Olasılık ve İstatistik - 1. Hafta
Giriş ve Temel Kavramlar

Prof. Dr. Rüya ŞAMLI


1
2
?
Y


Σ %

Olasılık ve İstatistik

- Olasılık: Belirsizlik durumlarının matematiksel ifadesi
- İstatistik: Verilerin toplanması, düzenlenmesi, analiz edilmesi

- Olasılık, istatistiğin matematiksel temelidir

Anakütle, Örneklem, Parametre, İstatistik

- Anakütle/Popülasyon (Population): İncelenen tüm bireyler
- Örneklem (Sample): Anakütleden seçilen alt grup
- Parametre: Anakütleye ait özellik (μ)
- İstatistik: Örneklemden hesaplanan değer ($\bar{x}$)

Anakütle ve Örneklem

Anakütle, Örneklem, Parametre, İstatistik

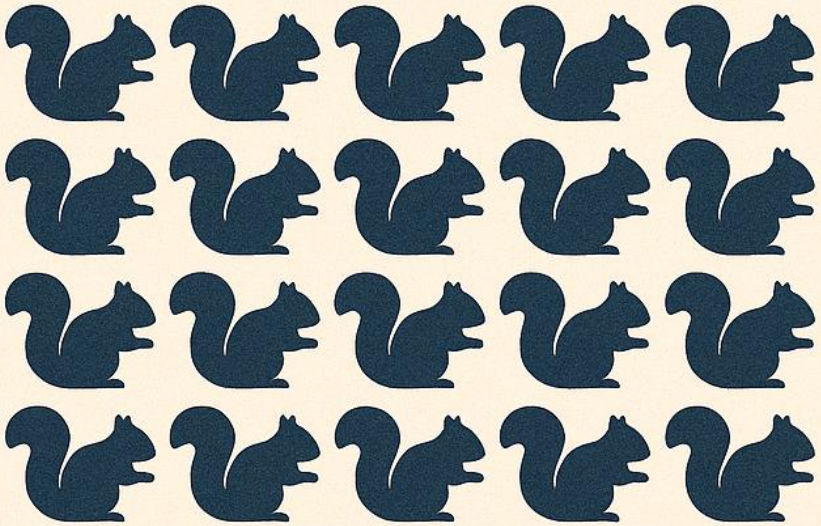
- **Anakütle/Popülasyon (Population)**
- Bir ormandaki **tüm sincaplar**.

- **Örneklem (Sample)**
- Bu ormandan rastgele seçilmiş **50 sincap**.
- Tüm sincapları incelemek çok zor ve zaman alıcı olacağından, içlerinden belirli bir grup seçip çalışmayı bu grup üzerinden yaparız.

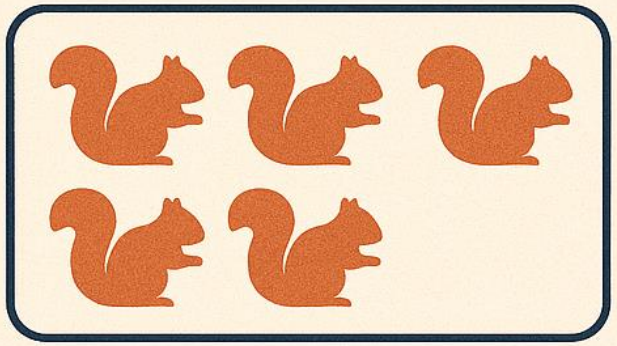
- **Parametre (Parameter)**
- Ormandaki **tüm sincapların ortalama kilosu**.
- Bu değer sabittir (ama genellikle bilinmez). Biz örneklem üzerinden tahmin etmeye çalışırız.

Anakütle, Örneklem, Parametre, İstatistik

- **Popülasyon:** Ormandaki tüm sincaplar
- **Örneklem:** Rastgele seçilen 50 sincap
- **Parametre:** Tüm sincapların gerçek ortalama kilosu



POPÜLASYON



ÖRNEKLEM

PARAMETRE

Tüm sincapların
ortalama kilosu

Veri Türleri

- Nitel (Qualitative): Kategorik veriler
- Nicel (Quantitative): Sayısal veriler
- - Kesikli (Discrete)
- - Sürekli (Continuous)

Veri Türleri

- **Nitel (Qualitative): Kategorik veriler**

- Kuşların tüy renkleri (kırmızı, mavi, yeşil…)
- Balıkların yaşadıkları su türü (tatlı su, tuzlu su)
- Köpeklerin ırkları (Golden Retriever, Kangal, Bulldog…)
- Kaplanların yaşam alanı (orman, savan, dağlık bölge)
- Arıların görevleri (işçi, kraliçe, erkek arı)

Veri Türleri

Nitel (Qualitative): Kategorik veriler

Veri Türleri

- **Nicel (Quantitative): Sayısal veriler**

- Kuşların kanat uzunluğu (cm)
- Balıkların ağırlığı (kg)
- Köpeklerin yaşı (yıl)
- Kaplanların bir günde kat ettiği mesafe (km)
- Arıların kovandan topladığı nektar miktarı (gram)

Veri Türleri

- **Kesikli (Discrete) Veriler**
- Sayılabilen değerlerdir.
- Kesikli veriler tam sayılarla ifade edilir ve sayılar arasında ara değerler yoktur.

- Bir kümesteki tavuk sayısı (12 tavuk, 13 tavuk…)
- Bir balıkçı ağındaki balık sayısı
- Bir köpeğin doğurduğu yavru sayısı
- Bir çiftlikteki inek sayısı
- Bir arı kovanındaki arı sayısı

