3. Numeričke metode opisa podataka

Vrste numeričkih opisnih mjera

Mjere – opisuju svojstva nekog skupa podataka.

- mjere srednje vrijednosti – opisuju središte razdiobe, odnosno položaj oko kojeg se gomilaju podaci.
- mjere raspršenja (disperzije) – mjere varijabilnost podataka.
- mjere položaja – opisuju relativni položaj nekog podatka u odnosu na ostale podatke.
- mjere asimetrije

Mjere srednje vrijednosti

Najčešća je **aritmetička sredina,** $\bar{x}$.

Aritetička sredina skupa od n podataka $x_1, x_2, \ldots, x_n$ definira se kao

$$\bar{x} = \frac{Suma\ svih\ podataka}{Broj\ podataka} = \frac{\sum_i x_i}{n}.$$

Sljedeća mjera srednje vrijednosti je **medijan, M**.
Medijan je vrijednost sa svojstvom da je pola podataka manje ili jednako njoj, a pola podataka veće ili jednako njoj.

Primjer 3.1. Nađite medijan za sljedeći skup podataka: 7, 4, 3, 5, 3.

Rješenje:
Najprije poredajmo podatke po veličini:

3, 3, 4, 5, 7.

Budući da raspolažemo s neparnim skupom podataka, izbor medijana je očit. To je broj točno u sredini (po veličini) uređenog skupa podataka, i.e.

$$M = 4.$$

∎

Primjer 3.2. Pretpostavimo da nam je zadan paran skup podataka: 5, 7, 3, 1, 4, 6. Nađite medijan.

Rješenje:

Poredajmo opet podatke po veličini:

$$1, 3, 4, 5, 6, 7.$$

Očito da izbor medijana u ovom slučaju nije jednoznačan. Svaki broj između 4 i 5 zadovoljava definiciju medijana.
Međutim, dogovor je da se u tom slučaju izabere aritmetička sredina dvaju srednjih podataka. Dakle,

$$M = \frac{4+5}{2} = 4.5\,.$$

∎

Medijan M se za skup od n podataka $x_1, x_2, \ldots, x_n$ definira na sljedeći način.

Za neparan n	–	medijan je podatak u sredini, odnosno podatak na rednom mjestu $(n+1)/2$ (pri čemu su podatci poredani po veličini).
Za paran n	–	medijan je jednak aritmetičkoj sredini podataka na rednom mjestu $n/2$ i $n/2+1$ (pri čemu su podatci poredani po veličini).

Sljedeća mjera srednje vrijednosti je **mod**.

Mod je vrijednost s najvećom frekvencijom. Ako su podatci grupirani po razredima (intervalima), mod definiramo kao središte razreda s najvećom frekvencijom, a taj razred nazivamo **modalnim razredom**.

Mod (za razliku od drugih spomenutih mjera) ima smisla i za kategorijalne podatke. Npr. rezultati prodaje ljetnih majica u trgovini *Kupi* prikazani su u sljedećoj tablici.

veličina	frekvencija
S	9
M	30
L	16
XL	40
XXL	13
$\sum$	108

Mod je XL.

Kako opisati prosječnog Hrvata?

Primjer 3.3. Prema podacima Statističkog zavoda, prosječan Hrvat je star 39 g, zaposlen je s plaćom od 5 300kn, banci je dužan 27 000 kn, ima srednju stručnu spremu i zove se Ivan.

U ovom primjeru su korištene mjere srednje vrijednosti. Koje?

Rješenje:

- Aritmetička sredina - starost, plaća, dug banci;
- Mod - zaposlenost, stupanj obrazovanja, ime.

