Uvod u matematičku statistiku

Pojam matematičke statistike.

Pojednostavljeno rečeno, matematička statistika je znanstvena disciplina koja iz poznavanja određenih svojstava **uzorka** donosi zaključke o svojstvima **populacije**.
Populacija je skup svih entiteta koje razmatramo. Na primjer:

(1) ako nas zanimaju izbori, populaciju čine svi glasači (ili svi glasači koji su izašli na izbore, odnosno svi glasački listići),
(2) ako nas zanima duljina studiranja, populaciju čine svi studenti nekog sveučilišta,
(3) ako nas zanimaju svojstva nekog proizvoda proizvedenog određenom tehnološkom metodom, populaciju čine svi proizvodi tako proizvedeni,
(4) ako nas zanima postotak vodika u Svemiru, populaciju čini skup svih atoma itd.

Općenito, zanima nas neko **statističko obilježje** populacije, na primjer visina, masa, politički stav itd..
Često smo u nemogućnosti razmatrati cijelu populaciju (na primjer ako je čine svi atomi), ili je to vrlo skupo (na primjer ako je čine svi proizvodi u nekom velikom pogonu), ili zbog vremenskog ograničenja (na primjer ako želimo odmah po zatvaranju birališta procijeniti rezultate izbora), ili je to besmisleno (na primjer ako nas zanima vijek trajanja žarulja koje je neka tvornica uputila na tržište) itd. U tim smo uvjetima prisiljeni procjenjivati svojstva populacije na osnovi svojstva nekoliko članova te populacije.

Uzorak je neki slučajno odabran podskup populacije, na primjer:
(1) 2% slučajno odabranih glasačkih listića,
(2) 300 slučajno odabranih studenata,
(3) 15 slučajno odabranih proizvoda itd.

Postavlja se pitanje koliko je opravdano na osnovi nekoliko primjera zaključivati o cijeloj populaciji i koliko možemo vjerovati takvom zaključku, koje je često podložno subjektivnim stavovima, pojedinačnim interesima i sl. Jedan od zadataka matematičke statistike jest izgradnja metoda kojim će se ovakvi problemi moći rješavati egzaktno. Mogu se izdvojiti tri glavna koraka:

1.**korak**. Biranje uzorka, priprema i provođenje pokusa (testa, ankete i sl.).
2. korak. Obrada dobivenih podataka.
3. korak. Vršenje procjena i donošenje odluka.

Prvim se korakom manje bavi matematička statistika, a više statistika u širem smislu. Zato ovdje nećemo obrađivati taj **vrlo važan** korak. Spomenimo samo dva pojma.
Već smo rekli da uzorak za nas znači **slučajno izabrani podskup** populacije. To znači da svaki član populacije ima jednak izgled da bude izabran u uzorak. Dakle, pod uzorkom smatramo **slučajni uzorak** (u literaturi se često to naziva **jednostavni slučajni uzorak**).
Na primjer, ako proizvođač izabere 10 svojih proizvoda i ponudi ih na uvid, radi svoje promidžbe, to u pravilu nije slučajan uzorak, već **pristran** (proizvođač u pravilu izabere najkvalitetnije proizvode za promidžbu).

Jednako tako, trgovac kojemu bi bilo dopušteno da izabere 10 proizvoda na osnovi kojih bi poslije pregovarao o cijeni, možda ne bi izabrao slučajan uzorak (već bi birao lošije proizvode kako bi mogao spustiti cijenu).
Slučajan uzorak **duljine** 10 mogao bi se izabrati na primjer tako da proizvođač dade na uvid serijske brojeve proizvoda, potom da se iz skupa brojeva slučajno izabere desetoračlani podskup i konačno, da uzorak čine upravo proizvodi s izabranim serijskim brojevima.
Kažemo da je uzorak **reprezentativan**, ako odražava svojstva cijele populacije. Na primjer, uzorak koji je vrlo malen (male duljine) u usporedbi s veličinom populacije, u pravilu nije reprezentativan. Tako, slučajno odabranih 20 birača nakon napuštanja biračkog mjesta za državne izbore nije reprezentativan uzorak. Nasuprot tome, uzorak u koji bismo ciljano uključili, proporcionalno prema udjelu u biračkom spisku, birače iz svake županije, iz gradske i ruralne sredine, prema spolu i dobi, prema nacionalnoj i vjerskoj pripadnosti, prema stručnoj spremi, prema imovinskom stanju, prema zaposlenosti, prema visini primanja i sl. bio bi u dobroj mjeri reprezentativan, ali ne bi bio slučajan.

Drugim korakom u većem se dijelu bavi **deskriptivna statistika,** koja je dio matematičke statistike, jer donekle uključuje matematičke tehnike. Deskriptivna se statistika bavi uzorkom, a o populaciji izravno ne govori ništa ili vrlo malo. Naravno, na osnovi svojstava uzorka, mogu se naslućivati svojstva populacije.

U trećem se koraku, na osnovi svojstava uzorka, procjenjuju svojstva populacije. Za prijelaz od uzorka na populaciju presudna je uloga **teorije vjerojatnosti**, koja je teoretski temelj matematičke statistike. Puko naslućivanje svojstava populacije iz svojstava uzorka je nesustavno i, u pravilu, subjektivno. Uz pomoć teorije vjerojatnosti ono se može izgraditi u egzaktnu znanstvenu metodu.

Deskriptivna statistika

Uređivanje i grupiranje podataka.

Često su podatci koje dobijemo mjerenjem napisani redoslijedom koji nam otežava jasnu predodžbu o njima. Jedan od načina da podatke učinimo jasnijima jest taj da se napišu u rastućem (ili padajućem) redoslijedu.

Primjer 1. Kontrolom slučajno odabranih 20 staklenka s kemikalijom, punjenih od jednog proizvođača, dobiveni su sljedeći podatci (u litrama):
1.97 1.95 2.02 1.99 1.95 2.03 2.00 1.96 1.98 2.00 2.01 1.99 1.98 1.97 1.97 1.94 1.94 2.04 2.02 1.93
U ovom primjeru **populaciju** čine sve staklenke te vrste punjene od tog proizvođača, a izabranih 20 staklenaka čine **uzorak**. Broj 20 je **duljina uzorka**.
Vidimo da se podatci vrte oko vrijednosti 2, a da bismo dobili jasniju predožbu o njima poredajmo ih prema veličini (od manjeg prema većem). Napomenimo da tim postupkom nećemo izgubiti ni jednu statistički važnu informaciju (naime uzorak je izabran slučajno). Dobijemo:

1.93 1.94 1.94 1.95 1.95 1.96 1.97 1.97 1.97 1.98 1.98 1.99 1.99 2.00 2.00 2.01 2.02 2.02 2.03 2.04

Sad nam je puno lakše uočavati svojstva podataka i odnose među njima. Na primjer, brzo uočavamo da je 1.93 **najmanji** (**minimalni**), a 2.04 **najveći** (**maksimalni**) podatak. Također, brzo uočavamo da se podatak 1.97 pojavljuje tri puta; kažemo da mu je **frekvencija** 3.
Nadalje, vidimo da frekvenciju 2 imaju podatci 1.94, 1.95, 1.98, 1.99, 2.00 i 2.02, a da preostali imaju frekvenciju 1.
Tu činjenicu pregledno zapisujemo pomoću **tablice frekvencija** kojoj su u prvom redku redom različiti podatci, a u drugom njihove frekvencije.

1.93	1.94	1.95	1.96	1.97	1.98	1.99	2.00	2.01	2.02	2.03	2.04
1	2	2	1	3	2	2	2	1	2	1	1

Vidimo da u ovom primjeru od ukupno 20 podataka ima 12 međusobno različitih.
Uočite da je zbroj frekvencija jednak ukupnom broju podataka (s ponavljanjem); u ovom primjeru je 1+2+2+1+3+2+2+2+1+2+1+1 = 20.

