Veri Madenciligi
COZUM HW1

Yrd. Doc. Dr. Cengiz Orencik

November 8, 2017

1. Elimizde analiz etmek istedigimiz verinin *yas* niteligini icerdigini dusunelim. Bu *yas* verisi kucukten buyuge sral halde su sekilde olsun: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   (a) Verinin ortalama ve ortanca (medyan) degerleri nedir?
   veri eleman says ($N = 27$)

   $$mean(ortalama) = \\ = \frac{13+15+16+16+19+20+20+\ldots+35+36+40+45+46+52+70}{27} \\ = 29.962$$

   toplam 27 veri var, tam ortadaki 14uncu eleman olan 25

   (b) Verinin mode degeri nedir? Hangi mode yapsnda oldugunu yorumlayin (unimodal, trimodal, vs.).
   Hem 25 hem de 35 verinin icinde 4er defa geciyor. Diger veriler 3 veya daha az sayida geciyor. Bu yuzden en sk gecen elemanlar 25 ve 35 mod degerleridir. Iki mod degeri oldugu icin bimodal yada genel olarak birden fazla mod degeri oldugu icin multimodal yapisindadir.

   (c) Verinin orta-aralik (midrange) degeri nedir?
   $avg(min, max) = \frac{13+70}{2} = 41.5$

   (d) Verinin (yaklask olarak) ilk ceyrek (Q1) ve ikinci ceyrek (Q3) degerleri nedir?
   27 elemani 4 parcaya bolmek istersek ortadaki 14 uncu eleman medyan. 1inci ile 13uncu arasi Q_1 degeri 7inci elman 20
   15inci ve 27inci arasi yani Q_3 degeri 21 inci eleman 35 olarak bulunur.

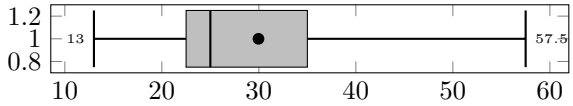
(e) Veri sapan deger (outlier) iceriyor mu? Aciklayin.

Once 1.5 IQR hesaplanir: $1.5 \times (35 - 20) = 22.5$
alt limit $Q_1 - 1.5IQR = 20 - 22.5 = -2.5$
ust limit $Q_3 + 1.5IQR = 35 + 22.5 = 57.5$

-1.75ten kucuk veya 57.5ten buyuk degerler sapan veri, dolayisiyla verimizdeki tek sapan deger 70.

(f) Veriyi kutu grafigi (boxplot) olarak ifade ediniz.

5 value representation $min(Q_1 - 1.5IQR), max(Q_3 + 1.5IQR), Q_1, Median, Q_3$



(g) Verinin standart sapmasini (σ) bulun.

$$\sigma^2 = \frac{\sum_i (x_i - 29.96)^2}{27} = 161.3$$

$standartdeviation(\sigma) = \sqrt{161.3} = 12.7$

2. Bir hastanede 18 rastgele secilmis yetiskin uzerinde yapilan testte yas ve yag oranlar uzerine asagidaki sonuc alinmistir:

age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
% fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) yas ve yag oran degerlerini iki ayri kutu grafigi (boxplot) olarak ifade ediniz.

yas zaten sirali, yag oranlarda siralanmali

sirali yag orani: $\{7.8, 9.5, 17.8, 25.9, 26.5, 27.2, 27.4, 28.8, 30.2, 31.2, 31.4, 32.9, 33.4, 34.1, 34.6, 35.7, 41.2, 42.5\}$

$median_{age} = avg(50, 52) = 51$

$median_{fat} = avg(30.3, 31.2) = 30.75$

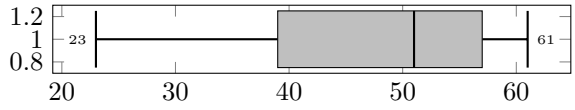
(yaklask bastan ve sondan 5inci eleman)

$$Q_{1_{age}} = 39 \; Q_{1_{fat}} = 26.5 \; Q_{3_{age}} = 57 \; Q_{3_{fat}} = 34.1$$

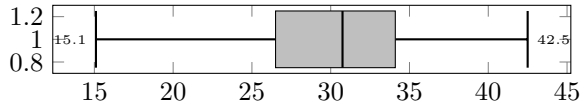
yas verisinde sapan deger yok ama yag orani verisinde var: 1.5IQR = (34.1 - 26.5) * 1.5 = 11.4

alt sinir $26.5 - 11.4 = 15.1$.

AGE



FAT RATE



(b) Korelasyon katsays r'yi hesaplayin. Bu iki nitelik pozitif mi yoksa negatif yonde mi birbirleriyle alakaldr?

$$r_{age,fat} = \frac{\sum_i (age_i - mean_{age})(fat_i - mean_{fat})}{N \sigma_{age} \sigma_{fat}}$$

$mean_{age} = 46.44, \; mean_{fat} = 28.78 \; \sigma_{age} = 12.84 \; \sigma_{fat} = 8.99 \; N = 18$

$$\frac{1700.33}{18 \times 12.84 \times 8.99} = 0.818$$

0.818 sifirdan buyuk ve 1 e cok yakin oldugu icin kuvvetli bir sekilde pozitif korelasyon vardir.