İstanbul
GEDİK
Üniversitesi

İstanbul GEDİK Üniversitesi

Veri Madenciliği

Öğr. Gör. Zeki ÇIPLAK

Veri Madenciliğinde Kullanılan Bazı Kavramlar

Veri Madenciliğinde Bazı Kavramlar

- ***Öznitelik ve Gözlem***

- ***Değişken Türleri***
  - *Kategorik Değişkenler*
  - *Sürekli Değişkenler*

- ***Öğrenme Türleri***
  - *Denetimli (Supervised)*
  - *Denetimsiz (Unsupervised)*
  - *Pekiştirmeli (Reinforcement)*

- ***İstatistiksel Kavramlar***

Öznitelik & Gözlem

Öznitelikler

Gözlemler

id	first_name	last_name	email	gender
1	Tiler	Willford	twillford0@blogtalkradio.com	Male
2	Brinn	McCamish	bmccamish1@goodreads.com	Female
3	Giff	Grivori	ggrivori2@wiley.com	Male
4	Franky	Ghiraldi	fghiraldi3@tinyurl.com	Male
5	Brad	Toun	btoun4@printfriendly.com	Male
6	Mordecai	Gleader	mgleader5@linkedin.com	Male
7	Tandi	Lineham	tlineham6@slashdot.org	Female
8	Marlin	Klimochkin	mklimochkin7@squidoo.com	Male
9	Jermaine	Delahunty	jdelahunty8@narod.ru	Female
10	Hedda	Caltun	hcaltun9@studiopress.com	Female

- **Öznitelik (Attribute):** Veri setindeki her bir verinin sahip olduğu özelliktir. Kısaca özellik (feature) veya değişken (variable) olarak da isimlendirilebilir.

- **Gözlem:** Veri setindeki her bir satır kayıttır. Nesne (object) olarak da isimlendirilebilir.

Değişken Türleri

Temel Değişken Tipleri

Veri Madenciliği projelerinde kullanılan temel değişken tipleri

- **Kategorik (Sınıflayıcı) Değişkenler**
  - *Nominal (İsimsel)*
  - *Binary (İkili)*
  - *Ordinal (Sıra Gösteren)*

- **Sürekli Değişkenler**
  - *Tam Sayılı (Integer)*
  - *Aralıklı-Ölçümlendirilmiş (Interval-Scaled)*
  - *Oranlı Ölçümlendirilmiş (Ratio-Scaled)*

Kategorik Değişkenler

- **Nominal (İsimsel) Değişken:** Üst-Ast ilişkisi olmayan, sayısal olarak ifade edilse bile, sadece etiket görevi gören nitel değişkendir. Sayısal değere sahip nominal değişken, **matematiksel amaçlı kullanılmaz**.

- **Nominal değişkenlere örnekler:**
  - ***Meslek*** *(Mühendis, Doktor, Avukat)*
  - ***Coğrafi Bölge*** *(Marmara, Ege)*
  - ***Göz Rengi*** *(Yeşil, Mavi, Siyah)*
  - ***Eşya*** *(Koltuk, Kanepe, Sandalye)*
  - ***Ülke Adı*** *(Türkiye, Azerbaycan)*
  - ***Takım Adı*** *(Bir, 90, Three)*

Kategorik Değişkenler

- **Binary (İkili) Değişken:** Nominal değişkenlerin özel bir formudur. Sadece iki değer alabilen değişkenlerdir.

- **Binary değişkenlere örnekler:**
  - ***Cinsiyet*** *(Erkek, Kadın)*
  - ***Sonuç*** *(Pozitif, Negatif)*
  - ***Değerlendirme*** *(Olumlu, Olumsuz)*
  - ***Tutum*** *(Doğru, Yanlış)*
  - ***Durum*** *(İyi, Kötü)*
  - ***Renk*** *(Siyah, Beyaz)*

Kategorik Değişkenler

- **Ordinal (Sıra Gösteren) Değişken:** Üst-Ast ilişkisi olan, sıraya dizildiğinde anlam ifade eden, derecelendirme yapılabilen nitel değişkendir.