Grupiranje u razrede. I nakon uređivanja često imamo poteškoća s podatcima, naročito ako ih ima puno; zato ih grupiramo u **razrede**. Da to ilustriramo, podatke iz ovog primjera grupirat ćemo u 6 razreda, svaki duljine 0.02. Ispišimo ih tako da razrede odvojimo:

1.93 1.94 1.94 | 1.95 1.95 1.96 | 1.97 1.97 1.97 1.98 1.98

1.99 1.99 2.00 2.00 | 2.01 2.02 2.02 | 2.03 2.04

Rezultate ćemo predočiti tablicom koja ima četiri stupca.
U prvom stupcu su redni brojevi razreda (od 1 do 6).
U drugom su stupcu granice razreda; na primjer 1.925-1.945 znači da su u prvom razredu podatci koji su između 1.925 i 1.945 (treću smo decimalu dodali da nam se ne dogodi da neki podatak upadne u dva razreda).
U trećem su stupcu **frekvencije razreda**, tj. broj podataka u pojedinim razredima (uočite da se te frekvencije razlikuju od **frekvencija podataka uzorka** koje smo prije spominjali); na primjer $f_1=3$, znači da su u prvom razredu tri podatka (1.93, 1.94, 1.94).
U četvrtom stupcu su relativne frekvencije $\frac{f_i}{n}$ razreda, gdje je *n* ukupan broj podataka; na primjer relativna frekvencija prvog razreda je $\frac{3}{20}$ jer je $f_1=3$, a $n = 20$.

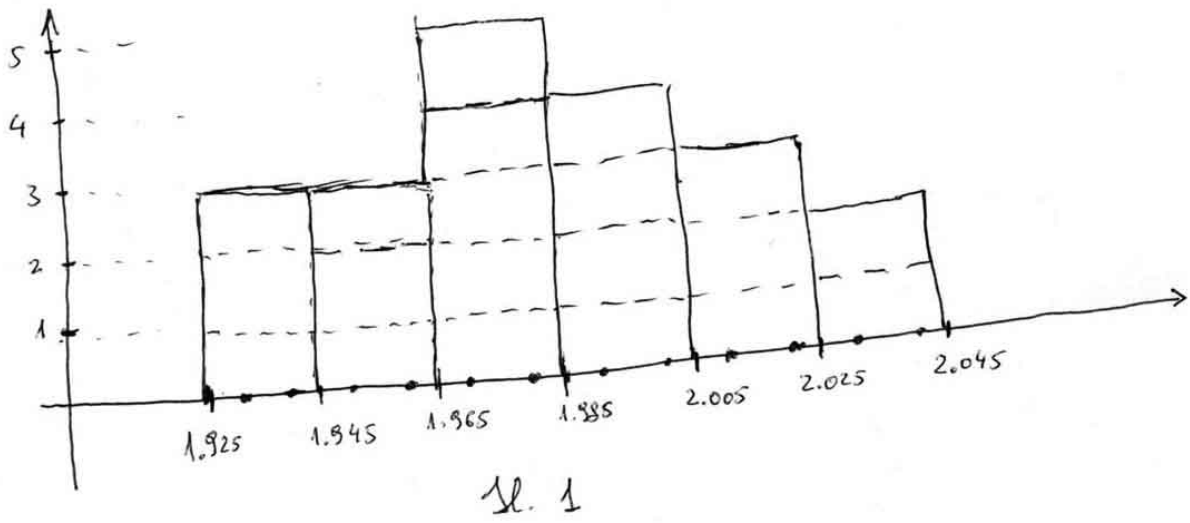
Redni broj razreda	Granice razreda	Frekvencija razreda f_i	Relativna frekvencija razreda
1.	1.925-1.945	3	3/20
2.	1.945 -1.965	3	3/20
3.	1.965-1.985	5	5/20

4.	1.985-2.005	4	4/20
5.	2.005-2.025	3	3/20
6.	2.025-2.045	2	2/20

Važna napomena. Iako nam ova tablica jasnije dočarava odnos među podatcima, u njoj su se neke informacije izgubile. Na primjer, iz nje očitavamo da su u prvom razredu 3 podatka, ali ne znamo koja su to tri podatka. Gubitak je to neznatniji što je uzorak veće duljine, a razredi uži. Za valjanost nekih statističkih zaključivanja često se traži da frekvencija svakog razreda bude barem 5, što ovo naše grupiranje ne zadovoljava.
Važnost grupiranja nije samo u dobivanju veće preglednosti, već i u uočavanju statističkih zakonitosti. Na primjer, frekvencije podataka iz Primjera 1 variraju, ali nakon grupiranja uočavamo sljedeću pravilnost: frekvencije razreda se povećavaju, potom smanjuju.
Uočite da su svi razredi bili iste duljine. Tako će, u pravilu, uvijek biti, iako, načelno, možemo vršiti i grupiranje u razrede različitih duljina.
Uočite, također, da je već tablicom frekvencija izvršeno jedno grupitranje podataka: u pojedini razred upadaju podatci koji su međusobno jednaki.

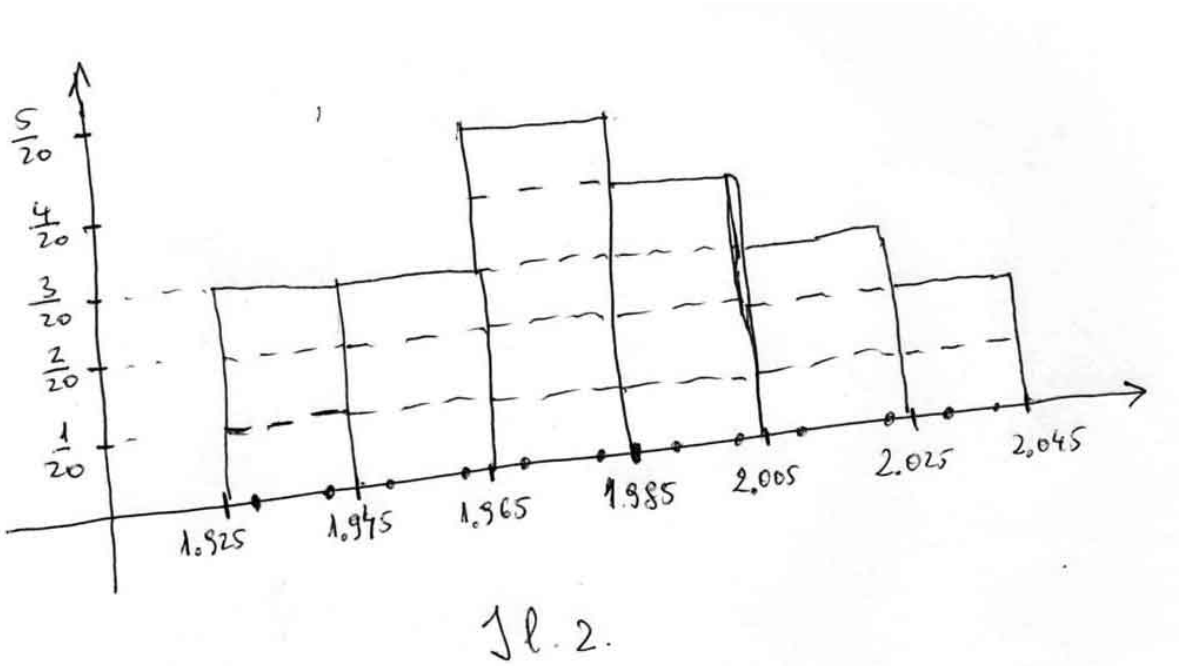
Grafičko predočavanje podataka. Histogram frekvencija i histogram relativnih frekvencija.

Grupirani podatci iz primjera mogu se grafički predočiti tako da na horizontalnu os koordinatnog sustava nanosimo količinu u litrama, a na vertikalnu frekvencije. Ako se iznad svakog razreda postavi pravokutnik kojemu je visina jednaka frekvenciji razreda dobije se **histogram frekvencija**. Na primjer, za podatke iz Primjera 1. dobije se (sl.1.):



sl. 1

Sličnu sliku dobili bismo ako bismo na vertikalnu os nanosili relativne frekvencije. To je **histogram relativnih frekvencija.** Iako ni jedan od njih nema prioritetno značenje (jedan se dobije iz drugoga promjenom skale na vertikalnoj osi), obično se histogram relativnih frekvencija naziva **distribucijom frekvencija** (sl.2.).



Sl. 2.

Svojstva histograma (jednog i drugog).

