olur.
 -

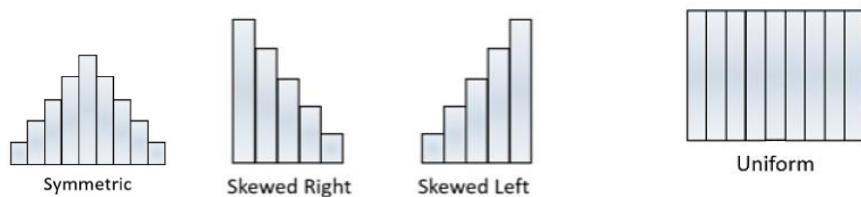
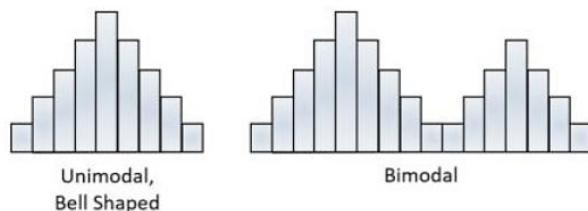


- **Yayılım (Spread):** Verilerin ne kadar geniş bir aralığa yayıldığını inceleriz.

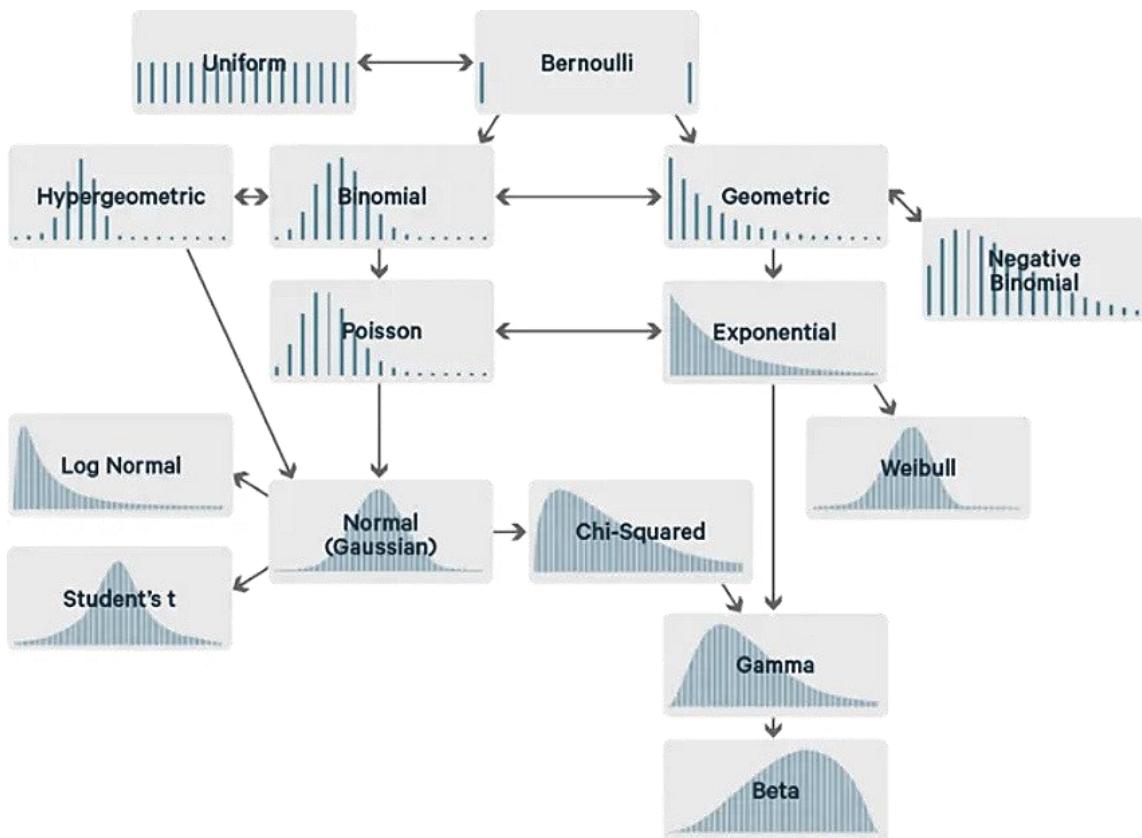


- **Şekil (Shape):** Verilerin dağılımının şekli (örneğin, simetrik veya çarpık).

- **Simetri (Symmetric):** Verilerin ortalamaya göre simetrik olup olmadığı.
- **Tepe Sayısı (Number of Peaks):** Dağılımda kaç tane tepe noktası olduğu.
- **Çarpıklık (Skewness):** Verilerin bir tarafa doğru çarpık olup olmadığı.
- **Düzgün Dağılım (Uniform):** Verilerin eşit olarak dağılması.

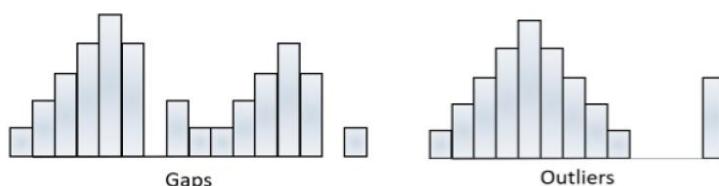


Olasılık Dağılımları (Probability Distributions)



Olağanüstü Özellikler

- **Tanım:** Verilerde beklenmedik veya alışılmadık özellikler.
- **Türleri:**
 - **Boşluklar (Gaps):** Verilerdeki eksik veya boş alanlar.
 - **Aykırı Değerler (Outliers):** Verilerdeki aşırı değerler.



Frekans Tablosu

- **Tanım:** Verilerin belirli aralıklarda kaç kez tekrarlandığını gösteren tablodur. Veriyi grafiksel olarak özetlemek için kullanılır.
- **Kullanım Alanı:** Kesikli (discrete) ve sürekli (continuous) veriler için uygundur.
- **Özellikler:**
 - Veriyi kategorilere ayırır ve her kategorinin frekansını gösterir.

- Frekans tablosu tek başına çok fazla bilgi vermez, ancak diğer grafikler için temel oluşturur.
- **Türleri:**
 - **Görelî Frekans (Relative Frequency):** Her bir aralığın toplam veriye oranıdır.
 - **Kümülatif Frekans (Cumulative Frequency):** Belirli bir aralığa kadar olan toplam frekansıdır.

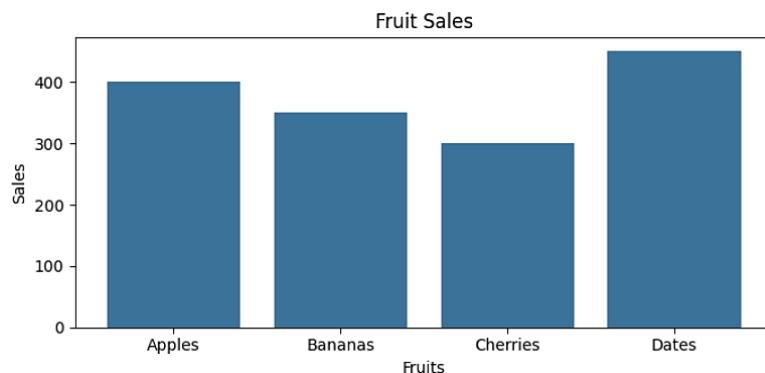
The figure shows two tables side-by-side. The left table is a frequency distribution table with columns for 'Sınıflar' (Classes) and 'Frekans, f' (Frequency). It has six rows with class intervals 1-4, 5-8, 9-12, 13-16, and 17-20, and frequencies 4, 5, 3, 4, and 2 respectively. Arrows point from the class intervals to the frequency values. The right table is a tally marks table with columns for 'Colour' and 'Tally marks'. It lists four colors: Black, Blue, Pink, and White, each with its corresponding tally marks (I, II, III, IV) and frequency (1, 5, 2, 4). A total row at the bottom shows 'Total = 12'.

Sınıflar	Frekans, f
1 → 4	4
5 → 8	5
9 → 12	3
13 → 16	4
17 → 20	2

Colour	Tally marks	Frequency
Black	I	1
Blue	IIII	5
Pink	II	2
White	IIII	4
		Total = 12

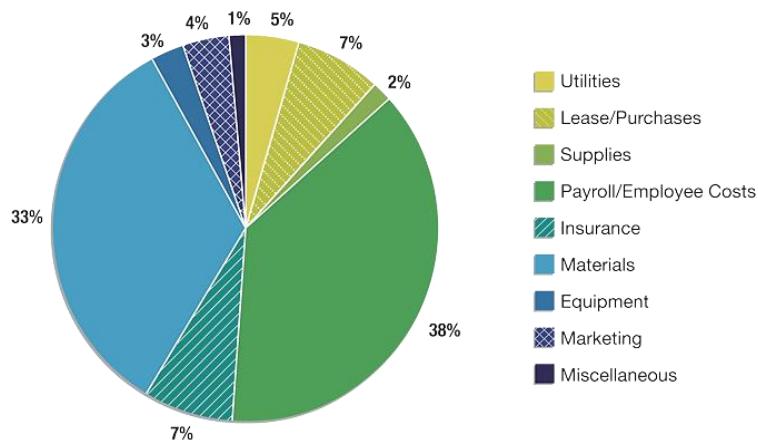
Çubuk Grafik (Bar Chart)

- **Tanım:** Kategorik verilerin çubuklarla temsil edildiği grafiktir. Her bar yüksekliği her niteliğin frekansını gösterir.
- **Örnek:** Farklı ürünlerin satış miktarlarını göstermek.



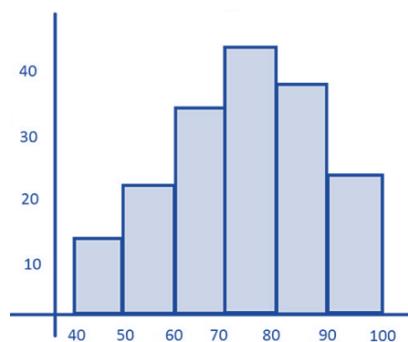
Pasta Grafik (Pie Chart)

- **Tanım:** Verilerin bir daire içinde dilimler halinde temsil edildiği grafiktir. Genelde nominal ve ordinal (kategorik) değişkenlerde kullanılır.
- **Örnek:** Bir şirketin gelirlerinin farklı kaynaklara göre dağılımı.



Histogram

- **Tanım:** Sürekli verilerin frekans dağılımını gösteren grafik.
- **Kullanım Alanı:** Varyasyon, çarpıklık (skewness) ve aykırı değerler (outliers) tespiti.
- **Özellikler:**
 - Sütunlar arasında boşluk yoktur.
 - Verinin yoğunlaştığı bölgeleri ve dağılım karakterini gösterir.
 - **Yorumlama:**
 - Standart sapma ve varyans hesaplanabilir.
 - Outlier'lar tespit edilebilir.
 - Medyan ve ortalama yorumlanabilir.
- **Örnek:** Öğrencilerin sınav notlarının dağılımı.

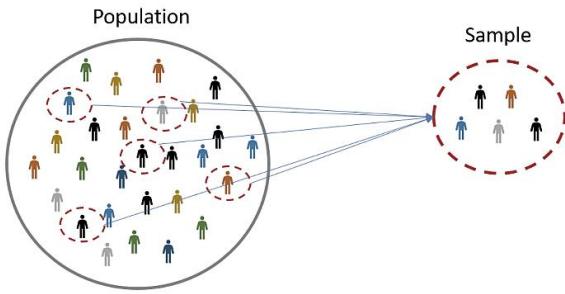


Çubuk Grafik (Bar Chart) vs Histogram

Kriter	Çubuk Grafik (Bar Chart)	Histogram
Kategoriler	Kategoriler vardır.	Kategoriler yoktur, aralıklar vardır.
Değişken Türü	Ayrık (discrete) değişkenler kullanılır.	Sürekli (continuous) değişkenler kullanılır.
Veri Türü	Kategorik veriler sunar.	Sayısal veriler sunar.
Barlar Arası Boşluk	Barlar arasında boşluk vardır.	Barlar arasında boşluk yoktur.
Grafik Gösterimi	Kategorik verilerin şematik karşılaştırmasını yapar.	Sürekli verilerin frekans dağılımını gösterir.
Kullanım Alanı	Kategorik verilerin karşılaştırılması için kullanılır.	Sayısal verilerin dağılımını göstermek için kullanılır.

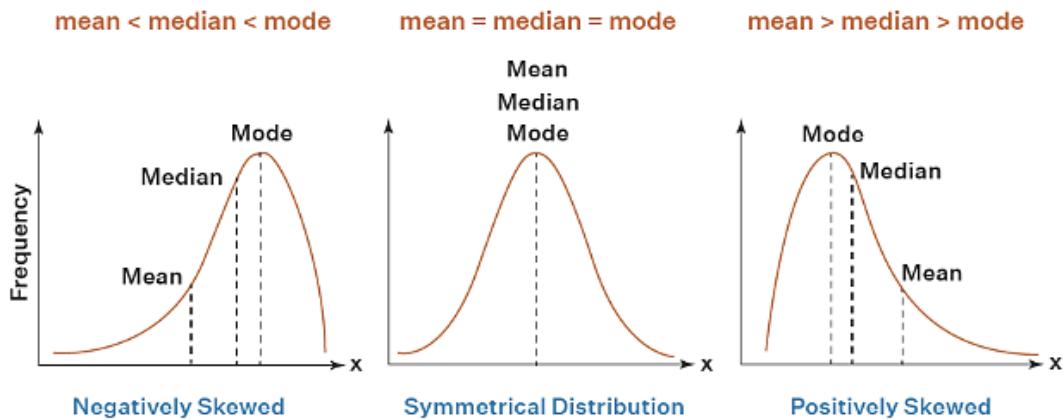
Popülasyonlar ve Örneklemeler

- **Tanım:** Popülasyon, tüm veri setini temsil ederken, örneklem popülasyondan çekilen bir alt kümedir.
- **Türleri:**
 - **Parametreler:** Popülasyonun özelliklerini tanımlar.
 - **İstatistikler:** Örneklemenin özelliklerini tanımlar.



Merkezi Eğilim (Measure of Centre)

- **Tanım:** Verilerin orta noktasını gösteren ölçütlerdir.
 - **Türleri:**
 - **Ortalama (Mean):** Verilerin aritmetik ortalamasıdır..
 - **Medyan (Median):** Verilerin ortanca değeridir.
 - **Mod (Mode):** Verilerde en sık tekrar eden değerdir. Tepe değeri diye de adlandırılır.

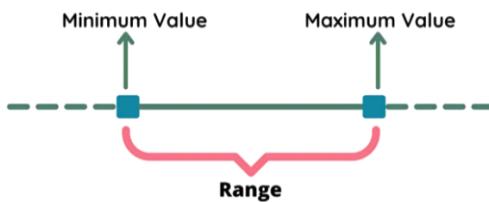


Ortalama (Mean) vs Medyan (Median)

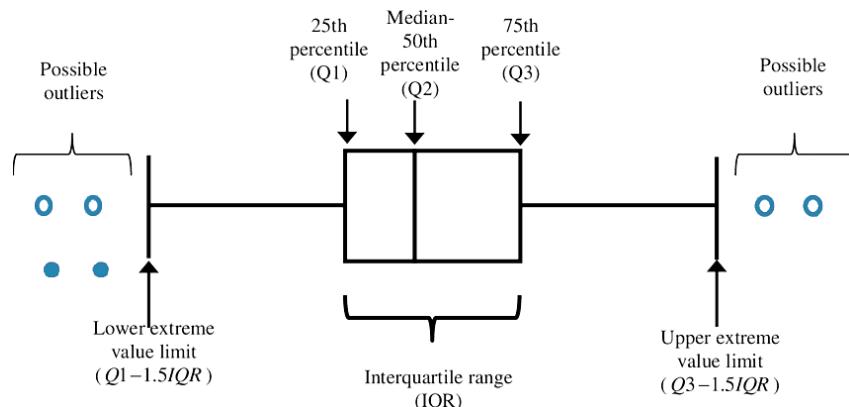
Kriter	Ortalama (Mean)	Medyan (Median)
Tanım	Verilerin toplamının veri sayısına bölünmesiyle bulunur.	Veriler sıralandığında ortadaki değerdir.
Aykırı Değerlere (Outliers) Duyarlılık	Aykırı değerlerden çok etkilenir.	Aykırı değerlerden etkilenmez.
Küçük Veri Setleri	Aykırı değerler varsa yaniltıcı olabilir.	Aykırı değerler varsa daha güvenilirdir.
Büyük Veri Setleri	Aykırı değerler yoksa daha iyi bir ölçütür.	Aykırı değerler yoksa ortalama kadar iyidir.
Kullanım Alanı	Simetrik dağılımlar ve aykırı değerlerin olmadığı durumlar.	Çarpık dağılımlar ve aykırı değerlerin olduğu durumlar.
Örnek	Öğrencilerin sınav notlarının ortalaması.	Maaş dağılımında medyan kullanılabilir.

Dispersion (Measure of Spread) – Dağılım Ölçüleri

- **Tanım:** Verilerin ne kadar geniş bir aralığa yayıldığını gösteren ölçütlerdir.
- **Türleri:**
 - **Aralık (Range):** Verilerin en büyük(maksimum) ve en küçük(minimum) değerleri arasındaki farktır.



- **Çeyrekler Arası Aralık (IQR):** Bir sayı grubunu dörde bölen değerdir. Verilerin ortadaki %50'sinin yayılımıdır.



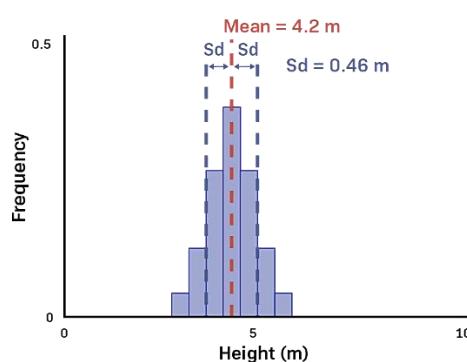
- **Standart Sapma (Standard Deviation):** Verilerin ortalamadan ne kadar uzaklaştığını gösterir. Veriler ne kadar çok yayılırsa, standart sapma o kadar büyük olur.

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

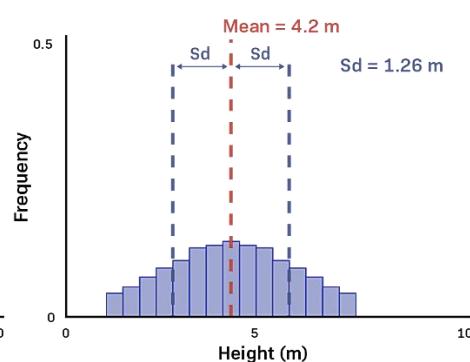
element mean
 standard deviation σ
 number of elements

<u>Sample</u> $S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$	<u>Population</u> $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
---	--

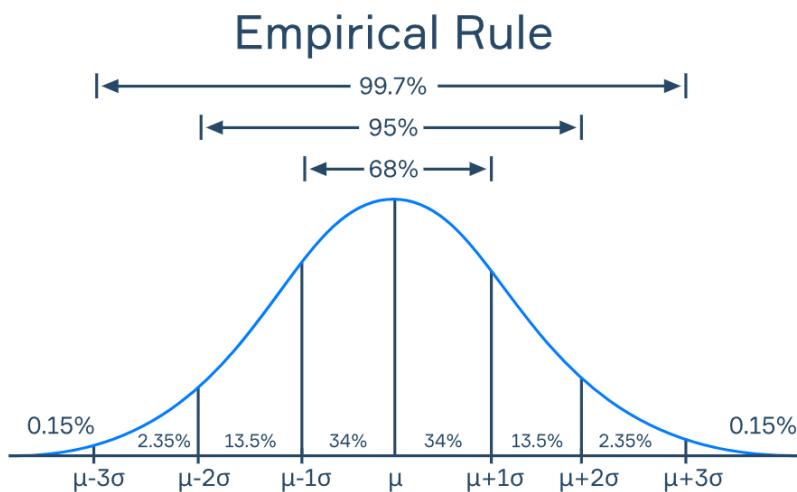
A. Data set with a smaller standard deviation



B. Data set with a larger standard deviation

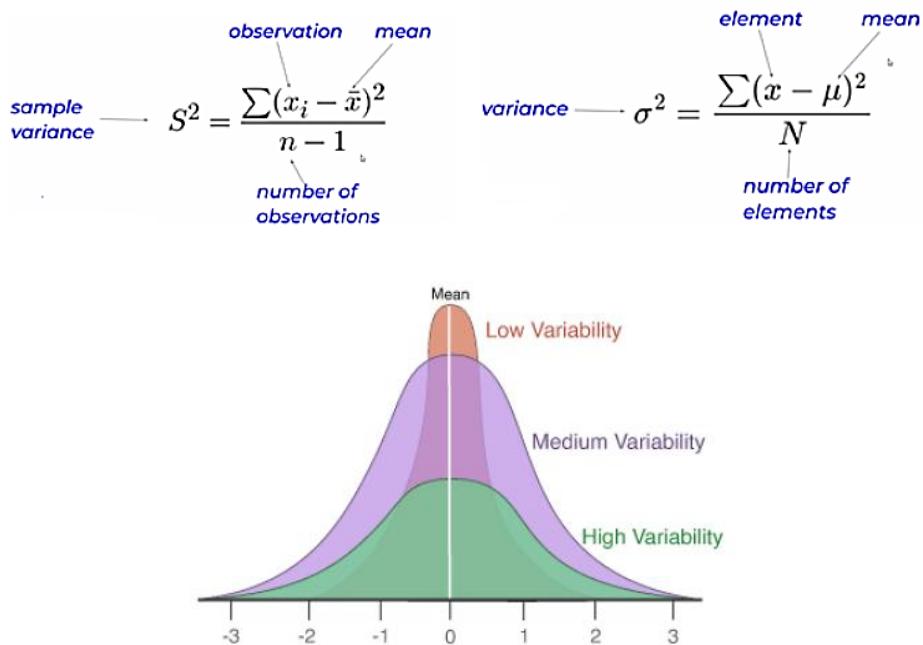


- **Ampirik Kural ()**: 3 sigma kuralı veya Normal dağılımda verilerin %68, %95 ve %99.7'sinin ortalamanın belirli standart sapma aralıklarında olduğunu belirtir.
- 1. % 68'de kural,
= (Ortalama standart sapma) ve (Ortalama + standart)
- 2. % 95'de kural,
= (Ortalama $2 \times$ standart sapma) and (Ortalama + $2 \times$ standart sapma)
- 3. % 99.7'de kural,
= (Ortalama $3 \times$ standart sapma) and (Ortalama + $3 \times$ standart sapma)



Varyasyon (Variation)

- **Tanım:** Varyans, ortalamadan farkların karelerinin ortalaması olarak tanımlanır. Verilerin ne kadar değişken olduğunu gösteren ölçütür.
- **Örnek:** Bir ürünün kalite kontrolünde varyasyonun düşük olması istenir.



