Numeričke karakteristike obilježja

Numeričke karakteristike obilježja

- Pokazatelji centralne tendencije
- Pokazatelji rasipanja oko centralnih vrijednosti
- Pokazatelji oblika raspodjele

Pokazatelji centralne tendencije

- Aritmetička sredina za negrupisane podatke

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Aritmetička sredina za grupisane podatke

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}$$

Primjer

- Aritmetička sredina je (1*5+3*6+…+3*10)/19=151/19=7,947

Ocjene	Apsolutne frekvencije
x_i	f_i
5	1
6	3
7	4
8	2
9	6
10	3

- Aritmetička sredina je (3*7+7*11+…+2*27)/30=474/30=15.8

Broj stabala	Apsolutne frekvencije
l_i	f_i
[5, 9)	3
[9, 13)	7
[13, 17)	9
[17, 21)	5
[21, 25)	4
[25, 29)	2

Medijana

- Za obilježje *X* medijana M_e je vrijednost takva da je jednak broj vrijednosti obilježja koja su manja od medijane i onih koja su veća od medijane.
- Populaciju čini 19 studenata, a registruju se njihove ocjene iz statistike. Dobijeni su podaci: 6, 7, 6, 8, 9, 7, 9, 9, 10, 9, 7, 10, 6, 9, 10, 5, 7, 8, 9. (M_e=8)

Neparan broj podataka: $M_e = x_{\frac{n+1}{2}}$;

Paran broj podataka: $M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$;

Medijalni interval

- L_{me}, donja granica medijalnog intervala
- K_{me}, redni broj medijalnog intervala
- f_{me}, frekvencija medijalnog intervala
- d_{me}, širina medijalnog intervala

$$M_e = L_{me} + \frac{\frac{N}{2} - \sum_{i=1}^{K_{me}-1} f_i}{f_{me}} \cdot \Delta$$

Primjer

- Obim populacije N=10+14+…+2=100

Prodato(kom)	Br. prodavnica
0,1 - 25	10
25,1 - 50	14
50,1 - 75	40
75,1 - 100	24
100,1 - 125	10
125,1 - 150	2

Medijalni interval

$$M_e = L_m + \frac{\frac{n}{2} - K_{m-1}}{f_m} \cdot \Delta = 50 + \frac{50 - 24}{40} \cdot 25 = 66,25$$

Mod

- Mod M_o obilježja X je vrijednost koja se najčešće pojavljuje u nizu podataka. Za podatke 0, 2, 1, 0, 0, 7, 3, 2, 1, 0 mod je $M_o = 0$. Skup podataka 6, 7, 6, 8, 9, 7, 10, 9, 10, 6, 10, 5, 7, 8, 9 ima četiri moda 6, 7, 9, 10. Skup podataka 1, 2, 3, 4, 5, 6 nema mod.
- Za grupisane podatke odredi se modalni interval – interval čija je frekvencija najveća

$$M_o = L_m + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \cdot \Delta$$

Primjer

Prodato(kom)	Br. prodavnica
0,1 - 25	10
25,1 - 50	14
50,1 - 75	40
75,1 - 100	24
100,1 - 125	10
125,1 - 150	2

← Modalni interval

$$M_o = L_m + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \cdot \Delta$$

$$= 50 + \frac{40 - 14}{(40 - 14) + (40 - 24)} \cdot 25 = 65,47619 \approx 65,48$$

Primjer

- Da li je New York bogata država? Prosječan prihod u državi New York je osmi po redu među američkim državama, blizu susjednog Connecticuta i New Jersyua, koje su prva i druga po redu. Medjutim, dok su medijane prihoda u državama Connecticut i New Jersy treća i peta po redu, New York je na 27. mjestu.

Kvantili

- Za obilježje *X* kvantil reda *p*, u oznaci M_p, dijeli niz vrijednosti obilježja y_1, y_2, ..., y_N u odnosu 100p% prema 100(1-p)%, odnosno 100p% je manje od M_p, a 100(1-p)% je veće od M_p.
  - Medijana je kvantil reda 0.5

- Kvartili
  - $Q_1 = M_{0.25}$
  - $Q_2 = M_{0.5}$
  - $Q_3 = M_{0.75}$
  - Intervali $(-\infty, Q_1)$, (Q_1, Q_2), (Q_2, Q_3), (Q_3, ∞) sadrže po 25% populacije

- Slično se definišu kvintili, decili itd.

