Bilişimde İstatistiksel Analiz

Ders 2

Ali Mertcan KOSE Msc.
amertcankose@ticaret.edu.tr

İstanbul Ticaret Üniversitesi

İSTANBUL TİCARET
ÜNİVERSİTESİ

İstatistik Nedir?

İstatistik Verinin toplanması, düzenlenmesi, özetlenmesi ve analiz edilmesi ile ilgili çalışmaların yapıldığı bir alandır.

İstatistik bir veri bilimidir.

- Çok sayıda bilginin özetlenmesini sağlar
- Eldeki az sayıda veriden yola çıkarak öngörü yapar(Bütüne ilişkin çıkarımda bulunur).

Kitle ve Örneklem

Kitle: Üzerinden araştırma yaptığımız birimler topluluğuna *"Kitle"* denir.

Örneklem: Çekildiği Kitleyi temsil ettiği düşünülen ve evrene göre daha az sayıda birey ya da gözlemden oluşan alt kümedir. Basit bir ifade ile kitlenin bir bölümüne *örneklem* denir. (Kitleyi iyi temsil etmesi gerekir!)

Değişkenlerin Ölçme Düzeyi

- *Sınıflandırılmış (Nominal) Değişkenler*
- *Sıralanmış (Ordinal) Değişkenler*
- *Eşit Aralıklı Değişkenler*
- *Oranlı Değişkenler*

Not

- *Kesikli Değişkenler (Sayılarak Elde edilen)*
- *Sürekli Değişkenler (Ölçülerek Elde edilen)*

Verilerin Özetlenmesi

Bir araştırmacı araştırmaya başlamadan önce araştırma evrenini oluşturmalıdır. Araştırma evrenini(Kitle) belirlerken, birimlerin ortak *nitelik*, *mekan* ve *zaman* da olması önemli bir husustur. Bir sonraki aşamada belirlediğimiz kitle üzerinden, Kitleyi iyi temsil edecek örneklem çekilmesi gerekir. Böylece örneklem seçildikten sonra gözlem birimlerinden veriler elde edilir. Bu veriler üzerinden herhangi bir düzenleme, özetleme veya başka bir oynama yapılmadığı zaman *ham veri(raw data)* elde edilir.

Verilerin Özetlenmesi

```
##   npreg glu bp skin  bmi   ped age type
## 1     5  86 68   28 30.2 0.364  24   No
## 2     7 195 70   33 25.1 0.163  55  Yes
## 3     5  77 82   41 35.8 0.156  35   No
## 4     0 165 76   43 47.9 0.259  26   No
## 5     0 107 60   25 26.4 0.133  23   No
## 6     5  97 76   27 35.6 0.378  52  Yes
```

npreg: Gebelik sayısı, **glu:** Plazma glukoz konsantrasyonu, **bp:** diyastolik kan basıncı, **skin:**cilt kat kalınlığı, **bmi:** Beden kitle indeksi, **ped:** diyabetik pedigre fonksiyonu, **age:** yaş, **type:** diyabet statüsü

Verilerin Özetlenmesi

Verilerin anlamlı bilgiler verebilmesi ve bu bilgilerin açık, kolay anlaşılır olması için, verilerin belirli bir düzende sunulması, bazı matematiksel işlemler yardımlarıyla özetlenmesi için bazı yöntemlerin uygulanması gerekir.

1. Tablo oluşturulması(Frekans dağılım tabloların hazırlanması)
2. Grafik çizilmesi
3. Tanımlayıcı istatistiklerin hesaplanması

Tanımlayıcı istatistikler

- Tanımlayıcı istatistikler hesaplandığı örneklemin özelliklerini tek bir değer ile özetler
- Tanımlayıcı istatistikler veri setleri hakkında daha fazla bilgi edinmemizi sağlar.
- Kitleye ait parametler üzerinden tahmin yapılan ölçülerdir.

Tanımlayıcı istatistikler

- Merkezi eğilim Ölçüleri
  - Mod
  - Medyan
  - Aritmetik Ortalama
  - Diğer Ortalamalar (Geometrik Ortalama, Harmonik Ortalama, Karesel Ortalama)
- Dağılış Ölçüleri (Yayılış, Saçılış)
  - Açıklık
  - Çeyrekler arası açıklık
  - Çeyrek ayrılış
  - Ortalama ayrılış
  - Varyans ve Standart Sapma
  - Değişim Katsayısı
  - Basıklık ve Çarpıklık

Merkezi Eğilim ölçüleri

Verinin ağırlık noktasını(merkezini) gösterir.