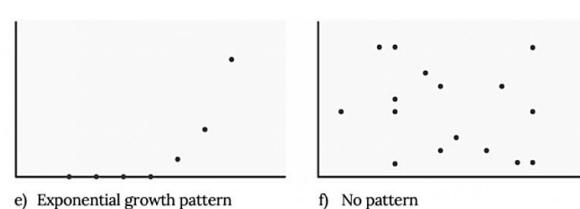
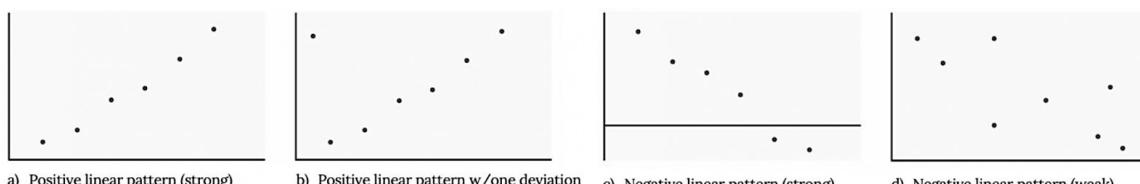
Özet Tablo

Kavram	Açıklama
Veri Görselleştirme	Verilerin grafiklerle temsil edilmesi.
Desenler	Verilerin dağılımında gözlemlenen örüntüler.
Olağandışı Özellikler	Verilerdeki beklenmedik özellikler (boşluklar, aykırı değerler).
Frekans Tablosu	Verilerin belirli aralıklarda kaç kez tekrarlandığını gösteren tablo.
Çubuk Grafik	Kategorik verilerin çubuklarla temsili.
Pasta Grafik	Verilerin bir daire içinde dilimler halinde temsili.
Histogram	Sürekli verilerin frekans dağılımını gösteren grafik.
Popülasyonlar ve Örneklemeler	Popülasyon ve örneklem arasındaki farklar.
Merkezi Eğilim	Verilerin orta noktasını gösteren ölçütler (ortalama, medyan, mod).
Yayılım	Verilerin ne kadar geniş bir aralığa yayıldığını gösteren ölçütler.
Varyasyon	Verilerin ne kadar değişken olduğunu gösteren ölçüt.

3. Bölüm : İlişki Analizi

Saçılım Grafiği (Scatter Plots)

- Tanım:** İki sürekli değişken arasındaki ilişkiyi göstermek için kullanılır. İlişkinin yönünü ve kuvvetini gösterir.
- Kullanım Alanı:** Korelasyon analizi ve doğrusal ilişki tespiti.
- Özellikler:**
 - Pozitif Korelasyon:** Noktalar sağ yukarı dağılır. İki değişken de artıysa pozitif bir ilişki vardır.
 - Negatif Korelasyon:** Noktalar sağ aşağı dağılır. Bir değişken artıp diğerinin azalıyorsa negatif bir ilişki vardır.
 - Güçlü/Zayıf İlişki:** Noktaların doğrusal düzeni ile ilişkilidir. Güçlü ilişkide lineer düzenli bir yönetim vardır. Zayıf ilişkide ise lineere yakın daha dağınık dizilmiş bir ilişki vardır.
- Örnek Görsel:**



- Belli bir düzen yoksa, grafik yuvarlak çemberimsi bir görünümde ise no pattern dizilim vardır. No pattern ilişki analiz edilemez.(ÖR: Yazın dondurma satışı artması ve ev fiyatları artması ilişkisi)
- Bir dağılım grafine baktığımızda genel deseni ve desenden ne kadar sapma olduğunu görmek isteriz. İki değişken arası ilişki , gücü, yönü bilgilerini scatter plot grafiğindeki dağılımlardan analiz ederiz.
- Scatter plotta noktalar bir doğru şeklinde hareket ediyorsa burada strong bir ilişki vardır. Pozitif ve negatif olması gücünü etkilemez. Eksi yönde de güçlü bir ilişki olabilir.

Dağılım Karakteri ve Tahmin Süreçleri

- **Dağılım Karakteri:**
 - Dağılım analizi ile veri setinin normal olup olmadığı analiz edilir.
 - Sektörel bilgiler (örn: trafik mühendisliğinde Poisson dağılımı) kullanılır.
- **Tahmin Süreçleri:**
 - Dağılım karakteri bilindiğinde, yeni durumlarda verinin nasıl davranışacağı tahmin edilebilir.
 - Bu, tahmin süreçlerini kolaylaştırır.

Özet Tablo

Grafik Türü	Kullanım Alanı	Özellikler
Frekans Tablosu	Kesikli ve sürekli veriler	Veriyi kategorilere ayırır, frekansları gösterir.
Çubuk Grafik	Kesikli veriler	Kategoriler arası karşılaştırma yapar.
Histogram	Sürekli veriler	Dağılım, çarpıklık, outlier tespiti.
Kutu Grafiği	Sürekli veriler	Çeyrekler, medyan, outlier tespiti.
Saçılım Grafiği	İki sürekli değişken	İlişki yönü ve gücünü gösterir.

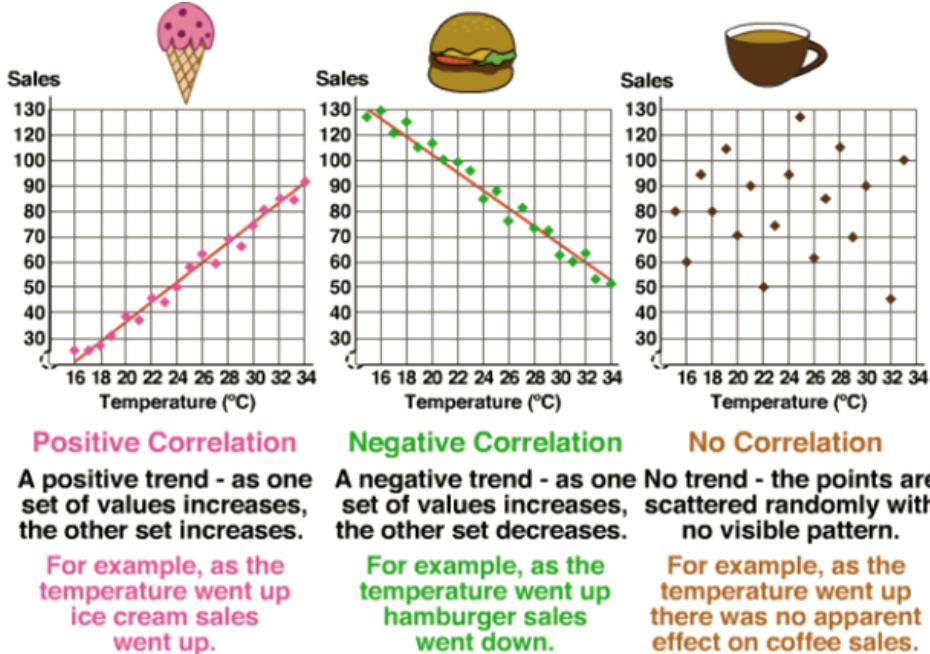
Line of Best Fit (En Uygun Çizgi)

- **Tanım:** Bir dağılım grafiği üzerine çizilen ve verilerin genel eğilimini en iyi şekilde temsil eden doğru(trend çizgisi).
- **Kullanım Alanı:** Veriler arasındaki doğrusal ilişkiyi modellemek ve geleceğe yönelik tahminler yapmak.
- **Formül:**

$$y=ax+b$$

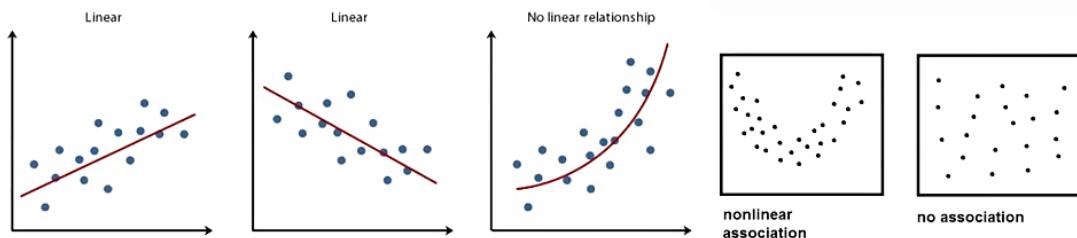
- **a (Eğim):** X değişkenindeki 1 birimlik artışın Y'yi ne kadar etkilediğini gösterir.
- **b (Intercept):** X=0 olduğunda Y'nin aldığı değer.

- **Örnek Görsel:**



Linearity (Doğrusallık)

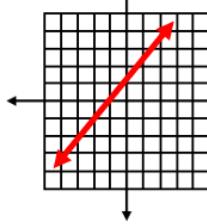
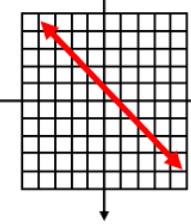
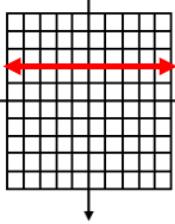
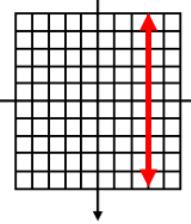
- **Lineer (Doğrusal):** İki değişken arasındaki ilişki düz bir çizgi ile ifade edilebiliyorsa, bu ilişki lineerdir.
- **Non-Lineer (Doğrusal Olmayan):** İlişki düz bir çizgi ile ifade edilemiyorsa, non-lineerdir.
- **Mükemmel Lineerlik Yoktur:** Korelasyon katsayısı (r) tam olarak 1.0 veya -1.0 olmaz, ancak bu değerlere yaklaşabilir.
- **Non-Lineer İlişkiler:** İleri seviye analizlerle ikiye bölünerek incelenebilir.



Slope (Eğim)

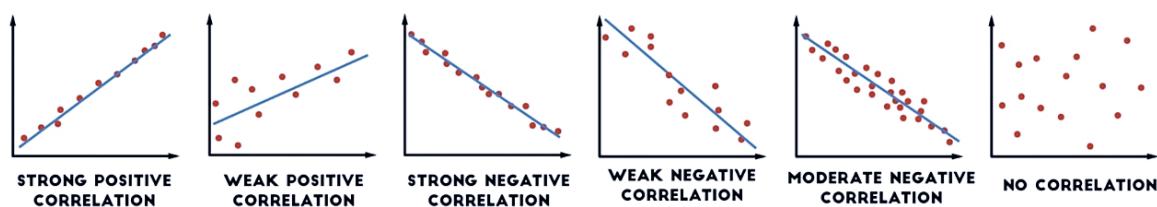
- **Tanım:** X değişkenindeki değişimin Y değişkenini nasıl etkilediğini gösterir.
- **Türleri:**
 - **Pozitif Eğim:** X artarken Y de artar.
 - **Negatif Eğim:** X artarken Y azalır.
 - **Tanımsız Eğim (Undefined):** Dikey doğru (X değişmezken Y değişir).
 - **Sıfır Eğim (Zero):** Yatay doğru (Y sabit, X değişir).

- Örnek Görsel:

1	2	3	4
Positive Slope	Negative Slope	Zero Slope	Undefined Slope
<ul style="list-style-type: none"> - Graph goes <u>up</u> from left to right - As x increases, y increases - The equation $y = mx + b$ has $m > 0$ 	<ul style="list-style-type: none"> - Graph goes <u>down</u> from left to right - As x increases, y decreases - The equation $y = mx + b$ has $m < 0$ 	<ul style="list-style-type: none"> - Graph goes <u>side to side</u> - Horizontal line - As x increases, y stays constant - The equation $y = b$ 	<ul style="list-style-type: none"> - Graph goes <u>up</u> and <u>down</u> - Vertical line - x stays constant, as y increases - The equation $x = b$
<u>Graph:</u>	<u>Graph:</u>	<u>Graph:</u>	<u>Graph:</u>
			

Strength (Güç)

- **Tanım:** Grafikteki dağılımın ne kadar güçlü bir ilişki gösterdiğini ifade eder.
- **Yorum:**
 - **Güçlü İlişki:** Noktalar doğruya yakın.
 - **Zayıf İlişki:** Noktalar dağınık.
- **Örnek Görsel:**



Unusual Features (Olağandışı Özellikler)

A. Clusters (Kümelenme)

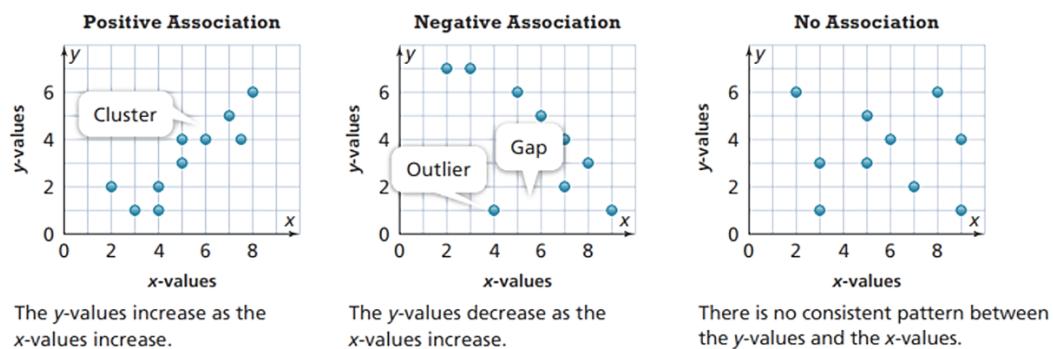
- **Tanım:** Benzer davranış gösteren verilerin bir arada toplanmasıdır.
- **Örnek:** E-ticaret verilerinde benzer alışveriş davranışları.

B. Gaps (Boşluklar)

- **Tanım:** Veride eksik veya boşluklar olması.
- **Çözüm:** Eksik veriler doldurulmalı veya analiz ayrı ayrı yapılmalıdır.
- **Örnek:** Covid dönemi giriş yapılmamıştır öncesi ve sonrası diye ayırarak analiz yapmalıyız.

C. Outliers (Aykırı Değerler)

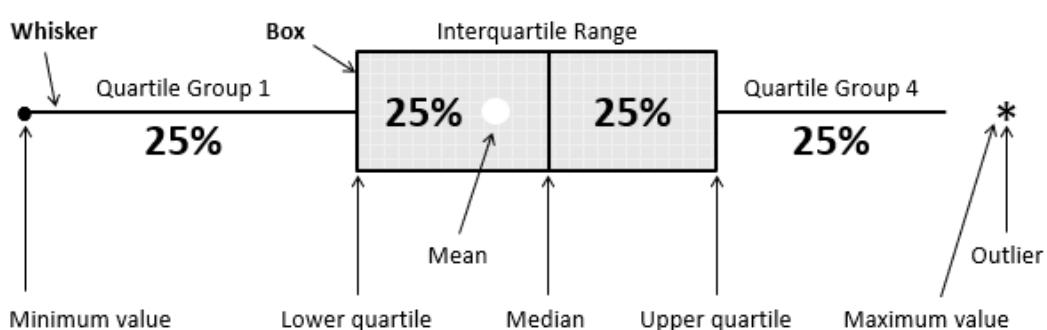
- **Tanım:** Diğer verilerden belirgin şekilde farklı olan değerler.
- **Etkisi:** Tahmin gücünü zayıflatır.
- **Çözüm:** Outlier'lar tespit edilip düzeltilmeli veya analizden çıkarılmalıdır.



NOT: Çöp data yoktur, datanın suyu sıkılana kadar üzerine gidilir.

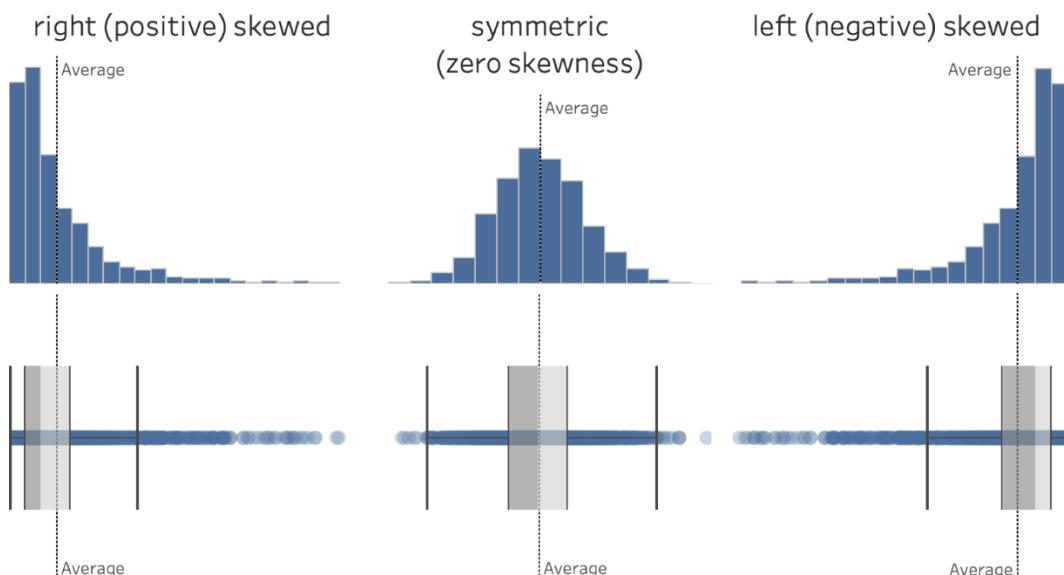
Box Plot (Kutu Grafiği)

- **Tanım:** Veri setinin dağılımını, çeyreklerini ve outlier'ları gösteren grafik.
- **Bileşenler:**
 - **Q1 (25. persentil):** Alt çeyrek.
 - **Q3 (75. persentil):** Üst çeyrek.
 - **Medyan:** Orta çizgi.
 - **IQR (Interquartile Range):** $Q3 - Q1$.
 - **Outlier'lar:** $Q1 - 1.5 \times IQR$ ve $Q3 + 1.5 \times IQR$ dışındaki değerler.
 - Kedi bıyığının dışı IQR'dır.
- **Örnek Görsel:**



Skew (Çarpıklık)

- **Tanım:** Veri dağılımının simetrik olmaması durumu.
- **Türleri:**
 - **Pozitif Çarpıklık:** Sağa çarpık (sağ kuyruk uzun).
 - **Negatif Çarpıklık:** Sola çarpık (sol kuyruk uzun).
- **Örnek Görsel:**

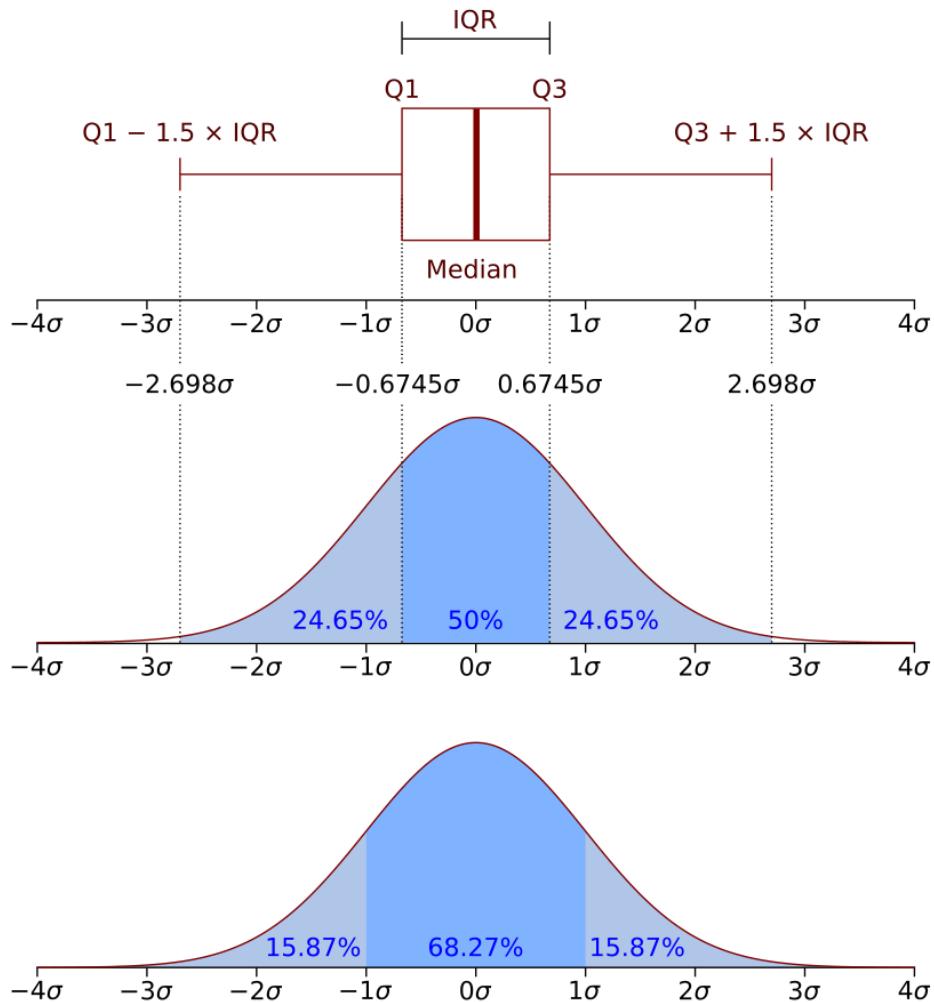


Box Plot - Min & Max Values

- **Tanım:** Veri setindeki minimum ve maksimum değerler.
- **Özellikler:**
 - Veriler sıralı olmalıdır.
 - Min ve max değerler outlier tespiti için kullanılmaz, IQR kullanılır.

