

BTK Akademi

Veri Bilimi İçin Temel İstatistik

—
Açıklamalı
Kılavuz

@yavuzcancolak

İçindekiler

Bölüm 1 – Temel İstatistik ve Betimsel Yöntemler

→ Sayfa 2

Bölüm 2 – Olasılık

→ Sayfa 18

Bölüm 3 – Dağılımlar

→ Sayfa 23

Bölüm 4 – Hipotez Testleri

→ Sayfa 30

Bölüm 5 – İlişki Analizi ve Tahmin Yöntemleri

→ Sayfa 34

Bölüm 6 – Çok Değişkenli Sonuç Çıkarımı

→ Sayfa 40

Bölüm 7 – Sınıflama Yöntemleri

→ Sayfa 44

Bölüm 8 – Kapanış

→ Sayfa 48

Bölüm 1

Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

Bu bölümde, istatistiğin temel kavramları ile verileri tanıma ve yorumlama sürecinde kullanılan betimsel analiz yöntemleri ele alınmaktadır.

Temel Tanımlar

1. İstatistik

Verilerin toplanması, düzenlenmesi, analiz edilmesi ve yorumlanması sağlayen bilim dalıdır. Veri bilimiyle entegre çalışır ve sayısal içgörü üretir.

2. Veri

Araştırmalarda gözlemlenen veya ölçülen ham bilgilerdir. Sayılar, kategoriler, semboller gibi farklı biçimlerde olabilir.

3. Gözlem Birimi

İstatistiksel veri toplamada ölçüm yapılan en küçük birimdir. Örneğin: bir birey, bir ürün, bir şirket.

4. Örneklem ve Anakütle

- **Anakütle (evren):** Hakkında bilgi edinmek istediğimiz tüm bireylerin oluşturduğu grup.
- **Örneklem:** Anakülteden seçilen, analiz edilen daha küçük gruptur.

Değişken Türleri

İstatistikte "değişken", ölçtüğümüz ya da gözlemlediğimiz özelliklerdir. Örneğin bir öğrencinin yaşı, cinsiyeti, sınav puanı ya da mezun olduğu okul birer değişkendir.

Değişkenleri 3 ana başlık altında inceleyebiliriz:

- Yapı ve Özelliklerine Göre Değişkenler
- Değerine Göre Değişkenler
- Neden-Sonuç İlişkisine Göre Değişkenler

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

Yapı ve Özelliklerine Göre Değişkenler

Bu gruplama, verinin nitelidine göre yapılır: sayı mı, kategori mi?

📍 Nitel Değişken (Kalitatif)

Sayısal olmayan, kategorik ifadelerdir. Gruplama ya da sınıflandırma yapar ama ölçüm içermez.

Örnekler:

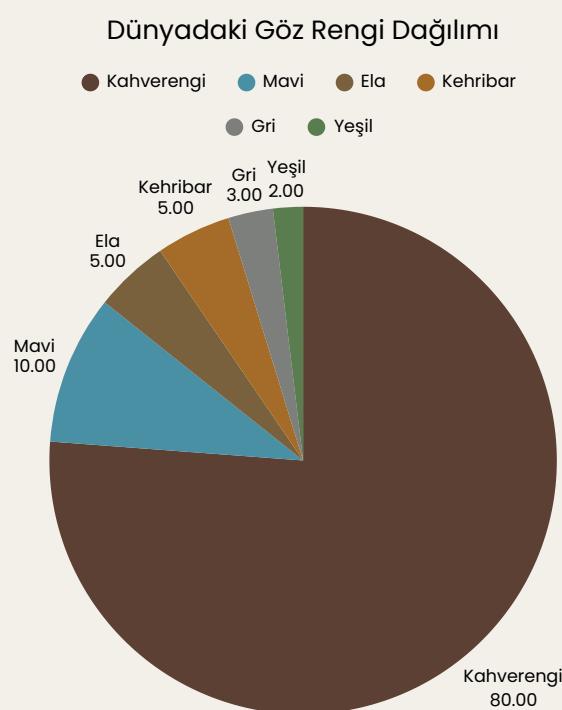
- Cinsiyet: Kadın / Erkek
- Medeni Durum: Evli / Bekar
- Göz Rengi: Kahverengi / Mavi / Yeşil
- Şehir: İstanbul / Ankara / İzmir

📍 Nicel Değişken (Kantitatif)

Sayısal olan, ölçülebilir değişkenlerdir. İşlem yapılabılır (toplama, ortalama alma vb.)

Örnekler:

- Yaş
- Gelir
- Sıcaklık
- Sınav Puanı



Değerine Göre Değişkenler

Nicel değişkenlerin sahip olduğu değerin sürekli mi yoksa sayılabilir mi olduğunu tanımlar.

📍 Sürekli Değişken

İki değer arasında sonsuz sayı olabilir. Ondalıklı değerlere sahiptir.

Örnekler:

- Boy: 172.5 cm, 172.55 cm...
- Sıcaklık: 22.1°C, 22.15°C...

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

✖ Süreksiz (Kesikli) Değişken

Sadece tam sayılar alır. Sayılabilir, araya değer girmez.

Örnekler:

- Öğrenci sayısı: 25 kişi (25.5 öğrenci olmaz)
- Kitap sayısı: 12 kitap

Neden-Sonuç İlişkisine Göre Değişkenler

Bu sınıflandırma, bir deneyde ya da analizde değişkenlerin rollerine göre yapılır.

✖ Bağımsız Değişken

Araştırmada kontrol edilen veya değiştiren değişkendir. Sebep rolündedir.

Örnekler:

- Uykusuzluk süresi → sınav başarısını etkiler mi?
- Gübre miktarı → bitki boyunu etkiler mi?

Burada uykusuzluk süresi ve gübre miktarı bağımsız değişkendir.

✖ Bağımlı Değişken

Bağımsız değişkene bağlı olarak değişen/ölçülen değişkendir. Sonuçtur.

Örnekler:

- Sınav Puanı
- Bitki Boyu

Burada uykusuzluk süresi ve gübre miktarı bağımsız değişkendir.

Uyku Süresi (x)



Sınav Başarısı (y)

Bağımsız

Bağımlı

Ölçek Türleri

Verileri sayılarla ifade ederken kullandığımız farklı ölçme düzeyleri vardır. Bu ölçme düzeyleri, veriye nasıl davranışımız gerektiğini belirler. Temel olarak 5 farklı ölçme düzeyi bulunur:

Sınıflama (Nominal) Ölçme Düzeyi

- Verileri **kategorilere ayırmak** için kullanılır.
- Kategoriler arasında **sıralama yoktur**.

Örnekler:

- Cinsiyet: Kadın / Erkek
- Kan grubu: A / B / AB / 0
- Göz rengi: Kahverengi / Mavi / Yeşil
- Şehirler: İstanbul / Ankara / İzmir

 Bu veriler sadece gruplamayı ifade eder; hangi grubun daha “fazla” ya da “önemli” olduğunu göstermez.

Sıralama (Ordinal) Ölçme Düzeyi

- Veriler **sıralanabilir**, ancak aralarındaki farklar bilinmez.
- Sayılar büyülüklük ifade eder ama ne kadar büyük olduğunu göstermez.

Örnekler:

- Yarışma dereceleri: 1. / 2. / 3.
- Eğitim düzeyi: İlkokul < Lise < Üniversite
- Acı seviyesi: Az / Orta / Çok

 “Kim önce geldi?” sorusuna yanıt verir, ama aralarındaki süreyi bileyemeyiz.