Primjer

- Odrediti kvartile za skup podataka: 0, 1, 1, 3, 0, 0, 0, 1, 2, 3, 4, 1, 1, 0, 0, 2, 3.
  - Poređati podatke po veličini: 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 4
  - Medijana 9. po redu, Q_2 =1
  - Prvi kvartil je sredina 4. i 5. podatka, Q_1 = 0
  - Treći i četvrti kvartil je sredina 13. i 14. podatka, Q_3 = 2.5

Kvantilni interval

- Kvantilni interval je prvi interval čija je kumulativna frekvencija veća ili jednaka od *N*p*
  - *L_p, donja granica kvantilnog intervala*
  - *F_{p-1}, kumulativna frekvencija intervala prije kvantilnog*
  - *f_p, frekvencija kvantilnog intervala*
  - *d_p, širina kvantilnog intervala*

$$P_i = L_p + \frac{i \cdot \frac{N}{100} - \sum_{i=1}^{K_p - 1} f_i}{f_p} \cdot \Delta$$

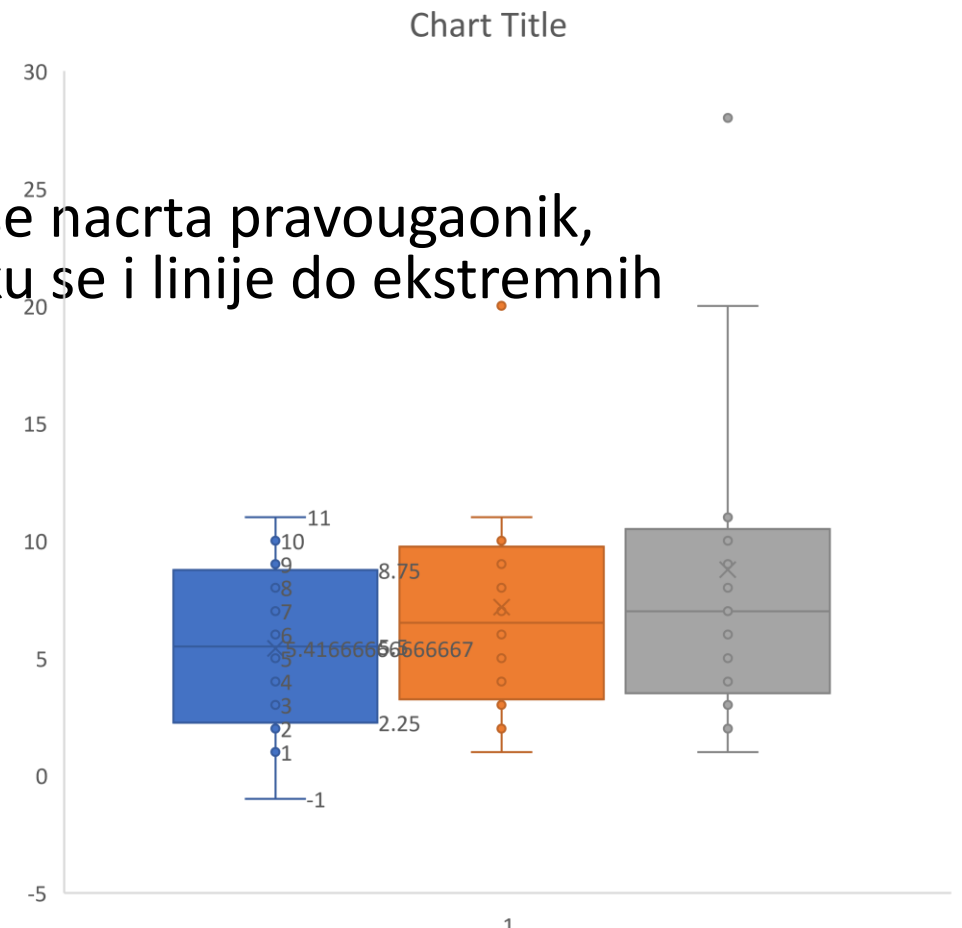
Primjer

- Odrediti kvartile
  - Za p=0.25, Np=575*0.25=143.75, pa prvi kvartil pripada interval [0,10), Q_1=0+((143.75-0)*10)/229=6.277
  - Q_2=12.773
  - Q_3=19.585

I_i	f_i	x_i	F_i
[0,10)	229	5	229
[10,20)	211	15	440
[20,30)	93	25	533
[30,40)	35	35	568
[40,50)	7	45	575
∑	575		