1. Svaki podatak doprinosi u histogramu n-ti dio površine, tj. 1/n od ukupne površine, gdje je *n* ukupan broj podataka (računajući i ponavljanje). To najbolje uočavamo tako što svakom podatku odgovara jedan mali pravokutnik u histogramu (tu je važno to da **svi razredi imaju jednaku širinu**).

2. Površine stupaca iznad pojedinih razreda proporcionalne su pripadajućim frekvencijama (odnosno relativnim frekvencijama). Naime iznad i-tog razreda ima f_i malih pravokutnika, pa i-ti stupac čini $\frac{f_i}{n}$ ukupne površine.

3. Stupac iznad i-tog razreda histograma čini $\frac{100 f_i}{n}\%$ površine cijelog histograma. To odmah proizlazi iz svojstva 2. Uočite da se traženi postotak dobije množenjem relativne frekvencije sa 100.

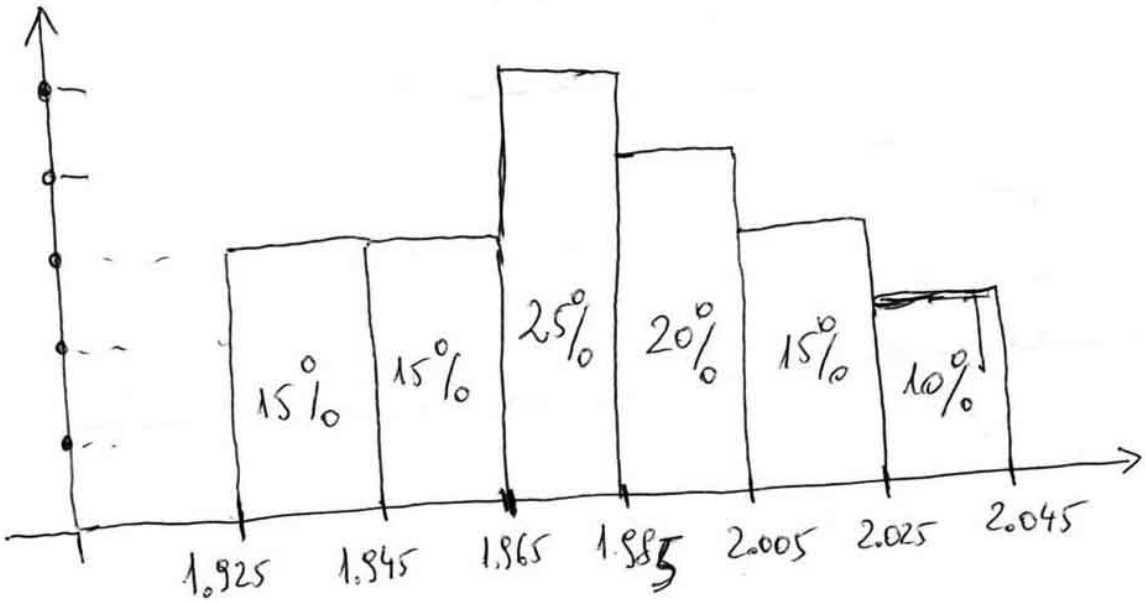
Primjer 2. Odredimo površinske udjele iznad pojedinih razreda u postotcima za podatke iz Primjera 1.

$100\cdot\frac{3}{20}=15$ (napomenimo da smo rezultat mogli dobiti i kao brojnik razlomka kad relativnu frekvenciju 3/20 zapišemo s nazivnikom 100, tj. $\frac{3}{20}=\frac{3\cdot 5}{20\cdot 5}=\frac{15}{100}$).

Dalje je:

$100\cdot(5/20)=25$; $100\cdot(4/20)=20$; $100\cdot\frac{3}{20}=15$ i $100\cdot(2/20)=10$.

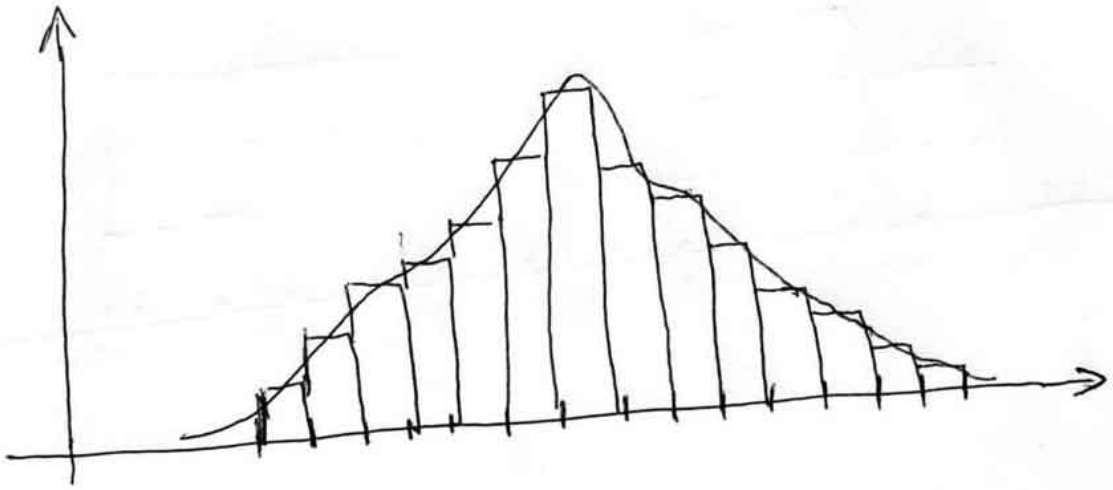
Rezultate smo predočili na slici 3.



Sl. 3

Kako treba tumačiti ove postotke? Prvo što nam pada na pamet jest to da će od svih staklenka ove vrste, punjenih od ovog proizvođača, odprilike 15% imati količinu iz prvog razreda, 15% iz drugoga, 25% iz trećega, 20% iz četvrtoga, 15% iz petoga i 10% iz šestoga. Naravno postavlja se pitanje s koliko sigurnosti možemo predviđati svojstva populacije iz poznatih svojstava uzorka. To je jedan od najvažnijih zadataka matematičke statistike.

Napomene. 1. Histogram ne ovisi samo o podatcima (uzorku) već i o izabranoj podjeli na razrede. Zato uz isti uzorak može biti više različitih podjela na razrede.
2. Ako je broj podataka vrlo malen, onda podjela na razrede nije potrebna, pa tako ni histogrami. Za broj podataka iz Primjera 1 prije bismo rekli da je malen (iako nije vrlo malen) nego da je velik. Iako za to nema jasnog opravdanja, obično se uzorci duljine veće od 30 smatraju dovoljno velikima.
3. Ako je broj podataka vrlo velik, a također i broj razreda, onda se histogram može dobro aproksimirati neprekinutom crtom kao na slici 4.



Sl. 4.

Aritmetička sredina i medijan skupa podataka (uzorka).

Aritmetička sredina uzorka.
Jedno od osnovnih pitanja o staklenkama kemikalije iz Primjera 1. jest kolika je **prosječna količina** kemikalije u njima. Naravno da iz poznatih podataka ne možemo znati kolika je **aritmetička sredina** populacije, međutim možemo izračunati aritmetičku sredinu uzorka (za koju vjerujemo da bi mogao biti blizu stvarnog prosjeka). Sjetimo se definicije aritmetičke sredine $\bar{x}$ brojeva $x_1, x_2,...,x_n$.

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Na primjer, aritmetička sredina brojeva 1,2,3,4,5 je broj $\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$.

Taj se izraz zove **aritmetička sredina uzorka**.