- **Ordinal değişkenlere örnekler:**
  - ***Rütbe*** *(Er, Onbaşı, Teğmen, Yüzbaşı, Binbaşı)*
  - ***İdari*** *(Memur, Şef, Müdür Yardımcısı, Müdür)*
  - ***Sıra*** *(Birinci, İkinci, Üçüncü)*
  - ***Durum*** *(Kötü, Orta, İyi)*
  - ***Hava*** *(Soğuk, Ilık, Sıcak)*
  - ***Eğitim*** *(İlkokul, Ortaokul, Lise, Üniversite)*

Sürekli Değişkenler

- **Tam Sayılı (Integer) Değişken:** Alacağı değerler 0, 1, 2 gibi tam sayı olabilen değişkendir. Kesikli değer alır. Bu değişkenler **matematiksel amaçla kullanılabilirler**.

- **Tam Sayılı değişkenlere örnekler:**
  - ***Toplam*** *(10, 200, 3000)*
  - ***Kilo*** *(73, 59, 68, 90)*
  - ***Yaş*** *(34, 25, 19, 42, 56, 78)*
  - ***Adet*** *(33, 20, 15, 76)*

Sürekli Değişkenler

- **Aralıklı Ölçümlenmiş (Interval) Değişken:** Ordinal **değişkenin özelliklerini içerir** ve değerler arasındaki farklar matematiksel olarak belirlenebilir. Kullanılan ölçüm için, belirli bir **yokluk anlamına gelmeyen sıfır** ölçme düzeyi bulunabilir.

- **Interval değişkenlere örnekler:**
  - ***Hava Durumu*** *(-10, 0, 5)*
  - ***Saat*** *(23, 00, 01)*
  - ***Yıl*** *(-3500, 0, 1923)*

Sürekli Değişkenler

- **Oranlı Ölçümlenmiş (Ratio) Değişken:** Interval değişkenlere benzemekle birlikte, bu değişkende **sıfır bir başlangıç noktasıdır** ve tüm ölçüm birimlerinde aynı anlamı taşır. Bu değişkenler **negatif değer alamaz**.

- **Ratio değişkenlere örnekler:**
  - ***Kilogram*** *(0, 1.5, 20, 35.7)*
  - ***Gram*** *(0, 100.2, 1500)*
  - ***İnç*** *(0, 13.5, 15.6, 21)*
  - ***Saniye*** *(0, 30, 60, 120)*
  - ***Dakika*** *(0, 1, 5, 10)*
  - ***Uzunluk*** *(0, 1.33, 25.45)*

Öğrenme Türleri

Öğrenme Türleri

1. *Denetimli Öğrenme (**Supervised**)*

2. *Denetimsiz Öğrenme (**Unsupervised**)*

3. *Pekiştirmeli Öğrenme (**Reinforcement**)*

Öğrenme Türleri

1. Denetimli Öğrenme (Supervised Learning):
Algoritmaya, Girdi ile Çıktı birlikte verilir.

Sınıflandırma (Classification) problemleri denetimli öğrenme türüne girer.

Sınıflandırmada, veriyi eğitmek için kullanılan veri setinin dışındaki bir verinin, **sınıfı tahmin edilmeye çalışılır**.

Sınıflar sınırlı sayıdadır ve **önceden belli**dir.

Bu yüzden, oluşturulan sınıflandırma modelinin başarısı, test verisi ile ölçülebilir.

Öğrenme Türleri

Sınıflandırma

Sarı değişkenler girdi, yeşil değişken ise çıktıdır.

Çanak Uzunluğu	Çanak Genişliği	Yaprak Uzunluğu	Yaprak Genişliği	Tür
7,9	3,8	6,4	2	virginica
7,7	3,8	6,7	2,2	virginica
7,7	2,6	6,9	2,3	virginica
7,2	3	5,8	1,6	virginica
7,1	3	5,9	2,1	virginica
7	3,2	4,7	1,4	versicolor
6,9	3,1	4,9	1,5	versicolor
6,9	3,2	5,7	2,3	virginica
6,9	3,1	5,4	2,1	virginica
6,9	3,1	5,1	2,3	virginica
6,8	2,8	4,8	1,4	versicolor
6,8	3	5,5	2,1	virginica
5,4	3,9	1,7	0,4	setosa
5,4	3,7	1,5	0,2	setosa
5,4	3,9	1,3	0,4	setosa
5,4	3,4	1,7	0,2	setosa
5,4	3,4	1,5	0,4	setosa

Öğrenme Türleri

2. Denetimsiz Öğrenme (Unsupervised Learning): Algoritmaya sadece Girdi verilir, Çıktı verilmez. Sınıflar algoritma tarafından tespit edilir.