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

Aralık (Interval) Ölçme Düzeyi

- Veriler arasında **eşit aralıklar** vardır.
- Ancak **mutlak sıfır noktası yoktur**.

Örnekler:

- Sıcaklık: 20°C, 30°C (ama 30°C, 20°C'nin 1.5 katı değildir!)
- Takvim yılları: 1990, 2000, 2020

 Toplama ve çıkarma yapılabilir ama oranlama yapılmaz.

Oran (Ratio) Ölçme Düzeyi

- Hem eşit aralıklar vardır hem de mutlak sıfır noktası bulunur.
- Matematiksel işlemlerin tamamı yapılabilir: toplama, çıkarma, çarpma, oranlama...

Örnekler:

- Ağırlık: 0 kg, 5 kg, 10 kg
- Yaş: 0 yaş, 20 yaş
- Gelir: 0 TL, 1000 TL

 10 kg, 5 kg'nin tam 2 katıdır. Çünkü sıfır noktası gerçekten.

Likert Ölçme Düzeyi

- Bireylerin görüşlerini **ölçmek** için kullanılır.
- Genellikle anketlerde kullanılır.

Örnek:

“Bu eğitimi faydalı buldunuz mu?”

- Katılmıyorum
- Kararsızım
- Katılıyorum

 Genellikle ordinal (sıralı) gibi değerlendirilir ama bazı analizlerde interval gibi de ele alınabilir.

Merkezi Eğilim Ölçüleri

Merkezi eğilim ölçüleri, bir veri setindeki değerlerin nasıl bir noktada **yöğunlaştığını** ya da hangi değerin **temsil edici** olduğunu anlamamıza yardımcı olur.

Bu ölçüler sayesinde:

- Verinin genel yapısını daha kolay yorumlarız.
- Aykırı (uç) değerleri fark edebiliriz.
- Kıyaslama yaparken hangi değerin temel alınacağını biliriz.

İki ana gruba ayrılır:

■ **Analitik Ortalamalar**

(Sayısal işlemlerle hesaplanır)

■ **Analitik Olmayan Ortalamalar**

(Sıralama ve frekansa dayanır)

◆ **Analitik Ortalamalar**

Aritmetik Ortalama

En sık kullanılan ortalama türüdür.

Tüm değerlerin toplamının, veri sayısına bölünmesiyle hesaplanır.

Formül:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Örnek:

Sınav notları: 60, 70, 90

$$(60 + 70 + 90)/3 = 220/3 = 73.3$$

Avantaj: Kolay hesaplanır.

Dezavantaj: Aykırı değerlerden çok etkilenir (örneğin: 5, 6, 100).

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

Geometrik Ortalama

Özellikle **oran, büyümeye hızı** ve **yüzdelik artışlar** içeren verilerde tercih edilir. Sayıların çarpımının kökü alınarak hesaplanır.

Formül:

$$GO = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Örnek:

Yıllık büyümeye oranları: %10, %20

Bunlar sırasıyla 1.10 ve 1.20 olarak yazılır.

$$GO = \sqrt{1.10 \times 1.20} = \sqrt{1.32} \approx 1.148$$

Bu sonuç %14.8'lik ortalama büyümeyi ifade eder.

Yani: Geometrik Ortalama \approx %14.8 büyümeye

Ağırlıklı Ortalama

Verilerin önem derecesi farklısa kullanılır.

Her değere bir **ağırlık katsayısı** verilir.

Formül:

$$\frac{x_1 w_1 + x_2 w_2 + \dots + x_n w_n}{w_1 + w_2 + \dots + w_n}$$

Örnek:

Vize (%40): 70

Final (%60): 90

$$(70 \times 0.4) + (90 \times 0.6) = 28 + 54 = 82$$

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

Harmonik Ortalama

Hız, oran gibi **zaman** ve **oran bazlı ölçümlerde** kullanılır.

Tüm sayıların terslerinin ortalamasının tersidir.

Formül:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Örnek:

Bir yolun yarısı 60 km/s, diğer yarısı 90 km/s hızla gidildi:

$$H = \frac{2}{\left(\frac{1}{60} + \frac{1}{90}\right)} \approx \mathbf{72} \text{ km/s}$$

◆ Analitik Olmayan Ortalamalar

Mod (Tepe Değer)

Veri setinde en sık tekrar eden değer.

Örnek:

Veri: 5, 6, 7, 5, 9

Mod = 5 (iki kez geçti)

 Mod her zaman tek sayı olmayıpabilir. İkili ya da çoklu mod da olabilir.

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

Medyan (Ortanca)

Veri seti sıralanır, **tam ortadaki değer** medyandır.

- Tek sayı varsa → Ortadaki sayı
- Çift sayı varsa → Ortadaki iki sayının ortalaması

Örnek (tek sayı):

Veri: 2, 4, 6 → Medyan = 4

Örnek (çift sayı):

Veri: 2, 4, 6, 8 → $(4+6)/2 = 5$

 Aykırı değerlerden etkilenmez, bu yüzden güvenilir bir merkez göstergesidir.

Kartiller

Veri setini %25'lik parçalara böler.

Bu sayede **dağılımın detayları** hakkında bilgi verir.

- **1. Kartil (Q1):** Alt %25
- **2. Kartil (Q2):** Medyan
- **3. Kartil (Q3):** Üst %25

Örnek:

Veri: 5, 10, 15, 20, 25, 30, 35

- $Q1 = 10$
- $Q2 = 20$
- $Q3 = 30$

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

🔍 Merkezi Eğilim Ölçülerinin Karşılaştırması (Özet Tablo)

Ölçü	Temel Dayanak	Aykırıya Duyarlılık	Hangi Durumlarda Tercih Edilir?
Aritmetik Ortalama	Sayısal hesap	Yüksek	Sayılar dengeliyse ve uç değer yoksa. Günlük hesaplamalar için ideal.
Medyan (Ortanca)	Değer sıralaması	Düşük	Aykırı (uç) değerler varsa, merkezi daha doğru yansıtır.
Mod (Tepe Değer)	En sık tekrar	Düşük	En çok hangi değer tekrar ediyorsa onu bulmak için.
Ağırlıklı Ortalama	Sayısal hesap + ağırlık	Orta	Bazı değerler daha önemliyse (ör. vize/final notu gibi).
Harmonik Ortalama	Oran hesapları	Yüksek	Hız, verim, oran gibi zaman bazlı ölçümlerde.

Bu ölçüler, tek bir merkezi değeri bulmak, birbirleriyle kıyaslamak, veri setini özetlemek için kullanılır. Bu nedenle tabloda yer almaları mantıklı.

✗ Tabloda Yer Almayanlar:

Merkezi eğilim ölçüsü olarak kabul edilmeyen ya da farklı yapıda olanlar:

Ölçü	Neden dahil edilmedi?
Geometrik Ortalama	Daha çok oran ve büyümeye hesaplarında kullanılır, doğrudan “merkez” ifadesiyle değil.
Kartiller (Q1, Q3 vs.)	Merkezi değil, dağılım ölçüsüdür; alt ve üst bölgelere odaklanır.



Frekans (Sıklık) Tabloları

Frekans (ya da sıklık), bir değerin veri setinde **kac kez tekrar ettiğini** ifade eder.

Örnek:

Bir sınıfındaki öğrencilerin sınav puanları:

70, 80, 90, 70, 60, 90, 70

Burada 70 puanı 3 kez, 90 puanı 2 kez görülmüştür.