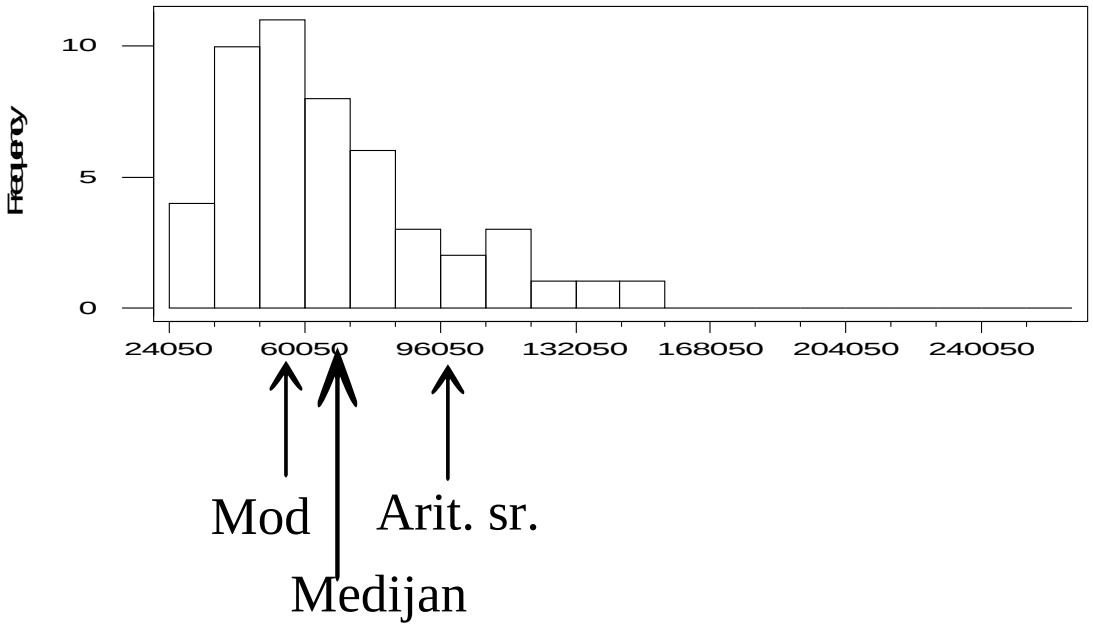
Promotrimo novouvedene mjere na sljedećem primjeru.

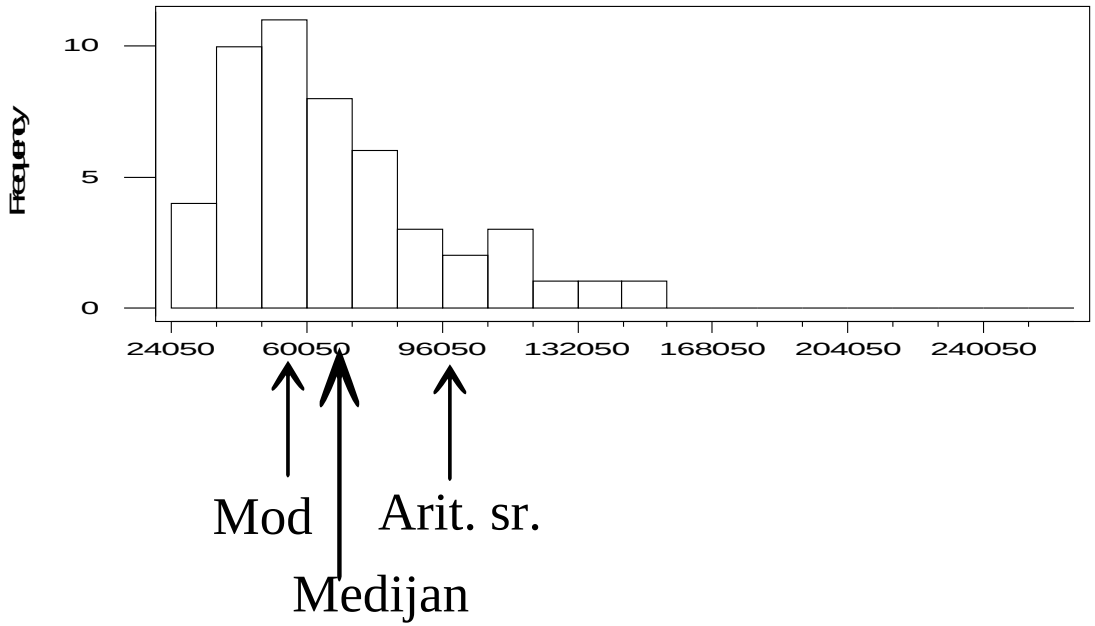
Primjer 3.4.