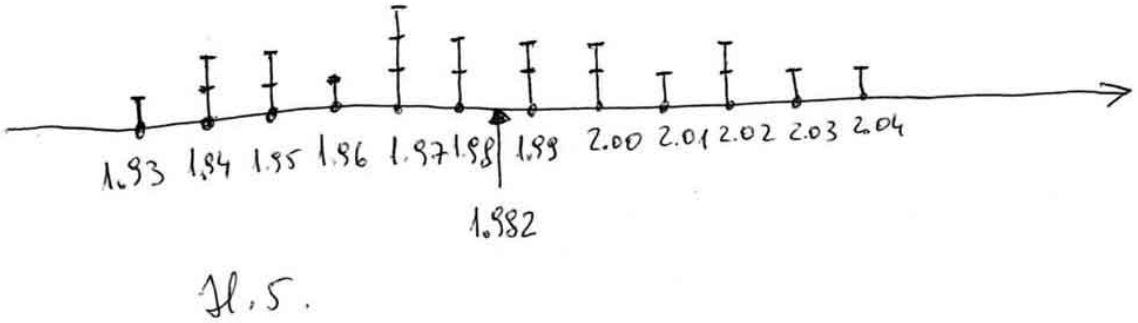
Primjer 3. Odredimo aritmetičku sredinu podataka iz Primjera 1.

$$\bar{x} = \frac{1.93 + 1.94 \cdot 2 + 1.95 \cdot 2 + 1.96 + 1.97 \cdot 3 + 1.98 \cdot 2 + 1.99 \cdot 2 + 2.00 \cdot 2 + 2.01 + 2.02 \cdot 2 + 2.03 + 2.04}{20}$$

$= 1.982$

Uočite da aritmetička sredina (prosječna vrijednost) ovaj put nije jednaka ni jednom od podataka, što je uglavnom i slučaj.
Intuitivno značenje aritmetičke sredine. Aritmetička sredina 1.982 jest upravo ona količina kojom bi trebalo napuniti svaku od 20 staklenka pa da ukupna količina bude kao u uzorku.

Fizikalna interpretacija aritmetičke sredine – težište sustava masa na pravcu
Aritmetička sredina ima jednostavno, ali vrlo važno fizikalno značenje. Ako zamislimo da su u točkama $x_1, x_2,...,x_n$ smještene jednake mase (za svaki podatak masa m), onda je u $\bar{x}$ koordinata težišta sustava tih masa. Naravno ako se neki podatak pojavljuje više puta onda, u tu koordinatu treba staviti toliko masa koliko se puta on pojavljuje, tj. kolika mu je frekvencija u uzorku.
Na slici 5. predočena je fizikalna interpretacija podataka iz Primjera 1 i interpretacija aritmetičke sredine kao težišta. Radi lakšeg dočaravanja, možemo zamisliti da je koordinatna os čvrsta, ali bez mase. Tada je u aritmetičkoj sredini (težištu) ravnoteža.



Sl. 5.

Intuitivno je jasno da se težište sustava masa ne mijenja ako se svaka masa proporcionalno poveća ili smanji (odnosno ako promijenimo mjernu jedinicu mase). Jednako tako, aritmetička sredina se ne mijenja ako se frekvencije pojavljivanja svakog podatka proporcionalno povećaju ili smanje.

Procjena aritmetičke sredine populacije.
Postavlja se pitanje možemo li dobro procijeniti aritmetičku sredinu populacije, ako znamo aritmetičku sredinu uzorka. Odgovor je pozitivan i vrijedi:
Aritmetička sredina uzorka je najbolja procjena aritmetičke sredine populacije.
Značenje izraza *najbolja* može se preciznije matematički definirati i o tome ćemo više reći poslije.

Medijan.
Jasno je da se aritmetička sredina brojeva nalazi između najmanjeg i najvećeg broja. Tako se 1.982 iz Primjera 3 nalazi između 1.93 i 2.04. Pogrješno bi bilo zaključiti da se lijevo i desno od aritmetičke sredine nalazi jednako mnogo podataka. Tako se u Primjeru 1, lijevo od aritmetičke sredine nalazi 11 podataka, a desno 9 (11 ih je manjih od aritmetičke sredine, a 9 većih). Zato se uvodi pojam medijana.

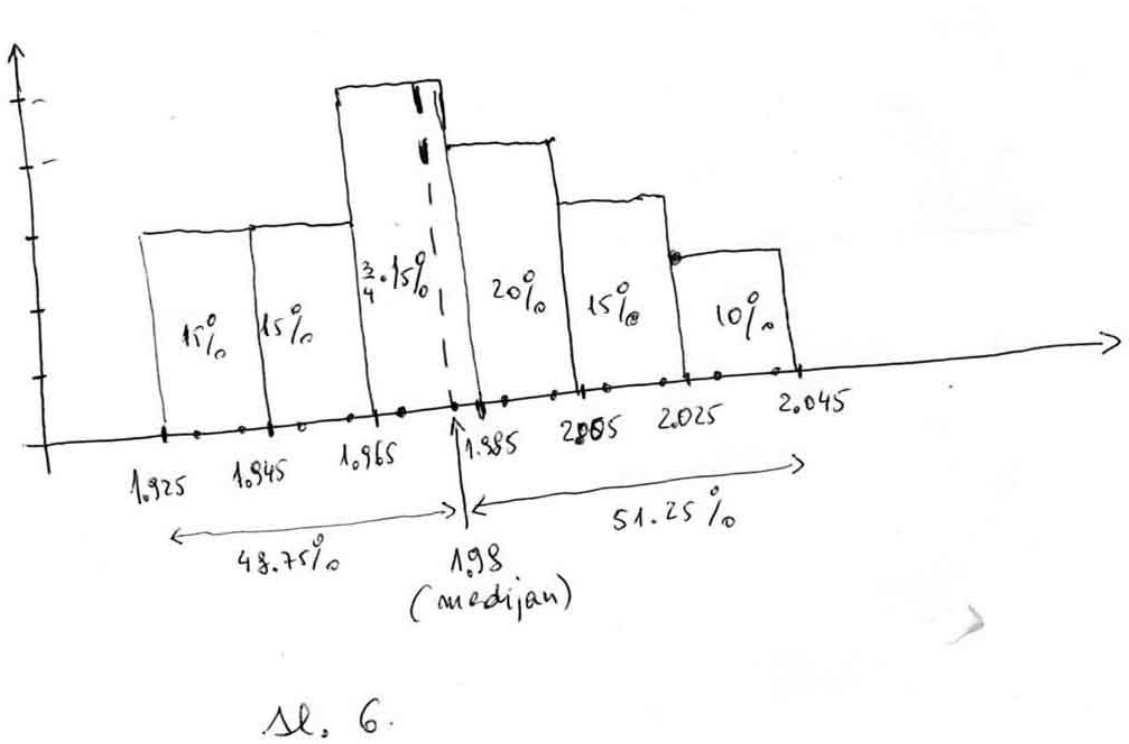
Medijan skupa podataka je srednji podatak ako ima neparno podataka, a aritmetička sredina dvaju srednjih ako ima parno podataka.

Na primjer, za podake 1,2, 4, 11, 13 medijan je 4,

a za podatke 1,2,4,7,11,13 medijan je $\frac{4+7}{2}=5.5$

Primjer 4. Odredimo medijan skupa podataka iz Primjera 1.
Tu ima 20 podataka pa je medijan aritmetička sredina 10-og i 11-og podatka, koja su oba 1.98. Zato je medijan
$\frac{1.98+1.98}{2}=1.98$.

Geometrijska i fizikalna interpretacija medijana. Medijan dijeli podatke na dva jednakobrojna dijela, jedan lijevo, a drugi desno od njega.
Također, medijan dijeli histogram na dva dijela odprilike jednakih površina (oko polovice je površine lijevo, a oko polovice desno od medijana). Na primjer, histogram iz Primjera 1, medijan dijeli na lijevi dio koji ima 48.75% i desni koji ima 51.25% ukupne površine (sl. 6.).



Sl. 6.

U interpretaciji sa sustavom masa na pravcu, lijevo i desno od medijana mase su jednake, što **ne znači** da je tu ravnoteža (jer su krakovi različiti); ravnoteža je u aritmetičkoj sredini.

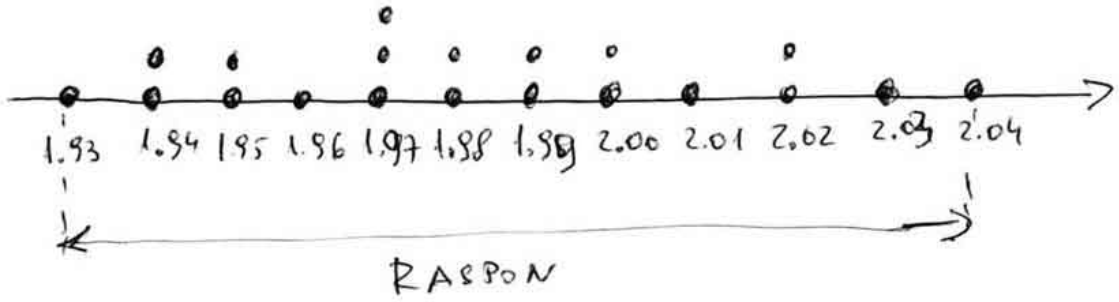
Mod. U nekim slučajevima važan je i **mod uzorka**; to je najfrekventniji podatak u uzorku. Na primjer, u Primjeru 1. mod je 1.97 jer ima najveću frekvenciju, broj 3. Slično skup podataka 1, 1, 2, 2, 3, 11, 64, ima dva moda, brojeve 1 i 2 (oba imaju frekvenciju 2).