Kümeleme (Clustering) problemleri, denetimsiz öğrenme türüne örnektir.

Kümelemede, veri setinin sınıfları belli değildir. Her bir verinin birçok özniteliği, **kümeleme algoritmaları ve benzerlik ölçüleri** kullanılarak gruplar oluşturulur.

Grup içi benzerlik maksimum, gruplar arası benzerlik minimum olacak şekilde gruplar meydana getirilir.

Öğrenme Türleri

Kümeleme

Çanak Uzunluğu	Çanak Genişliği	Yaprak Uzunluğu	Yaprak Genişliği
7,9	3,8	6,4	2
7,7	3,8	6,7	2,2
7,7	2,6	6,9	2,3
7,2	3	5,8	1,6
7,1	3	5,9	2,1
7	3,2	4,7	1,4
6,9	3,1	4,9	1,5
6,9	3,2	5,7	2,3
6,9	3,1	5,4	2,1
6,9	3,1	5,1	2,3
6,8	2,8	4,8	1,4
6,8	3	5,5	2,1
5,4	3,9	1,7	0,4
5,4	3,7	1,5	0,2
5,4	3,9	1,3	0,4
5,4	3,4	1,7	0,2
5,4	3,4	1,5	0,4

Öğrenme Türleri

3. Pekiştirmeli Öğrenme (Reinforcement Learning): Deneyimlere bağlı olarak, kendini geliştirme şeklindeki öğrenme türüdür. Ödül tabanlı olarak çalışır.

Veri Madenciliğinde kullanılmaz. Daha çok makine öğrenmesi ve yapay zeka projelerinde kullanılır.

İnsan öğrenmesine en yakın öğrenme türüdür.

Pekiştirmeli Öğrenme Örneği: https://youtu.be/spfpBrBjntg

Sınıflandırma & Kümeleme

*Hangisinin **Sınıflandırma**,*
*hangisinin **Kümeleme** problemi olduğunu belirtelim.*

1. Spam e-postaların tespiti.
2. İnternetten toplanan haberlerin kategorilerinin tespiti.
3. Kan, tansiyon, çeşitli tahliller vb. sağlık değerlerine bakılarak diyabetli hastaların tespiti.
4. Müşteri bilgilerinden faydalanılarak, müşteri gruplarının oluşturulması.
5. Belli bir bölgedeki çiçeklerin, yaprak ve çanak büyüklüklerine bakılarak hangi türe ait olduğunun tespiti.
6. Öğrenci notlarına bakılarak, çalışkan-orta-tembel öğrencilerin tespiti.

İstatistiksel Kavramlar

İstatistiksel Kavramlar

Populasyon: Aynı veya benzer özelliklere (değişkenlere) sahip birimlerin (verilerin) oluşturduğu topluluğa denir.

Örneklem: Populasyonu en iyi şekilde temsil eden alt topluluktur.

İstatistiksel Kavramlar

Frekans: Bir seride, hangi değerden kaç kez tekrar edildiğini gösteren bir ölçüdür.

Örnek serimiz aşağıdaki gibi olsun:
12, 13, 45, 12, 27, 12, 36, 13, 27, 12, 58, 27