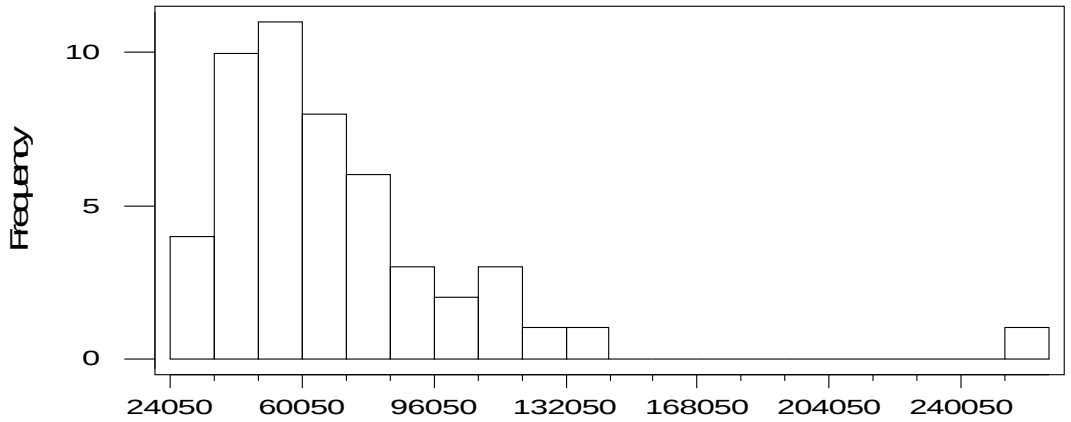
Zadan je sljedeći niz podataka.