IQR Kuralı

- **Tanım:** Veri setindeki aykırı değerleri (outlier) tespit etmek için kullanılan bir kurallıdır.
- **Kim Tarafından Önerildi?:** İstatistikçi **John Tukey** tarafından önerilmiştir.
- **Nasıl Çalışır?:**
 - **IQR (Interquartile Range):** $Q3 - Q1$ (Üst çeyrek - Alt çeyrek).
 - **Alt Sınır:** $Q1 - 1.5 \times IQR$
 - **Üst Sınır:** $Q3 + 1.5 \times IQR$
 - **Bu sınırların dışında kalan değerler outlier olarak kabul edilir.**
- **Örnek:**
 - $Q1 = 25$, $Q3 = 75$ ise $IQR = 50$.
 - Alt Sınır: $25 - 1.5 \times 50 = 25 - 75 = -50$
 - Üst Sınır: $75 + 1.5 \times 50 = 75 + 75 = 150$
 - Bu aralığın dışındaki değerler outlier'dır.



Özet Tablo

Konu	Kısa Açıklama
Line of Best Fit	Verilerin eğilimini en iyi temsil eden doğru.
Linearity	İlişkinin doğrusal olup olmadığı.
Slope	X'teki değişimin Y'yi nasıl etkilediğini gösterir.
Strength	İlişkinin gücünü ifade eder.
Clusters	Benzer davranış gösteren verilerin kümelenmesi.
Gaps	Verideki eksiklikler veya boşluklar.
Outliers	Diğer verilerden belirgin şekilde farklı olan değerler.
Box Plot	Veri dağılımını, çeyreklerini ve outlier'ları gösteren grafik.
Skew	Veri dağılımının simetrik olmaması durumu.
1.5 IQR Kuralı	Outlier tespiti için kullanılan kurallıdır.

Top 60 Statistics Interview Questions 2024

Question 11: What is the benefit of using box plots?

Answer: Boxplot is a visually effective representation of two or more data sets and facilitates quick comparison between a group of histograms.

Soru 11: Kutu grafikleri kullanmanın faydası nedir?

Cevap: Kutu grafiği, iki veya daha fazla veri kümelerinin görsel olarak etkili bir temsilidir ve bir grup histogram arasında hızlı karşılaştırma yapılmasını kolaylaştırır.

Question 12: How to detect outliers?

Answer: The best way to detect outliers is through graphical means. Apart from that, outliers can also be detected through the use of statistical methods using tools such as Excel, Python, SAS, among others. The most popular graphical ways to detect outliers include box plot and scatter plot. We can use IQR. Out of $[Q1-1.5IQR, Q3+1.5IQR]$ range show us outliers.

Soru 12: Aykırı değerler nasıl tespit edilir?

Cevap: Aykırı değerleri tespit etmenin en iyi yolu grafiksel yöntemlerdir. Bunun yanı sıra, aykırı değerler Excel, Python, SAS gibi araçlar kullanılarak istatistiksel yöntemlerle de tespit edilebilir. Aykırı değerleri tespit etmek için en popüler grafiksel yöntemler kutu grafiği ve saçılım grafiğidir. IQR'yi kullanabiliriz. $[Q1-1.5IQR, Q3+1.5IQR]$ aralığının dışı bize aykırı değerleri gösterir.

Kovaryans (Covariance)

Tanım: İki değişken arasındaki değişimin birlikte nasıl hareket ettiğini ölçer. İki veri arasındaki negatif ya da pozitif ilişki kovaryans ile söylenir.

- **Formül:**

Population Covariance

$$Cov(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

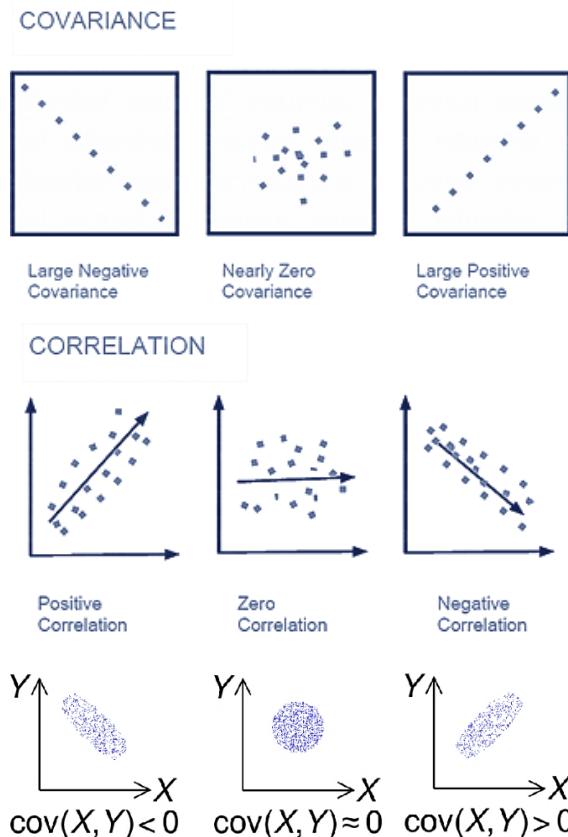
Sample Covariance

$$Cov(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

These are the formula for finding Population and Sample Covariance.

- x_i = data value of x
- y_i = data value of y
- \bar{x} = mean of x
- \bar{y} = mean of y
- N = number of data values.

- **Yorum:**
 - **Pozitif Kovaryans:** X artarken Y de artar.
 - **Negatif Kovaryans:** X artarken Y azalır.
 - **Sıfır Kovaryans:** İki değişken arasında ilişki yoktur.
- **Özellikler:**
 - Sadece ilişkinin **yönünü** gösterir, **güçünü** göstermez.
 - Değer aralığı sınırsızdır, bu nedenle yorumlamak zordur.



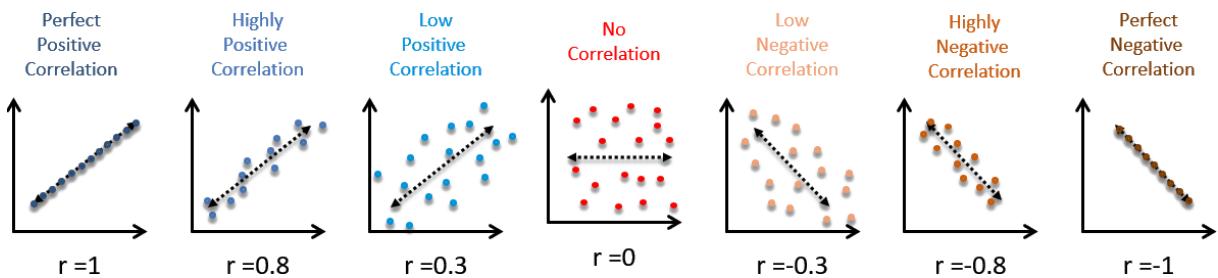
Korelasyon (Correlation)

- **Tanım:** İki değişken arasındaki ilişkinin hem **yönünü** hem de **güçünü** gösterir.
- **Formül (Pearson Korelasyon Katsayısı):**

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

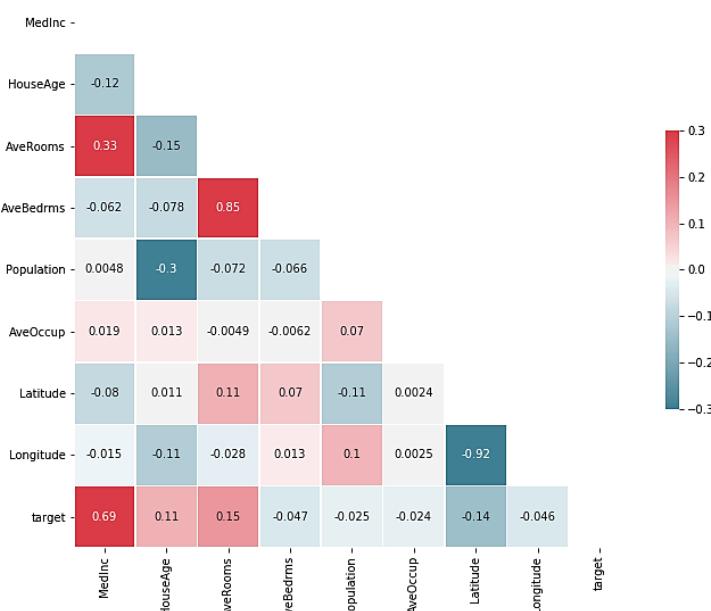
- **Değer Aralığı:** -1 ile +1 arasındadır. -1 ve +1 aynı güçtedir fark yönündür, biri artarken diğerinin azalması da tam tersi birisi azalırken diğerini artırır.
 - **+1:** Mükemmel pozitif ilişki.
 - **-1:** Mükemmel negatif ilişki.
 - **0:** İlişki yok.
- **Yorum:**
 - Korelasyon, kovaryansın -1 ile +1 arasında ölçeklendirilmiş halidir.
 - **Pozitif Korelasyon:** İki değişken birlikte artar veya azalır.

- **Negatif Korelasyon:** Bir değişken artarken diğerinin azalır.
- **Önemli Not:** Korelasyon **nedensellik** ifade etmez! İki değişken arasında korelasyon olması, birinin diğerine neden olduğu anlamına gelmez.



Heatmap ve Korelasyon Matrisi

- **Tanım:** Veri setindeki tüm değişkenler arasındaki korelasyonları görselleştirmek için kullanılır.
- **Kullanım Alanı:**
 - **Feature Seçimi:** Hangi değişkenlerin hedef değişkenle ilişkili olduğunu belirlemek için kullanıyoruz.
 - **Çoklu Doğrusal Bağlantı (Multicollinearity):** Aynı anda aynı etkiyi yapan değişkenleri tespit etmek.
- **Örnek Görsel:**



Pearson Korelasyon Katsayısı

- **Tanım:** Pearson Korelasyon Katsayısı (genellikle r ile gösterilir), iki sürekli değişken arasındaki ilişkili ve yönünü ölçen bir ölçütür.

- **Değer Aralığı:** -1 ile +1 arasındadır.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}}$$

- x_i ve y_i bireysel veri noktalarıdır,
- \bar{x} ve \bar{y} , x ve y değişkenlerinin ortalamalarıdır.

- **Örnek:**

- **$r = 0.85$:** Güçlü pozitif ilişki.
- **$r = -0.70$:** Orta düzeyde negatif ilişki.
- **$r = 0.10$:** Zayıf ilişki.

- **Pearson Korelasyon Katsayısı (1.0):** Bu değer, yaş ve maaş arasında mükemmel bir pozitif doğrusal ilişki olduğunu gösterir. Yani, yaş arttıkça maaş da aynı oranda artmaktadır.
- **P-değeri (0.0):** Bu değer, korelasyonun istatistiksel olarak anlamlı olduğunu gösterir. Yani, bu ilişki tesadüfi değildir.

Korelasyon Hesaplama (R Calculation)

- **Adımlar:**

1. Kovaryansı hesapla.
2. Her iki değişkenin standart sapmalarını hesapla.
3. Kovaryansı standart sapmaların çarpımına böl.

- **Formül:**

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

- **Örnek Senaryo:** Bir şirketin, çalışanlarının yaşları ve maaşları arasındaki ilişkiyi incelemek istedığını varsayıyalım. Elimizde şu veriler olsun:
- **Yaş (x):** [25, 30, 35, 40, 45, 50]
- **Maaş (y):** [5000, 5500, 6000, 6500, 7000, 7500]

```
import numpy as np
from scipy.stats import pearsonr

yas = [25, 30, 35, 40, 45, 50]
maas = [5000, 5500, 6000, 6500, 7000, 7500]

# Pearson korelasyon katsayısını ve p-değerini hesapla
korelasyon_katsayisi, p_degeri = pearsonr(yas, maas)
```

Sonuç: Pearson Korelasyon Katsayısı: 1.0 P-değeri:0.0

What types of variables are used for Pearson's correlation coefficient?

Answer: Variables (both the dependent and independent variables) used for Pearson's correlation coefficient must be quantitative. It will only test for the linear relationship between two variables.

Pearson korelasyon katsayısı için ne tür değişkenler kullanılır?

Cevap: Pearson korelasyon katsayısı için kullanılan değişkenler (hem bağımlı hem de bağımsız değişkenler) nicel olmalıdır. Sadece iki değişken arasındaki doğrusal ilişkiyi test eder.

Question 14: What is the difference between Covariance and Correlation?

Covariance:

- Signifies the direction of the linear relationship between two variables.
- In simple terms, It is a measure of variance between two variables.
- It can take any value from positive infinity to negative infinity.

Correlation:

- It measures the relationship between two variables, as well as the strength between these two variables.
- It can take any value from -1 to 1.

Soru 14: Kovaryans ve Korelasyon arasındaki fark nedir?

Kovaryans:

- İki değişken arasındaki doğrusal ilişkinin yönünü gösterir.
- Basitçe, iki değişken arasındaki varyansın bir ölçüsüdür.
- Pozitif sonsuzdan negatif sonsuza kadar herhangi bir değeri alabilir.

Korelasyon:

- İki değişken arasındaki ilişkiyi ve bu iki değişken arasındaki gücü ölçer.
- -1 ile 1 arasında herhangi bir değeri alabilir.

Kaynaklar ve ilgili içerik

Çoklu Doğrusal Bağlantı (Multicollinearity)

- **Tanım:** Bir modelde aynı etkiye yapan birden fazla değişkenin bulunması.
- **Etkisi:** Modelin verimini düşürür ve overfitting'e neden olabilir.
- **Çözüm:** Korelasyon matrisi kullanılarak yüksek korelasyonlu değişkenler tespit edilir ve gereksiz olanlar çıkarılır.

Özellik Mühendisliği (Feature Engineering)

- **Tanım:** Değişkenler üzerinde işlemler yaparak yeni özellikler oluşturmak veya mevcut özellikleri iyileştirmek.
- **Korelasyonun Rolü:**
 - Hedef değişkenle yüksek korelasyonlu özellikler seçilir.
 - Düşük korelasyonlu veya gereksiz özellikler elenir.

Özet Tablo

Konu	Kısa Açıklama
Kovaryans	İki değişken arasındaki değişimin yönünü gösterir.
Korelasyon	İki değişken arasındaki ilişkinin yönünü ve gücünü gösterir.
Heatmap	Değişkenler arasındaki korelasyonları görselleştirir.
Pearson Katsayısı	İki değişken arasındaki doğrusal ilişkinin gücünü ölçer.
Korelasyon Hesaplama	Kovaryans ve standart sapmalar kullanılarak hesaplanır.
Multicollinearity	Aynı etkiyi yapan değişkenlerin model verimini düşürmesi.
Feature Engineering	Değişkenler üzerinde işlemler yaparak model performansını artırma.

4. Bölüm : Regresyon Analizi

Lineer Regresyon Nedir?

- Tanım:** İki değişken arasındaki doğrusal ilişkiyi modellemek ve bu ilişkiye dayanarak geleceğe yönelik tahminler yapmak için kullanılan bir istatistiksel yöntemdir.
- Amaç:** Bağımsız değişken (X) ile bağımlı değişken (Y) arasındaki ilişkiyi anlamak ve bu ilişkiye kullanarak Y'yi tahmin etmek.

Temel Kavramlar

A. Bağımsız ve Bağımlı Değişkenler

- Bağımsız Değişken (Independent Variable - X):** Sebep veya girdi olarak kabul edilen değişken.
 - Örnek: Reklam harcamaları (X), Satış adedi (Y).
- Bağımlı Değişken (Dependent Variable - Y):** Sonuç veya çıktı olarak kabul edilen değişken.

B. Doğrusal İlişki

- Formül:**

$$y=ax+b$$

- a (Eğim - Slope):** X'teki 1 birimlik artışın Y'yi ne kadar etkilediğini gösterir.
- b (Intercept):** X=0 olduğunda Y'nin aldığı değer.

Lineer Regresyon Türleri

A. Basit Lineer Regresyon (Simple Linear Regression)

- Tanım:** Tek bir bağımsız değişken (X) ile bağımlı değişken (Y) arasındaki ilişkiyi modellemek.
- Örnek:** Reklam harcamaları (X) ile satış adedi (Y) arasındaki ilişki.

B. Çoklu Lineer Regresyon (Multiple Linear Regression)

- **Tanım:** Birden fazla bağımsız değişken (X_1, X_2, \dots) ile bağımlı değişken (Y) arasındaki ilişkiyi modelllemek.
- **Örnek:** Reklam harcamaları (X_1), ürün fiyatı (X_2) ve satış adedi (Y) arasındaki ilişki.

Lineer Regresyon Adımları

1. **Veri Toplama:** Bağımsız ve bağımlı değişkenlerin verileri toplanır.
2. **Model Kurma:** En uygun doğru (line of best fit) çizilir.
3. **Tahmin Yapma:** Model kullanılarak geleceğe yönelik tahminler yapılır.

En Küçük Kareler Yöntemi (Least Squares Method)

- **Tanım:** Gerçek değerler ile tahmin edilen değerler arasındaki farkların karelerinin toplamını minimize eden doğruya bulma yöntemi.
- **Formül:**

$$\sum(y_i - \hat{y}_i)^2$$

- y_i : Gerçek değerler.
- \hat{y}_i : Tahmin edilen değerler.

Lineer Regresyon Örneği

- **Örnek Veri:**
 - $X = [1, 2, 3, 4, 5]$
 - $Y = [2, 4, 6, 8, 10]$
- **Model:**

$$y=2x+0$$

- **Eğim (a):** 2
- **Intercept (b):** 0

Hata Terimi (Error Term)

Tanım: Hata terimi, modelin tahmin ettiği değerler ile gerçek değerler arasındaki sapma olarak da düşünülebilir. Her modelde genelleme yapma eğiliminde olduğundan lineer çizgi dışında kalan değerler errorları simge eder.

- **Önemli Not:** Hata terimleri mümkün olduğunca küçük olmalıdır.

Overfitting (Aşırı Uyum)

- **Tanım:** Overfitting, bir modelin eğitim verisine aşırı derecede uyum sağlama durumudur. Bu, modelin eğitim verisi üzerinde çok iyi bir performans gösterirken, yeni ve görülmemiş veriler üzerinde kötü performans sergiler.

- **Çözüm:**
 - Modelin karmaşıklığını azaltmak.
 - Daha fazla veri toplamak.
 - Cross-validation kullanmak. (Modelin performansını daha iyi değerlendirmek için çapraz doğrulama kullanmak.)

Lineer Regresyon ve Python

- **Python Kütüphaneleri:**
- **Spicy:**

```
# Gerekli kütüphaneleri içe aktar
import numpy as np
from scipy import stats

# Verileri tanımla
Ekran_sure = np.array([3, 5, 2, 0.5, 5, 3, 1, 4, 3, 4]) # Ekran süresi (saat)
AGNO = np.array([2.7, 2.3, 3.3, 3.4, 2.3, 3.6, 2.4, 3.3, 3.3, 2.6]) # AGNO (Akademik Genel Not Ortalaması)

# Doğrusal regresyon uygula
reg = stats.linregress(Ekran_sure, AGNO)

# Eğim (slope) ve kesim (intercept) değerlerini yazdır
print("Kesişim (Intercept - b): ", reg.intercept)
print("Eğim (Slope - a): ", reg.slope)

# Doğrusal regresyon denklemini yazdır
print(f'Doğrusal Regresyon Denklemi: Y = {reg.intercept:.2f} + ({reg.slope:.3f})X')
```

Kodun Çıktısı

```
Kesişim (Intercept - b): 3.293103448275862
Eğim (Slope - a): -0.1413793103448276
Doğrusal Regresyon Denklemi: Y = 3.29 + (-0.141)X
```

Açıklamalar

1. **Kesişim (Intercept - b):**
 - $b=3.29$: Bu, $X=0$ olduğunda Y 'nin alacağı değerdir. Yani, ekran süresi sıfır olduğunda AGNO'nun beklenen değeri **3.29**'dur.
2. **Eğim (Slope - a):**
 - $a=-0.141$: Bu, ekran süresindeki 1 saatlik artışın AGNO'yu **-0.141** puan etkilediğini gösterir. Yani, ekran süresi arttıkça AGNO düşme eğilimindedir.

3. Doğrusal Regresyon Denklemi:

- $Y=3.29+(-0.141)X$: Bu denklem, ekran süresi (X) ile AGNO (Y) arasındaki ilişkiyi ifade eder. Örneğin, ekran süresi 2 saat olduğunda AGNO:

$$Y=3.29+(-0.141)\times 2$$

$$Y=3.29-0.282$$

$$Y=3.008$$

olarak tahmin edilir.

- **Scikit-learn:**

```
from sklearn.linear_model import LinearRegression  
  
model = LinearRegression()  
  
model.fit(X, y)
```

- **Statsmodels:**

```
statsmodels.api as sm  
  
model = sm.OLS(y, X).fit()
```

- Eğer modelimizde bağımsız değişkenlerin (özelliklerin) seviyeli değerlerini (kare veya küp olmayan) kullanıyorsak, bunlara Lineer Regresyon modeli denir.
- Eğer bir modelde tek bir özellik (bağımsız değişken) varsa, buna BASIT LINEER REGRESYON MODELİ denir.

- $Y=aX+b$
- $Y=mX+n$
- $Y=\beta_0+\beta_1X_1$

- "Eğer bir modelde birden fazla özellik (bağımsız değişken) varsa, buna ÇOKLU LINEER REGRESYON MODELİ denir."

$$Y=aX+bZ+c$$

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\dots+\beta_kX_k$$

Bu denklemler, basit ve çoklu lineer regresyon modellerinin farklı temsil biçimlerini göstermektedir. Özette:

- Basit Lineer Regresyon modeli, tek bir bağımsız değişken XX kullanarak bağımlı değişken Y'yi tahmin eder.