Veri Türleri

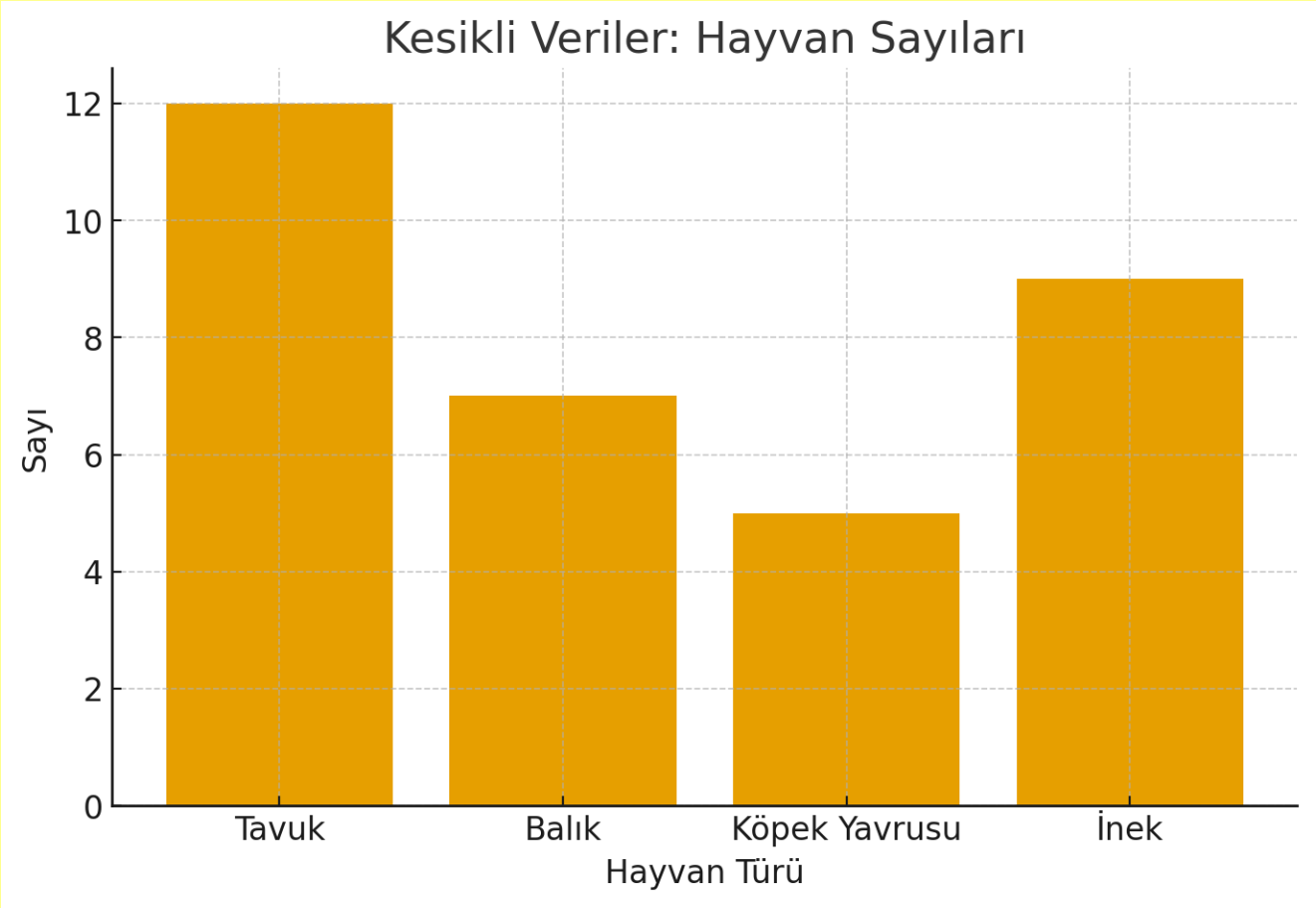
Nicel (Quantitative): Sayısal veriler

Veri Türleri

Kesikli (Discrete) Veriler

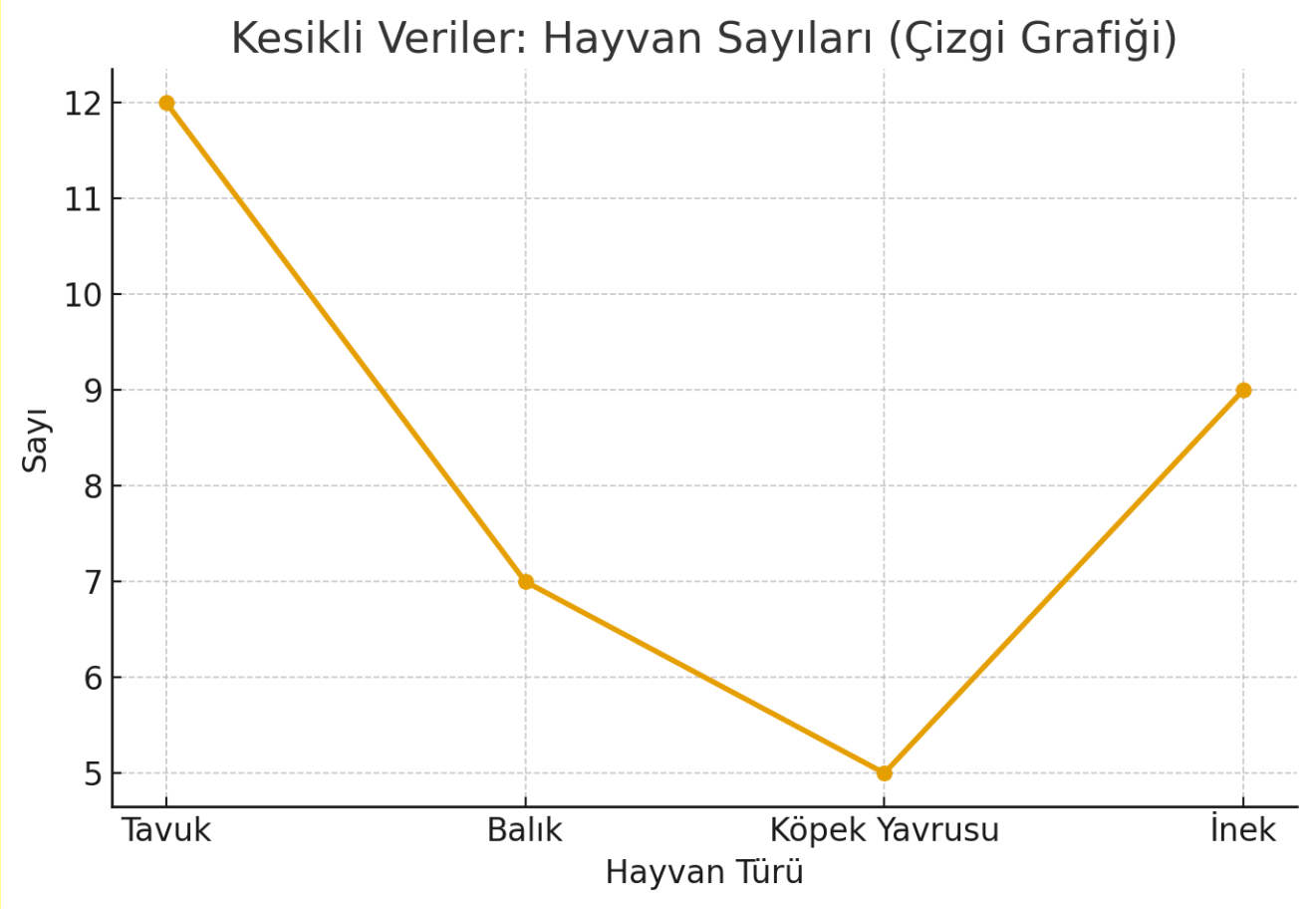
Veri Türleri

- **Kesikli (Discrete) Veriler**



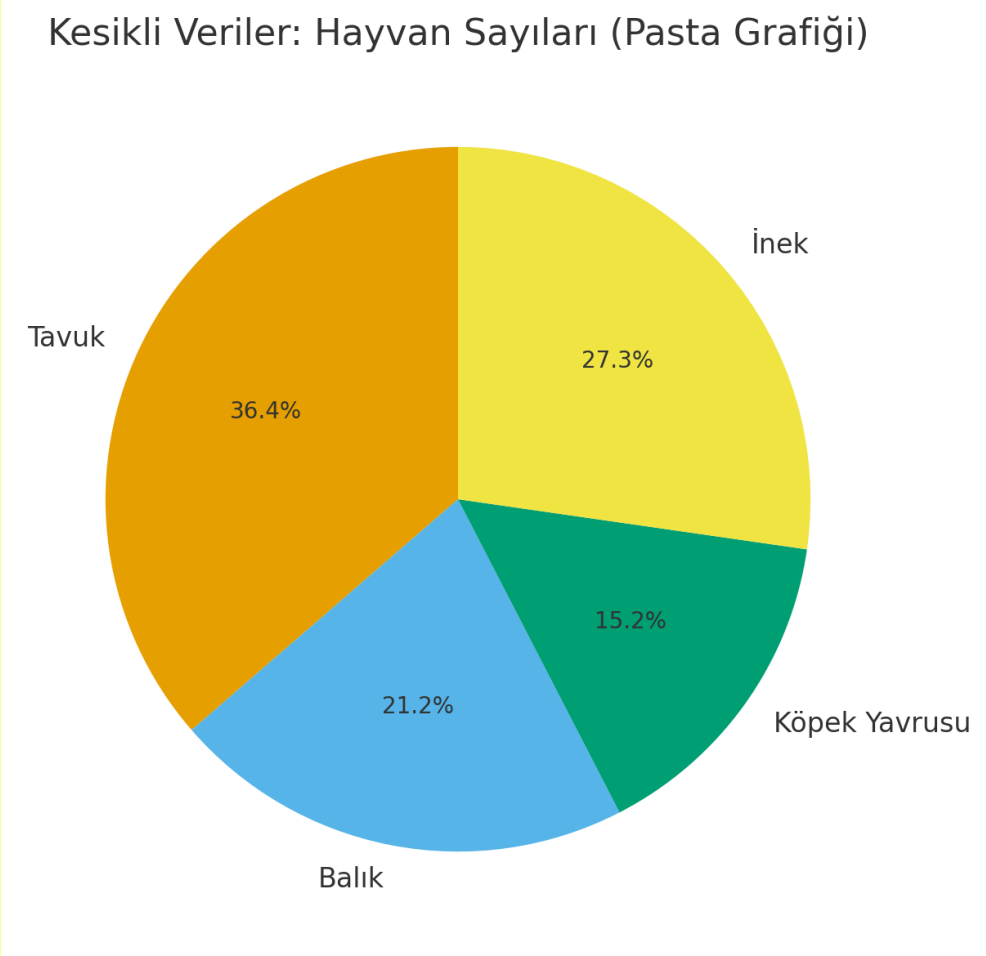
Veri Türleri