Mod (Tepe değeri): Frekansı en fazla olan değerdir.

Ham veri $\rightarrow$ En büyük frekansı veren değer.

Sınıflandırılmış veri $\rightarrow$ mod$=$ L $+ \frac{fs}{fs+f\ddot{o}} \times$ c

L: mod sınıfının alt değeri

fs: mod sınıfından sonra gelen frekans

fö: mod sınıfından önce gelen frekans

c: sınıf aralığı

örnek:

x_i: 20, 40,40,20,60,80,20

Merkezi Eğilim ölçüleri

Medyan(Ortanca): Veriler küçükten büyüğe sıralandıktan sonra tam ortaya düşen değerdir. Bu nedenle medyana ikinci çeyrek (Q_2) denir.

Ham veri $\rightarrow$ n tek ise, M $= X_{(n+1)/2}$ n çift ise, M $= \frac{(Xn/2)+(Xn+2/2)}{2}$

Sınıflandırılmış veri $\rightarrow$ M$=$ L $+ \frac{c}{f} \times (\frac{N}{2}$-d)

L: medyan sınıfı alt sınıfı

f: medyan sınıfı frekansı

c: sınıf aralığı

d: medyan sınıfından önce gelen sınıfların frekansı

örnek:

i: 1 2 3 4 5

x_i: 20 60 40 80 50

Merkezi Eğilim ölçüleri

Aritmetik Ortalama(μ):

$(\mu) = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$

Sınıflandırılmış veri $\rightarrow \frac{1}{N} \sum_{i=f}^{m} f_j \overline{x_j}$

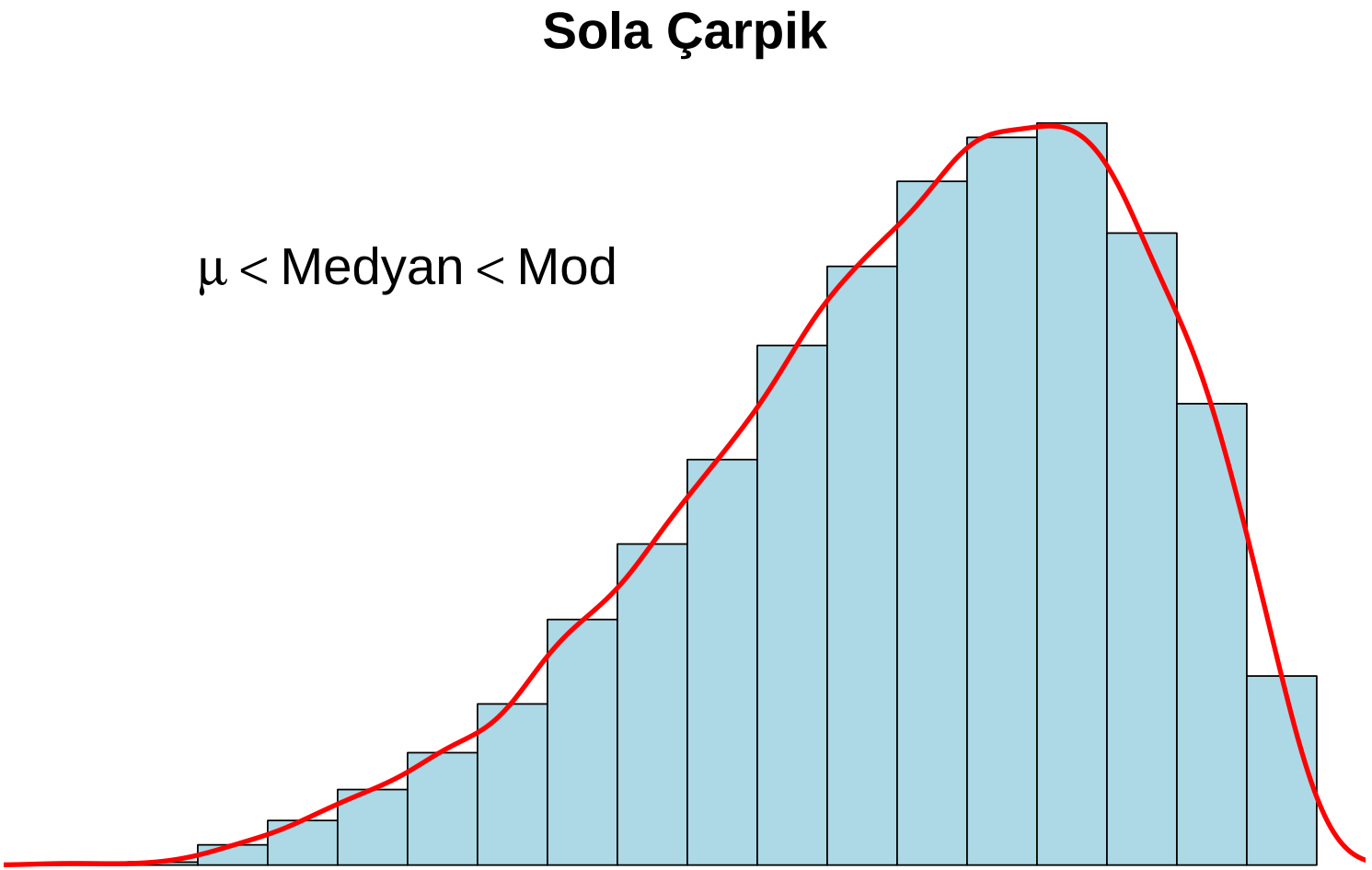
f(x) $\rightarrow \mu = \sum$ x $\times$ f(x) $\Longrightarrow$ Kesikli değişken