Bu sayıların her biri frekans değeridir.



Neden Frekans Tabloları Kullanılır?

- Büyük veri setlerinde verilerin dağılımını kolay görmek için
- Sınıflandırılmış veya gruplandırılmış verileri düzenli summak için
- Grafiksel gösterimlere (çubuk grafik, histogram) hazırlık olarak

✳️ Örnek Veri Seti (Yaş – Mezuniyet – Puan)

Yaş	Mezuniyet Durumu	Puan
18	Lise	50
28	Lisans	70
45	Yüksek Lisans	80
27	Lise	40
54	Yüksek Lisans	60
...

Bu tablo, sade bir veri listesidir. Ancak bu verileri **yorumlamak, gruplamak** ve **özetlemek** için frekans tablolarına ihtiyaç duyuyoruz.

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

👉 Sınıflandırma Nasıl Yapılır?

Veri türü sayısal (örneğin yaşı) olduğunda, veriler gruplara ayrılarak frekans hesaplanabilir.

👉 Örnek: Yaş Gruplama İşlemi

Veriler:

18, 18, 20, 27, 28, 30, 32, 40, 45, 54...

1. En büyük ve en küçük değer farkı:

$$54 - 18 = 36$$

2. Sınıf sayısı belirlendi (örnek: 3)

$$36 \div 3 = 12 \rightarrow \text{sınıf genişliği}$$

3. Gruplar oluşturulur:

- 18–29
- 30–41
- 42–53
- 54 ve üzeri

📘 Frekans Tablosu (Çaprazlama Örneği)

Aşağıdaki tablo; yaş grubu, mezuniyet durumu ve puan durumuna göre verileri çapraz gruplar halinde özetler.

Yaş Grubu	Lise	Lisans	Yüksek Lisans	50 Altı	50 ve Üzeri
18-29	12	4	0	4	12
30-41	0	4	1	1	4
42-53	1	3	3	2	5
54+	0	0	2	0	2

 **Bu tablo sayesinde:**

- Hangi yaş grubunda hangi mezuniyet düzeyi daha baskın?
- 50 puan barajını geçenlerin dağılımı nedir?
- Eğitim seviyesiyle başarı arasında ilişki var mı? gibi sorulara **görsel** ve **sayısal** yanıt verilebilir.

Grafiksel Gösterimler

Veri analizi sürecinde, verilerin daha kolay yorumlanabilmesi için farklı grafik türleri kullanılır. Hangi grafik türünün kullanılacağı, **veri tipine** ve **analiz amacına** göre değişir.

Histogram

Sayısal verilerin sınıflara bölünüp, her sınıfın frekansına göre sütunla gösterilmesidir.

Ne zaman kullanılır?

- Sürekli sayısal veri dağılımını görmek için
- Veri setinin şekli (örneğin simetrik mi?) incelenmek isteniyorsa

Örnek Grafik:

X eksen → yaşı aralığı (örn. 18–25), Y eksen → kişi sayısı

Sütun Grafiği

Kategorik verilerin frekanslarını veya yüzdelerini dikey/sütun şeklinde gösterir.

Ne zaman kullanılır?

- Cinsiyet, şehir, mezuniyet durumu gibi kategoriler karşılaştırıldığında

Örnek Grafik:

X eksen → şehirler, Y eksen → kişi sayısı

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

Çizgi Grafiği

Zaman serisi verilerini göstermek için kullanılır; noktalar çizgiyle birleştirilir.

Ne zaman kullanılır?

- Günlük sıcaklık, yıllık satış gibi zamanla değişen veriler için

Örnek Grafik:

X ekseni → yıllar, Y ekseni → gelir

Daire Grafiği (Pasta Grafiği)

Kategorik verilerin oranlarını dairesel dilimlerle gösterir.

Ne zaman kullanılır?

- Toplam içindeki dağılım oranlarını göstermek için

Örnek Grafik:

Şirket bütçesinin harcama kalemlerine göre dağılımı

Saçılım Grafiği (Scatter Plot)

İki değişken arasındaki ilişkiyi nokta bulutu şeklinde gösterir.

Ne zaman kullanılır?

- Korelasyon analizi yapmak istendiğinde
- Değişkenler arasındaki eğilim gözlemlenecekse

Örnek Grafik:

X ekseni → çalışma süresi, Y ekseni → sınav puanı

Q-Q Grafiği (Quantile-Quantile Plot)

Bir dağılımin, teorik bir dağılıma (genelde normal dağılım) ne kadar uyduğunu gösterir.

Ne zaman kullanılır?

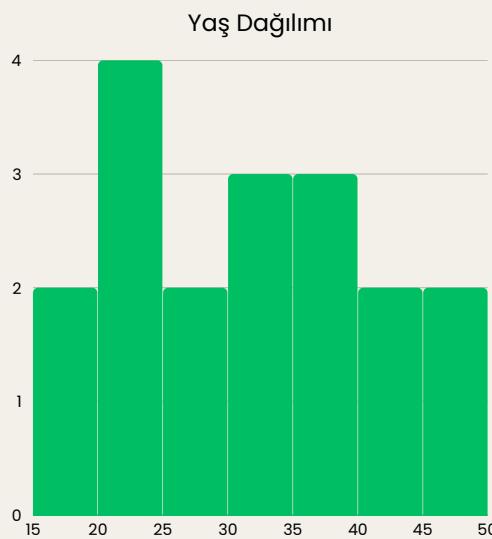
- İstatistiksel test öncesi veri yapısını değerlendirmek için

Örnek Grafik:

Veri noktalarının teorik eğriye ne kadar yakın olduğu incelenir.