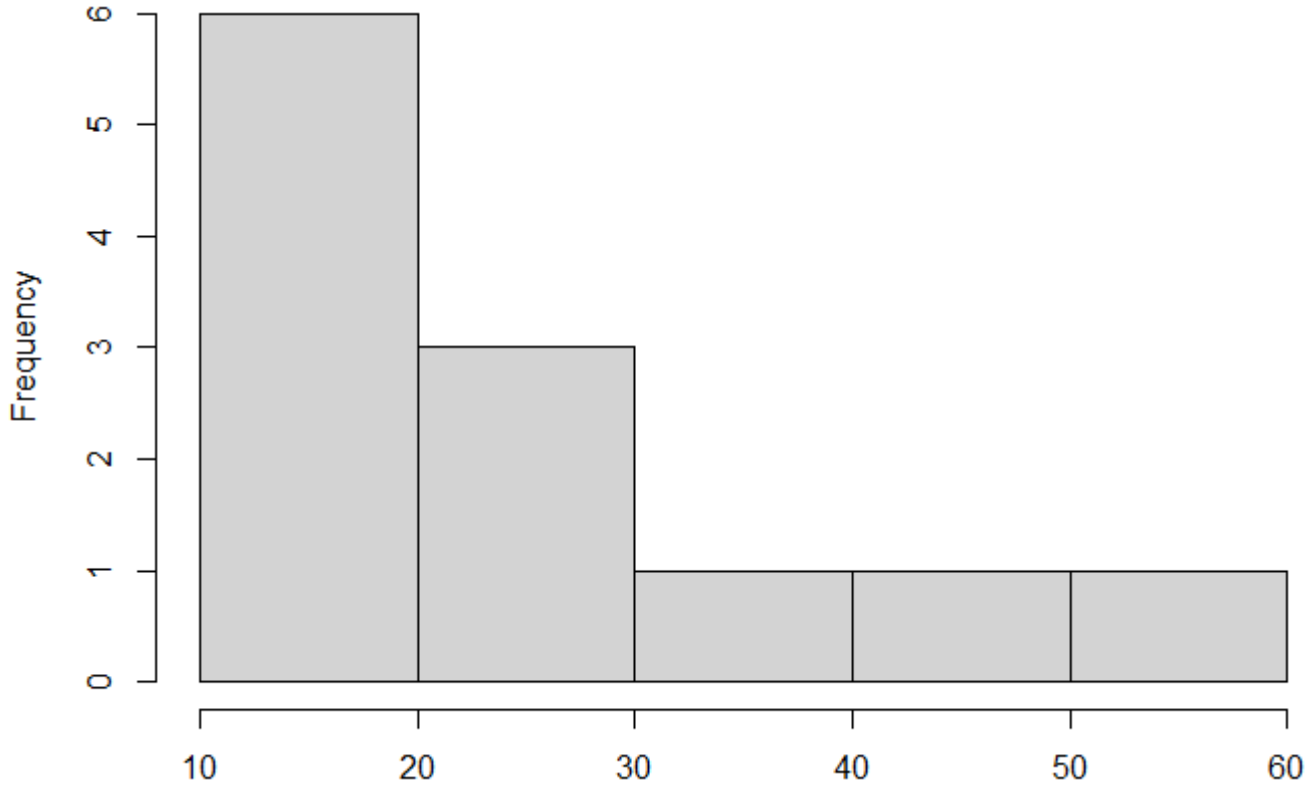
Bu serinin Frekans tablosu yandaki gibidir.

Serilerin frekans bilgileri kullanılarak, **Histogram Grafikleri** elde edilir.

Değer	Frekansı
12	**4**
13	**2**
27	**3**
36	**1**
45	**1**
58	**1**

İstatistiksel Kavramlar

Bir önceki sayfada frekans tablosu verilen serinin histogram grafiği de aşağıdaki gibidir.



İstatistiksel Kavramlar

Aritmetik Ortalama: Populasyon veya örneklem içerisindeki tüm gözlemlerin toplamının, toplam gözlem sayısına bölünmesi ile elde edilen niceliktir.

Medyan: Öncelikle tüm gözlemler, küçükten büyüğe doğru sıralanmalıdır. Toplam gözlem sayısı çift ise, **ortadaki iki değerin ortalaması** alınır. Toplam gözlem sayısı tek ise, medyan değeri **ortadaki değer**dir.

Aritmetik ortalama, Medyan değerine yakınsa; verilerin dağılımı için, simetriğe yakındır yorumu yapılabilir.

İstatistiksel Kavramlar

Örneklemimiz aşağıdaki değerlerden oluşsun:
1, 12, 9, 34, 26, 40

Aritmetik Ortalama = (1 + 12 + 9 + 34 + 26 + 40) / 6 = 20.333

Medyan değerini bulmak için önce seriyi sıralamak gerekir.
1, 9, 12, 26, 34, 40

Toplam gözlem sayısı 6 (yani çift) olduğu için medyan değeri, ortadaki iki değerin ortalamasıdır.

Medyan = (12 + 26) / 2 = 19

İstatistiksel Kavramlar

Mod: Veriler içinde en çok tekrar eden (**frekansı en büyük** olan) değer veya değerlerdir.