$\mu = \int$ x $\times$ f(x) $\Longrightarrow$ Sürekli değişken

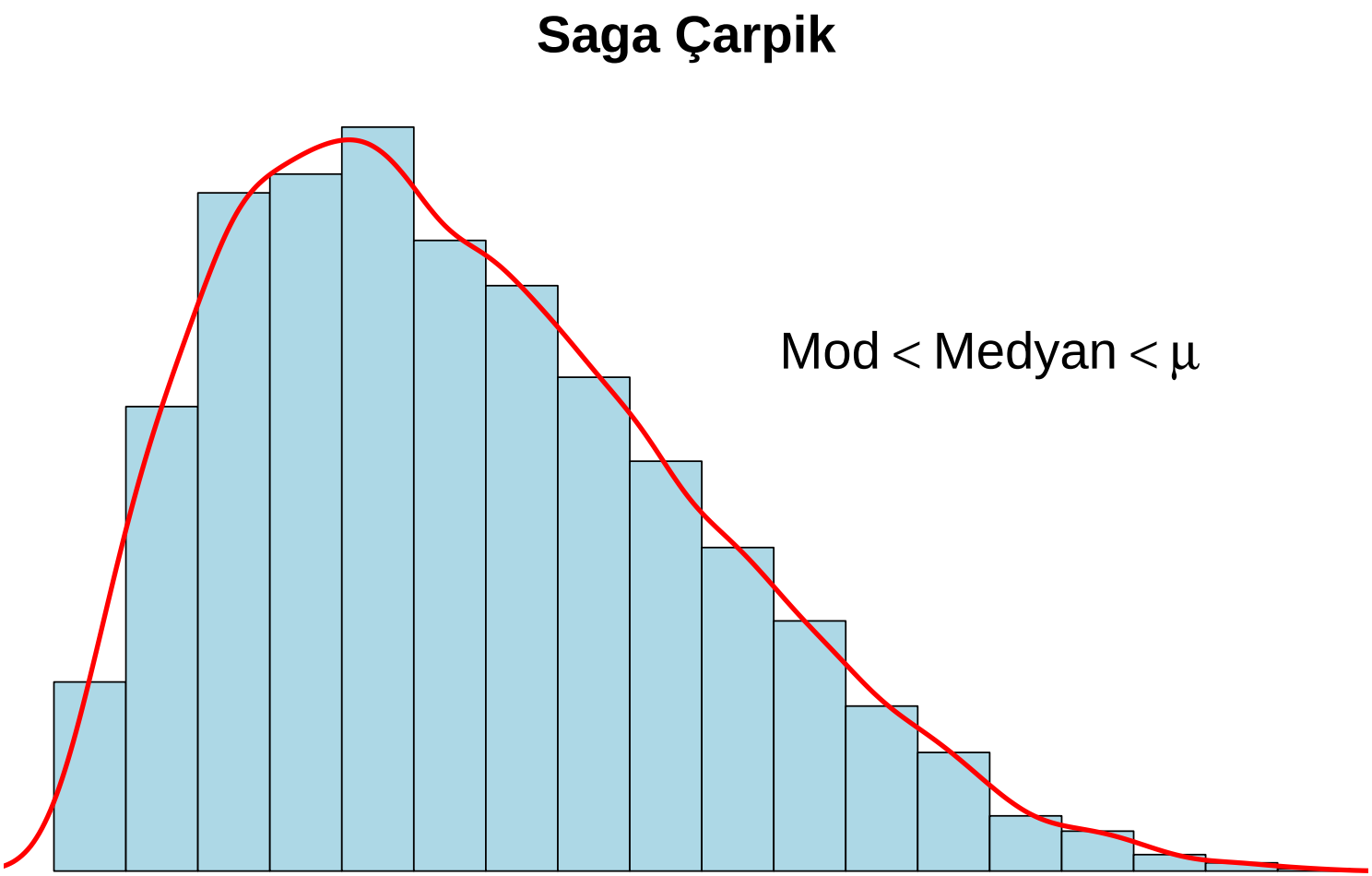
Not

A.O'nın dezavantajı: Uç değerden aşırı derece etkilenmektedir. Bu yüzden Aykırı değer varsa A.O'dan kaçınılır.