Mjere raspršenja podataka. Raspon, kvartili, varijanca i standardna devijacija.

Uočite da su za podatke iz Primjera 1 aritmetička sredina i medijan vrlo blizu (brojevi 1.982 i 1.98), što općenito ne mora biti. Na primjer, za podatke
1, 1, 2, 2, 3, 11, 64
medijan je 2,
a aritmetička sredina 12,
što je veće od svih podataka osim jednog. Razlog tome je zbijenost podataka iz Primjera 1 i njihova grupiranost oko aritmetičke sredine, što nije slučaj s ovima gore. Za opisivanje takvih fenomena uvode se mjere raspršenja podataka. Najjednostavnija mjera raspršenja podataka je **raspon** ili **rang**.

Raspon podataka $x_1, x_2,...,x_n$ poredanih prema veličini je razlika x_n-x_1 najvećeg i najmanjeg podatka.

Na primjer, raspon podataka 1,1,2,2,3,11,64 je 64-1=63,
a raspon podataka iz Primjera 1 je 2.04-1.93=0.11 (sl.7.). Tu su točkicama predočene frekvencije pojedinih podataka.

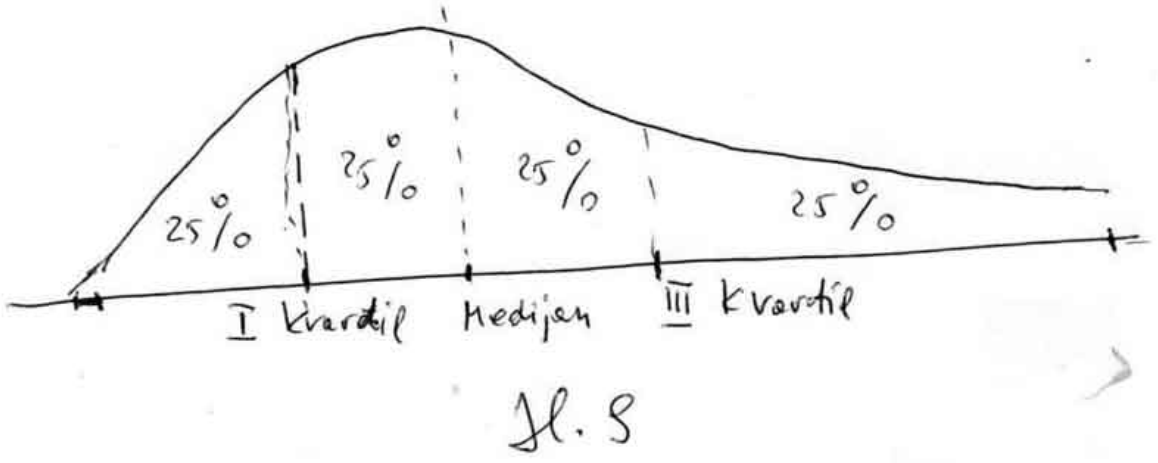


Sl.7.

Uočite da je raspon ujedno udaljenost na koordinatnom pravcu između najvećeg i najmanjeg podatka, odnosno to je duljina intervala određenog najmanjim i najvećim podatkom. Na slici 8. su kontinuirane aproksimacije histograma dviju serija podataka koji imaju jednake raspone, ali se očito bitno razlikuju. Naime, površine histograma drukčije su raspoređene.

Sl. 8.

Da bismo mogli opisivati razlike poput ovih na slici uvodimo pojam kvartila. Ima tri kvartila koji dijele histogram na četiri dijela, tako da svaki ima odprilike 25% površine (tim preciznije što je duljina uzorka veća). To predočujemo na kontinuiranoj aproksimaciji histograma (sl.9.).



Sl. 9

Prvi ili **donji kvartil** je broj od kojega je 25% podataka manje ili je njemu jednako.
Drugi je kvartil medijan.
Treći ili **gornji kvartil** je broj od kojega je 75% podataka manje ili je njemu jednako.

Napominjemo da 25% znači ¼, a 75% znači ¾, pa bi za primjenu ove definicije broj podataka trebao biti djeljiv s 4, a ni onda kvartili ne bi bili jednoznačno određeni. Zato se za računanje kvartila obično daju algoritmi. Opisat ćemo jedan od najuobičajenijih.

Primjer 5. Odredimo kvartile za podatke iz Primjera 1.

Tu je n=20, pa za prvi kvartil množimo 0.25·21 i dobijemo 5.25. Zato je prvi kvartil q_1 između petog podatka 1.95 i šestoga 1.96. Pravu vrijednost dobijemo kao
$q_1 = 1.95+0.25\cdot(1.96-1.95)=1.9525$
Koeficijent 0.25 u zadnjoj formuli dobili smo kao 5.25-5, što se slučajno poklopilo s 0.25 što se odnosi na prvi kvartil.
Već smo rekli da je drugi kvartil medijan što je bilo 1.98. Provjerimo to i algoritmom.
Za drugi kvartil množimo 0.5 s 21 i dobijemo 10.5, što znači da je drugi kvartil između 10-og i 11-og podatka. Kako su oba ta podatka 1.98, a 10.5-10=0.5 dobijemo
$q_2 = 1.98+0.5(1.98-1.98)=1.98$.
Za treći kvartil množimo 0.75 s 21 i dobijemo 15.75 pa je q_3 između 15-og podatka 2.00 i 16-og podatka 2.01. Kako je 15.75-15=0.75 dobijemo
$q_3 = 2.00+0.75(2.01-2.00)=2.0075$.

Slično kvartilima definiraju se i druge podjele, na primjer na **percentile**, kojima se histogram dijeli na 100 dijelova, svaki od kojih ima odprilike 1% površine. Mogu se razmatrati dijelovi s odprilike 10% površine itd. Općenito govorimo o **kvantilima**.

Varijanca i standardna devijacija.

Kvartili, percentili i, općenito, kvantili dobro opisuju variranje podataka unutar raspona, ali imaju jednu ozbiljnu slabost – ima ih puno: tri kvartila, 99 percentila itd.
Postavlja se pitanje postoji li neki broj ovisan o podatcima koji dobro opisuje variranje podataka. Odgovor je potvrdan, ima ih više, a najvažnija je varijanca. Varijanca je mjera rasipanja podataka oko aritmetičke sredine.
Odstupanje podatka x_i od aritmetičke sredine $\bar{x}$ mjeri se razlikom $x_i - \bar{x}$. Uočite da vrijedi:
ako je $x_i - \bar{x} > 0$ onda je podatak x_i veći od $\bar{x}$, tj. nalazi se desno od $\bar{x}$
ako je $x_i - \bar{x} < 0$ onda je podatak x_i manji od $\bar{x}$, tj. nalazi se lijevo od $\bar{x}$
ako je $x_i - \bar{x} = 0$ onda je $x_i = \bar{x}$.
Za ukupnu mjeru odstupanja nije dobro uzeti zbroj pojedinačnih odstupanja jer je to nula, tj. odstupanja se međusobno poništavaju. To ćemo potkrijepiti primjerom.

Primjer 6. Odredimo odstupanja podataka od aritmetičke sredine u Primjeru 1. i izračunajmo njihov zbroj.