- **Kesikli (Discrete) Veriler**



Veri Türleri

- **Kesikli (Discrete) Veriler**



Veri Türleri

- **Sürekli (Continuous) Veriler**

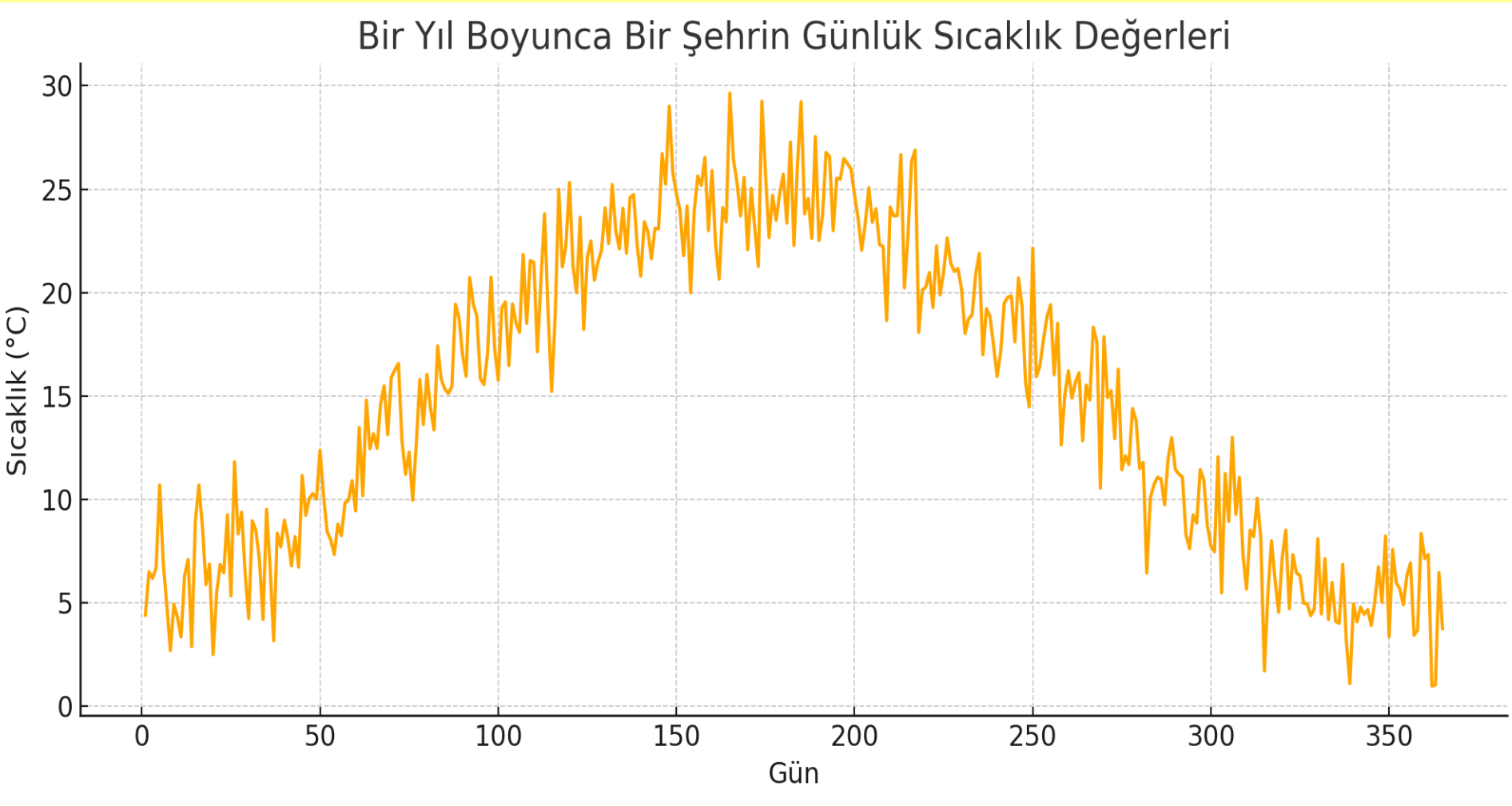
- **Sürekli (Continuous) veriler**, belli bir aralıkta sonsuz değer alabilen ve ölçümle ifade edilen nicel verilerdir.