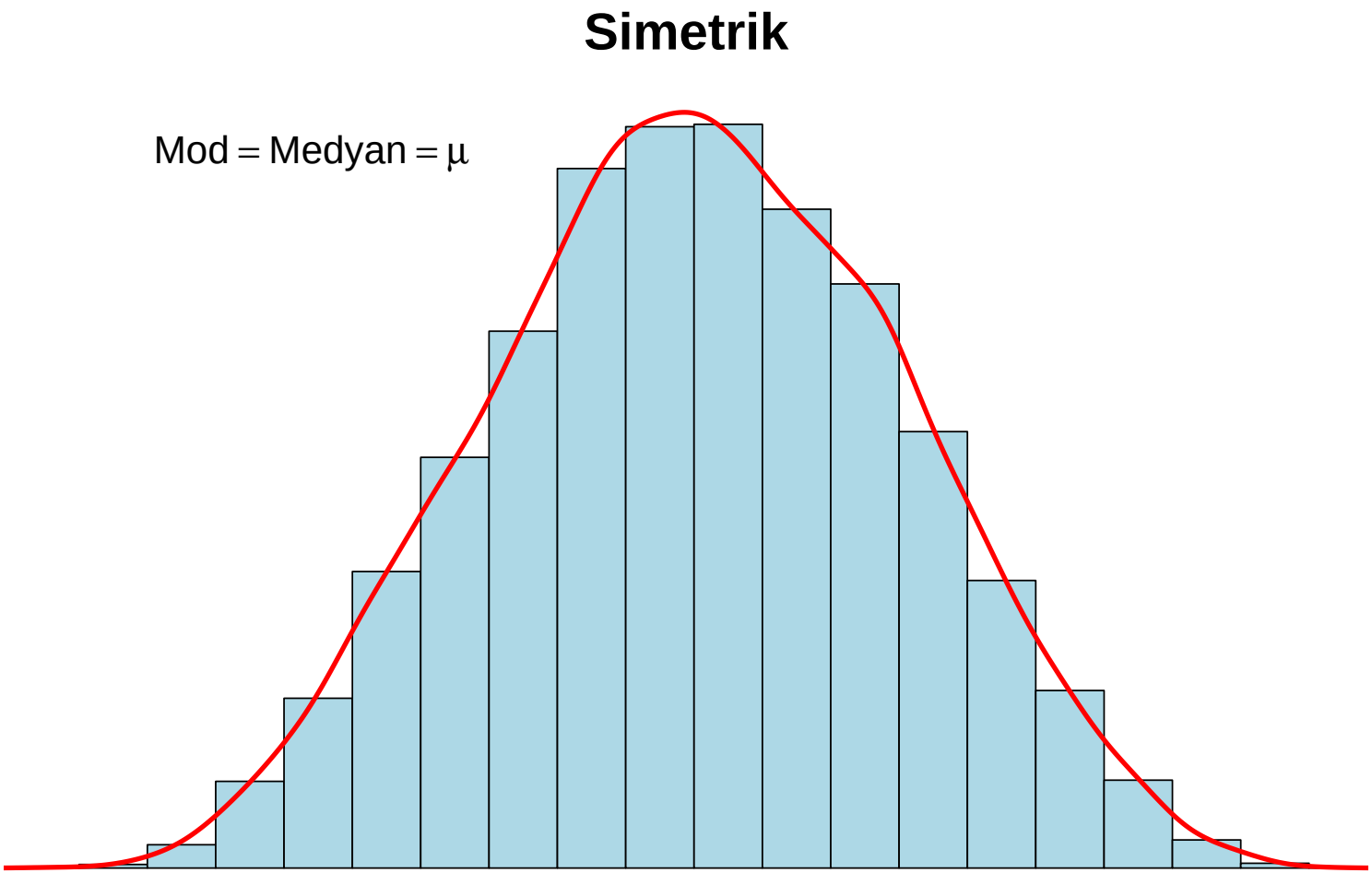
Merkezi Eğilim ölçüleri



Merkezi Eğilim ölçüleri



Merkezi Eğilim ölçüleri



Merkezi Eğilim ölçüleri

Diğer Ortalamalar

Geometrik Ortalama:

$G = \sqrt[N]{X_1.X_2...X_N}$

Birim zamanda kullanılan orana mutlak farktır.

Bu oran sadece oran ölçeğinde kullanılır.

Herhangi bir gözlem değeri 0 olamaz.

Merkezi Eğilim ölçüleri

Harmonik Ortalama:

$$H = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_N}}$$

Uç değerden en az etkilenendir.

$X_i > 0$

Oran ölçeğinde kullanılır.

Birim mesafe birim zamanda yapılan işlerin ort.'da kullanılır.

$A \geq G \geq H$

Karesel Ortalama:

$$K = \sqrt{\frac{\sum x^2}{N}}$$

$\sigma^2 = K^2 - \mu^2$

Dağılış ölçüleri

- Dağılış ölçüsü verilerin merkezden uzaklığını gösteren ölçülerdir.
- Verilerin birbirinden uzaklığını gösteren bir ölçüdür.
- Bir homojenlik ölçüsüdür.