- Çoklu Lineer Regresyon modeli, birden fazla bağımsız değişken X_1, X_2, \dots, X_k , $X_{-1}, X_{-2}, \dots, X_{-k}$ kullanarak bağımlı değişken Y 'yi tahmin eder.

Regrasyon Hesabı İçin Alternatif programlar:

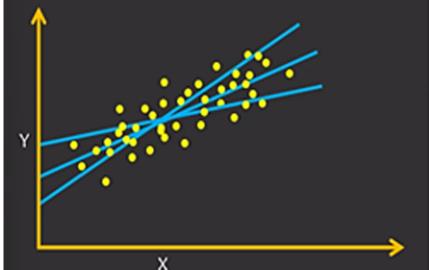


Hangi program kullanılrsa kullanılın hesaplamak için önemli olan denklemi yorumlamaktır.

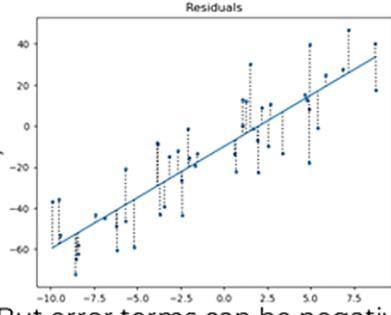
Özet Tablo

Konu	Kısa Açıklama
Lineer Regresyon	İki değişken arasındaki doğrusal ilişkiyi modellemek ve tahmin yapmak.
Bağımsız Değişken (X)	Sebep veya girdi olarak kabul edilen değişken.
Bağımlı Değişken (Y)	Sonuç veya çıktı olarak kabul edilen değişken.
En Küçük Kareler	Hata karelerinin toplamını minimize eden doğruya bulma yöntemi.
Overfitting	Modelin eğitim verilerine aşırı uyum sağlama.
Python Uygulaması	Scikit-learn veya Statsmodels kütüphaneleri ile model kurma.

Question: How can we be sure to our regression line is the best fit line?



Answer: Minimising the error.



- But error terms can be negative and positive. If we sum of them, result will be "0".

Grafik üzerindeki cevap: "Hataları minimuma indirmek" şeklinde belirtilmiş.

Alt kısmda, hataların negatif ve pozitif olabileceği ve bunların toplamının sıfır olacağı belirtilmiş.

Residual Term (Artık Terim = Hata)

Tanım:

Hata terimi (residual term), bir regresyon modelinin tahmin ettiği değerler ile gerçek gözlemler arasındaki farkı ifade eden bir kavramdır.

Amaç:

- Modelin ne kadar doğru olduğunu ve gözlemlerle ne kadar uyumlu olduğunu değerlendirmek.

Örnek Denklemi:

- Bir regresyon modelinde hata terimi, şu şekilde gösterilir:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i : Bağımlı değişken (gerçek gözlem).
- β_0 : Y kesim noktası (intercept).
- β_1 : Eğim katsayısı (slope).
- X_i : Bağımsız değişken.
- ϵ_i : Hata terimi (residual term).

Pearson's R Calculation (Pearson Korelasyon Katsayısı Hesaplama)

- Tanım:** İki değişken arasındaki doğrusal ilişkinin yönünü ve gücünü ölçen bir istatistiksel yöntemdir.
- Formül:**

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

- Cov(X, Y):** X ve Y arasındaki kovaryans.
- σ_X :** X'in standart sapması.
- σ_Y :** Y'nin standart sapması.
- Değer Aralığı:** -1 ile +1 arasındadır.
 - +1:** Mükemmel pozitif ilişki.
 - 1:** Mükemmel negatif ilişki.
 - 0:** İlişki yok.
- Yorum:**
 - Pozitif Korelasyon:** İki değişken birlikte artar veya azalır.
 - Negatif Korelasyon:** Bir değişken artarken diğeri azalır.
- Örnek:**
 - $X = [1, 2, 3, 4, 5]$
 - $Y = [2, 4, 6, 8, 10]$

- $r=1$ (Mükemmel pozitif ilişki).

Residual Term (Artık Terim = Hata)

- **Tanım:** Gerçek değerler ile modelin tahmin ettiği değerler arasındaki farktır.
- **Formül:**

$$(\text{Residual}) = y_i - \hat{y}_i$$

- y_i : Gerçek değer.
- \hat{y}_i : Tahmin edilen değer.

- **Yorum:**
 - Artık terimlerin küçük olması, modelin gerçek değerlere yakın tahminler yaptığını gösterir.
 - Artık terimlerin büyük olması, modelin hatalı tahminler yaptığını gösterir.

Coefficient of Determination (R^2 - Determinasyon Katsayısı)-ÖNEMLİ!

- **Tanım:** Bağımlı değişkendeki (Y) toplam varyansın ne kadarının bağımsız değişkenler (X) tarafından açıkladığını gösteren bir ölçütür.
- **Formül:**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- SSRES Sartıkların karelerinin toplamıdır (residual sum of squares).
- SSTOT Toplam kareler toplamıdır (total sum of squares).
- y_i gözlemlenen değerleri temsil eder.
- \hat{y}_i tahmin edilen değerleri temsil eder.
- \bar{y} gözlemlenen değerlerin ortalamasıdır.
- Toplama işaretü Σ tüm veri noktaları üzerinde toplamayı ifade eder.
- **Değer Aralığı:** 0 ile 1 arasındadır.
 - $R^2 = 1$: Model, bağımlı değişkendeki tüm varyansı açıklar (mükemmel uyum).
 - $R^2 = 0$: Model, bağımlı değişkendeki varyansı hiç açıklamaz.
 - $0 < R^2 < 1$: Model, bağımlı değişkendeki varyansın bir kısmını açıklar.

NOT: ML'de model eğitirken modelimi ne kadar iyi fit etti(modelledik) kavramını kullanırız. Modeli ne kadar iyi fit ettiğimizi R2 belirler.

R^2 'nin Yorumu

- **Modelin Uyum Kalitesi:**
 - R^2 değeri ne kadar yüksekse, modelin verilere o kadar iyi uyduğu söylenir.
 - Örneğin, $R^2 = 0.85$ ise, model bağımlı değişkendeki varyansın %85'ini açıklıyor demektir.

- **Bağımsız Değişkenlerin Etkisi:**
 - R^2 , bağımsız değişkenlerin (X) bağımlı değişken (Y) üzerindeki etkisini ölçer.
 - Yüksek R^2 , bağımsız değişkenlerin Y 'yi iyi açıkladığını gösterir.
- **Örnek**

Veri:

- $X = [1, 2, 3, 4, 5]$
- $Y = [2, 4, 6, 8, 10]$

- **Model:**

$$Y=2X$$

- **R^2 Hesaplama:**

- Tahminler: $Y=[2,4,6,8,10]$

$$SS_{\text{res}} = \sum (y_i - \hat{y}_i)^2 = 0$$

$$SS_{\text{tot}} = \sum (y_i - \bar{y})^2 = 40$$

$$R^2 = 1 - \frac{0}{40} = 1$$

Özet Tablo

Kavram	Açıklama
Pearson's R	İki değişken arasındaki doğrusal ilişkinin yönünü ve gücünü gösterir.
Residual Term	Gerçek değer ile tahmin edilen değer arasındaki fark.
R^2 (Determinasyon)	Bağımlı değişkendeki varyansın ne kadarının bağımsız değişkenlerle açıkladığını gösterir.
$R^2 = 1$	Mükemmel uyum (tüm varyans açıklanır).
$R^2 = 0$	Hiçbir varyans açıklanmaz.
$0 < R^2 < 1$	Varyansın bir kısmı açıklanır.

What is R-squared (R^2)?

Answer: R-squared, denoted as R^2 , is a statistical measure that represents the proportion of the variance in the dependent variable that is explained¹ by the independent variable(s) in a regression model.

R-kare (R^2) nedir?

Cevap: R-kare, R^2 olarak gösterilir, bir regresyon modelinde bağımlı değişkendeki varyansın bağımsız değişken(ler) tarafından açıklanan oranını temsil eden istatistiksel bir ölçütür.

What does an R-squared value of 0.75 mean?

Answer: An R-squared value of 0.75 means that **75%** of the variance in the dependent variable can be explained by the independent variable(s) in the model, and the remaining 25% is unexplained.

0,75'lik bir R-kare değeri ne anlama gelir?

Cevap: 0,75'lik bir R-kare değeri, bağımlı değişkendeki varyansın **%75'inin** modeldeki bağımsız değişken(ler) tarafından açıklanabileceği ve geri kalan %25'in açıklanamadığı anlamına gelir.

Can R-squared be negative?

Answer: No, R-squared cannot be negative. It will always fall within the range of **0 to 1**. A negative value would not make sense in the context of explaining variance.

R-kare negatif olabilir mi?

Cevap: Hayır, R-kare negatif olamaz. Her zaman **0 ile 1** aralığında olacaktır. Negatif bir değer, varyansı açıklama bağlamında mantıklı olmaz.

What is the range of possible values for R-squared?

Answer: R-squared values range from **0 to 1**. An R² of 0 indicates that the model does not explain any of the variance in the dependent variable, while an ¹R² of 1 means that the model explains all of the variance.

R-kare için olası değerlerin aralığı nedir?

Cevap: R-kare değerleri **0 ile 1** arasında değişir. 0'lık bir R², modelin bağımlı değişkendeki varyansın hiçbirini açıklamadığını gösterirken, 1'lik bir R² modelin varyansın tamamını açıkladığı anlamına gelir.

What are the limitations of R-squared?

Answer: R-squared has some limitations. It cannot determine causation; it doesn't reveal the significance of individual predictors ¹(features), and a high R-squared does not guarantee a good model fit if the model is overfitted. It's important to consider these limitations when using R-squared in analysis.

R-karenin sınırlamaları nelerdir?

Cevap: R-karenin bazı sınırlamaları vardır. Nedenselliği belirleyemez; bireysel tahmin edicilerin (özelliklerin) önemini ortaya çıkarmaz ve yüksek bir R-kare, model aşırı uyumluysa (overfitted) iyi bir model uyumunu garanti etmez. Analizde R-kare kullanırken bu sınırlamaları göz önünde bulundurmak önemlidir.

When should you use R-squared as an evaluation metric?

Answer: R-squared is commonly used in regression analysis to assess the ¹model's fit. It is useful when you want to understand how well your independent variables explain the variation in the dependent variable.

R-kareyi bir değerlendirme ölçüyü olarak ne zaman kullanmalısınız?

Cevap: R-kare, modelin uyumunu değerlendirmek için regresyon analizinde yaygın olarak kullanılır. Bağımsız değişkenlerinizin bağımlı değişkendeki varyasyonu ne kadar iyi açıkladığını anlamak istediğinizde kullanışlıdır.

5. Bölüm : Olasılık

Probability (Olasılık)- İhtimaliyet Prensibi

- **Tanım:** Bir olayın gerçekleşme şansını ifade eden 0 ile 1 arasında bir sayıdır.
 - **0:** Olayın gerçekleşme şansı yok.
 - **1:** Olay kesinlikle gerçekleşecektir.
- **Örnek:** Yazı-tura atarken tura gelme olasılığı 0.5'tir.

Neden Probabilistic Yaklaşım

- **Belirsizlik:** Hayatta her şey kesin değildir, belirsizlikler vardır. Olasılık, bu belirsizlikleri ölçme ve anlama yöntemi sağlar.
- **Tahmin Yapma:** Gelecekteki olayları tahmin etmek için olasılık kullanılır. Bu, belirsiz olaylar hakkında bilgi sahibi olmanın yoludur.
- **Veri Biliminde Önemi:** Veri analizi ve makine öğrenmesinde güvenilir tahminler yapmak için olasılık temel bir araçtır.

Örnekler:

- **Hava Durumu Tahminleri**

Meteorologlar, hava durumunu tahmin etmek için probabilistic modeller kullanır. Örneğin, %70 yağmur ihtimali, geçmiş verilere dayanarak hesaplanır.

- **Sağlık Riski Değerlendirmesi**

Doktorlar, bir hastanın belirli bir hastalığı geliştirme riskini probabilistic modeller kullanarak değerlendirir. Örneğin, sigara içen birinin kalp hastalığı riskinin hesaplanması.

- **Finans ve Yatırım**

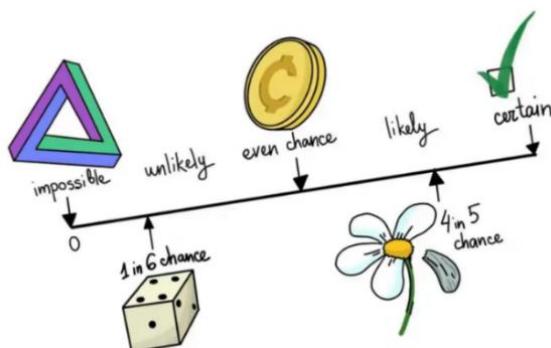
Finansal analistler, yatırım kararlarını probabilistic modellerle destekler. Örneğin, bir hissenin gelecekteki fiyatını tahmin etmek.

- **Makine Öğrenmesi**

Makine öğrenmesi modelleri, veri sınıflandırma ve tahmin problemlerinde probabilistic yaklaşımalar kullanır. Örneğin, bir e-posta'nın spam olup olmadığını belirlemek.

- **Optimizasyon ve Lojistik**

Şirketler, tedarik zinciri ve lojistik planlamada probabilistic modeller kullanır. Örneğin, ürünlerin teslimat süresini tahmin etmek.



Büyük Sayılar Yasası (Law of Large Numbers)

- Tanım:** Bir olayı ne kadar çok tekrarlarsanız, sonuçlar beklenen olasılığa o kadar yaklaşır.
 - Örnek:** Yazı-tura atarken, ne kadar çok atarsanız tura gelme oranı %50'ye yaklaşır.
 - Uygulama Alanları:** Kumarhaneler, borsa tahminleri, istatistiksel analizler.

What is the Law of Large Numbers?

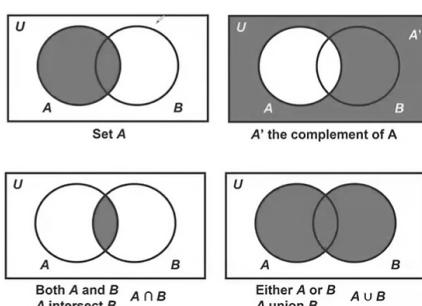
Answer: It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate.

Büyük Sayılar Yasası nedir?

Cevap: Aynı deneyi çok sayıda kez gerçekleştirmenin sonucunu açıklayan bir teoremdir. Bu teorem, frekans tarzı düşüncenin temelini oluşturur. Örneklemler ortalamalarının, örneklemler varyansının ve örneklemler standart sapmasının tahmin etmeye çalışıkları şeye yakınsadığını söyler.

Kümeler ve Olasılık (Union, Intersection, Complements)

- Birleşim (Union):** İki olaydan en az birinin gerçekleşme olasılığı.
 - Örnek:** $P(A \cup B)$
- Kesişim (Intersection):** İki olayın aynı anda gerçekleşme olasılığı.
 - Örnek:** $P(A \cap B)$
- Tümleyen (Complement):** Bir olayın gerçekleşmemeye olasılığı.
 - Örnek:** $P(A') = 1 - P(A)$



Permütasyon (Permutation)

- **Tanım:** Belirli bir sıraya göre dizilim yapmaktadır. Sıralama önemlidir.
- **Formül:**

$$P(n, k) = \frac{n!}{(n - k)!}$$

- n : Toplam eleman sayısı.
 - k : Seçilen eleman sayısı.
- **Örnek:** 3 farklı kitabı bir rafa kaç farklı şekilde dizebilirsınız?

$$P(3, 3) = \frac{3!}{(3 - 3)!}$$

Kombinasyon (Combination)

- **Tanım:** Belirli bir grup içinden sıra gözetmeksizin seçim yapmaktadır. Sıralama önemsizdir.
- **Formül:**

$$C(n, k) = \frac{n!}{k!(n - k)!}$$

- n : Toplam eleman sayısı.
 - k : Seçilen eleman sayısı.
- **Örnek:** 5 kişilik bir gruptan 2 kişi kaç farklı şekilde seçilir?

$$C(5, 2) = \frac{5!}{2!(5 - 2)!}$$

Büyük Sayılar Kanunu (Law of Large Numbers)

Tanım:

Büyük Sayılar Kanunu, bir deney veya gözlem tekrarlandıkça, gözlemlenen sonuçların teorik olasılığa yakınsadığını ifade eden istatistiksel bir prensiptir. Başka bir deyişle, bir olayın gerçekleşme sıklığı, deneme sayısı arttıkça beklenen olasılığa yaklaşır.

Temel Kavramlar:

1. **Teorik Olasılık:** Bir olayın ideal koşullarda gerçekleşme olasılığıdır. Örneğin, bir madeni para atıldığındaysa yazı gelme olasılığı teorik olarak %50'dir.
2. **Ampirik (Deneysel) Olasılık:** Bir olayın gerçekleşme sıklığının gözlemlenmesiyle elde edilen olasılıktır. Örneğin, 100 kez para atıldığındaysa 45 kez yazı gelirse, yazı gelme ampirik olasılığı %45'tir.
3. **Göreceli Frekans (Relative Frequency):** bir olayın belirli bir sayıda deneme veya gözlemede ne sıklıkta gerçekleştiğini gösteren bir ölçütür. Basitçe, bir olayın tüm denemelere oranını ifade eder. Formülü şu şekildedir:

$$\text{Relative Frequency (RF)} = \frac{\text{Number of Times Event Occurred}}{\text{Total Number of Trials}}$$

Örnekler:

- **Para Atma Örneği:** Bir madeni para atıldığında yazı gelme olasılığı teorik olarak %50'dir. Az sayıda atışta (örneğin 10 atış) yazı gelme oranı %40 ile %60 arasında değişebilir. Ancak atış sayısı arttıkça (örneğin 1000 atış) yazı gelme oranı %50'ye yaklaşır.
- **Kumarhane Örneği:** Kumarhanelerde rulet gibi oyunlarda belirli bir sayının gelme olasılığı sabittir. Çok sayıda oyun oynandığında, sonuçlar bu olasılığa yakınsar ve kumarhane uzun vadede kâr eder.

Tablo Örneği:

Atış Sayısı	Göreceli Frekans Aralığı	Yüzde
10	0.4 - 0.6	66
100	0.49 - 0.51	92
1,000	0.499 - 0.501	97
10,000	0.4999 - 0.5001	99

Uygulama Alanları:

- **Finans:** Yatırım stratejilerinde ve risk yönetiminde kullanılır.
- **Sigorta:** Sigorta şirketleri, büyük sayılar kanununu kullanarak riskleri hesaplar ve primleri belirler.
- **Kalite Kontrol:** Üretim süreçlerinde hata oranlarını tahmin etmek için kullanılır.

What is the Law of Large Numbers?

Answer: It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample means,¹ the sample variance and the sample standard deviation converge to what they are trying to estimate.

Büyük Sayılar Yasası nedir?

Cevap: Aynı deneyi çok sayıda kez gerçekleştirmenin sonucunu açıklayan bir teoremdir. Bu teorem, frekans tarzı düşüncenin temelini oluşturur. Örneklem ortalamalarının, örneklem varyansının ve örneklem standart sapmasının tahmin etmeye çalışıkları şeye yakınsadığını söyler.

Koşullu Olasılık (Conditional Probability)

- **Tanım:** Bir olayın gerçekleştiği bilindiğinde, başka bir olayın gerçekleşme olasılığı.
- **Formül:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- $P(A|B)$: B olayı gerçekleştiğinde A'nın olma olasılığı.
- $P(A \cap B)$: A ve B'nin birlikte gerçekleşme olasılığı.
- $P(B)$: B'nin gerçekleşme olasılığı.

Soru:

Bir okulda öğrencilerin %45'i fizikten, %35'i matematikten ve %25'i hem fizik hem de matematikten başarısızdır. Rasgele seçilen bir öğrenci için:

- a) Fizikten başarısız olduğu biliniyorsa, matematikten de başarısız olma olasılığı nedir?
- b) Matematikten başarısız olduğu biliniyorsa, fizikten de başarısız olma olasılığı nedir?

Cevap:

a) $P(M | F) = 0.55$

Fizikten başarısız olan bir öğrencinin matematikten de başarısız olma olasılığı **%55**'tir.

b) $P(F | M) = 0.71$

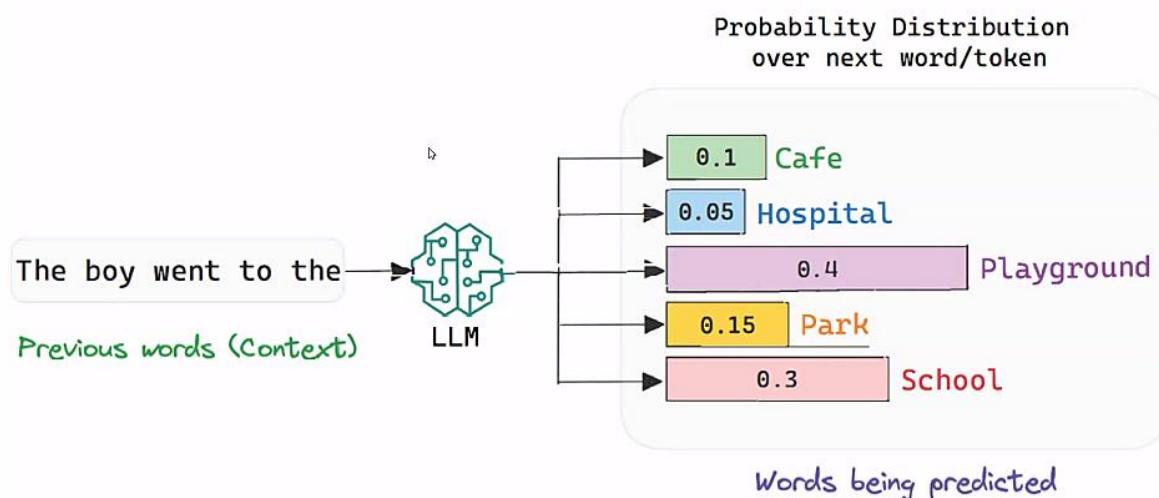
Matematikten başarısız olan bir öğrencinin fizikten de başarısız olma olasılığı **%71**'dir.