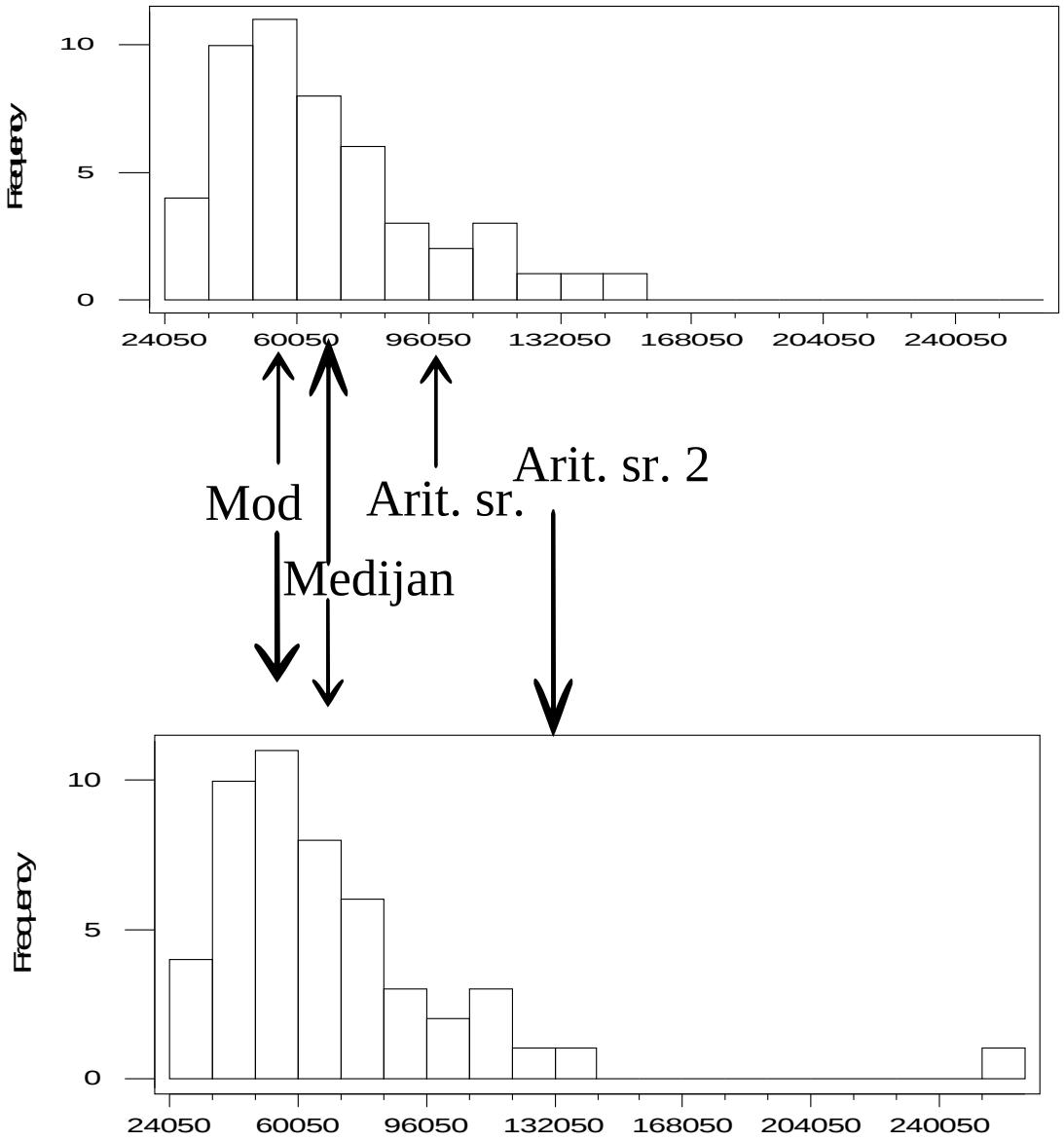
99000	45000	61500	78400	48500
123000	60000	155000	77000	56400
65700	50000	140000	49600	59500
115000	45500	112000	58500	25000
63000	70000	62000	46000	110000
76000	77100	61900	36500	25000
58000	45500	55000	38000	89500
87000	63500	31700	44900	90000
68000	51600	75300	40000	32000
50500	79000	47000	48000	103000

Zanima nas kako se ponašaju mjere srednje vrijednosti ako promijenimo nekoliko ekstremnih podataka. Zamijenit ćemo najveću vrijednost – $155\,000$, s još većom – $255\,000$.
Pogledajmo histogram za oba skupa podataka, te odredimo pripadne mjere srednje vrijednosti.


Frequency
10
5
0
24050
60050
96050
132050
168050
204050
240050
Mod
Arit. sr.
Medijan


Frequency
10
5
0
24050
60050
96050
132050
168050
204050
240050
Mod
Arit. sr.
Medijan


Frequency
10
5
0
24050
60050
96050
132050
168050
204050
240050


Frequency
10
5
0
24050
60050
96050
132050
168050
204050
240050
Mod
Arit. sr.
Arit. sr. 2
Medijan
Frequency
10
5
0
24050
60050
96050
132050
168050
204050
240050

Visoka vrijednost nekolicine podataka utjecala je na vrijednost aritmetičke sredine, koja je pomaknuta udesno, dok su se medijan i mod pokazali otpornim na utjecaj ekstremnih vrijednosti.

Važno!

Za podatke koji su izrazito asimetrični, bolja procjena srednje vrijednosti može biti medijan, jer manje ovisi o ekstremnim vrijednostima.

Kako god, za zvonolike, simetrične razdiobe, medijan i aritmetička sredina imaju približno jednaku vrijednost.

Mjere raspršenja

- mjere varijabilnost promatranog skupa podataka (koliko se podaci međusobno razlikuju).
Najjednostavnija je **raspon**, R.