Odstupanja su, redom:
1.93 – 1.982 = -0.052
1.94 – 1.982 = -0.042 dva puta
1.95 – 1.982 = -0.032 dva puta
1.96 – 1.982 = -0.022
1.97 – 1.982 = -0.012 tri puta
1.98 – 1.982 = -0.002 dva puta
1.99 – 1.982 = 0.008 dva puta
2.00 – 1.982 = 0.018 dva puta
2.01 – 1.982 = 0.028
2.02 – 1.982 = 0.038 dva puta
2.03 – 1.982 = 0.048
2.04 – 1.982 = 0.058

Zbroj odstupanja je
$\Sigma = -0.052-2\cdot0.042-2\cdot0.032-0.022-3\cdot0.012-$
$2\cdot0.002+2\cdot0.008+2\cdot0.018+0.028+2\cdot0.038+0.048+0.058$
$= -0.262+0.262$
$= 0.$
To vrijedi općenito, a ne samo u ovom primjeru, što se lako provjeri, a intuitivno je vrlo jasno. Naime, koliko ima tekućine ispod prosjeka, toliko mora biti i iznad prosjeka.

Napomena. U fizikalnoj interpretaciji gdje svakom podatku pridružujemo jednake mase, rezultat Primjera 6. vrlo je jasan. On upravo govori da je u aritmetičkoj sredini težište, tj. ravnoteža (doprinosi *sila puta krak sile* lijevo i desno od težišta su jednaki).

Suma apsolutnih odstupanja i prosječno apsolutno odstupanje.
Kao **dobra mjera rasipanja** podataka oko srednje vrijednost služi **suma apsulutnih vrijednosti odstupanja podataka od aritmetičke sredine**. Definira se kao:

$SAO := |x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|.$

Na primjer, koristeći rezultate iz Primjera 6, dobijemo za podatke iz Primjera 1:

SAO = 0.262+0.262 = 0.524.

To tumačimo kao da je ukupno odstupanje (na niže i na više) od prosječne vrijednosti 1.982 oko pola litre.
Ako taj rezultat podijelimo s 20 (brojem staklenka), dobit ćemo **prosječno apsolutno odstupanje** od prosjeka: $\frac{0.524}{20}$ =0.0262.
To znači da, u prosjeku, u svakoj staklenki ili ima za 0.0262 litara više ili 0.0262 litara manje kemikalije od 1.982 litara.

Općenito, **prosječno apsolutno odstupanje od aritmetičke sredine**, definiramo kao

$$\text{PAO}:= \frac{|x_1-\bar{x}|+|x_2-\bar{x}|+...+|x_n-\bar{x}|}{n}$$

Nedostatak izraza SAO i PAO jest taj da u definiciji sadrže apsolutnu vrijednost, koja nije baš pogodna za deriviranje. To i neki drugi **prirodni razlozi** utjecali su da se PAO zamijeni **standardnom devijacijom**, koju ćemo sad definirati.

Varijanca uzorka $(s')^2$ definira se kao **prosječno kvadratno odstupanje od prosjeka**:

$$(s')^2 := \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n}$$

Standardna devijacija uzorka s' je drugi korijen iz varijance uzorka:

$$s' := \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n}}$$

Primjer 7. Izračunajmo varijancu i standardnu devijaciju uzorka (podataka) iz Primjera 1.

$$(s')^2 = \frac{(-0.052)^2+2\cdot(-0.042)^2+...+0.048^2+0.058^2}{20} = \frac{0.01932}{20} = 0.000966$$

$s' = 0.0310805$
(i jedno idrugo su približne vrijednosti).

Uočite da je ispalo SAO<s'.
To vrijedi općenito, a ne samo u ovom slučaju.

Fizikalna interpretacija varijance – moment inercije oko težišta.
Sjetimo se da je moment inercije mase proporcionalan masi i kvadratno proporcionalan radijusu: $I = mr^2$. Sjetimo se također da se momenti inercije oko istog središta zbrajaju.
Zamislimo da smo **jediničnu masu** rasporedili po pravcu tako
da svaki podatak x_i uzorka opteretimo jednakom masom $m := \frac{1}{n}$, onda će ukupna inercija oko težišta (aritmetičke sredine) biti

$I=\frac{1}{n}(x_1-\bar{x})^2+\frac{1}{n}(x_2-\bar{x})^2+\ldots+\frac{1}{n}(x_n-\bar{x})^2$
$=(s')^2$.

Zaključujemo: **varijanca uzorka jednaka je momentu inercije oko težišta pripadajućeg sustava masa, pri čemu smo jediničnu masu jednoliko rasporedili na sve podatke – svakom po** $\frac{1}{n}$.

To je i bila jedna od motivacija za definiciju varijance uzorka. Naime, kako je moment inercije mjera disperzije (raspršenja) masa oko težišta, tako je i varijanca mjera disperzije podataka oko aritmetičke sredine.

Fizikalno je jasno da je od svih momenata inercije najmanji onaj oko težišta. Slično svojstvo minimalnosti ima i aritmetička sredina: to je realni broj od kojega je suma kvadrata odstupanja minimalna. To se može lako dokazati. Provjerite da je u Primjeru 1 suma kvadrata odstupanja od aritmetičke sredine manja od sume kvadrata odstupanja od medijana.

Procjena varijance populacije. **Korigirana varijanca i korigirana standardna devijacija**. Za razliku od aritmetičke sredine uzorka, koja je *najbolja* procjena aritmetičke sredine populacije, varijanca uzorka **nije najbolja procjena** varijance populacije. Pokazuje se da to svojstvo ima **korigirana varijanca uzorka** s^2, definirana kao:

$$s^2:=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$$

(razlikuje se po tome što u nazivniku, umjesto n ima n-1, a u oznaci što nema crtice).

Odatle se definira **korigirana standardna devijacija uzorka** s, kojom ćemo procjenjivati standardnu devijaciju populacije:

$$s:=\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}}.$$

Primjer 8. Izračunajmo korigiranu varijancu i korigiranu standardnu devijaciju uzorka (podataka) iz Primjera 1.

Koristeći se rezultatom Primjera 6, dobijemo

$s^2=\frac{0.01932}{19}=0.00101684$, te

$s=0.031888$

Vrijedi općenito, a može se provjeriti za ovaj uzorak:
SAO < s' < s.

Vidi se da vrijedi $s=\sqrt{\frac{n-1}{n}}s'$. Zato se za velike n devijacije s i s' praktično ne razlikuju. Međutim, za male n, razlika među njima je nezanemariva. U sljedećoj tablici dan je **približan** omjer (na tri decimale) između s i s' za neke n (kako se n povećava, tako se taj omjer približava broju 1).

n	2	3	4	5	6	7	8	9	10
$\frac{s}{s'}$	0.707	0.816	0.866	0.894	0.913	0.926	0.935	0.943	0.949

n	20	30	100	500
$\frac{s}{s'}$	0.975	0.983	0.995	0.999

Sad ćemo obrađene pojmove ilustrirati na jednom novom uzorku, ponešto različitom od onog u Primjeru 1.

Primjer 9. Mjerenjem vremena između dviju uzastopnih poruka pristiglih na neku adresu dobiveni su sljedeći podatci (u sekundama):
12, 8, 1, 7, 24, 4, 4, 6, 20, 10, 3, 2, 22, 23, 8, 6, 5, 25, 16, 3, 1, 14, 15, 18, 2, 6, 27, 19, 12, 4, 20, 14, 3, 13, 8, 15, 30, 5, 7, 16.

(I) Prebrojimo podatke. Vidimo da ih ima 40, dakle n = 40.

(II) Poredajmo podatke prema veličini (od manjeg prema većem):
1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6,6,6, 7, 7, 8, 8, 8, 10, 12, 12, 13, 14, 14, 15, 15, 16, 16, 18, 19, 20, 20, 22, 23, 25, 27, 30.
(III) Napravimo tablicu frekvencija:

1	2	3	4	5	6	7	8	10	12	13	14	15	16	18	19	20	22	23	24	25	27	30
2	2	3	3	2	3	2	3	1	2	1	2	2	2	1	1	2	1	1	1	1	1	1

Vidimo da frekvencije variraju iako imaju i opći trend prema opadanju. To bi još izrazitije bilo da smo stavili frekvencije 0 za brojeve od 1 do 30 koji se ne pojavljuju.

(IV) Grupirajmo podatke u razrede duljine 5:

0.5 - 5.5	5.5 – 10.5	10.5 – 15.5	15.5 – 20.5	20.5 – 25.5	25.5 – 30.5
11	9	7	6	4	2

Vidimo da, nakon ovakvog grupiranja, frekvencije razreda opadaju, što se dobro vidi i iz histograma. To je jedan od najvažnijih razloga grupiranja.