Aşağıdaki serinin modu 3'tür.
1, 64, 3, 72, 3, 90, 3, 85

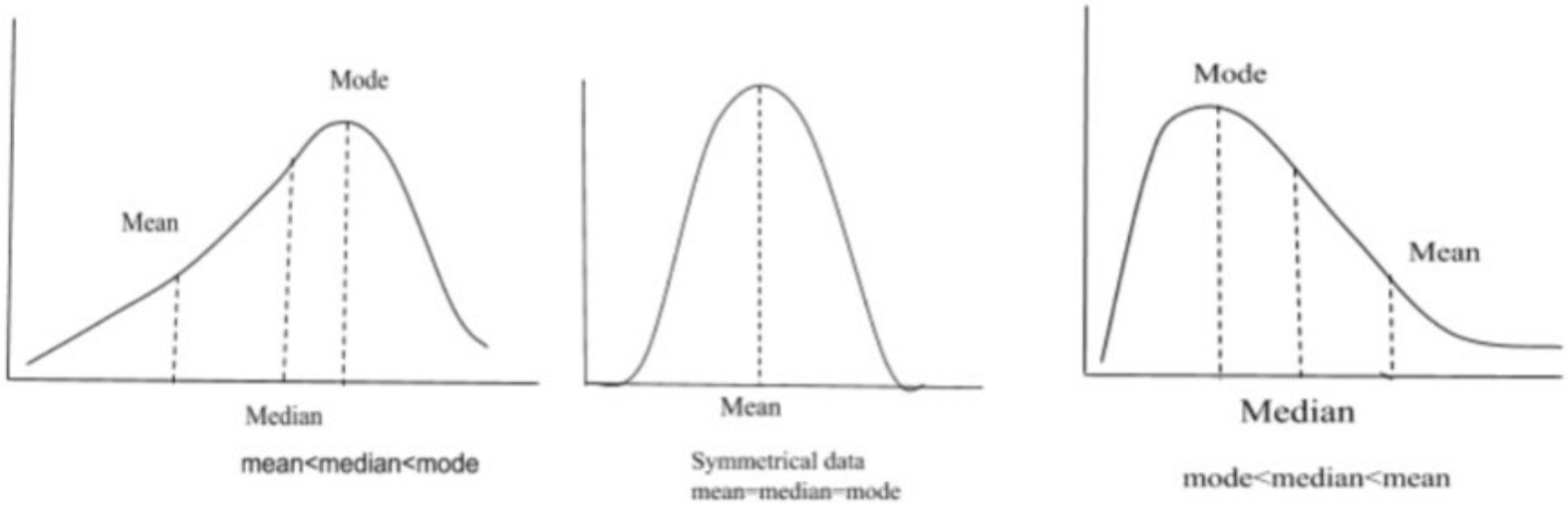
Aritmetik ortalama, medyan ve mod; bir serinin dağılımını anlamak için gerekli olmakla birlikte, tek başlarına yetersizdirler. Örneğin, aynı aritmetik ortalamaya sahip iki seri, bütünüyle farklı bir dağılım gösterebilir.

Aritmetik Ortalama (5, 10, 15, 20, 25, 30) = **17.5**
Aritmetik Ortalama (1, 4, 12, 13, 15, 60) = **17.5**

İstatistiksel Kavramlar

Aritmetik ortalama, medyan ve mod; birlikte değerlendirildiğinde ise daha faydalı olarak kullanılabilirler.

Aşağıdaki histogram grafikleri; aritmetik ortalama, mod ve medyanın değerine göre, **dağılımın nasıl olduğu** hakkında bilgiler verir.