Raspon skupa podataka se definira kao razlika najmanje i najveće vrijednosti:

$$R = x_{\max} - x_{\min}.$$

– što je raspon manji to je manje prostora unutar kojeg podatci mogu varirati.
– određena vrijednostima samo dva podatka, te stoga neosjetljiva na varijabilnost svih ostalih podataka.

Varijanca

Korisnija mjera je **varijanca**, s^2 – mjeri odstupanje svakog podatka od aritmetičke sredine:

$$x_i - \bar{x}.$$

Varijanca uzorka od n podataka se računa kao *prosjek* kvadrata gornjih odstupanja, i.e.

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

Što je varijanca veća, to je više odstupanja među podatcima.
Uz gornju definicju, često ćemo koristiti i sljedeću, ekvivalentnu formulu za varijancu:

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}. \tag{1}$$

Dokažimo formulu (2).

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \\
&= \frac{1}{n-1}\sum_i \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_i x_i^2 - 2\sum_i x_i\bar{x} + \sum_i \bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_i x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\
&= \frac{1}{n-1}\sum x_i^2 - n\bar{x}^2.
\end{aligned}
$$

Q.E.D.

Primjer 3.5. Odredite varijancu za sljedeći skup podataka: 3, 7, 2, 1, 8.

Rješenje:

Konstruirajmo sljedeću tablicu

	x	x^2
	3	9
	7	49
	2	4
	1	1
	8	64
$\sum$	21	127

Srednja vrijednost je $\bar{x} = 21/5 = 4.2$, te je stoga

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{127 - 5 \cdot 4.2^2}{4} = \frac{38.8}{4} = 9.7.$$

S ovom formulom imamo manji broj računskih operacija (reda $2n$) za razliku od originalne formule iz definicije (reda $3n$).

Ako među podacima ima više jednakih, pri čemu se vrijednost x_i pojavljuje s frekvencijom f_i, onda iz formule (2) slijedi

$$s^2 = \frac{1}{\sum_i f_i - 1}\left(\sum_i f_i x_i^2 - n\bar{x}^2\right),$$

pri čemu aritmetičku sredinu možemo računati kao

$$\bar{x} = \frac{1}{\sum_i f_i}\sum_i f_i x_i.$$

Primjer 3.6. Pet novčića smo bacali 1000 puta i zabilježili broj glava. Broj bacanja u kojima je palo 0, 1, 2, 3, 4, ili 5 glava zabilježen je u sljedećoj tablici:

broj glava	broj bacanja
0	38
1	144
2	342
3	287
4	164
5	25
$\sum$	1000

Odredite aritmetičku sredinu i varijancu.

Rješenje:

$$\bar{x} = 2.47, \qquad s^2 = 1.2443.$$

Standardno odstupanje ili devijacija

Varijanca
– mjeri se u kvadratima originalnih jedinica (podatak u cm, varijanca u cm^2).
– nema jasno tumačenje.
Korisnije su mjere izražene u jedinicama jednakim originalnim.

Standardno odstupanje ili devijacija, s, uzorka od n podataka se računa kao kvadratni korijen varijance:

$$s = \sqrt{s^2}$$
$$= \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}.$$

Standardna devijacija za podatke iz primjera 3.5 je
$s = \sqrt{s^2} = \sqrt{9.7} = 3.1.$

Tumačenje standardne devijacije

Cilj – odrediti intervale u koje upada najveći dio podataka.
Razmatrat ćemo intervale oblika

$$\bar{x} \pm ks, \quad k = 1, 2, 3.$$

Za zvonolike razdiobe imamo sljedeće *praktično* pravilo.

Ukoliko imamo zvonoliku razdiobu podataka s aritmetičkom sredinom $\bar{x}$, te standardnom devijacijom s, tada je omjer podataka unutar intervala $\bar{x} \pm ks,\ k = 1, 2, 3$ sljedeći:
$\bar{x} \pm s$ – obično između 60 i 80%. Postotak će biti blizu 70% za simetrične razdiobe, a veći (oko 90%) za izrazito asimetrične razdiobe,
$\bar{x} \pm 2s$ – oko 95%. Postotak će biti veći (blizu 100%) za izrazito asimetrične razdiobe,
$\bar{x} \pm 3s$ – blizu 100%.