(V) Odredimo, najmanji podatak, najveći podatak i raspon:
min = 1
max = 30
raspon = max – min = 30-1 = 29.

(VI) Odredimo medijan i aritmetičku sredinu i unaprijed procijenimo njihov odnos. Odredimo kvartile.
S obzirom da su podatci više grupirani na početak, medijan je manji od aritmetičke sredine. Kako je n = 40, medijan je aritmetička sredina 20-og i 21-og podatka. Dakle:

Medijan = $\frac{8+10}{2} = 9$

Aritmetička sredina, $\bar{x} = \frac{458}{40} = 11.45$ (zaista je medijan manji).

Prvi kvartil: $q_1 = 4.5$
Drugi kvartil (medijan): $q_2 = 9$
Treći kvartil: $q_3 = 17$

(VII) Odredimo varijancu i standardnu devijaciju te korigiranu varijancu i korigiranu standardnu devijaciju uzorka.

Varijanca: $(s')^2 = 63.1975$
Standardna devijacija: s' = 7.9497 (na 4 decimale)
Korigirana varijanca: $s^2 = 64.8179$ (na 4 decimale)
Korigirana standardna devijacija: s = 8.0510 (na 4 decimale).

Veličine koje smo odredili u Primjeru 9 jesu osnovne deskriptivno statističke veličine uzorka. Za njihovo računanje možemo se koristiti gotovim statističkim paketima. Na primjer, pomoću grafičkog kalkulatora te se veličine dobiju primjenom jedne naredbe.

Važna svojstva standardne devijacije – Čebiševljev teorem i empirijsko pravilo za zvonolike distribucije frekvencija.

Čebiševljev teorem. *Neka je $\bar{x}$ aritmetička sredina i s' standardna varijanca uzorka $x_1, x_2,\ldots,x_n$. Tada u intervalu $< \bar{x} - 2\cdot s',\ \bar{x} + 2\cdot s' >$ ima barem 75% podataka, a u intervalu $< \bar{x} - 3\cdot s',\ \bar{x} + 3\cdot s' >$ ima barem 88% podataka.*

Napomena. Budući da je s'<s, Čebiševljev teorem vrijedi i za korigiranu standardnu devijaciju.

Primjer 10. Provjerimo Čebiševljev teorem na uzorku iz Primjera 9.

Tu je 2s' = 15.89 (na dvije decimale) i $\bar{x}$ =11.45 pa je
$< \bar{x} - 2\cdot s',\ \bar{x} + 2\cdot s' > = <$ - 4.44, 27.34>
Vidimo da su njemu svi podatci osim podatka 30, pa je tvrdnja provjerena.

Primjer 11. Provjerimo Čebiševljev teorem na uzorku iz Primjera 1.

Tu je uzorak
1.93 1.94 1.94 1.95 1.95 1.96 1.97 1.97 1.97 1.98 1.98 1.99 1.99
2.00 2.00 2.01 2.02 2.02 2.03 2.04
Prema Primjeru 3, $\bar{x}$ = 1.982 .
Prema Primjeru 7, s' = 0.031 (na tri decimale), dakle 2s'=0.062.

Zato je $\bar{x}$ +2s'=2.044 i $\bar{x}$ -2s'=1.920. Vidimo da su svi zadani podatci između 1.920 i 2.044 (a Čebiševljev teorem garantira bar 75%).

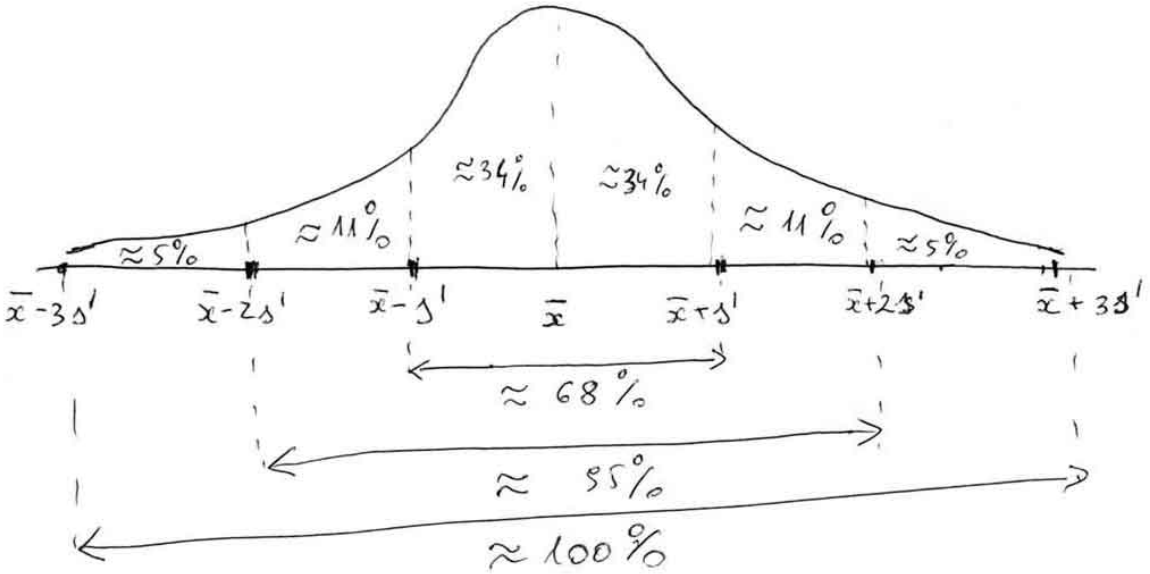
Empirijsko pravilo za zvonolike distribucije frekvencija.

Letimičan pogled na uzorke iz Primjera 1 i 9, odnosno na njihove histograme, upućuju na razlike među njima. Površine u histogramu iz Primjera 9 opadaju, dok za Primjer 1 vrijedi:

(N1) Površina je koncentrirana oko aritmetičke sredine.
(N2) Površina je približno simetrično raspoređena lijevo i desno od aritmetičke sredine
(N3) Površine rastu odprilike do aritmetičke sredine, potom padaju.

Uz ove uvjete histogram (odnosno pripadna krivulja) ima **zvonolik oblik.** Praksa pokazuje da takav oblik imaju histogrami distribucija kod **velikih uzoraka,** pri mjerenju mnogih statističkih fenomena (**statističkih obilježja**), poput mase, visine, postotka elementa koji se može nekom tehnološkom metodom izdvojiti iz neke rudače, grješaka pri mjerenju, kvocijenta inteligencije itd. Za takva statistička obilježja **uočeno je** sljedeće **empirijsko pravilo** (sl.10.):

U intervalu $< \bar{x} - s',\ \bar{x} + s' >$ ima oko 68% podataka, tj. oko 2/3 podataka (površine histograma)
U intervalu $< \bar{x} - 2 \cdot s',\ \bar{x} + 2 \cdot s' >$ ima oko 95% podataka (površine histograma)
U intervalu $< \bar{x} - 3 \cdot s',\ \bar{x} + 3 \cdot s' >$ su gotovo svi podatci (gotovo čitava površina).



Sl. 10

Napomenimo da empirijsko pravilo vrijedi samo približno i ne za sva statistička obilježja i, obično, samo za velike uzorke, dok je Čebiševljev teorem egzaktan (vrijedi bez ikakvih ograničenja). O tome ćemo više govoriti u teoriji vjerojatnosti.

Treba također napomenuti da postotci kod prebrojavanja podataka, neće biti jednaki onima pri procjeni površine histograma. Razlika među njima bit će znatnija za relativno male uzorke.

Za statistička obilježja za koje vrijedi empirijsko pravilo kažemo da su (približno) **normalno distribuirani**.