Veri Türleri

- **Sürekli (Continuous) Veriler**

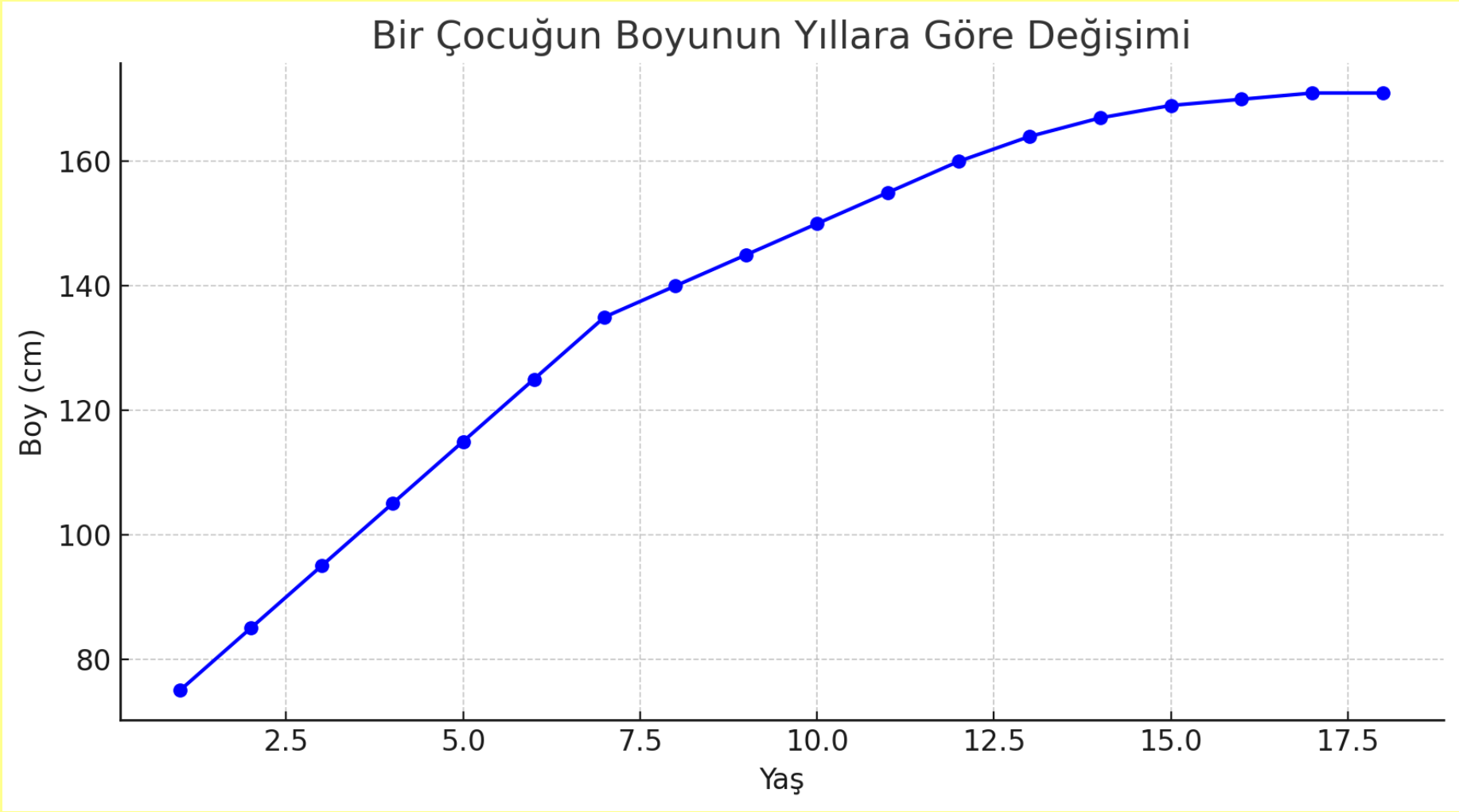
Veri Türleri

Sürekli (Continuous) Veriler



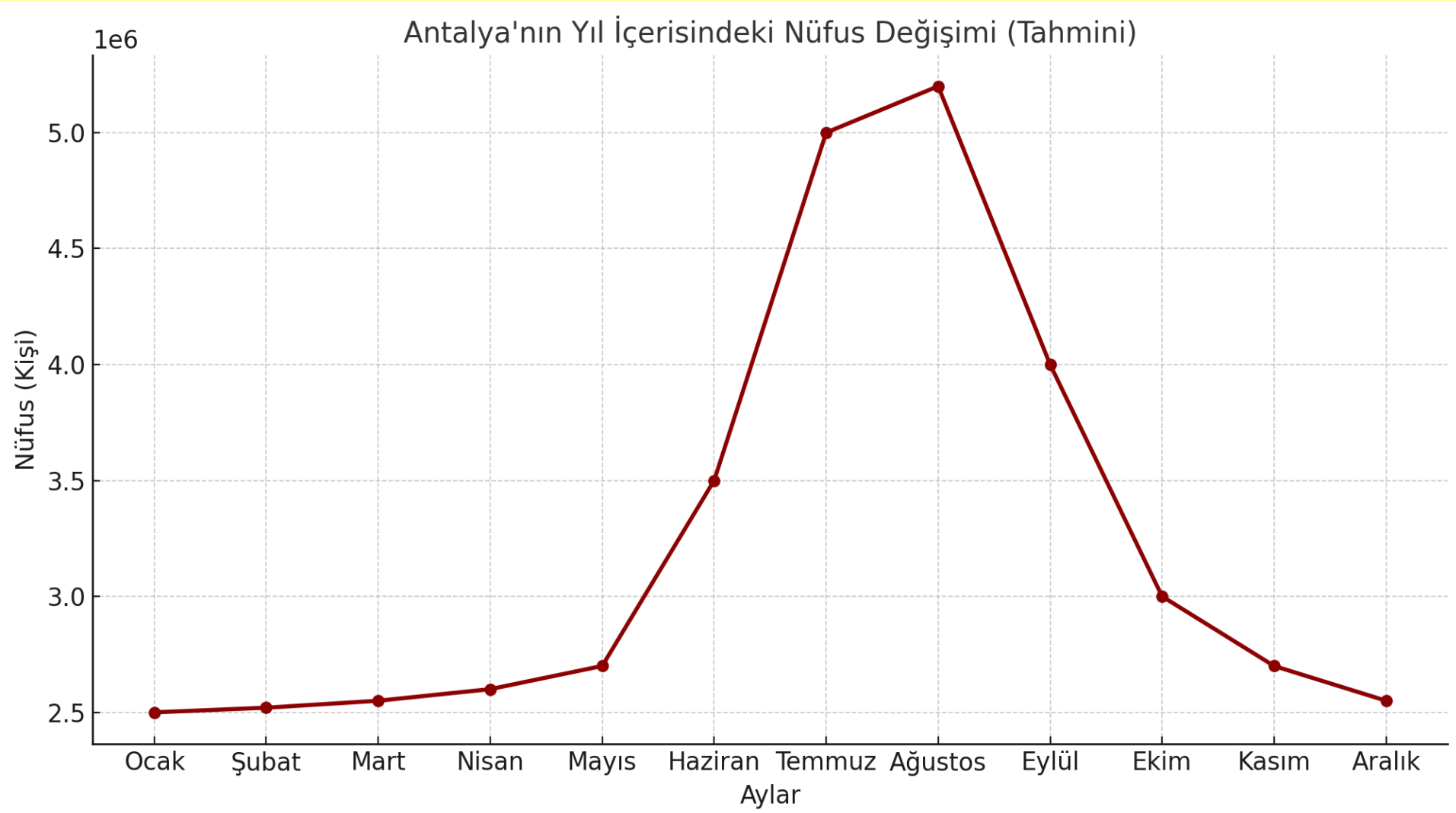
Veri Türleri

Sürekli (Continuous) Veriler



Veri Türleri

Sürekli (Continuous) Veriler



Ölçüm Ölçekleri

- Nominal: Sınıflandırma
- Ordinal: Sıralama
- Interval: Sıralama + fark, mutlak sıfır yok
- Ratio: Sıralama + fark + mutlak sıfır

Ölçüm Ölçekleri

1. Nominal (Sınıflandırma)

👉 Sadece kategoriler vardır, sıralama yoktur.

- Hayvan türleri: kedi, köpek, kuş
- Kan grupları: A, B, AB, 0
- Göz rengi: mavi, kahverengi, yeşil

Ölçüm Ölçekleri

2. Ordinal (Sıralama)

👉 Sıralama vardır ama aradaki farklar ölçülemez.

- Eğitim düzeyi: ilkokul < ortaokul < lise < üniversite
- Yarış sırası: 1., 2., 3.
- Anket yanıtları: "katılmıyorum", "kararsızım", "katılıyorum"

Ölçüm Ölçekleri

3. Interval (Sıralama + fark, mutlak sıfır yok)

👉 Farklar ölçülebilir ama gerçek sıfır noktası yoktur.

- Sıcaklık (°C veya °F): 20°C ile 30°C arasındaki fark 10°C'dir, ama 0°C "hiç sıcaklık yok" demek değildir.
- Takvim yılları: 2000 ile 2010 arasındaki fark 10 yıldır, ama "0 yılı" başlangıç değil.
- IQ test puanları

Ölçüm Ölçekleri

4. Ratio (Sıralama + fark + mutlak sıfır)

👉 Hem sıralama hem fark vardır, ayrıca mutlak sıfır anlamlıdır.