Tumačenje standardne devijacije

Na žalost gornje pravilo ne možemo korisititi za podatke koji nemaju zvonoliku razdiobu.
U tom slučaju koristimo konzervativnije pravilo – Čebiševljev teorem.
Teorem tvrdi da je postotak od ukupnog broja podataka unutar intervala $\bar{x} \pm ks$, pri čemu je k konstanta, barem $1 - 1/k^2$.

Čebiševljev teorem

Za proizvoljni skup podataka s aritmetičkom sredinom $\bar{x}$, te standardnom devijacijom s, omjer ukupnog broja podataka unutar intervala
$\bar{x} \pm 2s$ – je barem 75%,
$\bar{x} \pm 3s$ – je barem 89%.

Prednost – primjenljiv je na bilo koji skup podataka.
Nedostatak – konzervativan je, u smislu da daje samo donju ocjenu stvarnog omjera podataka u promatranom intervalu.

Dokaz

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum (x_i-\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{\{i:|x_i-\bar{x}|>ks\}} (x_i-\bar{x})^2 + \frac{1}{n-1}\sum_{\{i:|x_i-\bar{x}|\leqslant ks\}} (x_i-\bar{x})^2 \\
&> \frac{1}{n-1}\sum_{\{i:|x_i-\bar{x}|>ks\}} k^2 s^2 \\
&= \frac{1}{n-1} B_k\, k^2 s^2,
\end{aligned}
$$

gdje B_k označuje broj podataka sa svojstvom da je $|x-\bar{x}| > ks$. Stoga je

$$\frac{B_k}{n-1} < \frac{1}{k^2}.$$

Dokaz

Na taj način smo dokazali da je postotak broja podataka izvan intervala $\bar{x} \pm ks$ manji od $\frac{1}{k^2}$, te je stoga postotak od ukupnog broja podataka unutar tog intervala barem $1 - \frac{1}{k^2}$.
Uvrštavajući $k = 2$ i $k = 3$ dobivamo tvrdnje iz iskaza teorema.

Mjere položaja

– opisuju položaj podatka u odnosu na preostale podatke.

Percentili

Za $k \in [0, 100]$ definiramo k-**ti percentil,** P_k, kao vrijednost sa svojstvom da je $k\%$ podataka manje ili jednako njemu, a $(100 - k)\%$ podataka veće ili jednako od njega.

Neki važni percentili:

- 50-ti percentil – medijan, M
- 25-ti percentil – donji kvartil, Q_1
- 75-ti percentil – gornji kvartil, Q_3.

Za manje skupove podataka često je teško naći vrijednost koja premašuje, npr. točno 25% podataka. ($\{4, 5, 8\}$)
Međutim, u tom slučaju koristimo sljedeću proceduru koja nam daje dobru aproksimaciju percentila.

1. Poredajte podatke po veličini, od najmanjeg k najvišem.
2. Izračunajte $\frac{1}{4}(n+1)$ i zaokružite na najbližu cjelobrojnu vrijednost r (ukoliko je točno između dva cijela broja, zaokružite na više). Podatak na rednom mjestu r je donji kvartil.
3. Izračunajte $\frac{3}{4}(n+1)$ i zaokružite na najbližu cjelobrojnu vrijednost r (ukoliko je točno između dva cijela broja, zaokružite na niže). Podatak na rednom mjestu r je gornji kvartil.
4. Za naći k-ti percentil izračunajte $\frac{k}{100}(n+1)$ i zaokružite na najbližu cjelobrojnu vrijednost r (ukoliko je točno između dva cijela broja, zaokružite na više za $k < 50$, te na niže za $k > 50$). Podatak na rednom mjestu r je P_k.

Već smo spominjali da je prednost S-L prikaza što čuva originalne podatke, te ih prikazuje sortirane. To će nam pojednostavniti nalaženje percentila.