- **Örnek:** Fırtına çıktığında uçuşların iptal olma olasılığı.

LLM Modellerinde Koşullu Olasılık

- **Tanım:** Büyük dil modelleri (LLM), metin üretirken kelimelerin birbirini takip etme olasılıklarını kullanır.

- **Örnek:** "Bugün hava..." ifadesinden sonra "güneşli" kelimesinin gelme olasılığı.



Bayes Teoremi (Bayes' Theorem)

- **Tanım:** Yeni bilgiler geldiğinde, önceki olasılıkları güncellemek için kullanılır.
- **Formül:**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- $P(A|B)$: B olayı gerçekleştiğinde A'nın olma olasılığı.
- $P(B|A)$: A olayı gerçekleştiğinde B'nin olma olasılığı.
- $P(A)$: A'nın önceki olasılığı.
- $P(B)$: B'nin önceki olasılığı.

Örnek: Bayes Teoremi ile Hastalık Testi

- **Problem:** Bir hastalık testinin doğruluk oranı %99'dur. Hastalığın toplumda görülme sıklığı %1'dir. Test pozitif çıkan birinin gerçekten hasta olma olasılığı nedir?
- **Çözüm:**
 - $P(A) = 0.01$ (Hastalığın görülme olasılığı).
 - $P(B|A) = 0.99$ (Hasta olanlarda testin pozitif çıkma olasılığı).
 - $P(B|A') = 0.01$ (Sağlıklı olanlarda testin yanlış pozitif çıkma olasılığı).
 - **Bayes Teoremi ile:**

$$P(\text{Disease}|\text{Positive}) = \frac{0.99 \times 0.01}{(0.99 \times 0.01) + (0.01 \times 0.99)} = 0.5$$

- **Sonuç:** Test pozitif çıkan birinin gerçekten hasta olma olasılığı %50'dir.

Örnek: Bayes Teoremi Yangın

Verilenler:

1. Tehlikeli yangınlar çok az olur: %1
2. Duman genelde pikniklerde ve %10 seviyede görülür.
3. Yangınların %90'ında duman görülür.

Olasılıklar:

- $P(\text{Yangın}) = 0,01$ (Yangın olasılığı)
- $P(\text{Duman}) = 0,10$ (Duman olasılığı)
- $P(\text{Duman} | \text{Yangın}) = 0,90$ (Yangın olduğunda duman görülme olasılığı)

Çözüm: Bayes Teoremi, bir olayın olasılığını, diğer ilgili olayların koşullu olasılıklarını kullanarak hesaplamamıza olanak tanır. Formülü şu şekildedir:

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$

Bu durumda:

- A: Yangın olması
- B: Duman olması

Formüle uyarlayalım:

$$P(\text{Yangın}| \text{Duman}) = P(\text{Duman} | \text{Yangın}) \cdot P(\text{Yangın}) / P(\text{Duman})$$

Hesaplayalım:

$$P(\text{Fire}|\text{Smoke}) = \frac{P(\text{Smoke}|\text{Fire}) \cdot P(\text{Fire})}{P(\text{Smoke})} = \frac{0.90 \times 0.01}{0.10} = 0.09$$

Bu sonuç, duman görüldüğünde yangın çıkma olasılığının %9 olduğunu gösterir.

Özet olarak:

- Yangınların %1'i oluşur.
- Duman görülmeye olasılığı %10'dur.
- Yangın durumunda duman görülmeye olasılığı %90'dır.
- Bayes Teoremi ile, duman görüldüğünde yangın olma olasılığı %9'dur.

Bayes Teoremi kullanım alanları:

Makine Öğrenimi (Machine Learning):

- **Sınıflandırma (Classification):**
 - Naive Bayes sınıflandırıcıları, Bayes Teoremi'ni temel alır ve metin sınıflandırma, spam filtreleme gibi görevlerde kullanılır.
 - Olasılıksal modellerde, verilerin hangi sınıf'a ait olduğunu tahmin etmek için kullanılır.
- **Belirsizlik Altında Karar Verme:**
 - Modelin tahminlerindeki belirsizliği ölçmek ve bu belirsizliği karar verme süreçlerine dahil etmek için kullanılır.
 -

İstatistiksel Çıkarım (Statistical Inference):

- **Bayesçi İstatistik (Bayesian Statistics):**
 - Parametrelerin olasılık dağılımlarını güncellemek ve hipotezleri test etmek için kullanılır.
- **Anomali Tespiti (Anomaly Detection):**
 - Beklenmeyen veya sıra dışı olayları tespit etmek için olasılık modelleri oluşturulmasında kullanılır.
- **Parametre Tahmini (Parameter Estimation):**
 - Verilerden model parametrelerini tahmin etmek için kullanılır.
- **Zaman Serisi Analizi (Time Series Analysis):**
 - Gelecekteki değerleri tahmin etmek ve zaman içindeki değişiklikleri analiz etmek için kullanılır.

Diğer Uygulama Alanları:

- **Yapay Zeka (Artificial Intelligence):**
 - Olasılıksal akıl yürütme ve karar verme sistemlerinde kullanılır.
- **Doğal Dil İşleme (Natural Language Processing - NLP):**
 - Metin sınıflandırma, dil modelleme ve bilgi çıkarma gibi görevlerde kullanılır.
- **Tıp (Medicine):**
 - Hastalık teşhisi, risk değerlendirmesi ve genetik analizlerde kullanılır.
- **Finans (Finance):**
 - Risk yönetimi, portföy optimizasyonu ve dolandırıcılık tespiti gibi alanlarda kullanılır.
 -

Özetle:

Bayes Teoremi, olasılıksal akıl yürütme gerektiren her türlü problemde kullanılabilir. Özellikle, belirsizlik altında karar verme, olasılık tahminleri yapma ve verilerden bilgi çıkarma gibi durumlarda oldukça etkilidir.

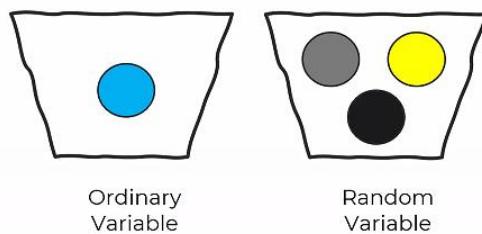
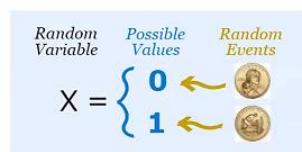
Özet Tablo

Kavram	Açıklama
Olasılık	Bir olayın gerçekleşme şansını ifade eder.
Rastgele Değişken	Bir deneyin sonucuna bağlı olarak değişen değerler.
Büyük Sayılar Yasası	Olay tekrarlandıkça sonuçlar beklenen olasılığa yaklaşır.
Permütasyon	Sıralama önemli olan dizilimler.
Kombinasyon	Sıralama önemsiz olan seçimler.
Koşullu Olasılık	Bir olayın gerçekleştiği bilindiğinde, başka bir olayın olasılığı.
Bayes Teoremi	Yeni bilgilerle olasılıkları güncellemek için kullanılır.
LLM ve Olasılık	Dil modelleri, kelimelerin birbirini takip etme olasılıklarını kullanır.

6. Bölüm: Rastgele Değişkenler ve Dağılımlar

Random Variables (Rassal Değişkenler)

- Tanım:** Bir istatistiksel deneyin sonucuna bağlı olarak değer alan değişkenlerdir. Random variablede olasılığa dayalıdır. Şansa dayalıdır. Kesin olarak bilmemez ama tahmin edilebilir. Örneğin, bir zar atıldığında gelen sayı veya bir ay içinde satılan araba sayısı.
 - Örnek:** Bir zar atıldığında gelen sayı (1, 2, 3, 4, 5, 6) bir rastgele değişkendir.



Random variable Discrete ve continuous olmak üzere ikiye ayrılır.

Kesikli (Discrete) Rastgele Değişkenler

- Tanım:** Kesikli rastgele değişkenler, **sonlu** veya **sayılabılır** sayıda değer alabilen değişkenlerdir. Bu değişkenler, belirli bir aralıkta sadece belirli noktalarda değer alır.
- Özellikler:**
 - Değerler arasında boşluklar vardır (örneğin, 1, 2, 3 gibi).

- Olasılık hesaplamaları için **olasılık kütle fonksiyonu (PMF - Probability Mass Function)** kullanılır.
- **Örnekler:**
 - Bir zar atıldığında gelen sayı (1, 2, 3, 4, 5, 6).
 - Bir sınıfındaki öğrenci sayısı.
 - Bir mağazaya bir günde gelen müşterileri sayısı.
- **Binomial Distribution (Binom Dağılımı)**

Tanım: İki olası sonucu olan (başarı/başarısızlık) deneyler için kullanılır. Tekrak eden denemeler vardır, çoklu deneylerde başarı sayısının olasılığını hesaplar.

Formül:

$$P(X = k) = C(n, k) \cdot p^k \cdot (1 - p)^{n-k}$$

P(X=k): n denemede k kez başarı elde etme olasılığı

C(n, k): n'nin k'lı kombinasyonu (n'den k eleman seçme sayısı)

p: Başarı olasılığı

n: Deneme sayısı

k: Başarı sayısı

Örnek: 10 kez para atıldığında 3 kez tura gelme olasılığı.

- **Bernoulli Distribution (Bernoulli Dağılımı)**

Tanım: Binom dağılımının özel bir hali, tek bir deney için kullanılır. İki olası sonucu vardır.

Örnek: Bir seferde para atma. Tura gelme olasılığı p , yazı gelme olasılığı $1-p$.

Örnek 2: Bir siyasi adayın önumüzdeki seçimi kazanıp kazanamaması durumu kazanamaması (iki ihtimal var seçim 1 kere yapılacak.).

- **Poisson Distribution (Poisson Dağılımı)**

Tanım: Belirli bir zaman aralığında nadir gerçekleşen olaylar için kullanılır.

Formül:

$$P(X) = \frac{\lambda^X \cdot e^{-\lambda}}{X!}$$

Örnek : Bir şehirde ortalama büyük sel sayısı yılda 3 kere oluyorsa gelecek yıl bu şehirde 4 kere sel gelme olasılı nedir?

Verilenler:

- Ortalama sel sayısı (λ) = 3
- Aranan değer (k) = 4

$$P(X = 4) = \frac{e^{-3} \cdot 3^4}{4!} = \frac{e^{-3} \cdot 81}{24}$$

Hesaplama:

$$e^{-3} \approx 0.0498$$

$$P(X = 4) = \frac{0.0498 \cdot 81}{24} = \frac{4.0338}{24} \approx \boxed{0.1681}$$

Sonuç: Gelecek yıl şehirde 4 sel olma olasılığı yaklaşık **%16.8**'dır.

Örnek: Bir hastaneye bir günde ortalama 5 hasta geliyorsa, bir gün içinde 8 hasta gelme olasılığı nedir?

- $\lambda=5$
- $k=8$

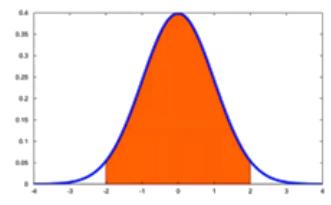
$$P(X = 8) = \frac{e^{-5} \cdot 5^8}{8!} = \frac{0.0067 \cdot 390625}{40320} \approx 0.065$$

Yani yaklaşık **%6.5** olasılıkla o gün 8 hasta gelir.

Sürekli (Continuous) Rastgele Değişkenler

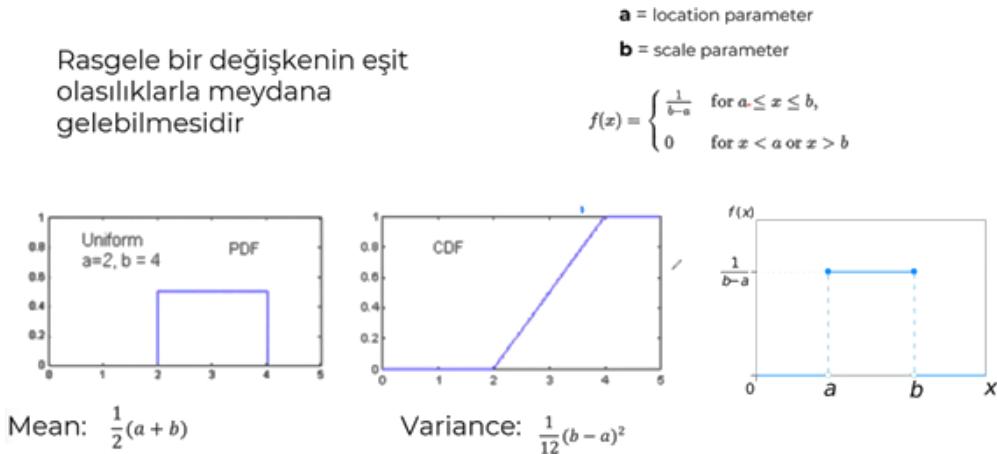
- **Tanım:** Sürekli rastgele değişkenler, **sonsuz** sayıda değer alabilen ve belirli bir aralıkta herhangi bir değeri alabilen değişkenlerdir. Bu değişkenler, sürekli bir ölçekte değerler alır. Bu dağılımlarda nokta atışı bir değer yoktur bir aralık söz konusudur.
- **Özellikler:**
 - Değerler arasında boşluk yoktur (örneğin, 1.5, 2.3, 3.7 gibi).
 - Olasılık hesaplamaları için **olasılık yoğunluk fonksiyonu (PDF - Probability Density Function)** kullanılır.
 - **Probability Density Function (PDF)**
 - $Y; X$ random değişkenin bir fonksiyonudur
 - Y ; tüm X değerleri için 0'a eşit veya büyütür
 - Eğri altındaki kalan alan 1'e eşittir.
- **Örnekler:**
 - Bir öğrencinin boy uzunluğu (örneğin, 1.70 m, 1.75 m, 1.80 m).
 - Bir arabanın hızı (örneğin, 50 km/s, 60.5 km/s).
 - Bir ürünün ağırlığı (örneğin, 1.2 kg, 1.25 kg).
- **Uniform Distribution (Uniform Dağılımı)**

Tanım: Tüm değerler eşit olasılıkla gerçekleşir.



$$\text{Alan} = \text{genişlik} \cdot \text{Uzunluk} = 1 \cdot 1 = 1$$

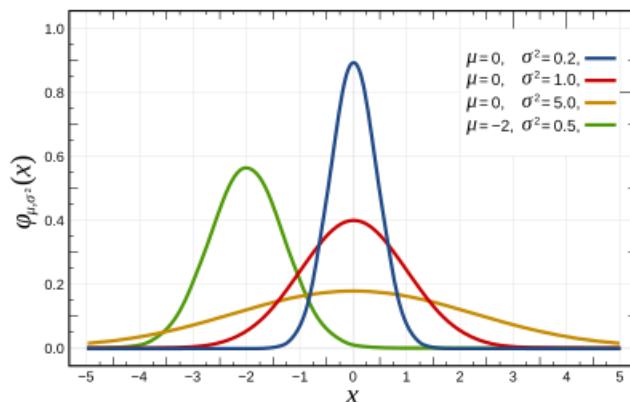
Örnek: Bir otobüs durağında, otobüsün her 10 dakikada bir geldiğini ve geliş saatlerinin tamamen rastgele olduğunu varsayıyalım. Bu durumda, otobüsü beklediğiniz sürenin 0 ile 10 dakika arasında herhangi bir değer alma olasılığı eşittir. Bu, sürekli uniform dağılıma bir örnektir.



- **Normal Distribution (Normal Dağılımı) -Gauss**

Tanım: Doğal olayların çoğu bu dağılıma uyar. Simetrik ve çan eğrisi şeklindedir. Mean, median, mode eşittir.

Örnek: İnsanların boy uzunlukları.



Eğri altında kalan alanlar 1'e eşittir.

3 Sigma Kuralı Nedir? (68-95-99.7 Kuralı olarak da bilinir)

3 Sigma Kuralı, normal dağılım gösteren bir veri setinde, verilerin ne kadarının ortalama değer etrafında belirli aralıklarda yer aldığı gösteren bir istatistiksel kurallıdır.

Normal Dağılım ve Sigma (σ)

Normal Dağılım: Çan eğrisi şeklinde simetrik bir dağılımdır. Birçok doğal olay ve veri seti normal dağılım gösterir.

Sigma (σ): Standart sapma olarak da bilinir ve verilerin ortalama değerden ne kadar yayıldığıni ölçer.

3 Sigma Kuralı'nın Anlamı

$\mu \pm 1\sigma$: Verilerin yaklaşık %68'i ortalama değerin (μ) 1 standart sapma (1σ) saği ve solu arasındaki aralıktır yer alır.

$\mu \pm 2\sigma$: Verilerin yaklaşık %95'i ortalama değerin 2 standart sapma saği ve solu arasındaki aralıktır yer alır.

$\mu \pm 3\sigma$: Verilerin yaklaşık %99,7'si ortalama değerin 3 standart sapma saği ve solu arasındaki aralıktır yer alır.

3 Sigma Kuralı'nın Kullanım Alanları

- **Kalite Kontrol:** Üretim süreçlerinde, ürünlerin belirli toleranslar içinde kalıp kalmadığını kontrol etmek için kullanılır.
- **Finans:** Risk yönetimi ve portföy analizinde kullanılır.
- **Mühendislik:** Tasarım ve analiz süreçlerinde kullanılır.
- **Veri Analizi:** Aykırı değerleri (outliers) tespit etmek ve verilerin dağılımını anlamak için kullanılır.

Örnek:

Diyelim ki bir fabrikada üretilen vidaların uzunlukları normal dağılım gösteriyor. Ortalama uzunluk 10 cm ve standart sapma 0,1 cm olsun.

Vidaların yaklaşık %68'i 9,9 cm ile 10,1 cm arasında olacaktır.

Vidaların yaklaşık %95'i 9,8 cm ile 10,2 cm arasında olacaktır.

Vidaların yaklaşık %99,7'si 9,7 cm ile 10,3 cm arasında olacaktır.

Özet

3 Sigma Kuralı, normal dağılım gösteren verilerin ortalama değer etrafında nasıl dağıldığını anlamamıza yardımcı olur. Bu kural, birçok alanda karar verme ve analiz süreçlerinde önemli bir araçtır.

What is Normal Distribution?

Normal Distribution is a probability distribution that is symmetric about the mean. It is also known as Gaussian Distribution. The distribution appears as a Bell-shaped curve which means the mean is the most frequent data in the given data set.

In Normal Distribution:

- Mean = Median = Mode
- Total area under the curve is 1.

Normal Dağılım nedir?

Normal Dağılım, ortalama etrafında simetrik olan bir olasılık dağılımıdır. Gauss Dağılımı olarak da bilinir. Dağılım, Çan Eğrisi şeklinde görünür, bu da ortalamanın verilen veri setindeki en sık veri olduğu anlamına gelir.

Normal Dağılımda:

- Ortalama = Medyan = Mod
- Eğrinin altındaki toplam alan 1'dir.

What do you understand by the term Normal Distribution?

Normal distribution, also known as the Gaussian distribution, is a bell-shaped frequency distribution curve. Most of the data values in a normal distribution tend to cluster around the mean.

Normal Dağılım teriminden ne anlıyorsunuz?

Gauss dağılımı olarak da bilinen Normal dağılım, çan şeklinde bir frekans dağılımı eğrisidir. Normal bir dağılımdaki veri değerlerinin çoğu ortalama etrafında kümelenme eğilimindedir.

What is the empirical rule?

The Empirical Rule is often called the 68-95-99.7 rule or Three Sigma Rule. It states that on a Normal Distribution:

- 68% of the data will be within one Standard Deviation of the Mean.
- 95% of the data will be within two Standard Deviations of the Mean.
- 99.7% of the data will be within three Standard Deviations of the Mean.

Ampirik kural nedir?

Ampirik Kural, genellikle 68-95-99.7 kuralı veya Üç Sigma Kuralı olarak adlandırılır. Normal Dağılımda şunları belirtir:

- Verilerin %68'i ortalamanın bir Standart Sapması içinde yer alır.
- Verilerin %95'i ortalamanın iki Standart Sapması içinde yer alır.
- Verilerin %99,7'si ortalamanın üç Standart Sapması içinde yer alır.

What is a bell-curve distribution?

A bell-curve distribution is represented by the shape of a bell and indicates normal distribution. It occurs naturally in many situations especially while analyzing financial data. The top of the curve shows the mode, mean and median of the data and is perfectly symmetrical. The key characteristics of a bell-shaped curve are -

- The empirical rule says that approximately 68% of data lies within one standard deviation of the mean in either of the directions.
- Around 95% of data falls within two standard deviations and
- Around 99.7% of data fall within three standard deviations in either direction.

Çan eğrisi dağılımı nedir?

Çan eğrisi dağılımı, bir çan şekli ile temsil edilir ve normal dağılımı gösterir. Özellikle finansal verileri analiz ederken birçok durumda doğal olarak ortaya çıkar. Eğrinin tepesi, verilerin modu, ortalaması ve medyanını gösterir ve tamamen simetiktir. Çan şeklindeki bir eğrinin temel özellikleri şunlardır:

- Ampirik kural, verilerin yaklaşık %68'inin ortalamanın bir standart sapması içinde her iki yönde de yer aldığı söyler.
- Verilerin yaklaşık %95'i iki standart sapma içinde yer alır ve
- Verilerin yaklaşık %99,7'si her iki yönde de üç standart sapma içinde yer alır.

What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

Normal Dağılım teriminden ne anlıyorsunuz?

Veriler genellikle sola veya sağa doğru bir eğilimle farklı şekillerde dağılır veya tamamen karmaşık olabilir. Ancak, verilerin sola veya sağa doğru herhangi bir eğilim olmadan merkezi bir değer etrafında dağılıma ve çan şeklinde bir eğri şeklinde normal dağılıma ulaşma olasılığı vardır.