- Kilo (kg): 0 kg demek hiç ağırlık yoktur.
- Boy (cm): 0 cm = hiç uzunluk yok.
- Gelir (₺): 0 gelir = hiç para yok.
- Yaş (yıl): 0 yaş = doğum anı.

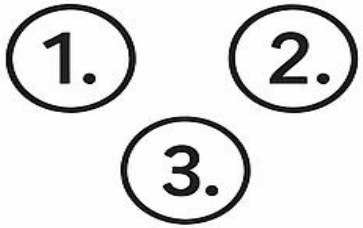
NOMINAL

Sınıflandırma

hayvan türleri

ORDINAL

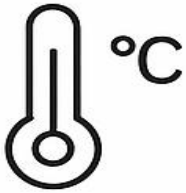
Sıralama



yarış sırası

INTERVAL

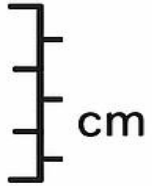
Sıralama + fark
mutlak sıfır yok



sıcaklık (°C)

RATIO

Sıralama + fark +
mutlak sıfır



boy (cm)

Ölçüm Ölçekleri Tablosu

Ölçüm Ölçekleri

Ölçek	Özellikler	Örnek
Nominal	Sadece sınıflandırma	Programlama dili: Python, C++, Java
Ordinal	Sıralama var, fark anlamsız	Başarı: düşük, orta, yüksek
Interval	Sıralama + fark, mutlak 0 yok	Sıcaklık (°C)
Ratio	Sıralama + fark + mutlak 0	İşlem süresi (ms), bellek (GB)

Bilgisayar Mühendisliğinde Kullanım Alanları

- Makine öğrenmesi: Model eğitimi, hata oranı
- Veri tabanı: Sorgu optimizasyonu
- Ağ güvenliği: Trafik analizi
- Performans analizi: Yazılım hız testleri

Bilgisayar Mühendisliğinde Kullanım Alanları

- **Makine öğrenmesi**

- **Naive Bayes Sınıflandırıcı:** Bir e-postanın spam olup olmadığını belirlerken, her kelimenin spam olma olasılığını hesaplar. Örnek: "bedava", "kazan", "hemen" kelimeleri yüksek spam olasılığına işaret edebilir.