Box & Whiskers grafik

- Ovaj grafik formira se na sljedeći način
  - Odredi se medijana
  - Odrede se prvi i treći kvartil
  - Odrese se najveća i najmanja vrijednost
  - Na liniju se ucrataju svi ovi podaci, oko kvartila se nacrta pravougaonik, medijana se označi vertikalnom linijom, a povuku se i linije do ekstremnih vrijednosti
- Dat je skup podataka:
  - V1: -10 1 2 3 4 5 6 7 8 9 10 11
  - V2: 1 2 3 4 5 6 7 8 9 10 11 20
  - V3: 1 2 3 4 5 6 7 8 9 10 11 20 28



Pokazatelji rasipanja, disperzije

- Raspon populacije je razlika između najveće i najmanje vrijednosti obilježja
- Interkvartilni razmak, IQR = Q3 – Q1
- Srednje apsolutno odstupanje
- Disperzija, varijansa, srednje kvadratno odstupanje
- Standardna devijacija, kvadratni korijen iz disperzije
- Koeficijent varijacije, $\sigma/m*100\%$

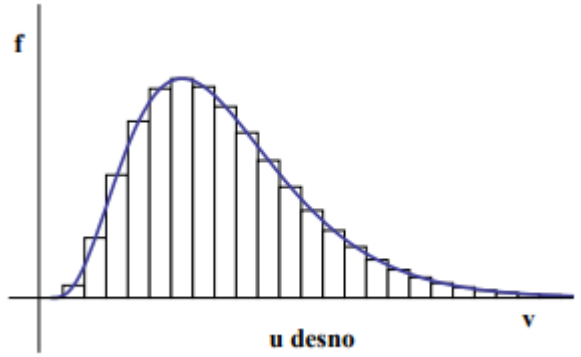
Primjer

l_i	f_i	x_i	$x_i f_i$	$(x_i-m)^2 f_i$	$\lvert x_i-m \rvert f_i$						6	14
[5, 9)	3		7	21	232.32	26.4		R	23			13	6
[9, 13)	7		11	77	161.28	33.6		D	4.48			9	25
[13, 17)	9		15	135	5.76	7.2		σ^2	30.29333			14	11
[17, 21)	5		19	95	51.2	16		σ	5.503938			10	13
[21, 25)	4		23	92	207.36	28.8		V	34.83505			18	21
[25, 29)	2		27	54	250.88	22.4						13	12
∑	30			474	908.8	134.4						11	23
												5	12
			m	15.8								15	22
												17	12
												20	16
												28	22
												19	14
												16	17

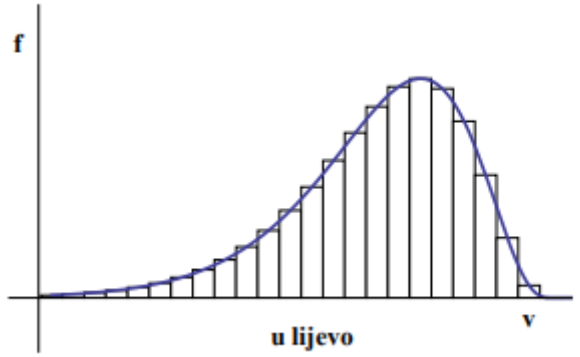
R	23	MAX(L1:M15)-MIN(L1:M15)
D	4.48	F8/B8
σ^2	30.29333	E8/B8
σ	5.503938	SQRT(I4)
V	34.83505	I5/D10*100

Pokazatelji oblika raspodjele

- $m \leq M_e \leq M_o$



- $M_o \leq M_e \leq m$



Pravilo tri sigme

- Čebiševljeva teorema - procenat od ukupnog broja podataka unutar intervala $m \pm k\sigma$, pri čemu je k konstanta, barem $1-1/k^2$
  - Za proizvoljni skup podataka s aritmetičkom sredinom m, te standardnom devijacijom σ, procenat ukupnog broja podataka unutar intervala $m \pm 2\sigma$ je barem 75%, $m \pm 3\sigma$ je barem 89%
  - Za podatke iz normalne raspodjele

σ	68,26894921371%
2σ	95,44997361036%
3σ	99,73002039367%
4σ	99,99366575163%
5σ	99,99994266969%
6σ	99,99999980268%
7σ	99,99999999974%

z-score

- Standardizacija vrijednosti obilježja, $z = (x - m)/\sigma$
- Ako neka vrijednost ima veliki, pozitivan z-score znači da je veća od većine vrijednosti tog obilježja
- Ako neka vrijednost ima mali, negativan z-score znači da je manja od većine vrijednosti tog obilježja
- Ako je z-score = 0, onda je ta vrijednost aritmetička sredina
- Vrijednosti obilježja koje imaju z-score između -2 i -1 nalaze se u intervalu $(m-2\sigma, m-\sigma)$