- Dağılış ölçüsü 0'dan büyük değer alır.
- Dağılış ölçüsü 0'a yaklaştıkça homojenlik artar.
- Bir sabitin dağılış ölçüsü 0'dır.

Dağılış ölçüleri

1. Açıklık

R = X_N (En büyük) - X_1 (En küçük)

- Gözlem sayısının az olduğu durumlarda kullanılması önerilir.
- $N \leq 5$ ise kullanılır.

2. Çeyrekler Arası Açıklık

Ç.A.= Q_3 - Q_1

$Q_1 = \frac{N}{4}$

$Q_3 = \frac{3N}{4}$

3. Çeyrek Ayrılış

Q = $\frac{Q_3 - Q_1}{2}$

Dağılış ölçüleri

4. Ortalama Ayrılış

O.A.= $\frac{1}{N}\sum^{N}(|x_i - \mu|)$

Sınıflandırılmış veri $\rightarrow \frac{1}{N}\sum^{N} f_j\,|\overline{x_j} - \mu|$

5. Varyans ve Standart Sapma

$\sigma^2 = \frac{1}{N}\sum^{N}(x_i - \mu)^2 \rightarrow$ Kitle için

$s^2 = \frac{1}{n-1}\sum^{n}(x_i - \overline{x})^2 \rightarrow$ Örneklem için

sınıflandırılmış veri $\sigma^2 = \rightarrow \frac{1}{N}\sum^{N} f_j(\overline{x_i} - \mu)^2$