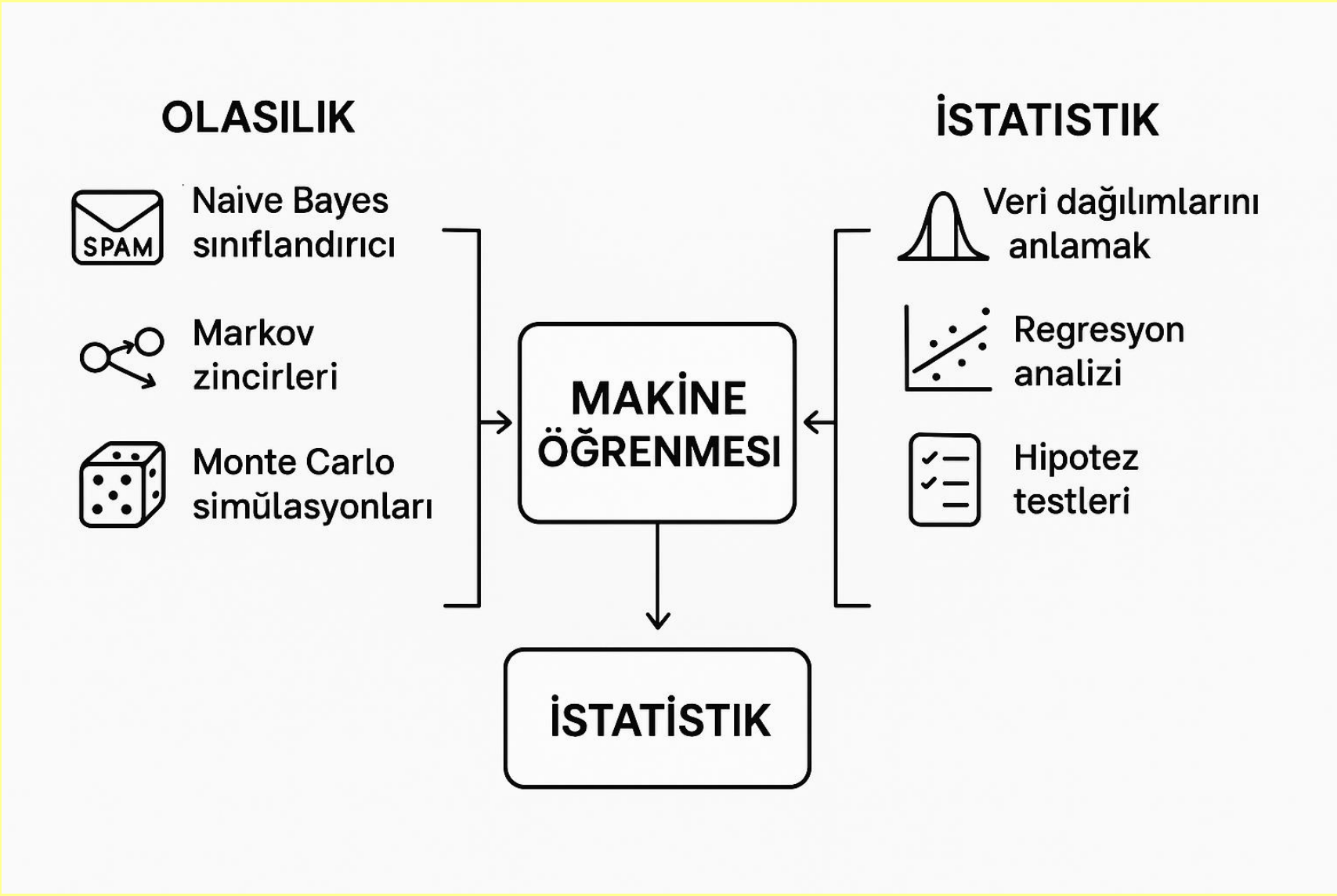
- **Markov Zincirleri:** Bir kullanıcının bir web sitesindeki sonraki tıklamasını tahmin etmek için olasılıklardan yararlanılır.

- **Monte Carlo Simülasyonları:** Belirsizlik içeren durumlarda (örneğin finansal risk analizi) rastgele örnekleme yaparak sonuçların dağılımı hesaplanır.

Bilgisayar Mühendisliğinde Kullanım Alanları

- **Makine öğrenmesi**
- **Veri Dağılımlarını Anlamak:** Öğrencilerin sınav notlarının ortalama, medyan ve standart sapmasına bakarak başarı seviyesini ölçmek.
- **Regresyon Analizi:** Ev fiyatlarını tahmin etmek için: oda sayısı, metrekare, konum gibi değişkenler ile fiyat arasındaki istatistiksel ilişki kurulur.
- **Hipotez Testleri:** Yeni bir ilaç ile mevcut tedavinin etkisini kıyaslamak için istatistiksel anlamlılık testleri yapılır.

Bilgisayar Mühendisliğinde Kullanım Alanları



Bilgisayar Mühendisliğinde Kullanım Alanları

- **Veri tabanı: Sorgu optimizasyonu**

- **Sorgu Optimizasyonu (Query Optimization)**
- Veri tabanı yönetim sistemleri (ör. PostgreSQL, Oracle), sorgu planını seçerken tabloların boyutları ve olasılıksal tahminler üzerinden karar verir.
- Örnek: SELECT * FROM müşteriler WHERE yaş > 30 sorgusunda, yaş > 30 koşulunu sağlayan kayıtların oranı olasılık tahminiyle hesaplanır.
- **Veri Madenciliği / Tahminleme**
- Veri tabanında kayıtlı müşterilerin satın alma davranışları üzerinden olasılıksal modeller kurulur.👉 Örnek: Bir müşterinin belirli bir ürünü satın alma olasılığını Naive Bayes ile tahmin etmek.