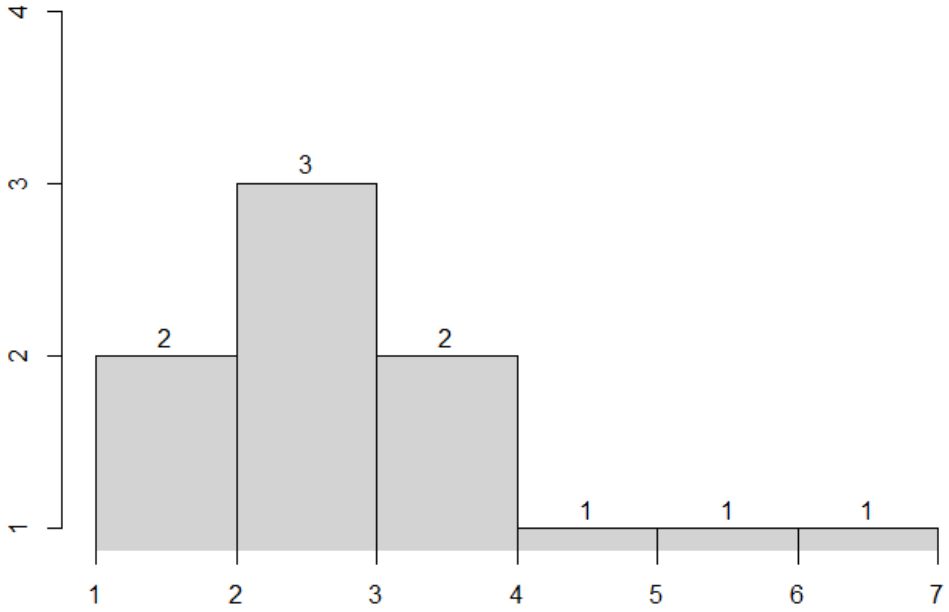
İstatistiksel Kavramlar

Verilerimiz aşağıdaki değerlerden oluşursa:
1, 2, 3, 3, 3, 4, 4, 5, 6, 7

Mod = 3 ; **Ortalama** = 3.8 ; **Medyan** = 3.5 bulunur.
Buna göre, **Ortalama > Medyan > Mod** demektir.

O halde vektörümüzdeki değerler, sağdan çarpık demektir.

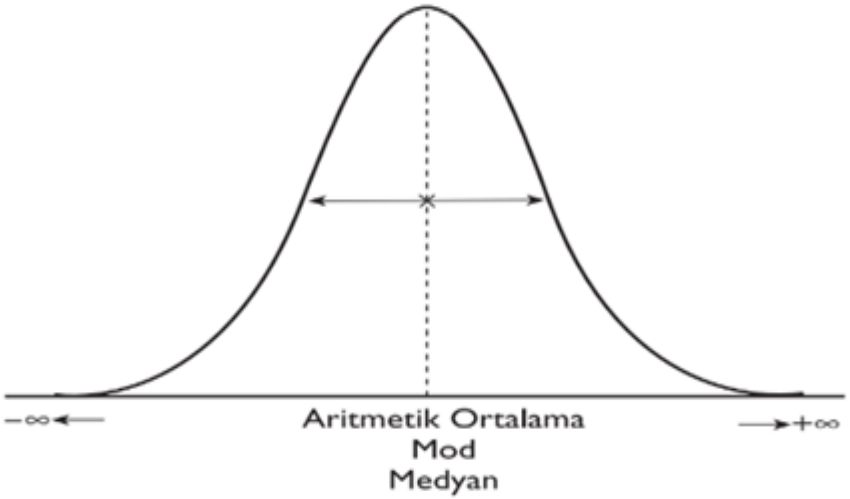
Histogram grafiği yandaki gibi olacaktır.