Standart sapma $= \sqrt{s^2}$ veya $\sqrt{\sigma^2}$

6. Değişim Katsayısı

c= $\frac{\sigma}{\mu} \times 100 \rightarrow$ Kitle için

$\hat{c}$= $\frac{s}{\overline{x}} \times 100 \rightarrow$ Örneklem için

Dağılış ölçüleri

Not

Gözlem değerleri arasında önem farklılığı varsa(olasılık) ağırlıklı ortalama kullanılır.
σ^2 = Kitle varyansı
s^2 = Örneklem varyansı
$s_{\overline{x}}$ = İstatistiğin varyansının karekökü

Çarpıklık ve Basıklık

Çarpıklık $(\alpha_3) = \frac{\frac{1}{N}\sum(x_i-\mu)^3}{\sigma^3}$

$\alpha_3 >$0 sağa çarpık

$\alpha_3 < 0$ sola çarpık

$\alpha_3 = 0$ simetrik

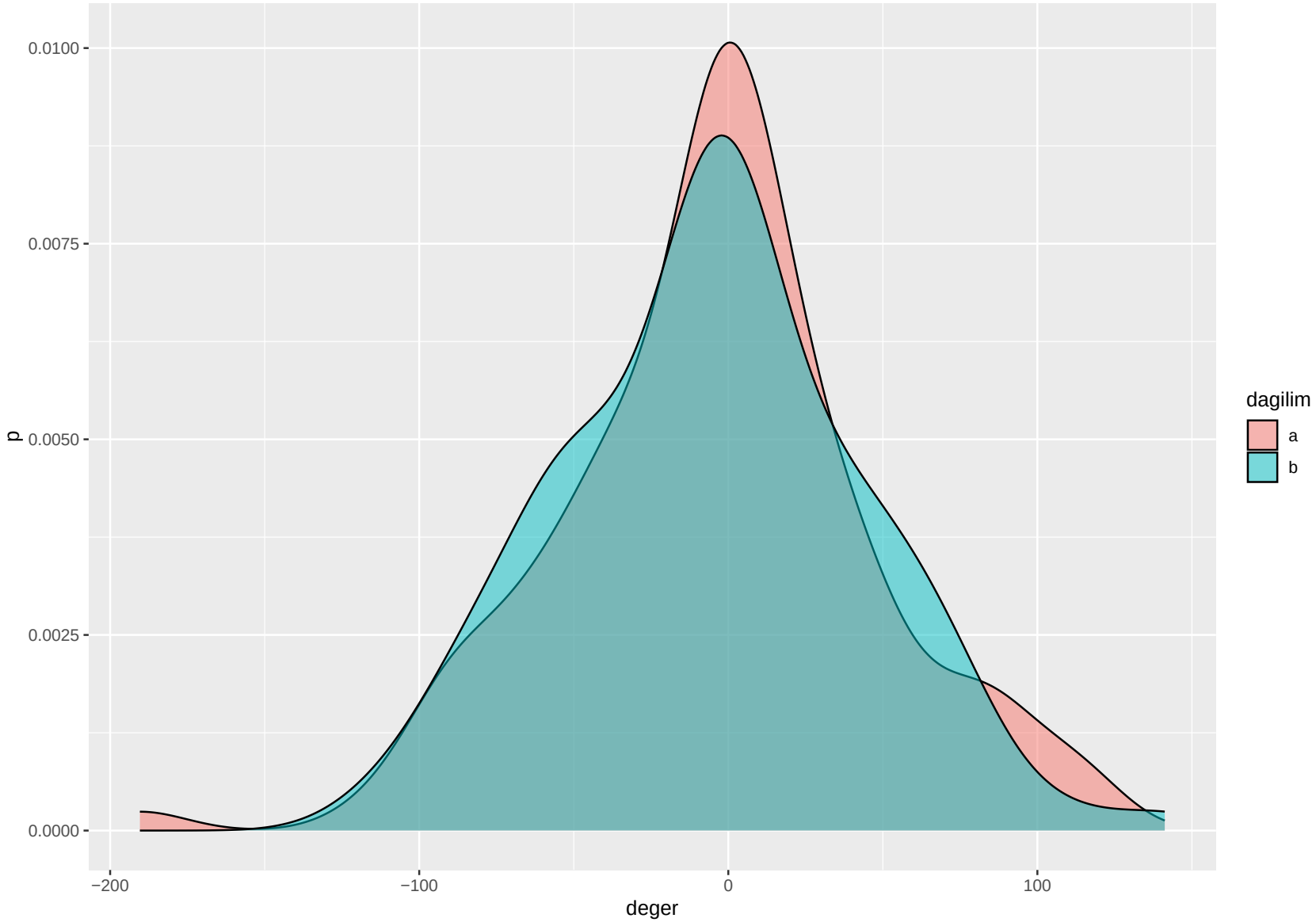
Basıklık $(\alpha_4) = \frac{\frac{1}{N}\sum(x_i-\mu)^4}{\sigma^4}$