1. Temel İstatistik Kavramları ve Betimsel Analiz Yöntemleri

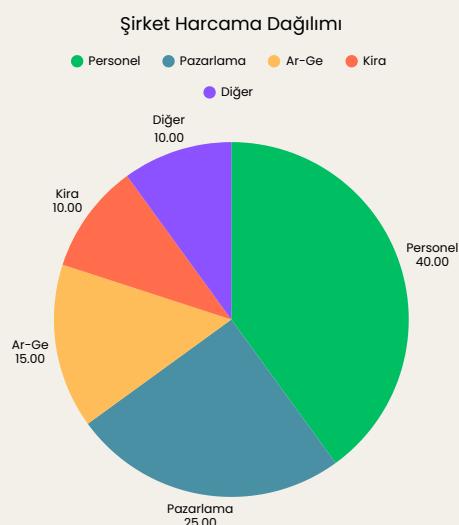
🎯 Örnek Grafik Gösterimleri



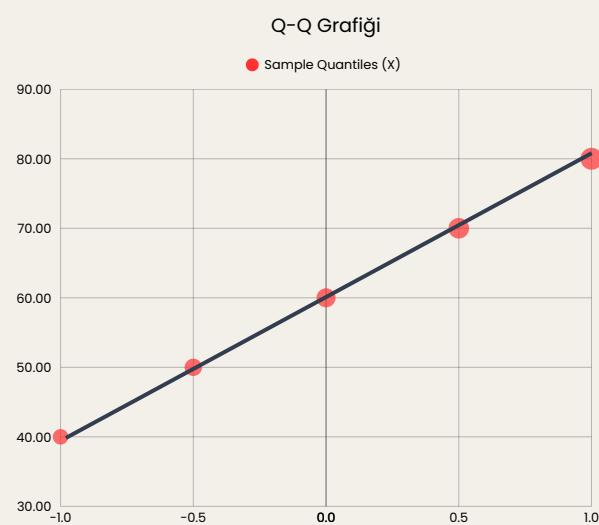
📍 Histogram | Histogram Grafiği



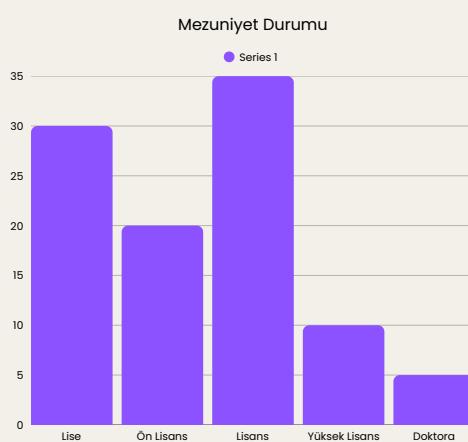
📍 Line Chart | Çizgi Grafiği



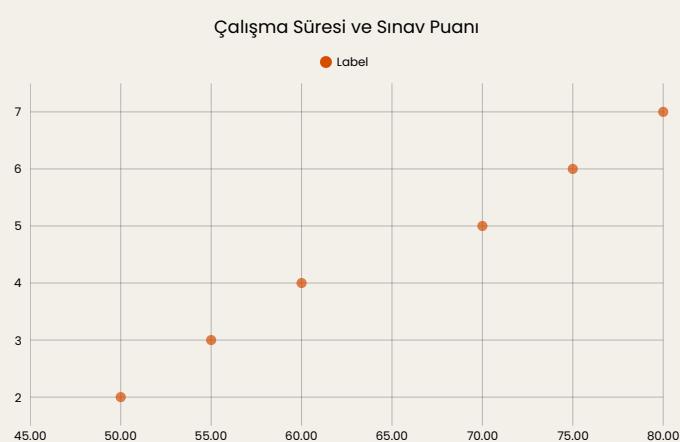
📍 Pie Chart | Pasta Grafiği



📍 Q-Q Plot | Q-Q Grafiği



📍 Bar Chart | Sütun Grafiği



📍 Scatter Plot | Saçılım Grafiği

Bölüm 2

Olasılık

Bu bölümde, olasılık kavramının temelleri, temel kurallar ve Bayes teoremi gibi önemli yapılar sade örneklerle ele alınmaktadır.

2. Olasılık

Olasılık ?

Olasılık, belirsizlik içeren durumları anlamamıza ve kararlarımıza sayısal bir temel kazandırmamıza yardımcı olur. Günlük hayatın her alanında, oyunlardan tıbb'a kadar geniş bir kullanım alanına sahiptir.

Olasılık, bir olayın gerçekleşme ihtimalini ifade eder. Değerler her zaman 0 ile 1 arasındadır:

- 0 = imkansız
- 1 = kesin
- Aradaki tüm değerler = olasılık derecesi

” Basit Örnek:

Bir madeni para atıldığında yazı gelme olasılığı nedir?

→ Olası sonuçlar: {yazı, tura}

→ Yazı gelme olasılığı = $1 / 2 = 0.5$

” Başka bir örnek:

Bir zar atıldığında “4” gelme olasılığı nedir?

→ Olası sonuçlar: {1, 2, 3, 4, 5, 6}

→ “4” gelme olasılığı = $1 / 6 \approx 0.167$

! Temel Olasılık Kavramları

Olasılık hesaplamalarında bazı temel kavramlar vardır:

- Örnek Uzay (S): Tüm olası sonuçlar kümesi
- Olay (A): Belirli bir durumu ifade eden alt küme
- Olasılık (P): Olayın gerçekleşme ihtimali

” Örnek 1:

Bir zar atıldığında tek sayı gelme olasılığı nedir?

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$A = \{1, 3, 5\}$$

$$P(A) = 3 / 6 = 0.5$$

2. Olasılık

” Örnek 2 – Birleşim ve Kesişim:

A: Çift sayılar (2, 4, 6)

B: 1'den büyük sayılar (2, 3, 4, 5, 6)

$$A \cup B \text{ (birleşim)} = \{2, 3, 4, 5, 6\}$$

$$A \cap B \text{ (kesişim)} = \{2, 4, 6\}$$

” Koşullu Olasılık ($P(A|B)$):

“Bir kartın kırmızı olduğu biliniyorsa, kupa gelme olasılığı nedir?”

- Deste: 52 kart
- Kırmızı kartlar: 26 kart (13 karo + 13 kupa)
- Bu 26 kırmızı kart arasında 13 tanesi kupadır.

Dolayısıyla:

$$P(\text{Kupa} | \text{Kırmızı}) = \frac{13}{26} = 0.5$$

💡 Bayes Teoremi

Bayes Teoremi, bir olay hakkında yeni bilgi edinildikten sonra o olayın olasılığını güncellemeyi sağlar.

📌 Formül:

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

📘 Anlamı:

- $P(A|B)$: B olduktan sonra A'nın olasılığı
- $P(B|A)$: A olduğunda B'nin olasılığı
- $P(A)$: A'nın önceden bilinen olasılığı
- $P(B)$: B'nin genel olasılığı

2. Olasılık

11 Örnek – Hastalık Testi:

Bir hastalığın toplumda görülmeye oranı %1.

Testin doğruluğu %99.

Bir kişinin testi pozitif çıktıysa, gerçekten hasta olma ihtimali nedir?

İlk İzlenim:

“Test pozitifse, bu kişi kesin hastadır.” gibi düşünebiliriz çünkü test çok güvenilir görünüyor (%99). Ancak hastalık çok nadir görüldüğü için bu çıkarım yaniltıcı olabilir. Burada Bayes Teoremi devreye girer.

Sayılarla Açıklama:

Toplam 10.000 kişilik bir toplum düşünelim.

◆ Gerçekten hasta olanlar (%1):

- 100 kişi hasta.
- Bu 100 hastanın %99'u testte pozitif çıkar → 99 kişi.

◆ Sağlıklı olanlar (%99):

- 9.900 kişi sağlıklı.
- Bu 9.900 kişiden %1'i yanlışlıkla pozitif çıkar → 99 kişi.

Sonuç:

Toplam pozitif çıkan kişi sayısı =

 Gerçek pozitifler (hasta olup testi pozitif çıkan) = 99 kişi

 Yanlış pozitifler (sağlıklı olup testi pozitif çıkan) = 99 kişi

Toplam pozitif = $99 + 99 = 198$ kişi

Bu 198 kişiden sadece 99'u gerçekten hasta.

Olasılık Hesabı:

$$P(\text{Hasta} \mid \text{Pozitif}) = \frac{99}{198} = 0,5 = \%50$$

2. Olasılık

Yorumu:

Yani test çok güvenilir olsa bile, hastalığın toplumda çok nadir görülmesi nedeniyle pozitif test sonucu alan birinin gerçekten hasta olma olasılığı sadece %50'dir.

Bu örnek, öncül olasılıkların yani "hastalığın toplumdaki yaygınlığının" kararları nasıl etkilediğini net bir şekilde gösterir.

Özette:

Kavram	Açıklama
Olasılık (P)	Bir olayın olma ihtimali
Örnek Uzay (S)	Tüm olasılıkların kümesi
Olay (A, B)	Belirli sonuçların alt kümesi
Koşullu Olasılık	B olayına bağlı olarak A'nın olasılığı
Bayes Teoremi	Yeni bilgiyle olasılığın güncellenmesi

Bölüm 3

Dağılımlar

Bu bölümde, bir veri setinin nasıl dağıldığını anlamaya yarayan ölçüler ve grafiklerle birlikte, temel dağılım türleri, çarpıklık-basıklık kavramları, serbestlik derecesi ve merkezi limit teoremi gibi konular ele alınmaktadır.