Bilgisayar Mühendisliğinde Kullanım Alanları

- **Veri tabanı: Sorgu optimizasyonu**
- **İndeks ve İstatistik Tabloları**
- Çoğu veri tabanı sistemi, tablolar için istatistiksel özetler (ortalama, minimum, maksimum, dağılım) tutar.
- Bu sayede sorgular daha hızlı çalışır çünkü sistem verinin nasıl dağıldığını bilir.

- **Veri Kalitesi ve Anomali Tespiti**
- Tablo sütunlarındaki değerlerin ortalama ve standart sapması alınarak sıra dışı kayıtlar bulunabilir.
- Örnek: Normalde 20–60 yaş arası olan bir müşteri tablosunda "150 yaşında" bir kayıt bulunursa istatistiksel olarak anomali kabul edilir.

- **Veri Özetleme**
- Örneğin, bir satış tablosunda ortalama satış miktarı, medyan, standart sapma gibi özet istatistikler hesaplanır.

Bilgisayar Mühendisliğinde Kullanım Alanları

- **Ağ güvenliği: Trafik analizi**

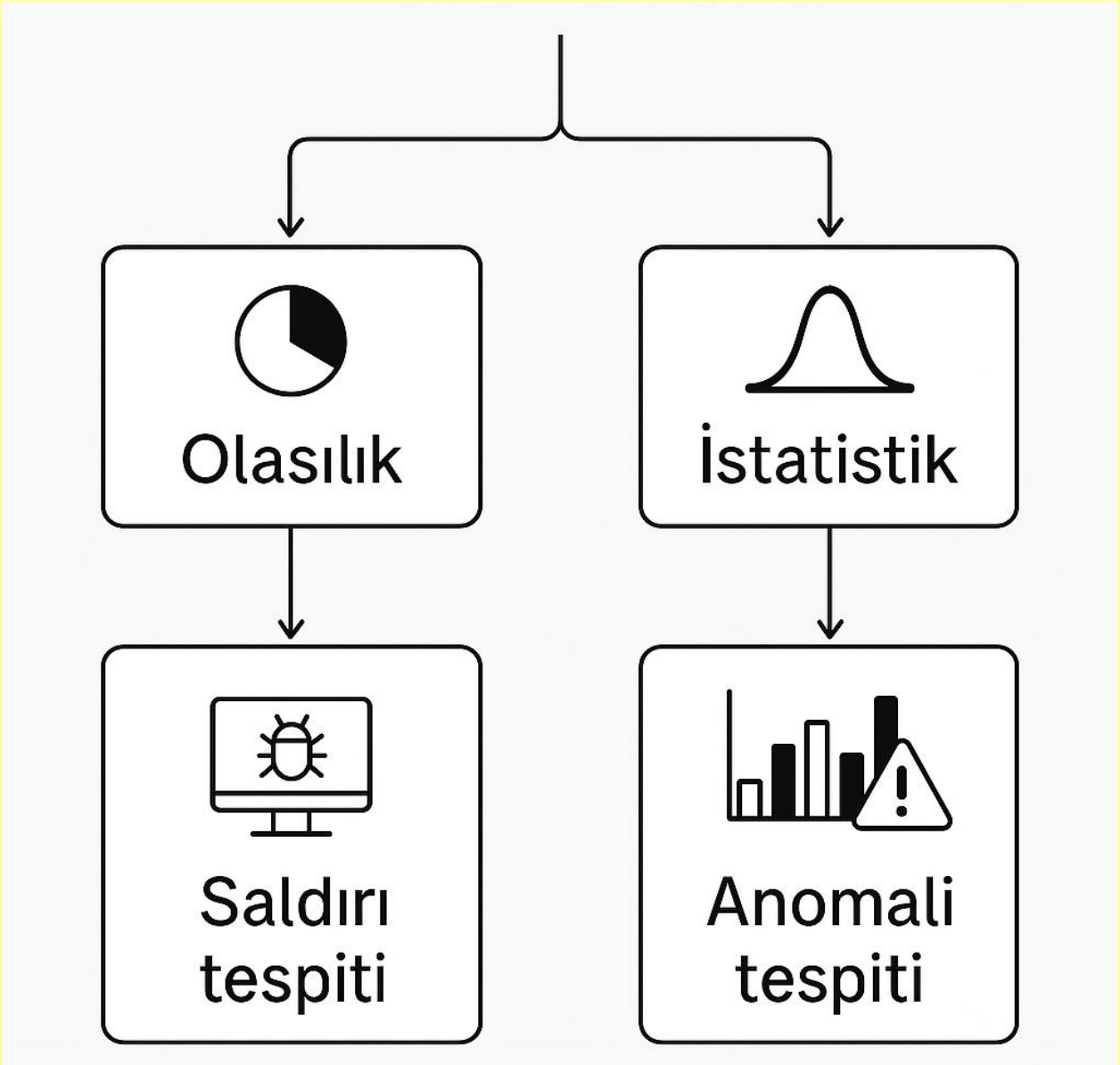
- **Saldırı Tespiti (Intrusion Detection)**
- Belirli bir IP'den gelen paketlerin saldırı olma olasılığı hesaplanır.
- Örnek: Bir IP'den saniyede 1000'den fazla paket geliyorsa, bunun DDoS olma olasılığı çok yüksektir.
- **Bayesçi Yöntemler**
- Ağ trafiğini "normal" veya "anormal" olarak sınıflandırmak için Naive Bayes kullanılır.
- Örnek: Paket boyutu, port numarası, protokol bilgisine göre saldırı olasılığı çıkarılır.
- **Markov Zincirleri**
- Trafik akışının sıradaki durumunu (örneğin HTTP isteği sonrası beklenen yanıt) olasılıkla tahmin eder.

Bilgisayar Mühendisliğinde Kullanım Alanları

- **Ağ güvenliği: Trafik analizi**

- **Ortalama ve Standart Sapma ile Anomali Tespiti**
- Normalde 100–200 ms arası olan ağ gecikmesinde aniden 2000 ms görülürse, bu istatistiksel olarak anormaldir.
- **Frekans Dağılımları**
- Paket türlerinin (TCP, UDP, ICMP) dağılımı incelenerek olağandışı bir artış olup olmadığı analiz edilir.
- **Korelasyon Analizi**
- Farklı portlardaki trafik hacimleri arasındaki ilişki incelenir. Aynı anda artış varsa koordineli saldırı şüphesi oluşur.
- **Zaman Serisi Analizi**
- Trafik hacmi belirli aralıklarla izlenir ve olağan dışı artışlar (ör. spam trafiği) tespit edilir.