Primjer 3.7. Nađite medijan, donji i gornji kvartil, te 90-ti percentil za skup podataka iz primjera 2.2.

Rješenje:

Podatke smo prikazali pomoću S-L prikaza:

Stablo	List
3	67
4	25
5	00,75,95
6	30,50,60,82
7	10,20,49,60,70
8	20,43,99
9	45,50
10	16,60,90
11	20
12	95
13	
14	80

Q_1 : $\frac{1}{4}(n+1) = \frac{26}{4} = 6.5 \simeq 7$

Stoga je Q_1 podatak na 7. mjestu, tj. $Q_1 = x_7 = 650$.

M: n- neparan

$\frac{n+1}{2} = \frac{26}{2} = 13$

M je podatak na 13. mjestu, tj. $M = x_{13} = 760$.

Q_3 : $\frac{3}{4}(n+1) = 3 \cdot \frac{26}{4} = 19.5 \simeq 19$

Q_3 je podatak na 19. mjestu, tj. $Q_3 = x_{19} = 950$.

P_{90} : $\frac{9}{10}(n+1) = 23.4 \sim 23.$

$P_{90} = x_{23} = 1120$.

z varijabla ili obilježje

Standardizirana ili z varijabla (obilježje) definira se kao omjer odstupanja od srednje vrijednosti i standardne devijacije, tj.

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Što predstavlja negativna vrijednost z varijable? A što pozitivna?
Negativna vrijednost znači da je podatak manji od srednje vrijednosti, a pozitivna da je veći od nje.
Ako se podatak nalazi unutar intervala $\bar{x} \pm ks$, što možemo reći o pripadnoj vrijednosti z varijable?

$$(x_i \in [\bar{x} - ks, \bar{x} + ks]) \Longleftrightarrow (|z_i| \leqslant k).$$

Metode određivanja sumnjivih vrijednosti

Netipična ili sumnjiva vrijednost – podatak koji po veličini odskače od ostatka populacije (relativno malen ili velik).
Takve podatke nalazimo iz niza razloga:

- Podatak je krivo izmjeren ili zapisan.
- Podatak pripada drugoj populaciji.
- Podatak je ispravan, ali opisuje veooma rijedak događaj.

Efikasna metoda za određivanje sumnjivih vrijednosti je z varijabla.

Primjer 3.8. Promatramo uzorak cijena nekretnina sa srednjom vrijednosti $\bar{x} = 1\,064\,050$kn, te standardnom devijacijom $s = 854\,414$kn. Jedan podatak iznosi $11\,460\,000$kn. Je li on netipičan?

Rješenje:

Izračunajmo pripadnu vrijednost z varijable

$$z_i = \frac{11\,460\,000 - 1\,064\,050}{854\,414} = 12.17\,.$$

I Čebiševljev teorem, kao i praktično pravilo nam kažu da bi skoro svi podaci trebali upasti u interval $\bar{x} \pm 3s$, dakle njihovo z obilježje mora po apsolutnoj vrijednosti biti manje od 3.

S obzirom da je vrijednost z varijable od 12.17 malo vjerojatna, to je riječ o netipičnom podatku.

U takvom slučaju moramo provjeriti podatak, utvrditi da li pripada populaciji koju proučavamo, te odlučiti da li ga zadržati ili isključiti. ∎

Korištenje z varijable za odredivanje sumnjivih vrijednosti

Podatci čija pripadna vrijednost z varijable je po apsolutnom iznosu veća od 3 smatraju se netipičnim.

Dijagram pravokutnika

Drugi način određivanja sumnjivih vrijednosti je pomoću dijagrama pravokutnika (*eng. box-whiskers plot*, *kutija-brk prikaz*).
Zasniva se na veličini koju nazivamo **interkvartilnim rasponom**, I_Q, i jednaka je razlici gornjeg i donjeg kvartila:
$I_Q = Q_3 - Q_1$.