3. Dağılımlar

🎯 Dağılım Nedir?

Bir veri setindeki değerlerin nasıl yayıldığını ve ne kadar değişkenlik gösterdiğini tanımlar.

🧠 Basitçe:

- Ortalama etrafında sıkışık mı?
- Geniş bir aralığa mı yayılmış?
- Uç değerler (aykırı değerler) var mı?

📏 Dağılım Ölçüleri

Dağılım ölçüleri, verilerin merkezi eğilimden ne kadar uzaklaştığını gösterir. İşte en yaygın ölçüler:

1. Ranj (Genişlik)

En büyük değer – En küçük değer

📌 Örnek:

Veri: [3, 7, 8, 10, 15] → Ranj = 15 – 3 = 12

2. Varyans (Variance)

Verilerin ortalamaya göre karesel uzaklıklarının ortalamasıdır.

Formül:

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

Bu formül, örneklem varyansını temsil eder ve şu şekilde açıklanabilir:

- s^2 : Varyans
- n : Gözlem sayısı
- x_i : Her bir veri noktası
- \bar{x} : Örneklem ortalaması
- Σ : Toplama işlemi

3. Dağılımlar

📌 Basit Örnek:

Veriler: [4, 6, 8]

Ortalama: 6

Adımlar:

- $(4-6)^2 = 4$
- $(6-6)^2 = 0$
- $(8-6)^2 = 4$
- Toplam: 8 →
- Varyans = $8 / (3-1) = 4$

3. Standart Sapma (Standard Deviation)

📎 Varyansın kareköküdür.

Formül:

$$s = \sqrt{s^2}$$

Açıklama:

- s : Standart sapma
- s^2 : Varyans

Yani standart sapma, varyansın kareköküdür.

📌 Yukarıdaki örneğe göre:

$$\text{Standart sapma} = \sqrt{4} = 2$$

4. Çarpıklık (Skewness)

Verinin simetrik olup olmadığını gösterir.

- Sağa çarpık → Ortalama > Medyan
- Sola çarpık → Ortalama < Medyan
- Simetrik → Ortalama ≈ Medyan

📌 Örnek: Gelir dağılımı genellikle sağa çarpıktır.

3. Dağılımlar

5. Basıklık (Kurtosis)

Verinin tepe noktasının sivrilğini belirtir.

- Yüksek basıklık → Sivri tepe
- Düşük basıklık → Düz tepe

Kesikli ve Sürekli Dağılımlar

- Kesikli Dağılım (Discrete Distribution)
 Belirli sayılabilir değerler alır.
 Örnek: Zar atıldığından çıkan sayılar (1–6), sınavda doğru cevap sayısı
- Sürekli Dağılım (Continuous Distribution)
 Sonsuz sayıda değer alabilir.
 Örnek: Boy uzunluğu (170.2 cm, 170.23 cm...), sıcaklık, ağırlık

Dağılım Türleri

Bernoulli Dağılımı

Sadece iki olasılığa sahip deneyler için kullanılır:

Başarı (1) veya Başarisızlık (0)

Kullanım Alanı: Tek bir deneme sonucunun değerlendirilmesi.

Örnek: Üretim hattından rastgele seçilen bir ürünün hatalı çıkma olasılığı.
Hatalısa 1, değilse 0 olarak değerlendirilir.

Binom Dağılımı

Birden fazla Bernoulli denemesinin toplam sonucudur.

Denemeler bağımsızdır ve başarı olasılığı sabittir.

- Formül:

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

3. Dağılımlar

Açıklama:

- n : Deneme sayısı
- k : Başarı sayısı
- p : Başarı olasılığı
- (nk) : Kombinasyon (n 'in k 'li kombinasyonu)

Örnek: 10 üründen 2'sinin hatalı olma olasılığı nedir?

Her ürün bağımsız olarak kontrol edilir ve hata olasılığı sabittir.

Poisson Dağılımı

Belirli bir zaman dilimi veya alanda nadir gerçekleşen olaylar için kullanılır. Olayların birbirinden bağımsız olduğu varsayıılır.

Formül:

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

Açıklama:

- X : Belirli bir zaman diliminde meydana gelen olay sayısı
- k : Gerçekleşmesi istenen olay sayısı
- λ : Belirli bir zaman aralığındaki ortalama olay sayısı
- e : Euler sayısı (yaklaşık 2.718)

Örnek: Bir gün içerisinde gelen e-posta sayısı.

Ortalama 5 e-posta geliyorsa, bugün 7 e-posta gelme olasılığı hesaplanabilir.

3. Dağılımlar

✓ Normal Dağılım (Gaussian)

🔍 Tanım:

Normal dağılım, simetrik bir çan eğrisi şeklinde olup, ortalama, medyan ve mod aynı noktadadır. Yani en çok gözlemlenen değer ortadadır, veriler bu değerin etrafında dengeli biçimde dağılır.

📌 Temel Özellikler:

- Verilerin büyük kısmı merkezin etrafında yer alır.
- Merkezden uzaklaştıkça (uç değerler) görülme olasılığı azalır.
- Dağılım simetiktir; sağ ve sol tarafı birbirinin aynısı gibidir.

💡 Basit Örnek:

Bir sınıfta 1000 öğrencinin boyaları ölçüldü.

Elde edilen veriler incelediğinde büyük çoğunluğun boyu 170–180 cm aralığında çıktı.

Bu, normal dağılıma örnektir çünkü:

- En fazla kişi bu aralıktaki yoğunlaşmış.
- Daha kısa (160 cm) veya daha uzun (190 cm) boyalar daha az sayıda.
- Boyalar ortalamaya göre simetrik biçimde dağılmış.

🧠 **Kısaca:** Normal dağılım, doğada ve insan davranışlarında çok sık görülen bir dağılım türüdür. Sınav notları, boy uzunlukları, zeka testleri gibi birçok ölçüm bu dağılıma yakınlık gösterir.

✓ Uniform Dağılım

🔍 Tanım:

Tüm değerlerin eşit olasılıkla seçildiği durumlarda kullanılır.

Dağılım düzgündür.

Örnek: 1 ile 10 arasında rastgele sayı üretimi.

- Her sayının gelme olasılığı %10'dur.
- Zar atımı da bir uniform dağılım örneğidir. 🎲

3. Dağılımlar

✓ Ki-Kare Dağılımı



Tanım:

Kategorik veriler arasındaki ilişkiyi test etmek için kullanılır.
Özellikle çapraz tablolar (kontenjans) ile analiz yapılır.

Örnek: Yapılan bir anket sonucu;

- cinsiyet ile ürün tercihi arasında istatistiksel olarak anlamlı bir ilişki var mı?
- gibi soruların yanıtı için kullanılır.