Bilgisayar Mühendisliğinde Kullanım Alanları



Bilgisayar Mühendisliğinde Kullanım Alanları

- **Performans analizi: Yazılım hız testleri**

- **Yanıt Süresi Dağılımı**
- Bir web servisinin yanıt süresinin belirli bir eşik altında olma olasılığı hesaplanır.
- Örnek: “Bir API isteğinin 200 ms’den kısa sürede tamamlanma olasılığı %85.”

- **Kuyruk Teorisi (Queueing Theory)**
- Sunucuların gelen istekleri karşılama ihtimalini modellemek için olasılık dağılımları (Poisson, Exponential) kullanılır.

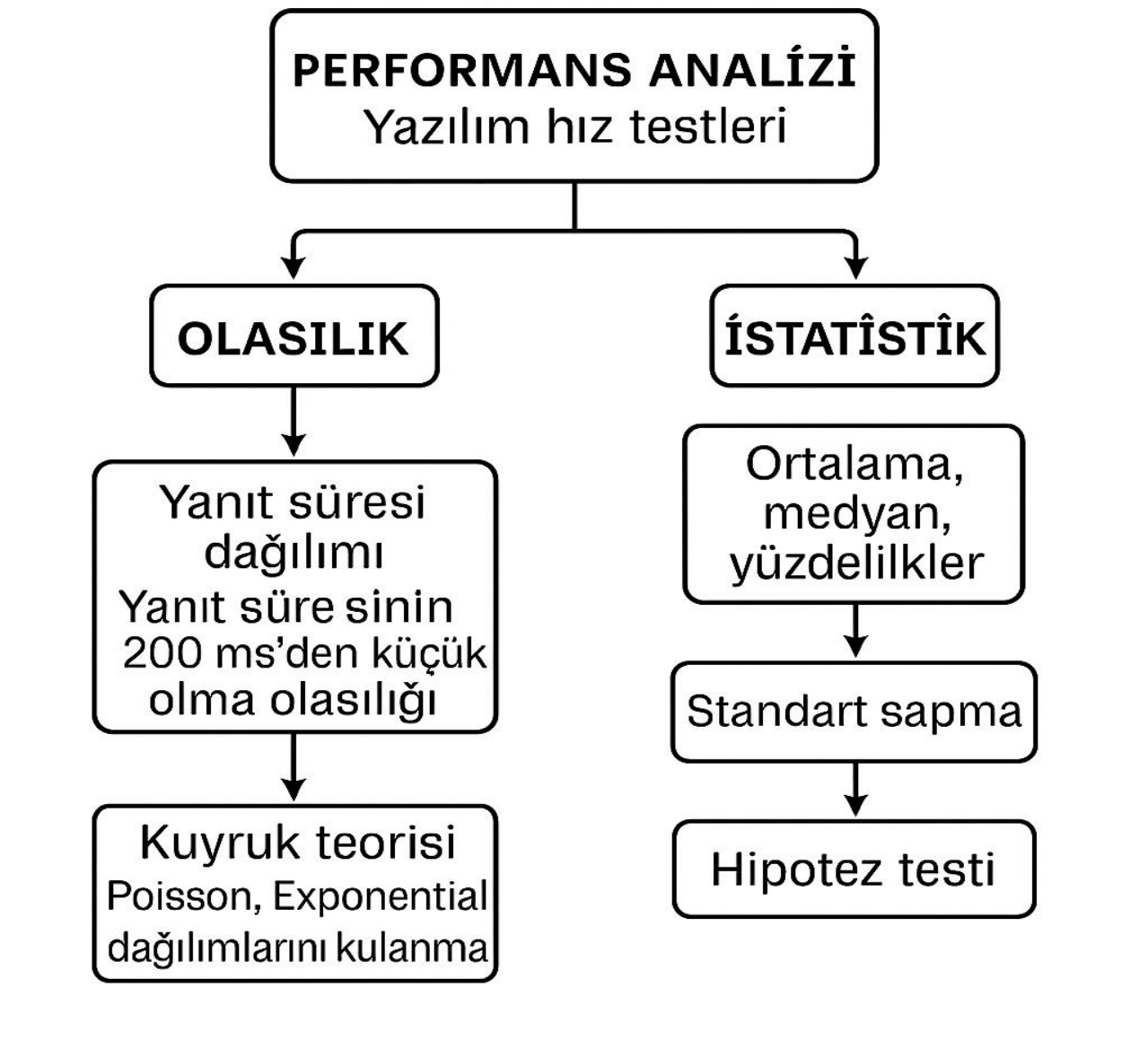
- **Aşırı Yük Senaryoları**
- Belirli sayıda eşzamanlı kullanıcı geldiğinde sistemin çökme olasılığı tahmin edilir.

Bilgisayar Mühendisliğinde Kullanım Alanları

- **Performans analizi: Yazılım hız testleri**

- **Ortalama, Medyan, Yüzdelikler**
- Performans testlerinde sadece ortalama değil, medyan ve %95 yüzdelik gibi ölçüler raporlanır.
- Örnek: Ortalama yanıt süresi 180 ms, ama %95 kullanıcı 250 ms'den kısa sürede yanıt alıyor.
- **Standart Sapma ve Varyans**
- Yanıt sürelerinin ne kadar değişken olduğunu ölçer.
- Düşük standart sapma = daha kararlı sistem.
- **Hipotez Testi**
- Yeni bir algoritmanın eskisine göre daha hızlı olup olmadığı istatistiksel olarak test edilir.
- **Regresyon Analizi**
- Donanım kaynakları (CPU, RAM) ile yazılım performansı arasındaki ilişki incelenir.

Bilgisayar Mühendisliğinde Kullanım Alanları



Örnek Problemler

- 1. Bir sınıfta 60 öğrenci var. 10 Python, 25 C++, 25 Java kullanıyor.
- - Veri türü nedir?
- - Ölçüm ölçeği nedir?

- 2. Bir yazılım firmasında aylık hata sayısı kaydediliyor.
- - Kesikli mi, sürekli mi?
- - Ortalama/standart sapma hesaplanabilir mi?

Küçük Ödev

- 1. En az 3 soruluk küçük bir anket tasarlayın.
- - Her sorunun veri türünü ve ölçüm ölçeğini belirleyin.
- 2. Topladığınız verilerden tablo oluşturun ve yorumlayın.