İstatistiksel Kavramlar

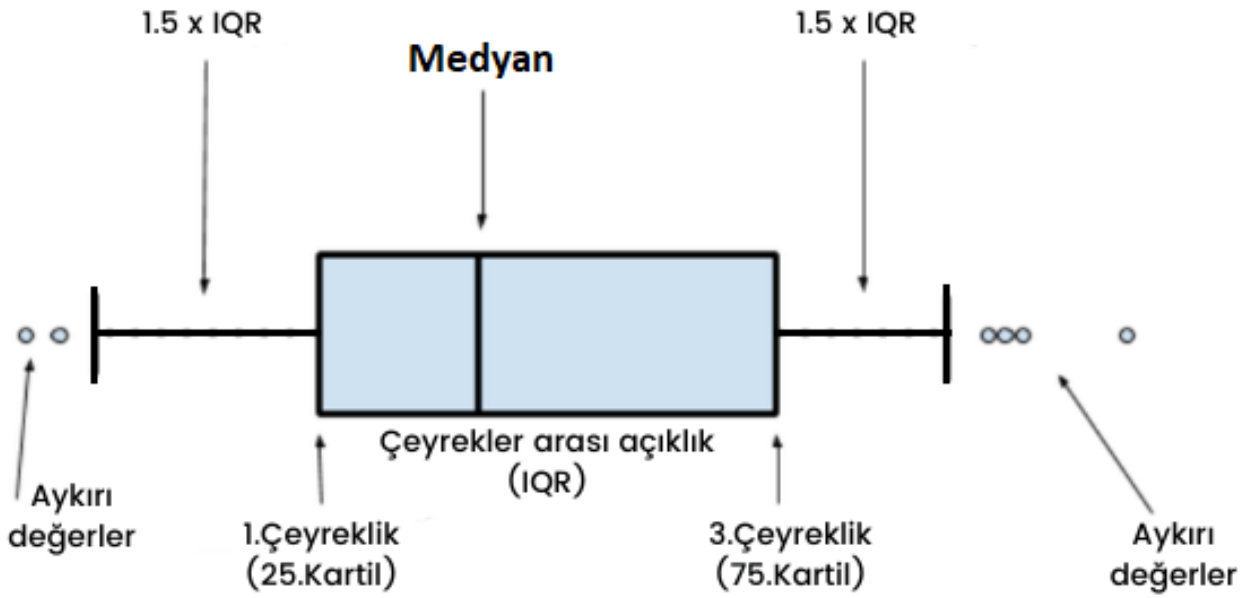
Normal Dağılım: Bir serideki değerlerin **aritmetik ortalaması, mod veya medyan değerleri birbirine eşit** ve dağılımın merkezinde ise, bu seri normal dağılım göstermiştir denir.

Bu sayılan değerlerin, birbirine yakın olmaları durumunda, normal dağılıma yakın bir dağılım görülebilmektedir. Normal dağılımın grafiği çan şeklindedir.



İstatistiksel Kavramlar

Kartiller (Çeyreklikler): Seri küçükten büyüğe doğru sıralandıktan sonra, seriyi **dört parça**ya ayıran değerlerdir. (Quantile)



%25 ve %75'e karşılık gelen değerler, Q1 ve Q3 olarak belirlenir. Bu değerler eşik değerin hesaplanmasında kullanılır. Eşik değerinin **dışındaki veriler, aykırı gözlem** kabul edilir. (IQR = Q3 – Q1)

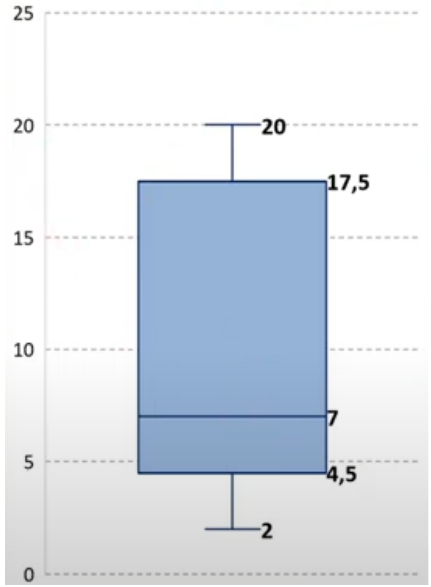
İstatistiksel Kavramlar

Şekildeki örnekte **IQR** = Q3 – Q1 = 17.5 – 4.5 = **13** bulunur.

Alt eşik değeri = 4.5 – 13 = **– 8.5**
Üst eşik değeri = 17.5 + 13 = **30.5** olur.

Serideki tüm değerler, eşik değerlerin arasında olduğu için, aykırı değerin olmadığı görülmüş olur.

(Uygulama dersinde, R ile örnekler yapılacaktır.)



Medyan

2 3 6 6 7 12 15 20 20

Q1 4.5

Q3 17.5

İstatistiksel Kavramlar

Değişim Aralığı (Range): Serinin en büyük gözlem değeri ile en küçük gözlem değeri arasındaki farktır.

Standart Sapma: Serideki her bir değerin, ortalamaya göre ne kadar saptığının genel bir ölçüsüdür. Matematiksel bakımdan en kuvvetli değişkenlik ölçüsüdür.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

İstatistiksel Kavramlar

Standart sapma küçüldükçe; değişkenliğin azaldığı, dolayısıyla gözlem değerlerinin birbirine daha çok yaklaştığı söylenebilir.