🧭 Özet Tablo

Dağılım Türü	Açıklama	Örnek	Tipik Kullanım
Bernoulli	İki olasılıklı	Yazı/Tura	Tek deneme olasılığı
Binom	Çoklu Bernoulli	10 denemede başarı sayısı	Sınırlı tekrar sayısı
Poisson	Zaman içinde nadir olay	1 saatte gelen mail	Olay yoğunluğu
Normal	Simetrik eğri	Boy, ağırlık	Doğal ölçümler
Uniform	Eşit olasılıklı	0–1 arası sayı	Rastgele seçim
Ki-Kare	Varyans testi	Gözlenen-fark testi	Istatistiksel analiz

Bölüm 4

Hipotez Testleri

Bu bölümde, istatistiksel analizlerin temel taşlarından biri olan hipotez testlerini sade, temel dille açıklayacağız. Hedefimiz: “Veriler gerçekten anlamlı mı, yoksa bu farklar sadece tesadüf mü?” sorusuna yanıt bulmaktır.

4. Hipotez Testleri

✓ 4.1 Hipotez Testine Giriş

🔍 Tanım:

Hipotez testi, bir önermenin (hipotezin) doğru olup olmadığını anlamaya yarayan istatistiksel bir test yöntemidir.

Temel Kavramlar:

- H_0 (Null Hipotez / Yokluk Hipotezi):
Her şeyin “normal” olduğu varsayımdır. Örnek: “İki grup arasında fark yoktur.”
- H_1 (Alternatif Hipotez):
 H_0 'ın tersi, yani bir farkın, etkinin ya da ilişkinin olduğunu ileri süren varsayımdır.

Örnek: “İki grup arasında anlamlı fark vardır.”

🧠 Unutma:

Hipotez testleri, doğrudan H_1 'i kanıtlamaz.

Ama elimizdeki verilere göre H_0 'ı reddedip reddetmeyeceğimize karar veririz.

✓ 4.2 Tip I ve Tip II Hatalar

💡 Hipotez testi sonucunda alınabilecek kararlar 4 farklı duruma yol açar:

Gerçek Durum

H_0 Kabul Edildi

H_0 Reddedildi

H_0 Doğru

✓ Doğru karar

✗ Tip I Hata (α)

H_0 Yanlış

✗ Tip II Hata (β)

✓ Doğru karar

4. Hipotez Testleri

◆ Tip I Hata (α):

H_0 doğru olduğu halde yanlışlıkla reddedilmesi.

→ "Aslında gruplar arasında fark yok, ama biz var dedik."

◆ Tip II Hata (β):

H_0 yanlış olduğu halde reddedilememesi.

→ "Aslında gruplar arasında fark var, ama biz fark yok dedik."

◆ Güven Düzeyi ($1 - \alpha$):

Test sonucunda H_0 doğrulanken doğru karar verme olasılığıdır.

→ Genellikle %95 olarak alınır. ($\alpha = 0.05$)

◆ Test Gücü ($1 - \beta$):

H_0 yanlışken doğru şekilde reddetme olasılığıdır.

→ Testin duyarlığını gösterir.

✓ 4.3 P-Degeri (p-value) Nedir?

🔍 Tanım:

P-değeri, gözlemlenen farkın sadece tesadüfen oluşmuş olma olasılığıdır. Yani, "Bu fark şansa mı oldu, yoksa gerçekten anlamlı mı?" sorusuna cevap verir.

Nasıl yorumlanır?

- Eğer $p < 0.05 \rightarrow H_0$ reddedilir. → "İstatistiksel olarak anlamlı fark var."
- Eğer $p \geq 0.05 \rightarrow H_0$ reddedilemez. → "Fark istatistiksel olarak anlamlı değil."

◆ Örnek:

Bir ilacın etkisini test ettiniz.

Sonuç: $p = 0.03 \rightarrow$ Bu farkın tesadüfen oluşma olasılığı %3 → Anlamlıdır → H_0 reddedilir.

4. Hipotez Testleri

✓ 4.4 Hipotez Testi Türleri

Hipotez testleri, veri türüne ve karşılaştırma şekline göre ikiye ayrılır:

Hipotez Testi Türü	Hipotez Testi Türü
 Parametrik Yöntemler	Veriler normal dağılıyorsa, ölçüm düzeyi eşit aralıklıysa
 Parametrik Olmayan Yöntemler	Veriler normal dağılmıyorsa, sıralı ya da kategorikse

- ❖ Parametrik testlerde genellikle sayısal veriler (örneğin sınav puanları) kullanılır.
- ❖ Parametrik olmayan testler daha çok sıralama, kategorik veri ya da küçük örneklemeler için uygundur.

Bölüm 5

İlişki Analizi ve Tahmin Yöntemleri

Veriler arasındaki ilişkileri çözümlemek ve geleceği öngörmek için istatistiğin sunduğu güçlü araçlara birlikte göz atıyoruz.

5. İlişki Analizi ve Tahmin Yöntemleri

Veri biliminin en önemli amaçlarından biri, veriler arasındaki ilişkileri anlamak ve bu ilişkileri kullanarak geleceğe dair tahminler yapabilmektir. Bu bölümde iki temel başlık üzerinde duracağız:

- 1. İlişki Analizi:** Değişkenler arasında bağ var mı? Varsa yönü ve gücü nedir?
- 2. Tahmin Yöntemleri:** Bu bağı kullanarak gelecekteki değerleri tahmin edebilir miyiz?

Kovaryans – İlişkinin Yönünü Anlamak



Tanım:

Kovaryans, iki değişkenin birlikte nasıl değiştığını ölçen sayısal bir değerdir.

- İki değişken aynı yönde değişiyorsa (biri artarken diğer de artıyorsa veya ikisi de azalırken), kovaryans pozitif olur.
- Ters yönde değişiyorlarsa kovaryans negatif olur.
- Kovaryans değeri 0'a yakınsa, anlamlı bir ilişki yoktur.



Formül:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Formül Açıklaması:

- X_i ve Y_i : $i.$ gözlemdeki değerler
- \bar{X} ve \bar{Y} : X ve Y 'nin ortalamaları
- n : gözlem sayısı
- $(X_i - \bar{X})(Y_i - \bar{Y})$: Her gözlemede X 'in ortalamadan sapması ile Y 'nin ortalamadan sapmasının çarpımı

Eğer sapmalar aynı işaretliyse (ikisi de ortalamanın üstünde veya ikisi de altında) çarpım pozitif çıkar \rightarrow pozitif ilişki.

Sapmalar ters işaretliyse çarpım negatif çıkar \rightarrow negatif ilişki.

5. İlişki Analizi ve Tahmin Yöntemleri

💡 Örnek (Kovaryans)

Bir öğretmen, öğrencilerin çalışma saati ile sınav puanı arasındaki ilişkiyi inceliyor.

Öğrenci	Çalışma Saati (X)	Sınav Puanı (Y)
A	5	70
B	7	85
C	3	60

- Çalışma saati arttıkça puanlar da artıyor.
- Kovaryans pozitif → iki değişken aynı yönde hareket ediyor.

Korelasyon – İlişkinin Yönü ve Gücü

📌 Tanım:

- Korelasyon, kovaryansı **standartlaştırarak** yorumlanabilir hale getirir.
- Kovaryansta birim problemi vardır (örneğin saat × puan gibi karışık bir birim).
- Korelasyon bu problemi ortadan kaldırır ve sonucu **-1 ile +1 arasında** verir.

⚠️ Formül:

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

Formül Açıklaması:

- **Cov(X, Y):** Kovaryans
- **σ_x :** X'in standart sapması
- **σ_y :** Y'nin standart sapması
- Kovaryans değeri iki değişkenin standart sapmalarına bölünerek **ölçeksiz** hale getirilir.