Ek Bilgiler:

- **Çan Eğrisi (Normal Dağılım) Özellikleri:**
 - Simetrik: Ortalama etrafında simetrik bir şeke sahiptir.
 - Ortalama, Medyan ve Mod Eşittir: Dağılımin merkezi ve en sık değeri aynıdır.
 - 68-95-99.7 Kuralı: Verilerin belirli standart sapma aralıklarında ne kadarının yer aldığı gösterir.
 - Finansal Verilerde Sıklıkla Görülür: Özellikle varlık getirileri, risk ölçümü gibi finansal verilerde normal dağılım sıkça görülür.
- **Neden Önemli?**
 - Verilerin dağılımını anlamak ve tahminler yapmak için önemlidir.
 - İstatistiksel analizlerde ve makine öğreniminde sıkça kullanılır.
 - Risk yönetimi ve karar verme süreçlerinde kullanılır.

Z Dağılımı (Standart Normal Dağılım)

Tanım: Ortalaması 0 ve standart sapması 1 olan normal dağılımdır.

Kullanım Alanları:

- Popülasyon standart sapması (σ) bilindiğinde.
- Örneklem büyüklüğü (n) büyük olduğunda (genellikle $n > 30$).

Özellikler:

- Simetrik ve çan eğrisi şeklindedir.
- Verilerin %68'i ortalamanın ± 1 standart sapma, %95'i ± 2 standart sapma, %99.7'si ± 3 standart sapma içinde yer alır.

Formül:

$$z = \frac{x - \mu}{\sigma}$$

- x : Veri noktası
- μ : Ortalama
- σ : Standart sapma

Örnek 1: Z Dağılımı

Veri: Bir sınıfındaki öğrencilerin boy uzunlukları ortalaması 170 cm, standart sapması 10 cm.

Soru: 180 cm'den uzun öğrencilerin oranı nedir?

Çözüm:

$$z = \frac{180 - 170}{10} = 1$$

- Z tablosundan $z=1$ için olasılık: 0.8413
- 180 cm'den uzun öğrencilerin oranı: $1 - 0.8413 = 0.15871$ (%15.87)

T Dağılımı (Student's T Dağılımı)

Tanım: Örneklem büyüklüğü küçük olduğunda ve popülasyon standart sapması bilinmediğinde kullanılan bir dağılımdır.

Kullanım Alanları:

- Popülasyon standart sapması (σ) bilinmediğinde.
- Örneklem büyüklüğü (n) küçük olduğunda (genellikle $n < 30$).

Özellikler:

- Normal dağılıma benzer ancak kuyrukları daha kalındır.
- Örneklem büyülüğu arttıkça normal dağılıma yaklaşır.
- Serbestlik derecesi (degrees of freedom - df) ile tanımlanır:

Formül:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- \bar{x} : Örneklem ortalaması
- μ : Popülasyon ortalaması
- s : Örneklem standart sapması
- n : Örneklem büyülüğu

Örnek 2: T Dağılımı

Veri: Bir örneklemde 10 öğrencinin boy uzunlukları ortalaması 172 cm, örneklem standart sapması 8 cm.

Soru: Popülasyon ortalamasının 170 cm olduğu varsayırlrsa, t değeri nedir?

Çözüm:

$$t = \frac{172 - 170}{8/\sqrt{10}} = \frac{2}{2.53} \approx 0.79$$

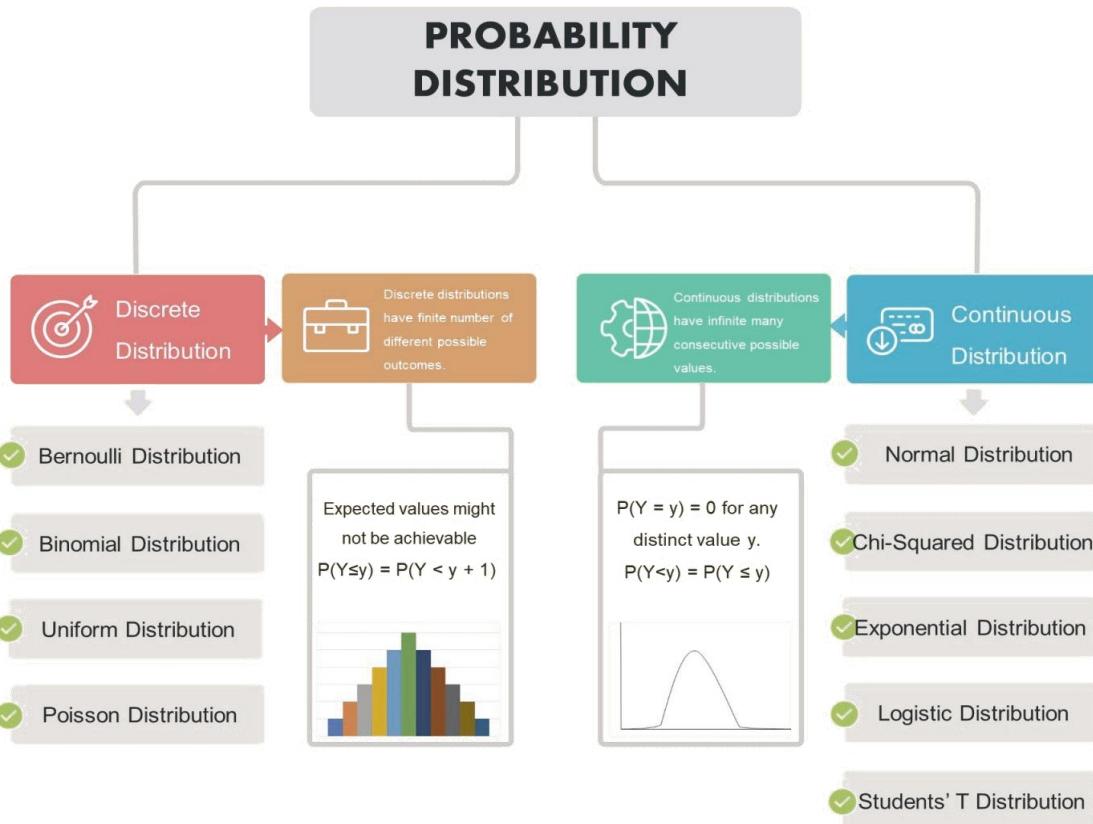
- **Serbestlik derecesi:** $df=10-1=9$
- T tablosundan $t=0.79$ ve $df=9$ için olasılık bulunur.

Olasılık Dağılımları Neden Önemlidir?

1. **Olasılık Tahminleri:** Olayların gerçekleşme olasılıklarını tahmin etmeye yardımcı olur.
2. **İstatistiksel Analiz:** Verilerin dağılımını anlamak ve analiz etmek için kullanılır.
3. **Makine Öğrenmesi:** Birçok makine öğrenmesi modeli, olasılık dağılımlarına dayalı varsayımlarla çalışır.

Types of Probability Distribution

Characteristics, Examples, & Graph



Python kütüphanelerinde fonsiyon olarak aşağıdaki kavamlar kullanılır:

1. PMF (Olasılık Kütle Fonksiyonu - Probability Mass Function):

- **Ne İşe Yarar?**
 - PMF, sadece ayrık (kesikli) rastgele değişkenlerle ilgilenir. Yani, sadece belirli değerleri alabilen değişkenlerle (örneğin, zar atışındaki sayılar, bir madeni paranın yazı veya tura gelmesi gibi).
 - PMF, bu ayrık değişkenin tam olarak belirli bir değeri alma olasılığını hesaplar.
 - **Örnek:**
 - Bir zar attığınızda 3 gelme olasılığı nedir? İşte PMF bu soruyu cevaplar.
- $$PMF(X = 3) = \frac{1}{6}$$
- (Eğer zar hilesizse).
- **Özet:**
 - PMF, "Tam olarak bu değerin olasılığı nedir?" sorusuna cevap verir.

2. CDF (Kümülatif Dağılım Fonksiyonu - Cumulative Distribution Function):

- **Ne İşe Yarar?**
 - CDF, hem ayrık hem de sürekli rastgele değişkenlerle ilgilenir.

- CDF, bir rastgele değişkenin belirli bir değerden küçük veya eşit olma olasılığını hesaplar.
- **Örnek:**
 - Bir zar attığınızda 3 veya daha küçük bir sayı gelme olasılığı nedir? İşte CDF bu soruyu cevaplar.

$$CDF(X \leq 3) = PMF(1) + PMF(2) + PMF(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

- **Özet:**
 - CDF, "Bu değer veya daha küçük değerlerin olasılığı nedir?" sorusuna cevap verir.

Temel Farklar:

- **PMF:**
 - Tek bir değerin olasılığı.
 - Ayrık değişkenler için.
- **CDF:**
 - Bir değer ve daha küçük değerlerin olasılığı.
 - Hem ayrık hem de sürekli değişkenler için.

Python'da örnek çözüm:

Zar Atma

```
from scipy.stats import binom

# Zar atma örneği (6 yüz, her yüzün olasılığı 1/6)
n = 6 # Deneme sayısı (zarın yüz sayısı)
p = 1/6 # Başarı olasılığı (herhangi bir yüzün gelme olasılığı)

# PMF: Tam olarak 3 gelme olasılığı
pmf_3 = binom.pmf(k=3, n=n, p=p)
print(f"PMF (X = 3): {pmf_3}") # Çıktı: 0.0536

# CDF: 3 veya daha küçük bir değer gelme olasılığı
cdf_3 = binom.cdf(k=3, n=n, p=p)
print(f"CDF (X ≤ 3): {cdf_3}")
```

Özet Tablo:

Kavram	Açıklama
Random Variables (Rassal Değişkenler)	Bir deneyin sonucuna bağlı olarak değişen değerler. Örnek: Zar atma sonucu.
Discrete Probability Distributions (Kesikli Olasılık Dağılımları)	Kesikli rastgele değişkenlerin olasılık dağılımları. Örnek: Zar atma.
Binomial Distribution (Binom Dağılımı)	İki sonuçlu (başarı/başarısızlık) deneyler için kullanılır. Örnek: Para atma.
Bernoulli Distribution (Bernoulli Dağılımı)	Binom dağılıminin özel hali, tek bir deney için kullanılır. Örnek: Tek para atma.
Poisson Distribution (Poisson Dağılımı)	Belirli bir zaman aralığında nadir gerçekleşen olaylar için kullanılır. Örnek: Bir günde gelen müşteri sayısı.
Continuous Probability Distributions (Sürekli Olasılık Dağılımları)	Sürekli rastgele değişkenlerin olasılık dağılımları. Örnek: Boy uzunluğu.
Uniform Distribution (Uniform Dağılımı)	Tüm değerler eşit olasılıkla gerçekleşir. Örnek: Zarın her yüzünün gelme olasılığı.
Normal Distribution (Normal Dağılımı)	Doğal olayların çoğu bu dağılıma uyar. Simetrik ve çan eğrisi şeklindedir. Örnek: İnsanların boy uzunlukları.
Standard Distribution (Standart Dağılım)	Ortalaması 0, standart sapması 1 olan normal dağılım. Örnek: Standart normal dağılım tablosu.
T Distribution (T Dağılımı)	Küçük örneklemelerde kullanılır, normal dağılıma benzer ancak kuyrukları daha kalındır. Örnek: Küçük örneklemelerde ortalama hesaplamaları.
Z Distribution (Z Dağılımı)	Büyük örneklemelerde ve popülasyon standart sapması bilindiğinde kullanılır. Örnek: Büyük örneklemelerde hipotez testleri.

7. Bölüm : Örneklem Dağılımları ve Güven Aralıkları

Örneklem Dağılımları (Sample Distributions)

- Tanım:** Örneklem dağılımı, bir popülasyondan çekilen örneklemelerin istatistiklerinin (örneğin, ortalama, standart sapma) dağılımını ifade eder.
- Özellikler:**
 - Örneklem dağılımı, örneklem büyüklüğüne ve popülasyon dağılımına bağlıdır.
 - Örneklem büyütüldüğünde, örneklem dağılımı popülasyon dağılımına daha çok benzer.
- Örnek:** Bir popülasyondan 100 farklı örneklem çekilirse, her bir örneklenin ortalaması farklı olacaktır. Bu ortalamaların dağılımı, örneklem dağılımını oluşturur.

Weibull Dağılımı ve Örneklemeler

- **Weibull Dağılımı:** Genellikle güvenilirlik analizlerinde ve yaşam süresi modellerinde kullanılan bir sürekli olasılık dağılımıdır. Şekil parametresi (k) ve ölçek parametresi (λ) ile tanımlanır.
- **Örneklemeler:** Bir popülasyondan çekilen farklı örneklemeler, farklı ortalama ve standart sapma değerlerine sahip olabilir. Bu, örneklemelerin popülasyonu ne kadar iyi temsil ettiğini gösterir.

Örneklem Hatası (Sample Error)

- **Tanım:** Örneklem hatası, örneklemelerin birbirine göre tutarlığını ve popülasyon parametrelerini ne kadar iyi tahmin ettiğini ölçer.
- **Özellikler:**
 - Örneklem hatası küçükse, örneklemeler birbirine yakındır ve tahminler daha doğrudur.
 - Örneklem hatası büyükse, örneklemeler birbirinden uzaktır ve tahminler daha az güvenilirdir.
- **Hesaplama:** Örneklem hatası, örneklemelerin standart sapmalarına ve örneklem büyütüklüğüne bağlıdır.

Örneklem Dağılımı ve Weibull Dağılımı

- **Weibull Dağılımı Grafiği:** Weibull dağılımı grafiğinde, farklı parametrelerle (örneğin, farklı k ve λ değerleri) oluşturulmuş 6 farklı sekme bulunabilir. Her bir sekme, farklı bir örneklem dağılımını temsil eder.
- **Ortalamaların Farklılığı:** Her bir sekmedeki örneklemelerin ortalamaları farklı çıkabilir. Bu, örneklemelerin popülasyonu ne kadar iyi temsil ettiğini gösterir.

Örneklem Hatası ve Standart Hata (Standard Error)

- **Standart Hata:** Örneklem ortalamasının popülasyon ortalamasına ne kadar yakın olduğunu gösterir. Standart hata, örneklem hatasının bir ölçüsüdür.
- **Formül:**

$$\text{Standart Hata} = \frac{\sigma}{\sqrt{n}}$$

- σ : Popülasyon standart sapması
- n : Örneklem büyütüğü
- **Örneklem Hatası:** Örneklemelerin birbirine göre standart sapmaları farklı ise, örneklem hatası yüksek olur. Bu, tahminlerin daha az güvenilir olduğunu gösterir.

Örnek: Weibull Dağılımı ve Örneklem Hatası

- **Popülasyon:** Bir ürünün عمر süresi Weibull dağılımına uygun olarak dağılmıştır.

- **Örneklemeler:** Popülasyondan 6 farklı örneklem çekilmiştir. Her bir örneklemin ortalama ve standart sapması farklıdır.
- **Weibull Dağılımı Grafiği:** Grafikte 6 farklı sekme bulunmaktadır. Her bir sekme, farklı bir örneklem dağılımını temsil eder.
- **Ortalamaların Farklılığı:** Her bir sekmedeki örneklemelerin ortalamaları farklı çıkmıştır. Bu, örneklemelerin popülasyonu ne kadar iyi temsil ettiğini gösterir.
- **Standart Hata Hesaplama:**

- **Örneklem 1:** Ortalama = 100, Standart Sapma = 10, Örneklem Büyüklüğü = 30

$$\text{Standart Hata} = \frac{10}{\sqrt{30}} \approx 1.83$$

- **Örneklem 2:** Ortalama = 105, Standart Sapma = 15, Örneklem Büyüklüğü = 30

$$\text{Standart Hata} = \frac{15}{\sqrt{30}} \approx 2.74$$

- **Örneklem Hatası:** Örneklem 1'in standart hatası daha küçük olduğu için, bu örneklemin tahminleri daha doğrudur. Örneklem 2'nin standart hatası daha büyük olduğu için, tahminleri daha az güvenilirdir.

Sonuç

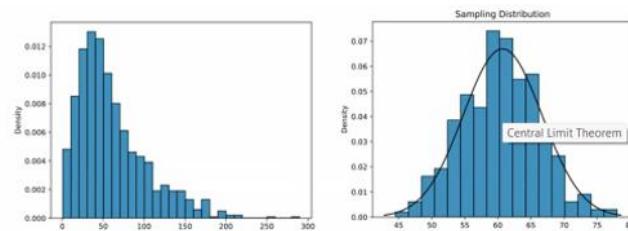
- **Weibull Dağılımı:** Farklı parametrelerle oluşturulmuş örneklemelerin dağılımını gösterir.
- **Örneklem Hatası:** Örneklemelerin birbirine göre tutarlığını ve tahminlerin doğruluğunu ölçer.
- **Standart Hata:** Örneklem ortalamasının popülasyon ortalamasına ne kadar yakın olduğunu gösterir. Standart hata küçükse, tahminler daha doğrudur.

2. Merkezi Limit Teoremi (Central Limit Theorem - CLT)

- **Tanım:** Büyük örneklemelerde, örneklem ortalamalarının dağılımı normal dağılıma yaklaşır.
 - **Özellikler:** Örneklem büyülüğu (n) arttıkça, örneklem ortalamalarının dağılımı normal dağılıma yaklaşır.
 - Popülasyon dağılımı normal olmasa bile, örneklem büyülüğu yeterince büyükse (genellikle $n > 30$), CLT geçerlidir.
- **Örnek:** Bir popülasyondan 50 örneklem çekilirse, bu örneklemelerin ortalamalarının dağılımı normal dağılıma yaklaşır.

What is the Central Limit Theorem?

The Central Limit Theorem (CLT) states that, given a sufficiently large sample size from a population with a finite level of variance, the sampling distribution of the mean will be normally distributed regardless of if the population is normally distributed.



Merkezi Limit Teoremi nedir?

Merkezi Limit Teoremi (CLT), sonlu bir varyans düzeyine sahip bir popülasyondan yeterince büyük bir örneklem büyüklüğü verildiğinde, ortalamanın örnekleme dağılımının, popülasyonun normal dağılım gösterip göstermediğine bakılmaksızın normal dağılım göstereceğini belirtir.

Örnekleme Dağılımı: (Grafik açıklaması)

- Görselde, sol tarafta orijinal popülasyon dağılımı (Somon Ağırlığı) gösterilmektedir. Bu dağılım, normal dağılıma benzememektedir.
- Sağ tarafta ise, örneklem ortalamalarının dağılımı (Örnekleme Dağılımı) gösterilmektedir. Görüldüğü gibi, örneklem ortalamalarının dağılımı, orijinal popülasyon dağılımından bağımsız olarak, normal dağılıma yakınsamaktadır.

What is the Central Limit Theorem?

The Central Limit Theorem (CLT) states that, given a sufficiently large sample size from a population with a finite level of variance, the sampling distribution of the mean will be normally distributed regardless of if the population is normally distributed.

Merkezi Limit Teoremi nedir?

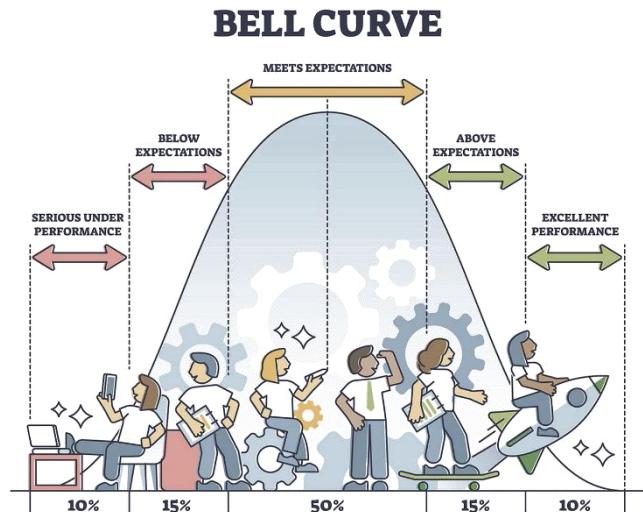
Merkezi limit teoremi, bir popülasyon ortalaması (μ) ve standart sapması (σ) olduğunda ve popülasyondan yerine koyma ile büyük rastgele örnekler aldiğinizda, örneklem ortalamalarının dağılımının, popülasyonun normal veya çarpık olup olmadığına bakılmaksızın yaklaşık olarak normal dağılım göstereceğini belirtir. Örneklem büyüklüğü yeterince büyük ($n > 30$) olduğunda geçerlidir.

Normal Dağılım Avantajları

- Analiz ve Yorumlama Kolaylığı:** Normal dağılım, istatistiksel analiz ve yorumlama için basitleştirilmiş bir çerçeve sunar.
- Parametrik Testlerin Kullanımı:** Normal dağılıma uygun verilerde, güçlü ve yaygın olarak kullanılan parametrik testler uygulanabilir.
- Merkezi Limit Teoremi:** Büyük örneklem boyutlarında, birçok dağılım normal dağılıma yakınsar, bu da istatistiksel çıkarımı kolaylaştırır.
- Hata Terimlerinin Dağılımı:** Regresyon analizinde, hata terimlerinin normal dağılması, modelin geçerliliği için önemlidir.
- Anormalliklerin ve Outlier Tespit:** Normal dağılımdan sapmalar, anormallikleri veya aykırı değerleri (outliers) tespit etmeyi kolaylaştırır.
- Öngörü ve Tahmin:** Normal dağılım, olasılık temelli öngörü ve tahmin modellerinin oluşturulmasını destekler.