Standart sapma büyüdükçe; değerler arası değişkenlik artar ve değerler birbirinden uzaklaşmaya başlar.

Bir başka ifadeyle; belli bir serideki değerlerin, **ortalamadan sapma değerlerinin ortalaması**dır.

Varyans: Standart sapmanın karesidir. Standart Sapma, değerlerin ortalamadan ne kadar saptığını ifade ederken; varyans, değerlerin değişkenliğini tanımlar.

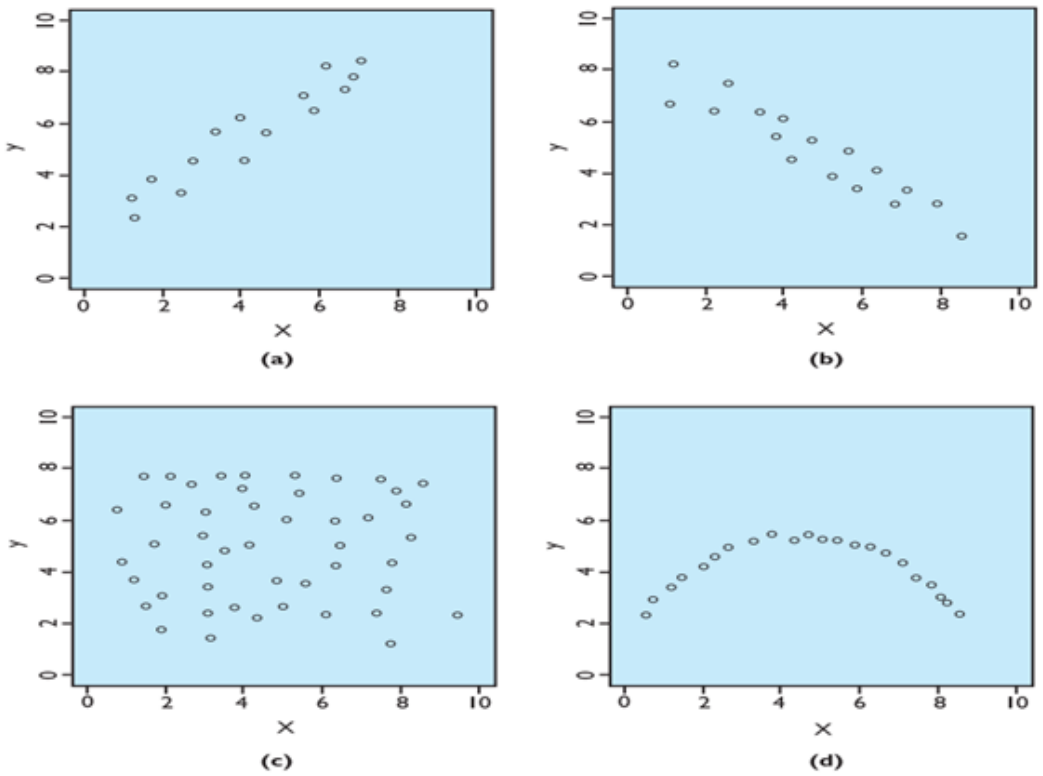
İstatistiksel Kavramlar

Kovaryans: İki değişken arasındaki ilişkinin değişkenlik ölçüsüdür. Sonucun **pozitif olması artan bir doğrusal ilişki**yi, **negatif olması azalan bir doğrusal ilişki**yi ve sıfır civarında olması ilişkinin olmadığını gösterir.

Korelasyon: İki değişken arasındaki ilişkiyi gösteren, ilişkinin boyutunu, anlamlı bir ilişki olup olmadığını, ayrıca ilişkinin yönünü ve şiddetini ifade eden istatistiksel bir ölçüdür. 1 ile -1 arası değer alır.

Korelasyon, **değişkenlerin arasındaki ilişki**yle ilgilenir. Kovaryans ise, **değişkenler arasındaki ilişkinin değişkenliği** ile ilgilenir.

İstatistiksel Kavramlar



(a) ve **(b)**'de değişkenler arasındaki ilişki doğrusaldır. **(c)**'de değişkenler arasında herhangi bir ilişki olmadığı görülür. **(d)**'de ise değişkenler arası ilişki eğriseldir.