5. İlişki Analizi ve Tahmin Yöntemleri

Korelasyon Katsayısının Yorumlanması

- **+1** → Mükemmel pozitif ilişki
- **0** → İlişki yok
- **-1** → Mükemmel negatif ilişki
- **0.7 ile 1 arası** → Güçlü ilişki
- **0.3 ile 0.7 arası** → Orta düzey ilişki
- **0 ile 0.3 arası** → Zayıf ilişki

Örnek (Korelasyon)

Bir şirket, çalışanların **haftalık çalışma saatı** ile **üretim miktarı** arasındaki ilişkiyi ölçüyor.

- Korelasyon katsayısı **r = 0.89** bulunuyor.
- Bu, güçlü ve pozitif bir ilişki demektir: çalışma süresi arttıkça üretim de artma eğiliminde.

⚠ Uyarı: Korelasyon nedensellik göstermez. Çalışma süresinin artması üretimi artırabilir, ancak üretimi artıran başka faktörler (makine sayısı, iş deneyimi) de olabilir.

Regresyon Analizi – Tahmin Yapmak

Kovaryans ve korelasyon ilişkiyi ölçer ama **tahmin yapmaz**. Tahmin için **regresyon analizi** kullanılır.

Basit Doğrusal Regresyon

 **Tanım:**

Bir bağımlı değişken (Y) ile tek bir bağımsız değişken (X) arasındaki ilişkiyi modelleyen yöntem.

5. İlişki Analizi ve Tahmin Yöntemleri

► Formül:

$$Y = a + bX$$

- **a**: Sabit terim ($X=0$ iken Y değeri)
- **b**: Eğim katsayısı (X 'te 1 birim değişim, Y 'yi ne kadar değiştirir?)
- **X**: Bağımsız değişken
- **Y**: Tahmin edilen değer

💡 Örnek:

Bir dondurma firması, sıcaklık arttıkça satışların nasıl değiştigini öğrenmek istiyor.

- **Y**: Günlük satış (adet)
- **X**: Günlük sıcaklık ($^{\circ}\text{C}$)
- Model:

$$\text{Satış} = 50 + 8 \times \text{Sıcaklık}$$

Bu formüle göre sıcaklık her 1°C arttığında satış 8 adet artıyor.

Çoklu Doğrusal Regresyon

📌 Tanım:

Birden fazla bağımsız değişkenin Y üzerindeki etkisini inceler.

► Formül:

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

💡 Örnek:

Bir evin fiyatını tahmin etmek için:

- **X₁**: Metrekare
- **X₂**: Oda sayısı
- **X₃**: Merkeze uzaklık (km)
- **X₄**: Binanın yaşı

Model:

$$\text{Fiyat} = 100000 + 1500X_1 + 20000X_2 - 5000X_3 - 1000X_4$$

Metrekare ve oda sayısı arttıkça fiyat yükselir, uzaklık ve yaşı arttıkça fiyat düşer.

5. İlişki Analizi ve Tahmin Yöntemleri

Regresyon Modelinin Değerlendirilmesi

■ Temel Metrikler:

- **R² (Determinasyon Katsayısı):** Modelin Y'yi ne kadar açıkladığını gösterir (0–1 arası).
- **MAE (Ortalama Mutlak Hata):** Tahminlerin ortalama hata miktarı.
- **MSE / RMSE:** Hataların kare ortalaması / karekökü.

💡 Örnek:

$R^2 = 0.92$ ise model, bağımlı değişkendeki değişimin %92'sini açıklıyor.

Lojistik Regresyon – Olasılık Tahmini

📌 Tanım:

Bağımlı değişken kategorik (ör. Evet/Hayır) olduğunda kullanılır. Sonuç olarak **olasılık** verir.

💡 Örnek:

Bir banka, müşterinin kredi kartı başvurusunun onaylanma ihtimalini tahmin ediyor.

- X: Gelir, kredi puanı, yaşı, borç durumu
- Model sonucu: Olasılık = 0.76 (%76 onay ihtimali)

■ Performans Ölçütleri:

- **Accuracy:** Doğru tahmin oranı
- **Precision:** Pozitif tahminlerin doğruluk oranı
- **Recall:** Gerçek pozitiflerin ne kadarının bulunduğu
- **F1 Skoru:** Precision & Recall dengesi
- **ROC-AUC:** Modelin genel başarı ölçüsü

🎯 Bölüm Özeti

- **Kovaryans:** İlişkinin yönünü verir.
- **Korelasyon:** İlişkinin yönünü ve gücünü verir.
- **Regresyon:** İlişkiden yararlanarak tahmin yapar.
- **Lojistik Regresyon:** Tahminin olasılık değerini verir.

Bu bilgiler, hem veri analizi hem de iş kararlarında güçlü bir temel sağlar.

Bölüm 6

Çok Değişkenli Sonuç Çıkarımı

Karmaşık veri setlerini anlamlandırmak için boyut indirgeme ve gizli yapı keşfi tekniklerini inceliyoruz.

6. Çok Değişkenli Sonuç Çıkarımı

Çok Değişkenli Sonuç Çıkarma

Veri setlerinde bazen çok fazla değişken bulunur. Bu değişkenlerin tümünü kullanmak:

- Analizi zorlaştırır
- Hesaplama maliyetini artırır
- Bazı değişkenlerin birbirini tekrar eden bilgileri barındırmamasına neden olur

Bu noktada boyut indirgeme teknikleri devreye girer. Boyut indirgeme, verideki değişken sayısını azaltarak:

- Daha basit modeller kurmamızı
- Daha hızlı hesaplama yapmamızı
- Görselleştirmeyi kolaylaştırmamızı sağlar.

PCA – Temel Bileşenler Analizi

💡 Tanım:

PCA, bir veri setindeki **en önemli bilgiyi** taşıyan yeni değişkenler (bileşenler) oluşturur.

Bu bileşenler, orijinal değişkenlerin **doğrusal kombinasyonlarıdır** ve birbirleriyle **ilişkisizdir**.

⚠️ Adımlar:

1. Veriyi Standartlaştırma

- Değişkenler farklı ölçeklerdeyse (örneğin cm ve kg), PCA öncesi standartlaştırılır.

2. Kovaryans Matrisi Oluşturma

- Değişkenler arası ilişkiyi görmek için.

3. Özdeğer ve Özvektör Hesabı

- Özvektörler → yeni eksenler (temel bileşenler)
- Özdeğerler → her bileşenin taşıdığı bilgi miktarı (varyans)

4. Bileşen Seçimi

- Toplam varyansın büyük kısmını açıklayan ilk birkaç bileşen seçilir.

5. Yeni Veri Seti Oluşturma

- Orijinal veriler, seçilen bileşenlere dönüştürülür.

6. Çok Değişkenli Sonuç Çıkarımı

💡 Örnek (PCA):

Bir öğrenci başarısını etkileyen 10 farklı değişken ölçülmüş (matematik puanı, fen puanı, ders çalışma süresi, devamsızlık, vb.). Bu değişkenlerin bir kısmı birbirine çok benzer bilgi taşıyor olabilir.

PCA ile:

- Birinci bileşen: Akademik başarıyı temsil eden birleşik skor
- İkinci bileşen: Disiplin/katılım faktörünü temsil eden skor
- elde edilebilir.
- Bu şekilde analiz, 10 değişken yerine sadece 2 bileşen ile yapılabilir.

Faktör Analizi

📌 Tanım

Faktör analizi, gözlenen değişkenler arasındaki ilişkileri açıklamak için **gizli (gözlenemeyen) faktörler** bulur.

Temel amacı, değişkenleri gruplara ayırarak her grubun arkasındaki ortak sebebi ortaya çıkarmaktır.