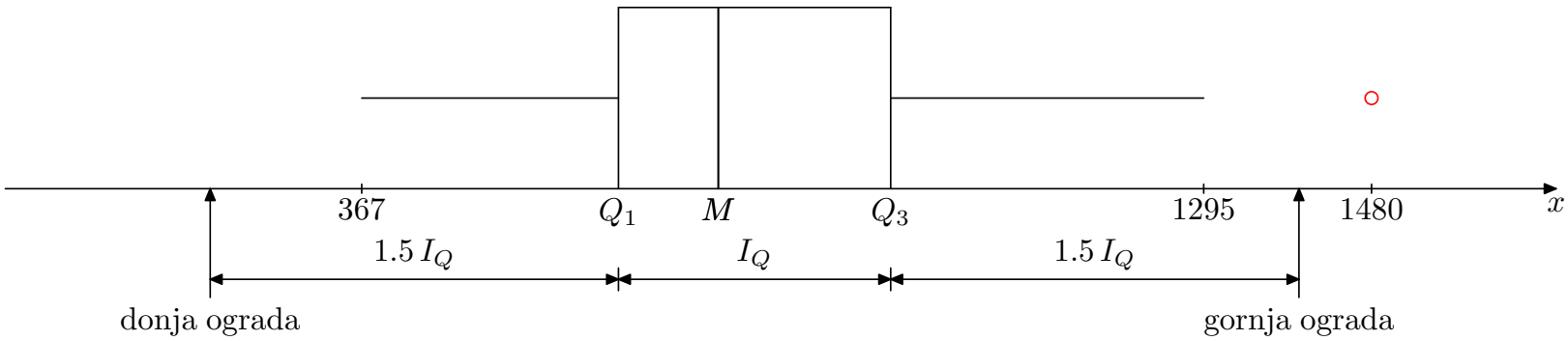
Primjer 3.9.

Konstrukcija dijagrama pravokutnika
Napravite kutija-brk prikaz za podatke iz primjera 3.7.

Podatci su dani pomoću S-L prikaza na desnoj strani.

Stablo	List
3	67
4	25
5	00,75,95
6	30,50,60,82
7	10,20,49,60,70
8	20,43,99
9	45,50
10	16,60,90
11	20
12	95
13	
14	80

1. Nađite Q_1, Q_3, i I_Q.
   Iz primjera 3.7 znamo da je $Q_1 = 650$, $Q_3 = 950$, te je stoga
   $$I_Q = 950 - 650 = 300.$$
2. Nacrtajte realnu os, te na njoj naznačite kvartile.
3. Nacrtajte kutiju nad nacrtanom osi, s donjim vrhovima u točkama Q_1 i Q_3. Kutiju prepolovite uspravnom crtom na mjestu medijana.
4. Naznačite ograde koje leže za $1.5I_Q$ lijevo od donjeg kvartila (donja ograda), te za $1.5I_Q$ desno od gornjeg kvartila (gornja ograda).
5. Naznačite najmanji podatak između donje ograde i kvartila. Povucite brk od kutije do tog podatka. Slično, povucite brk s desne strane kutije do najvećeg podatka između kvartila i gornje ograde.
6. Podaci koji se nalaze izvan ograda su sumnjivi podaci. Njih ćemo pojedinačno naznačiti pomoću kružića $\circ$.



Tumačenje dijagrama:

- polovica podataka (50%) se nalazi unutar kutije,
- gotovo svi podaci se nalaze unutar intervala određenog širinom brkova,
- podatci koji se nalaze izvan ograda su posebno naznačeni i predstavljaju netipične vrijednosti.

Zadatak 3.1. Kocku smo bacali 20 puta i zabilježili smo sljedeće rezultate:

6	3	3	6	3	5	6	1	4	6
3	5	5	2	2	2	2	3	2	3.

a) Odredite aritmetičku sredinu, mod i medijan uzorka.

b) Odredite varijancu i standardnu devijaciju uzorka.

c) Odredite raspon uzorka.

d) Odredite donji i gornji kvartil, te interkvartil uzorka.

e) Nacrtajte dijagram pravokutnika ("box and whisker plot").