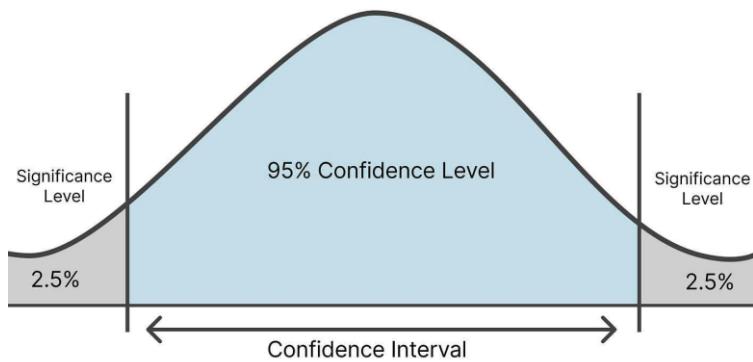
Çan Eğrisi (Bell Curve)

- Çan Eğrisi:** Normal dağılımın grafiksel temsilidir.
- Merkez:** Verilerin çoğu ortalama (mean), medyan (median) ve mod (mode) etrafında toplanır.
- Simetri:** Eğri, ortalama etrafında simetrikir.
- Dağılım:**
 - %10: Beklentilerin altında (Below Expectations)
 - %15: Ciddi şekilde bekлentilerin altında (Serious Under Expectations)
 - %50: Beklentileri karşılıyor (Meets Expectations)
 - %15: Beklentilerin üzerinde (Above Expectations)
 - %10: Mükemmel performans (Excellent Performance)



3. Güven Aralığı (Confidence Interval - CI)

- Tanım:** Popülasyon parametresinin belirli bir güven düzeyinde (örneğin, %95) bulunma olasılığını gösteren bir aralıktır.
- Özellikler:**
 - Güven aralığı, tahminin ne kadar güvenilir olduğunu gösterir.
 - Güven düzeyi (örneğin, %95), popülasyon parametresinin bu aralıkta olma olasılığını ifade eder.
- Örnek:** Bir anket sonucunda %50 oranında bir değer elde ederseniz ve hata payınız $\pm 3\%$ ise, gerçek değerin %47 ile %53 arasında olması muhtemeldir.



What is the difference between Point Estimates and Confidence Interval?

Point Estimation:

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

Confidence Interval: A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by $1 - \alpha$, where α is the level of significance.

Nokta Tahminleri ve Güven Aralığı arasındaki fark nedir?

Nokta Tahmini: Nokta Tahmini, bize bir popülasyon parametresinin tahmini olarak belirli bir değer verir. Momentler Yöntemi ve Maksimum Olabilirlik tahmin edici yöntemleri, popülasyon parametreleri için Nokta Tahminleri türetmek için kullanılır.

Güven Aralığı: Güven aralığı, popülasyon parametresini içermeye olasılığı olan bir değer aralığı verir. Güven aralığı genellikle tercih edilir, çünkü bu aralığın popülasyon parametresini içermeye olasılığının ne kadar olduğunu bize söyler. Bu olasılık veya olasılık, Güven Düzeyi veya Güven katsayıları olarak adlandırılır ve $1 - \alpha$ ile temsil edilir, burada α önem düzeyidir.

4. Hata Payı (Margin of Error - MOE)

- **Tanım:** Bir örneklemden elde edilen bir istatistiksel tahminin ne kadar doğru olduğunu gösterir.
- **Özellikler:**
 - Hata payı, güven aralığının genişliğini belirler.
 - Hata payı küçükse, tahmin daha doğrudur.
- **Örnek:** Bir anket sonucunda %50 oranında bir değer elde ederseniz ve hata payınız $\pm 3\%$ ise, gerçek değerin %47 ile %53 arasında olması muhtemeldir.

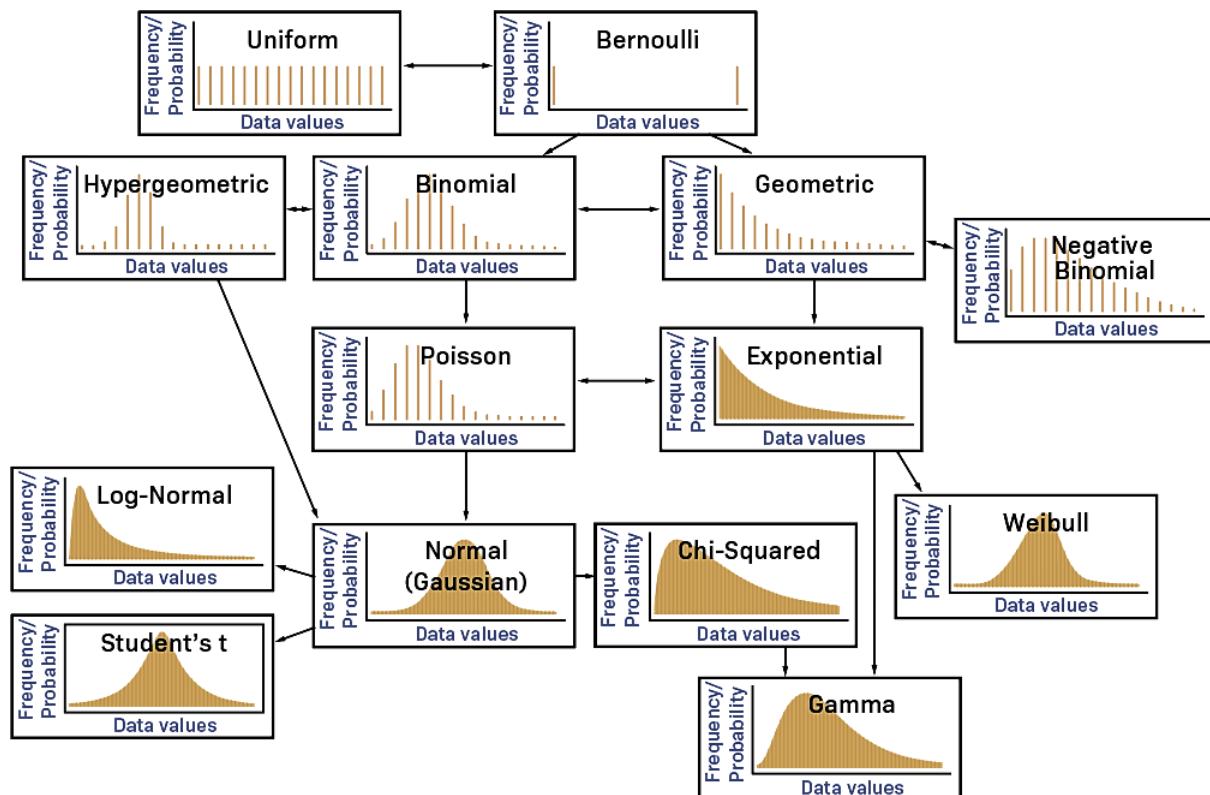
Nokta Tahminleri (Point Estimates)

- Tanım:** Bir popülasyon parametresinin (örneğin, ortalama, varyans) tek bir değer olarak tahmin edilmesidir.
- Özellikler:**
 - Popülasyon parametresinin en iyi tahmini olarak kabul edilir.
 - Tek bir değer verir, bu nedenle belirsizlik içermez.
- Örnek:** Örneklem ortalaması (\bar{x}), popülasyon ortalamasının (μ) nokta tahminidir.

Özet Tablo

Kavram	Açıklama
Örneklem Dağılımları (Sample Distributions)	Örneklemelerin istatistiklerinin dağılımı. Örnek: Örneklem ortalamalarının dağılımı.
Merkezi Limit Teoremi (Central Limit Theorem)	Büyük örneklemelerde, örneklem ortalamalarının dağılımı normal dağılıma yaklaşır.
Güven Aralığı (Confidence Interval)	Popülasyon parametresinin belirli bir güven düzeyinde bulunma olasılığını gösteren aralık. Örnek: %95 güven aralığı.
Hata Payı (Margin of Error)	Bir örneklemden elde edilen bir istatistiksel tahminin ne kadar doğru olduğunu gösterir. Örnek: $\pm 3\%$ hata payı.
Nokta Tahminleri (Point Estimates)	Popülasyon parametresinin tek bir değer olarak tahmin edilmesi. Örnek: Örneklem ortalaması.

Dağılım türüne göre grafik tipleri:



What is Cluster Sampling? Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Küme Örnekleme nedir? Küme örneklemesi, geniş bir alana yayılmış hedef popülasyonu incelemek zorlulığında ve basit rastgele örnekleme uygulanamadığında kullanılan bir tekniktir. Küme Örneği, her örnekleme biriminin bir eleman koleksiyonu veya kümesi olduğu bir olasılık örneğidir.

Örneğin, Bir araştırmacı Japonya'daki lise öğrencilerinin akademik performansını araştırmak istiyor. Japonya'nın tüm nüfusunu farklı kümelere (şehirlere) ayırbilir. Ardından araştırmacı, basit veya sistematik rastgele örnekleme yoluyla araştırmasına bağlı olarak bir dizi küme seçer.

What is sampling? How many sampling methods do you know?

"Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined."

Örnekleme nedir? Kaç örnekleme yöntemi biliyorsunuz?

"Veri örneklemesi, incelenen daha büyük veri setindeki desenleri ve eğilimleri belirlemek için veri noktalarının temsili bir alt kümesini seçmek, manipüle etmek ve analiz etmek için kullanılan bir istatistiksel analiz tekniğidir."

8. Bölüm : Hipotez Testleri

Anlamlılık Testleri (Significance Test)

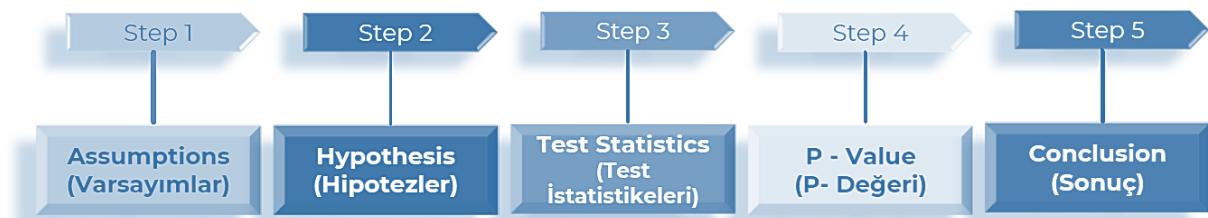
Anlamlılık testleri, bir örneklem üzerinden elde edilen sonuçların genel popülasyona genellenip genellenmeyeceğini belirlemek için kullanılan istatistiksel yöntemlerdir. Bu testler, gözlemlenen farkların gerçek bir etkiyi mi yoksa şansa bağlı bir varyasyonu mu temsil ettiğini değerlendirmemizi sağlar.

Temel Amaç:

- **Genelleme Yapmak:** Örneklemden elde edilen sonuçların popülasyon için geçerli olup olmadığını test etmek.
- **Farkın Anlamlılığını Değerlendirmek:** Gözlemlenen farkın gerçek bir fark mı yoksa ihmali edilebilir bir fark mı olduğunu belirlemek.

Örnek:

1. **"Kilo verdim ama bu kilo kaybı istatistik olarak anlamlı mı?"**
 - Bu örnek, bireysel bir deneyimin (kilo verme) istatistiksel olarak anlamlı olup olmadığını sorgulamaktadır.
 - Anlamlılık testi, kilo kaybının tesadüfi bir dalgalanma mı yoksa gerçek bir etki mi olduğunu belirlemeye yardımcı olur.
2. **"Farklı diyetler uyguladım, her birinden ayrı ayrı kilolar verdim.. Peki bunlar istatistiksel olarak anlamlı mı?"**
 - Bu örnek, birden fazla deneyimin (farklı diyetlerle kilo verme) istatistiksel olarak anlamlı olup olmadığını sorgulamaktadır.
 - Anlamlılık testi, her bir diyetin kilo verme üzerindeki etkisinin tesadüfi olup olmadığını veya diyetler arasında anlamlı bir fark olup olmadığını belirlemeye yardımcı olur.
3. **"3 kişi A diyetisyenine gitti, kilo verdi" ve "3 kişi B diyetisyenine gitti, kilo verdi. Bu 2 diyetisyen ile kilo verenlerin birbirlerine göre farkları istatistiksel olarak anlamlı mı?"**
 - Bu örnek, iki farklı grubun (A ve B diyetisyenlerine gidenler) kilo verme sonuçlarını karşılaştırmaktadır.
 - Anlamlılık testi, iki grup arasında anlamlı bir fark olup olmadığını belirlemeye yardımcı olur.



Hipotez Testi (Hypothesis Test)

- **Amaç:** Bir örneklem(sample) üzerinden elde edilen sonuçların popülasyona genellenip genellenemeyeceğini test etmek.
- **Temel Adımlar:**

1. Assumptions (Varsayımlar)

- Bu adımda, testin geçerliliği için gerekli olan varsayımlar kontrol edilir.
- Örneğin, verilerin normal dağılıp dağılmadığı, örneklerin bağımsız olup olmadığı gibi varsayımlar değerlendirilir.
- Bu varsayımların ihlal edilmesi, test sonuçlarının güvenilirliğini etkileyebilir.

2. Hipotezleri Formüle Et:

- **Null Hipotezi (H_0):** "Değişim yok" veya "mevcut durum geçerli" varsayımları.
Örnek: "Denizdeki kurşun oranı 10 ppm'dir."
- **Alternatif Hipotez (H_1):** Araştırmacının iddiasını temsil eder.
Örnek: "Denizdeki kurşun oranı 10 ppm'den fazladır."

3. **Test Seçimi:** Veri tipine (sayısal/kategorik) ve varsayımlara göre uygun test seçilir (z-testi, t-testi, ki-kare vb.). Örneklem verileri kullanılarak istatistiksel bir değer hesaplanır.

4. **P-Değeri Belirle:** Null hipotezin doğru olduğu varsayıldığında, gözlemlenen sonuçların veya daha uç değerlerin olasılığı hesaplanır.

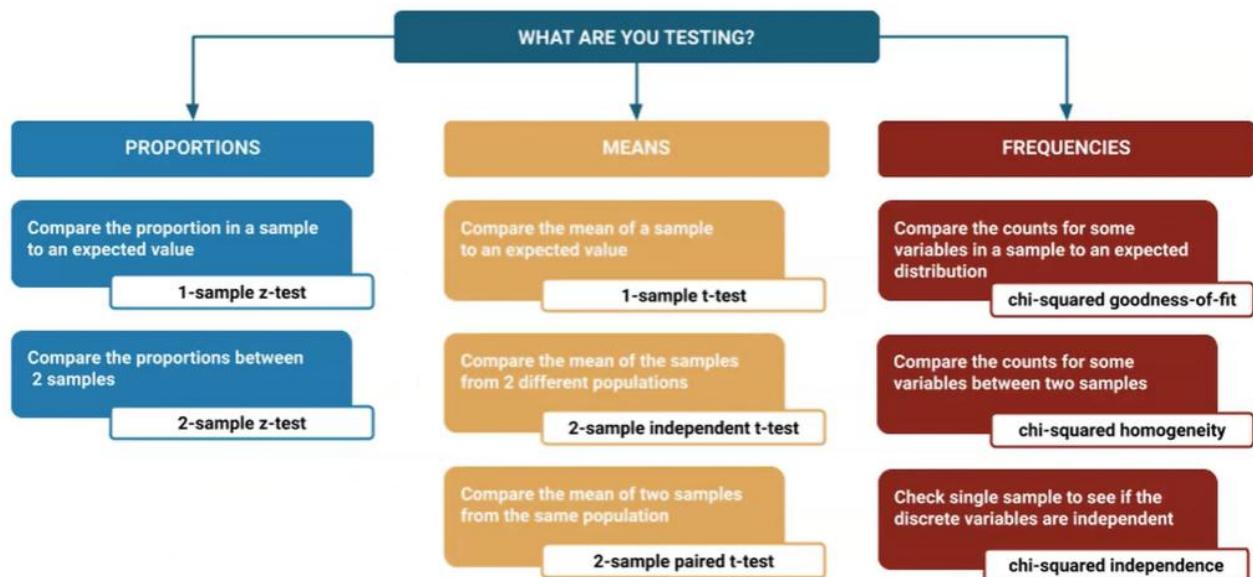
- Düşük p-değeri, null hipotezin yanlış olma olasılığının yüksek olduğunu gösterir.

5. **Sonuç:** P-değeri ile α (anlamlılık düzeyi) karşılaştırılır.

- $P < \alpha \rightarrow H_0$ reddedilir. (**Reject**) Bu, alternatif hipotezin desteklendiği anlamına gelir.
- $P \geq \alpha \rightarrow H_0$ reddedilemez. (**Fail to Reject**) Bu, verilerin null hipotezi çürütecek kadar güçlü kanıt sağlamadığı anlamına gelir.

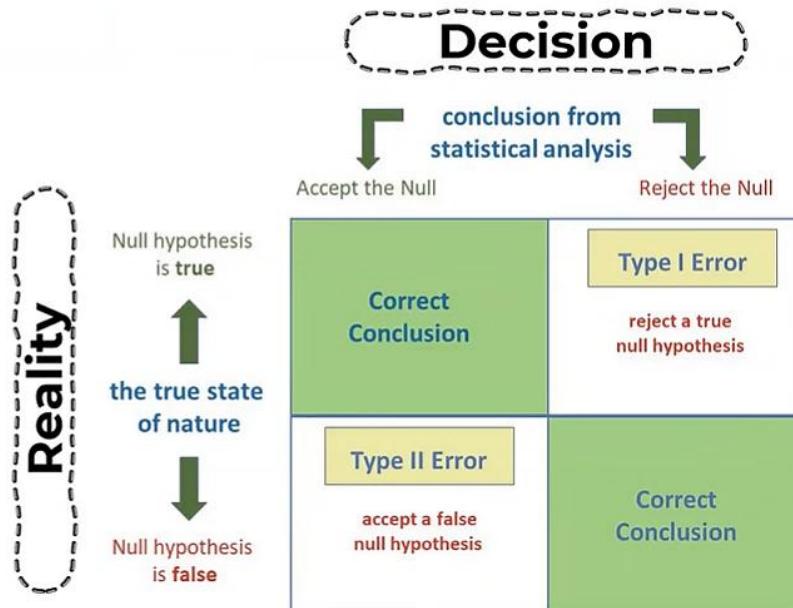
Simple Hypothesis Testing

Choosing a simple test for comparing differences in populations



Tip I – II Hatalar (Type I – II Error)

- **Tip I Hata (α):** H_0 doğruyken yanlışlıkla reddedilmesi.
 - Örnek: "İlaç etkisizken, etkili olduğu sonucuna varmak."
- **Tip II Hata (β):** H_0 yanlışken kabul edilmesi.
 - Örnek: "İlaç etkiliyken, etkisiz olduğu sonucuna varmak."



Tek – İki Kuyruk Testleri (One – Two Tail Tests)

- **Tek Kuyruk Testi:** Alternatif hipotez belirli bir yön belirtir (örneğin "büyüktür" veya "küçüktür").
 - Örnek: "Yeni ilaç, mevcut ilaçtan daha etkilidir."
- **İki Kuyruk Testi:** Alternatif hipotez "eşit değildir" şeklinde dir.
 - Örnek: "İki grup arasında fark vardır."
 - **P-değeri Hesaplama:** Çift kuyruk testlerinde p-değeri iki yönde de dikkate alınır.

Örnek Senaryo: Denizdeki Kurşun Seviyesi

Problem:

- Bir denizde olması gereken kurşun seviyesi 10 ppm dir (μ_0).
- Popülasyon normal dağılıma uygun olmakla beraber standart sapma $\sigma = 1.5$ 'dir.
- 40 farklı örnek alındı ve ortalama kurşun seviyesi 10.5 ppm (sample mean) bulundu.
- Bu ortalamadaki fark, $\alpha = 0.05$ (%95 güven) için, kurşun seviyesinin popülasyon ortalamasından anlamlı şekilde daha büyük olup olmadığını gösteriyor mu?

Çözüm:

Adım 1: Varsayımlar

- Normal Dağılım: Popülasyonun kurşun seviyeleri normal dağılım göstermektedir. Bu, z-testi kullanmamıza olanak tanır.
- Bağımsız Örnekler: 40 farklı örnek bağımsız olarak alınmıştır. Bu, örneklem ortalamasının güvenilirliğini artırır.
- Bilinen Standart Sapma: Popülasyon standart sapması ($\sigma = 1,5$) bilinmektedir. Bu, z-testi kullanmamıza olanak tanır.
- Büyük Örneklem ($n > 30$): örneklem 40 olduğundan büyük örneklem kabul edilir.

Adım 2: Hipotezler

- Sıfır Hipotezi (H_0): Denizdeki kurşun seviyesi, olması gereken 10 p/m seviyesindedir. ($\mu = 10$)
- Alternatif Hipotez (H_1): Denizdeki kurşun seviyesi, 10 p/m seviyesinden anlamlı şekilde daha yüksektir. ($\mu > 10$)

Adım 3: Test İstatistiği

- Popülasyon standart sapması bilindiği için z-testi kullanacağız.
- $z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$
- $z = (10,5 - 10) / (1,5 / \sqrt{40})$
- $z \approx 2,108$

Adım 4: p-değeri

- z-tablosundan veya bir istatistiksel hesaplayıcıdan, $z = 2,108$ için p-değerini buluruz.
- p-değeri $\approx 0,0175$

Adım 5: Sonuç

- p-değerini anlamlılık düzeyi ($\alpha = 0,05$) ile karşılaştırırız.
- $p\text{-değeri } (0,0175) < \alpha (0,05)$
- Bu nedenle, sıfır hipotezini reddederiz.

Adım 6. Sonuç:

Sonuç olarak, denizdeki kurşun seviyesi, %95 güvenle, olması gereken 10 p/m seviyesinden anlamlı şekilde daha yüksektir.

Ek Bilgiler:

- Bu bir tek yönlü (sağa kuyruklu) hipotez testidir, çünkü alternatif hipotez ortalamanın belirli bir değerden büyük olduğunu iddia etmektedir.
- p-değeri, sıfır hipotezi doğrulanen sonuçların veya daha uç sonuçların elde edilme olaslığını gösterir.
- Anlamlılık düzeyi (α), sıfır hipotezini yanlışlıkla reddetme olasılığını gösterir.

Anlamlılık Düzeyi (α) ve P-Degreri

- **α (Anlamlılık Düzeyi):** Genellikle **0,05** seçilir (%5 hata payı).
 - **Neden 0,05?** Geleneksel bir standarttır, ancak araştırma alanına göre değişebilir (örneğin tıpta 0,01 kullanılabilir).

- **P-Değeri:**
 - **Hakem Rolü:** Sonucun şansa bağlı olma ihtimalini gösterir.
 - **Örnek:** $P = 0.03$ ise, H_0 reddedilir ($\alpha = 0.05$ için).

Örnek Senaryo

Problem: Bir ilaç firması, yeni ürettiği ilaçın ateş düşürmede mevcut ilaçtan daha etkili olduğunu iddia ediyor.