▶ Adımlar:

1. Korelasyon Matrisi İncelemesi
 - Birbirine yüksek korelasyon gösteren değişkenler belirlenir.
2. Faktör Sayısının Belirlenmesi
 - Kaiser kriteri ($\text{özdeğer} > 1$) veya Scree Plot grafiği kullanılır.
3. Faktörlerin Çıkarılması
 - İstatistiksel yöntemlerle gizli faktörler elde edilir.
4. Faktör Yüklerinin Yorumlanması
 - Hangi değişkenin hangi faktörle güçlü ilişkili olduğu belirlenir.
5. Rotasyon (Dönüştürme)
 - Faktörlerin daha anlamlı yorumlanabilmesi için döndürme (Varimax vb.) uygulanır.

6. Çok Değişkenli Sonuç Çıkarımı

💡 Örnek (Faktör Analizi)

Bir anket çalışmasında 20 farklı soruya verilen yanıtlar inceleniyor.

Analiz sonucunda:

- Faktör 1 → "Müşteri Memnuniyeti" (hız, ilgi, kalite ile ilgili sorular)
- Faktör 2 → "Fiyat Algısı" (fiyat, indirim, uygunluk ile ilgili sorular) olarak tanımlanabilir.
- Böylece onlarca soru, 2 temel faktör ile özetlenmiş olur.

🎯 PCA vs Faktör Analizi

Özellik	PCA	Faktör Analizi
Amaç	Boyut İndirgeme	Gizli yapıları ortaya çıkarma
Kullanılan bilgi	Toplam varyans	Ortak varyans
Yorumu	Yeni değişkenler	Gizli faktörler
Bağımsızlık	Bileşenler bağımsızdır	Faktörler bağımsız olmayabilirler

❤️ Bölüm Özeti

- **PCA:** Bilgiyi kaybetmeden değişken sayısını azaltmak için
- **Faktör Analizi:** Değişkenleri gruplara ayırarak altta yatan yapıları bulmak için kullanılır.

Her ikisi de çok değişkenli veri setlerinde daha anlaşılır ve verimli analiz yapmayı sağlar.

Bölüm 7

Sınıflama Yöntemleri

Verileri doğru kategorilere ayırmak, tahmin ve analiz süreçlerinin temel adımlarından biridir. Bu bölümde, sınıflandırma ve kümeleme yöntemlerini öğrenerek verilerden anlamlı gruplar oluşturmayı keşfeliyoruz.

7. Sınıflama Yöntemleri

Veri bilimi uygulamalarında çoğu zaman amacımız, gözlemleri önceden **tanımlanmış sınıflara** ayırmaktır.

Örneğin:

- Bir e-postanın spam olup olmadığını belirlemek
- Bir hastanın belirli bir hastalığa sahip olup olmadığını tahmin etmek
- Bir müşterinin ürünü satın alıp almayacağını öngörmek

Bu tür problemler **sınıflama yöntemleri** ile çözülür.

Sınıflama Nedir?

📌 Tanım:

Sınıflama, etiketli veriler (yani her gözlemin hangi sınıf'a ait olduğu bilinen veriler) kullanılarak yeni gözlemlerin sınıfını tahmin etme sürecidir.

Örnek:

Bir banka, geçmiş müşteri verilerine göre kredi başvurusu yapan bir kişinin “Onay” veya “Red” sınıfına gireceğini tahmin etmek istiyor.

- Bağımsız değişkenler: Gelir, kredi puanı, yaş
- Bağımlı değişken: Başvuru sonucu (Onay / Red)

Diskriminant Analizi

📌 Tanım

Diskriminant analizi, gözlemleri ait oldukları grplara ayırmak için kullanılan istatistiksel bir yöntemdir.

Amaç, sınıflar arasındaki ayrimı maksimize eden bir fonksiyon bulmaktır.

▶ Türleri

1. Doğrusal Diskriminant Analizi (LDA):

- Sınıflar arasındaki ayrimı doğrusal bir sınırla yapar.
- Değişkenlerin varyans-kovaryans yapısının aynı olduğu varsayıılır.

2. Karesel Diskriminant Analizi (QDA):

- Sınıflar arasında doğrusal değil, eğrisel ayrim yapılabılır.
- Varyans-kovaryans yapısının farklı olmasına izin verir.

7. Sınıflama Yöntemleri

💡 Örnek

Bir üniversite, öğrencilerin ders başarı durumlarını (“Başarılı” / “Başarısız”) tahmin etmek istiyor.

- Bağımsız değişkenler: Devam oranı, ödev notları, sınav notları
- Diskriminant fonksiyonu, öğrencilerin hangi sınıfı gireceğini tahmin eder.

Kümeleme Yöntemleri

📌 Tanım:

Kümeleme, verileri benzerliklerine göre **önceden sınıf bilgisi olmadan** gruplandırma yöntemidir.

Sınıflama ile farkı: Kümelemede sınıf etiketleri bilinmez, algoritma bunları kendisi oluşturur.

⚠️ Yaygın Kümeleme Algoritmaları

1. K-Means:

- Belirli sayıda küme (k) seçilir.
- Veri noktaları, en yakın küme merkezine atanır.
- Merkezler güncellenir ve işlem tekrarlanır.

2. Hiyerarşik Kümeleme:

- Tüm gözlemler tek tek başlar ve benzer olanlar birleştirilerek kümeler oluşturulur (veya tersi).

💡 Örnek (Kümeleme)

Bir perakende şirketi, müşterilerini alışveriş alışkanlıklarına göre segmentlere ayırmak istiyor.

- **Veriler:** Harcama miktarı, alışveriş sikliği, kategori tercihleri
- **K-Means ile analiz yapıldığında:**
 - Küme 1: Sık alışveriş yapan yüksek harcama grubu
 - Küme 2: Orta sıklıkta alışveriş yapan düşük harcama grubu
 - Küme 3: Seyrek alışveriş yapan özel gün müşterileri

7. Sınıflama Yöntemleri

⌚ Bölüm Özeti

- **Sınıflama:** Etiketli veri ile yeni gözlemleri doğru sınıfaya yerleştirme süreci
- **Diskriminant Analizi:** Sınıflar arasındaki ayrimı istatistiksel olarak optimize etme
- **Kümeleme:** Etiketsiz verileri benzerliklerine göre grupperleme

Bu yöntemler, hem tahmin hem de veri keşfi aşamalarında önemli rol oynar.

Bölüm 8

Kapanış

Kısa bir teşekkür ve
yönlendirici kapanış ile
yolculuğu noktalıyoruz.

8. Kapanış

Genel Değerlendirme

Bu rehberde, temel istatistik kavramlarını veri bilimi perspektifinden ele aldık. Ölçüm tekniklerinden hipotez testlerine, dağılımlardan sınıflama yöntemlerine kadar birçok konuyu sade bir dille inceledik.

İstatistik, yalnızca geçmişi analiz etmek değil; geleceği daha doğru tahmin etmek ve veriye dayalı kararlar almak için güçlü bir araçtır.

Teşekkürler

- Bu çalışma, BTK Akademi – Veri Bilimi için Temel İstatistik eğitiminden esinlenerek hazırlanmıştır.

Kaynaklar

- BTK Akademi – Veri Bilimi için Temel İstatistik
- Kişisel notlar ve ek açıklamalar

Hazırlayan

- Yavuzcan ÇOLAK
Veri Odaklı Mühendis & İçerik Üreticisi