$\alpha_4 > 3$ sivri

$\alpha_4 < 3$ basık

$\alpha_4 = 3$ simetrik

Çarpıklık ve Basıklık



Dağılış ölçüleri ile İlgili Örnekler

Örnek: 10, 10, 5 15, 10 **örneklem** verisinin,

1. Mod
2. Medyan
3. A.o.
4. standard sapması ve varyansını bulunuz.

Dağılış ölçüleri ile İlgili Örnekler

Çözüm:

Mod (en çok tekrar eden) = 10

Medyan = 5 10 10 10 15 (Küçükten büyüğe sıralanır) sonrasında $(\frac{n+1}{2}) = 6/2 = 3$ ve 3. değer 10

A.o= 5+10+10+10+15/5 = 10

A.o= Medyan = Mode olduğu için simetrik bir dağılım.

varyans $\frac{\sum(x-\mu)^2}{n-1} = \frac{(5-10)^2+(10-10)^2...+(15-10)^2}{5-1} = 12{,}5$

Standard sapma $= \sqrt{(12,5)} = 3{,}53$

Dağılış ölçüleri ile İlgili Örnekler

Örnek: 2,9,3,7,5,6,10 dizine ait aşağıdaki değerleri bulunuz.

1. Açıklık
2. Çeyrekler arası genişlik
3. Çeyrek Ayrılış
4. Ortalama
5. Varyans ve standart sapma, Çarpıklık ve Basıklık.
6. Harmonik Ortalama
7. Geometrik Ortalama
8. Kareli Ortalama

Dağılış ölçüleri ile İlgili Örnekler

Çözüm:

2,3,5,6,7,9,10

Açıklık= R = X_N (En büyük) - X_1 (En küçük)

10-2 =8

Çeyrekler arası genişlik = Q_3 - Q_1

9-3=6

Çeyrek ayrılış = $\frac{Q_3 - Q_1}{2}$

$6/2 = 3$

A.Ortalama = $\frac{1}{N} \sum_{i=1}^{N} x_i$

$(2+3+5+\ldots+10)/7 = 6$

Dağılış ölçüleri ile İlgili Örnekler

Geometrik Ortalama $= \sqrt[N]{X_1.X_2...X_N}$

$\sqrt[7]{2 * 3 * ... * 10}$ =5,27

Harmonik Ortalama= $\frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+...+\frac{1}{X_N}}$

$\frac{7}{\frac{1}{2}+\frac{1}{3}+...+\frac{1}{10}}$ =4,50(A>G>H)

Kareli Ortalama $= \sqrt{\frac{\sum x^2}{N}}$

$\sqrt{\frac{2^2+3^2+...+10^2}{7}} = \sqrt{43,42} = 6,59$

Standart sapma $= \sqrt{\frac{1}{N}\sum^N (x_i - \mu)^2}$

$\sqrt{\frac{1}{7}(2-6)^2 + (3-6)^2 + ... + (10-6)^2} = 2,72$

Dağılış ölçüleri ile İlgili Örnekler

Varyans= $\frac{1}{N}\sum^{N}(x_i-\mu)^2$

$\frac{1}{N}\sum^{N}(x_i-\mu)^2 = 7{,}42$

Çarpıklık= $\frac{\frac{1}{N}\sum(x_i-\mu)^3}{\sigma^3}$

0

Basıklık = $\frac{\frac{1}{N}\sum(x_i-\mu)^4}{\sigma^4}$

1,75 $\rightarrow$ 1,75-3= -1,25

Dağılış ölçüleri ile İlgili Örnekler

Örnek:

Kan Basıncı	Hasta Sayısı
115-124	4
125-134	5
135-144	5
145-154	7
155-164	5
165-174	4
175-184	5

Dağılış ölçüleri ile İlgili Örnekler

Yukarıdaki tabloda yer alan veri seti üzerinden

a) Mod
b) Medyan
c) Aritmatik Ortalama
d) Varyans ve Standart Sapma
e) Açıklık, Çeyrekler arası Açıklık, Çeyrek ayrılış ve değişim katsayısını bulunuz.

Dağılış ölçüleri ile İlgili Örnekler

Çözüm:

Mod: Mod$= \mathrm{L} + \frac{fs}{fs+f\ddot{o}} \times \mathrm{c} = 145 + \frac{5}{10} \times 10 = 150$

Medyan: Medyan $= \mathrm{L} + \frac{c}{f} \times (\frac{N}{2}\text{-d}) = 145 + \frac{10}{7} \times (\frac{35}{2}\text{-14}) =$ 150

Dağılış ölçüleri ile İlgili Örnekler

frekans	xi	xi.fj	fj(xi-xort)^2
4	119,5	478	3668,89
5	129,5	647,5	2057,55
5	139,5	697,5	528,97
7	149,5	1046,5	0,57
5	159,5	797,5	471,83
4	169,5	678	1554,61
5	179,5	897,5	4414,69
Toplam:35		**Toplam:5242,5**	**Toplam:12697,14**

Dağılış ölçüleri ile İlgili Örnekler

A.O: $\frac{1}{N}\sum_{i=f}^{m} f_j\overline{x_j}$

$= \frac{1}{35}$ (119,5 $\times$ 4 + 129,5 $\times$ 5 + 139,5 $\times$ 5 + 149,5 $\times$ 7 + 159,5 $\times$ 5 +169,5 $\times$ 4 +179,5 $\times$ 5) $= 897{,}5/35 = 149{,}78$

Varyans: $\frac{1}{N}\sum^{N} f_j(\overline{x_i} - \mu)^2 = 12697\ /35 = 362{,}77$

Standart sapma $\sqrt{\sigma^2} = \sqrt{362,77} = 19{,}04$

Dağılış ölçüleri ile İlgili Örnekler

Açıklık: R = X_N (184) - X_1 (115) =69

Çeyrekler arası Açıklık: Q_1= L + $\frac{c}{f}$ × ($\frac{N}{4}$-d)

= 125 + $\frac{10}{5}$ × ($\frac{35}{4}$-4) = 134,5

Q_3 = L + $\frac{c}{f}$ ×($\frac{3N}{4}$-d)

= 165 + $\frac{10}{4}$ × 3 × ($\frac{35}{4}$-26) = 165,62

Q_3 - Q_1 = 165,62 - 134,5 = 31,12

Çeyrek Ayrılış: $\frac{Q_3-Q_1}{2}$ = 31,12/2 = 15,56

Değişim Katsayısı $\frac{\sigma}{\mu}$ × 100 = $\frac{19,04}{149,978}$ * 100 = 12,70

ÖDEV 1

Hasta id:	1	2	3	4	5	6	7	8	9	10
Tedavi Süresi:	12	11	12	6	11	11	8	5	5	5

Yukarıda yer alan veri seti üzerinden

a) Mod
b) Medyan
c) Aritmatik Ortalama
d) Varyans ve Standart Sapma
e) Açıklık, Çeyrekler arası Açıklık, Çeyrek ayrılış ve değişim katsayısını bulunuz.

ÖDEV 2

Kolesterol_Duzeyi	Erkek_sayisi
80-119	13
120-159	150
160-199	442
200-239	299
240-279	115
280-319	34
320-359	9
360-399	5

ÖDEV 2

Yukarıdaki tabloda yer alan veri seti üzerinden

a) Mod
b) Medyan
c) Aritmatik Ortalama
d) Varyans ve Standart Sapma
e) Açıklık, Çeyrekler arası Açıklık, Çeyrek ayrılış ve değişim katsayısını bulunuz.