- H_0 : "Yeni ilaç ile mevcut ilaç arasında fark yoktur."
- H_1 : "Yeni ilaç daha etkilidir." (Tek kuyruk testi)
- **Test:** t-testi (popülasyon varyansı bilinmiyor).
- **Sonuç:** $P = 0.02$ ve $\alpha = 0.05 \rightarrow H_0$ reddedilir.

. T Testi ve Z Testi

- **T Testi:**
 - Popülasyon standart sapması bilinmediğinde ve küçük örneklemelerde ($n < 30$) kullanılır.
 - Örnek: İki grubun ortalamalarını karşılaştırmak.
- **Z Testi:**
 - Popülasyon standart sapması bilindiğinde ve büyük örneklemelerde ($n \geq 30$) kullanılır.
 - Örnek: Bir örneklemenin popülasyon ortalaması ile karşılaştırılması.

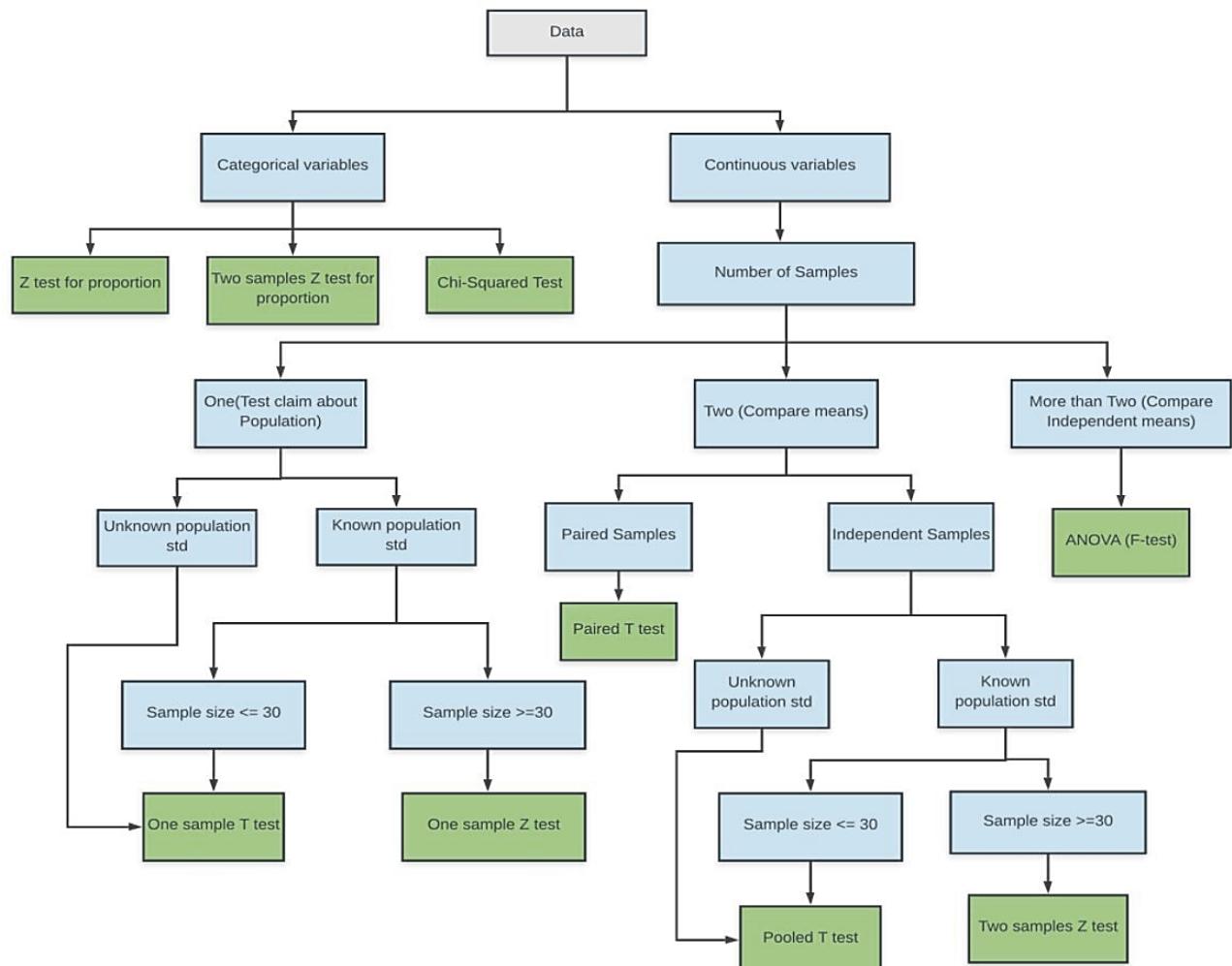
Özet Tablo

Kavram	Açıklama
H_0 (Null Hipotez)	Değişim olmadığını varsayan hipotez.
H_1 (Alternatif Hipotez)	Araştırmacının iddiasını temsil eden hipotez.
P-değeri	Null hipotezin doğru olduğu durumda gözlemlenen sonucun olasılığı.
α (Anlamlılık Düzeyi)	Kabul edilebilir hata payı (genellikle 0.05).
Tek/Çift Kuyruk Testi	Alternatif hipotezin yönüne göre belirlenir.
T Testi	Küçük örneklemelerde ve popülasyon standart sapması bilinmediğinde kullanılır.
Z Testi	Büyük örneklemelerde ve popülasyon standart sapması bilindiğinde kullanılır.

Sonuç

Anlamlılık testleri, bilimsel iddiaları istatistiksel kanıtlarla desteklemek için kritik bir araçtır. Doğru hipotez kurulumu, uygun test seçimi ve p-değerinin yorumlanması, sonuçların güvenilirliğini belirler. Bu yöntemler, verilerden anlamlı sonuçlar çıkarmak ve karar verme süreçlerine rehberlik etmek için kullanılır.

Hipotez Testlerinde İzleyeceğimiz Yol:



9. Bölüm : İleri Hipotez Testleri

Bağımsız Örneklemeler T Testi (Independent Samples T Test)

- Amaç:** İki bağımsız grup arasındaki ortalama farkın anlamlı olup olmadığını test etmek.
- Örnek:** Erkek ve kadın öğrencilerin matematik puanları arasında anlamlı bir fark var mı?
- Adımlar:**

1. Varsayımlar:

- Veriler normal dağılıma uymalı.
- Grupların varyansları benzer olmalı (varyans homojenliği).

2. Hipotezler:

- H_0 : İki grup arasında ortalama fark yoktur.
- H_1 : İki grup arasında ortalama fark vardır.

3. Test İstatistiği:

T istatistiği hesaplanır.

4. P-Değeri:

Hesaplanan T değerine karşılık gelen p-değeri bulunur.

5. Sonuç:

$P < \alpha$ ise H_0 reddedilir.

Bağımlı T Testi – Eşleştirilmiş T Testi (Paired T Test)

- **Amaç:** Aynı grubun iki farklı durumdaki ortalamaları arasındaki farkın anlamlı olup olmadığını test etmek.
- **Örnek:** Bir ilaçın hastalar üzerindeki etkisini ölçmek için tedavi öncesi ve sonrası değerler karşılaştırılır.
- **Adımlar:**
 1. **Varsayımlar:**
 - Farklar normal dağılıma uymalı.
 2. **Hipotezler:**
 - H_0 : Ortalama fark sıfırdır.
 - H_1 : Ortalama fark sıfırdan farklıdır.
 3. **Test İstatistiği:** T istatistiği hesaplanır.
 4. **P-Değeri:** Hesaplanan T değerine karşılık gelen p-değeri bulunur.
 5. **Sonuç:** $P < \alpha$ ise H_0 reddedilir.

Tek Yönlü ANOVA (One Way ANOVA)

- **Amaç:** Üç veya daha fazla grup arasındaki ortalama farkın anlamlı olup olmadığını test etmek.
- **Örnek:** Üç farklı ilaç türünün hastalar üzerindeki etkisi karşılaştırılır.
- **Adımlar:**
 1. **Varsayımlar:**
 - Gruplar normal dağılıma uymalı.
 - Grupların varyansları homojen olmalı (varyans homojenliği).
 2. **Hipotezler:**
 - H_0 : Tüm grupların ortalamaları eşittir.
 - H_1 : En az bir grup ortalaması diğerlerinden farklıdır.
 3. **Test İstatistiği:** F istatistiği hesaplanır.
 4. **P-Değeri:** Hesaplanan F değerine karşılık gelen p-değeri bulunur.
 5. **Sonuç:** $P < \alpha$ ise H_0 reddedilir.

ANOVA Tablosu ve Test İstatistikleri

- **Regresyon Kareler Toplamı (SSR):** Model tarafından açıklanan varyans.
- **Hata Kareler Toplamı (SSE):** Model tarafından açıklanamayan varyans.
- **Toplam Kareler Toplamı (SST):** Toplam varyans ($SST = SSR + SSE$).
- **Regresyon Ortalama Kareler (MSR):** $SSR / \text{serbestlik derecesi}$.
- **Hata Ortalama Kareler (MSE):** $SSE / \text{serbestlik derecesi}$.
- **F İstatistiği:** MSR / MSE .

Kategorik Veri Analizi

- **Amaç:** Kategorik değişkenler arasındaki ilişkiyi incelemek.
- **Örnek:** Sigara içme alışkanlığı ile akciğer kanseri arasında bir ilişki var mı?

Ki-Kare Testi (Chi-Square Test)

1. **Varsayımlar:**
 - Gözlenen frekanslar beklenen frekanslarla karşılaştırılır.
2. **Hipotezler:**
 - H_0 : İki değişken arasında ilişki yoktur.
 - H_1 : İki değişken arasında ilişki vardır.
3. **Test İstatistiği:** Ki-kare istatistiği hesaplanır.
4. **P-Değeri:** Hesaplanan Ki-kare değerine karşılık gelen p-değeri bulunur.
5. **Sonuç:** $P < \alpha$ ise H_0 reddedilir.

Örnek Senaryolar ve Çözümler

Örnek 1: Bağımsız T Testi

- **Problem:** Erkek ve kadın öğrencilerin matematik puanları arasında anlamlı bir fark var mı?
- **Adımlar:**
 1. **Varsayımlar:** Veriler normal dağılıyor ve varyanslar homojen.
 2. **Hipotezler:**
 - H_0 : Erkek ve kadın öğrencilerin puan ortalamaları eşittir.
 - H_1 : Ortalamalar eşit değildir.
 3. **Test İstatistiği:** $T = 2.45$ (hesaplanan değer).
 4. **P-Değeri:** $P = 0.015$.
 5. **Sonuç:** $P < 0.05 \rightarrow H_0$ reddedilir. Erkek ve kadın öğrencilerin puanları arasında anlamlı bir fark vardır.

Örnek 2: ANOVA

- **Problem:** Üç farklı ilaç türünün hastalar üzerindeki etkisi karşılaştırılıyor.
- **Adımlar:**
 1. **Varsayımlar:** Gruplar normal dağılıyor ve varyanslar homojen.
 2. **Hipotezler:**
 - H_0 : Üç ilaç türünün etkisi aynıdır.
 - H_1 : En az bir ilaç türünün etkisi farklıdır.
 3. **Test İstatistiği:** $F = 4.67$ (hesaplanan değer).
 4. **P-Değeri:** $P = 0.012$.
 5. **Sonuç:** $P < 0.05 \rightarrow H_0$ reddedilir. En az bir ilaç türünün etkisi diğerlerinden farklıdır.

Örnek 3: Ki-Kare Testi

- **Problem:** Sigara içme alışkanlığı ile akciğer kanseri arasında bir ilişki var mı?
- **Adımlar:**
 1. **Varsayımlar:** Gözlenen frekanslar beklenen frekanslarla karşılaştırılır.
 2. **Hipotezler:**
 - H_0 : Sigara içme ile akciğer kanseri arasında ilişki yoktur.
 - H_1 : İlişki vardır.
 3. **Test İstatistiği:** Ki-kare = 9.82 (hesaplanan değer).
 4. **P-Değeri:** $P = 0.002$.

5. **Sonuç:** $P < 0.05 \rightarrow H_0$ reddedilir. Sigara içme ile akciğer kanseri arasında anlamlı bir ilişki vardır.

Hangi Testi Kullanmalıyım?

Veri Tipi	Grup Sayısı	Test
Sürekli (Sayısal)	2	Bağımsız T Testi
Sürekli (Sayısal)	2 (Eşleşmiş)	Bağımlı T Testi
Sürekli (Sayısal)	≥ 3	ANOVA
Kategorik	Herhangi	Ki-Kare Testi

Sonuç

Bu testler, veriler arasındaki ilişkileri analiz etmek ve anlamlı sonuçlar çıkarmak için kullanılır. Doğru test seçimi ve uygun hipotez kurulumu, sonuçların güvenilirliğini belirler.

What do you understand by Hypothesis Testing?

In Statistics, Hypothesis Testing is mainly used to see if a certain experiment generates meaningful results. It helps assess the statistical significance of insight by finding the odds of the results occurring by chance. In Hypothesis Testing, the first thing is to know the null hypothesis and then specify it. After that, the p-value is calculated, and if the null hypothesis is true, the other values are also determined. The alpha value specifies the significance, and you can adjust it accordingly. If the p-value is less than the alpha value, the null hypothesis is rejected, but the null hypothesis is accepted if the p-value is greater than the alpha value. If the null hypothesis is rejected, it indicates that the results obtained are statistically significant.

Hipotez Testini Nasıl Anlarsınız?

İstatistikte, Hipotez Testi temel olarak bir deneyin anlamlı sonuçlar üretip üretmediğini görmek için kullanılır. Sonuçların şans eseri ortaya çıkma olasılığını bularak içgörünün istatistiksel anlamlılığını değerlendirmeye yardımcı olur. Hipotez Testinde ilk adım, sıfır hipotezini (null hypothesis) belirlemektir. Ardından p-değeri hesaplanır ve sıfır hipotezinin doğru olduğu varsayılarak diğer değerler belirlenir. Alfa değeri anlamlılık düzeyini belirtir ve buna göre ayarlanabilir. Eğer p-değeri alfa değerinden küçükse, sıfır hipotezi reddedilir; ancak p-değeri alfa değerinden büyükse sıfır hipotezi kabul edilir. Sıfır hipotezinin reddedilmesi, elde edilen sonuçların istatistiksel olarak anlamlı olduğunu gösterir.

What is the relationship between the significance level and the confidence level in Statistics?

In Statistics, the significance level is the probability of getting a completely different result from the condition where the null hypothesis is true. On the other hand, the confidence level is used as a range of similar values in a population.

We can specify the similarity between the significance level and the confidence level by the following formula:

$$\text{Significance level} = 1 - \text{Confidence level}$$

İstatistikte Anlamlılık Düzeyi ile Güven Düzeyi Arasındaki İlişki Nedir?

İstatistikte, anlamlılık düzeyi (significance level), sıfır hipotezinin doğru olduğu durumdan tamamen farklı bir sonuç elde etme olasılığıdır. Öte yandan, güven düzeyi (confidence level), bir popülasyondaki benzer değerlerin aralığı olarak kullanılır.

Anlamlılık düzeyi ile güven düzeyi arasındaki ilişki şu formülle açıklanabilir:

$$\text{Anlamlılık Düzeyi} = 1 - \text{Güven Düzeyi}$$

What is p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.

Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis. p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,

High P values: your data are likely with a true null.

Low P values: your data are unlikely with a true null.

p-değeri nedir?

İstatistikte bir hipotez testi yaptığınızda, p-değeri sonuçlarınızın gücünü belirlemenize yardımcı olabilir. p-değeri, 0 ile 1 arasında bir sayıdır. Değerine göre sonuçların gücünü gösterir. Test edilen iddiaya Sıfır Hipotezi (Null Hypothesis) denir.

Düşük p-değeri (≤ 0.05), sıfır hipotezine karşı güçlü bir kanıt olduğunu gösterir, bu da sıfır hipotezini reddedebileceğimiz anlamına gelir. Yüksek p-değeri (≥ 0.05), sıfır hipotezini desteklediğini gösterir, bu da sıfır hipotezini kabul edebileceğimiz anlamına gelir. 0.05 p-değeri, hipotezin iki yöne de gidebileceğini gösterir. Başka bir deyişle,

Yüksek P değerleri: verileriniz sıfır hipotezinin doğru olduğu bir durumla uyumludur.

Düşük P değerleri: verileriniz sıfır hipotezinin doğru olduğu bir durumla uyumsuzdur.

How do you assess the statistical significance of an insight?

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

Bir İçgörünün İstatistiksel Anlamılılığını Nasıl Değerlendirirsiniz?

İstatistiksel anlamılılığı belirlemek için hipotez testi yaparsınız. İlk olarak, sıfır hipotezini (null hypothesis) ve alternatif hipotezi (alternative hypothesis) belirlersiniz. İkinci adımda, sıfır hipotezinin doğru olduğu varsayılarak test sonuçlarının gözlemlenme olasılığı olan p-değerini hesaplarsınız. Son olarak, anlamlılık düzeyini (alfa) belirler ve eğer p-değeri alfadan küçükse, sıfır hipotezini reddedersiniz. Başka bir deyişle, sonuç istatistiksel olarak anlamlıdır.

What is the significance of p-value?

- **p-value typically ≤ 0.05**

This indicates that there is strong evidence against the null hypothesis, so you reject the null hypothesis.

- **p-value typically > 0.05**

This indicates that there is weak evidence against the null hypothesis, so you accept the null hypothesis.

p-değerinin anlamı nedir?

- **p-value typically ≤ 0.05**

Bu, sıfır hipotezine karşı güçlü bir kanıt olduğunu gösterir; bu nedenle sıfır hipotezini reddedersiniz.

- **p-value typically > 0.05**

Bu, sıfır hipotezine karşı zayıf bir kanıt olduğunu gösterir, bu nedenle sıfır hipotezini kabul edersiniz.

What is the Null and Alternate Hypothesis?

A null and alternate hypothesis is used in statistical hypothesis testing.

Null Hypothesis (Sıfır Hipotezi):

- It states that the population parameter is equal to the assumed value.
- It is an initial claim based on previous analysis or experience.

Alternate Hypothesis (Alternatif Hipotez):

- It states that population parameters are equal or different to the assumed value.
- It is what you might believe to be true or want to prove true.

Sıfır ve Alternatif Hipotez Nedir?

Sıfır ve alternatif hipotez, istatistiksel hipotez testinde kullanılır.

Sıfır Hipotezi (Null Hypothesis):

- Popülasyon parametresinin varsayılan değere eşit olduğunu belirtir.
- Önceki analizlere veya deneyimlere dayanan bir başlangıç iddiasıdır.

Alternatif Hipotez (Alternate Hypothesis):

- Popülasyon parametrelerinin varsayılan değere eşit veya farklı olduğunu belirtir.
- Doğru olduğuna inandığınız veya doğru olduğunu kanıtlamak istediğiniz şeydir.

What is Hypothesis Testing?

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

There are 3 steps in Hypothesis Testing:

1. State Null and Alternate Hypothesis
2. Perform Statistical Test
3. Accept or reject the Null Hypothesis

Hipotez Testi Nedir?

Hipotez testi, bir örneklemden alınan verileri kullanarak bir popülasyon parametresi veya popülasyon olasılık dağılımı hakkında sonuçlar çıkarmak için kullanılan bir istatistiksel çıkarım yöntemidir.

Hipotez testinde 3 adım vardır:

1. Sıfır ve Alternatif Hipotezi Belirleme
2. İstatistiksel Testi Uygulama
3. Sıfır Hipotezini Kabul veya Reddetme

What do you understand by Hypothesis Testing?

In Statistics, Hypothesis Testing is mainly used to see if a certain experiment generates meaningful results. It helps assess the statistical significance of insight by finding the odds of the results occurring by chance. In Hypothesis Testing, the first thing is to know the null hypothesis and then specify it. After that, the p-value is calculated, and if the null hypothesis is true, the other values are also determined. The alpha value specifies the significance, and you can adjust it accordingly. If the p-value is less than the alpha value, the null hypothesis is rejected, but the null hypothesis is accepted if the p-value is greater than the alpha value. If the null hypothesis is rejected, it indicates that the results obtained are statistically significant.

Hipotez Testini Nasıl Anlarsınız?

İstatistikte, Hipotez Testi temel olarak bir deneyin anlamlı sonuçlar üretip üretmediğini görmek için kullanılır. Sonuçların şans eseri ortaya çıkma olasılığını bularak içgörünün istatistiksel anlamlılığını değerlendirmeye yardımcı olur. Hipotez Testinde ilk adım, sıfır hipotezini (null hypothesis) belirlemektir. Ardından p-değeri hesaplanır ve sıfır hipotezinin doğru olduğu varsayılarak diğer değerler belirlenir. Alfa değeri anlamlılık düzeyini belirtir ve buna göre ayarlanabilir. Eğer p-değeri alfa değerinden küçükse, sıfır hipotezi reddedilir; ancak p-değeri alfa değerinden büyükse sıfır hipotezi kabul edilir. Sıfır hipotezinin reddedilmesi, elde edilen sonuçların istatistiksel olarak anlamlı olduğunu gösterir.

What are a p-value and its role in Hypothesis Testing?

P-value is the probability that a random chance generated the data or something else that is equal or rare.

P-values are used in hypothesis testing to decide whether to reject the null hypothesis or not.

- **p-value < alpha – value**

Means results are not in favor of the null hypothesis, reject the null hypothesis.

p-değeri nedir ve Hipotez Testindeki rolü nedir?

p-değeri, verilerin rastgele bir şans eseri veya eşit ya da daha nadir bir şey tarafından üretilme olasılığıdır.

p-değerleri, hipotez testinde sıfır hipotezini reddedip reddetmeyeceğimize karar vermek için kullanılır.

- **p-değeri < alfa değeri**

Sonuçların sıfır hipotezini desteklemediği anlamına gelir, bu nedenle sıfır hipotezi reddedilir.

✖ These results are based on 300+ statistics interview questions from 50+ companies.

Top Statistics Concepts in Data Science Interviews