Već smo vidjeli da su svi podatci iz Primjera 1. u intervalu $< \bar{x} - 2\cdot s',\ \bar{x} + 2\cdot s' >$.
Također, vidi se da je $< \bar{x} - s',\ \bar{x} + s' > = <1.951,\ 2.013 >$,
pa zaključujemo da je u tom intervalu 11 od 20 podataka, što je oko 55%. Da smo gledali površinu u histogramu, to bi bilo odprilike 60% površine, što je još uvijek nešto manje od 68%.
Tako, iako je u tom primjeru statističko obilježje bila količina kemikalije u staklenkama, što bi trebalo biti normalno distribuirano, uočava se odudaranje od empirijskog pravila. Razlog tome je relativno mala veličina uzorka. U sljedećem ćemo primjeru podatke iz Primjera 1 upotpuniti s još 20 novih podataka (20 novih staklenka).

Primjer 12. Mjerenjem količine kemikalije u dodatnih 20 staklenka iz Primjera 1. dobiveni su sljedeći podatci i unešeni u tablicu frekvencija (sad uzorak ima duljinu 40)..

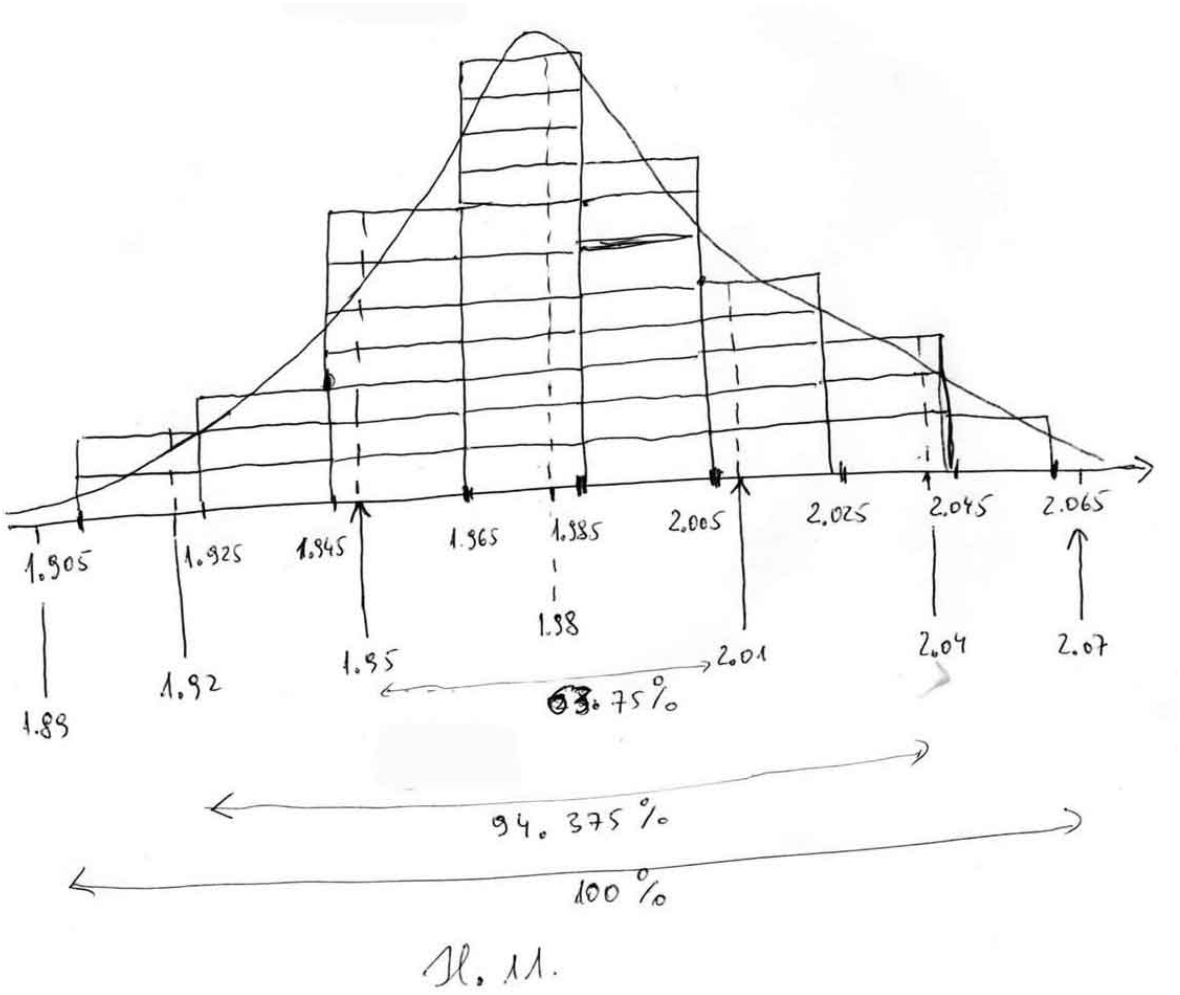
x_i	1.91	1.92	1.93	1.94	1.95	1.96	1.97	1.98	1.99	2.00	2.01	2.02	2.03	2.04	2.05
f_i	1	1	1	2	3	4	5	6	4	4	3	2	2	1	1

Provjerimo empirijsko pravilo o normalnoj distribuiranosti količine kemikalije u staklenkama.

Tu je $\bar{x} = 1.98$ i $s' = 0.030455$ (na šest decimala). Zato, na dvije decimale imamo:
$s'=0.03$, $2s'=0.06$, $3s'=0.09$
Vidimo, da je, na dvije decimale:
$< \bar{x} - s',\ \bar{x} + s' > = <1.95,\ 2.01>$
$< \bar{x} - 2\cdot s',\ \bar{x} + 2\cdot s' > = <1.92,\ 2.04>$,
$< \bar{x} - 3\cdot s',\ \bar{x} + 3\cdot s' > = <1.89,\ 2.07>$.

Budući da je broj podataka relativno malen, **za naslućivanje stanja u cijeloj populaciji**, realnije je razmatrati površine u histogramu, na primjer uz duljinu razreda 0.02, nego prebrojavati podatke uzorka u pojedinim intervalima. Tako dobijemo (vidi sliku 11.):

U intervalu $< \bar{x} - s',\ \bar{x} + s' >$ je 63.75% površine histograma.
U intervalu $< \bar{x} - 2\cdot s',\ \bar{x} + 2\cdot s' >$ je 94.375% površine histograma
U intervalu $< \bar{x} - 3\cdot s',\ \bar{x} + 3\cdot s' >$ je 100% površine histograma.



Sl. 11.

To se u velikoj mjeri slaže s empirijskim pravilom, što nam je oslonac za vjerovanje da je količina kemikalije u staklenkama (približno) normalno distribuirana.

Dvije osnovne vrste statističkih obilježja: kontinuirana i diskretna statistička obilježja.

U Primjeru 1. mjerili smo količinu kemikalije u litrama i rezultate mjerenja zapisivali na dvije decimale. Jasno je da smo, uz pomoć preciznijih uređaja, mjerenja mogli provoditi na tri, četiri ili više decimala. Načelno, rezultati mjerenja mogu biti sve precizniji, tako da za njihove zapisivanje trebamo sve realne brojeve. Zato kažemo da je količina **kontinuirana** ili da ima **kontinuirano statističko obilježje.** Slično je s masom, visinom, obujmom, vremenom itd. Na primjer, u Primjeru 9. mjeri se vrijeme između dviju poruka, pa je riječ o kontinuiranom statističkom obilježju, iako su podatci bili cijeli brojevi. Naime, rezultate smo pisali u sekundama, a da smo imali precizniji uređaj, koji registrira desetinke sekunda, rezultati bi (vjerojatno) bili decimalni brojevi.
Drugo važno statističko obilježje jest **diskretno statističko obilježje**. Ono u pravilu nastaje u pokusima u kojima nešto prebrojavamo. To ćemo ilustrirati primjerom.

Primjer 13. Da bismo dobili predodžbu o broju poruka koje pristignu na neku adresu tijekom fiksiranog vremenskog intervala (od 8 do 10 sati prije podne), kontrolirali smo tu adresu u 60 takvih intervala. Dobili smo sljedeće podatke.
4, 3,3, 0, 0, 5, 7, 1, 5, 1, 2, 3, 6, 2, 2, 2, 4, 0, 3, 3, 1, 4, 5, 6, 2, 1, 3, 2, 2, 0 , 5, 1, 2, 1, 3, 3, 3, 3, 4, 6, 8, 4, 2, 2, 1, 0, 4, 5, 2, 5, 1, 0, 2, 3, 3, 0, 1, 4, 2 , 5

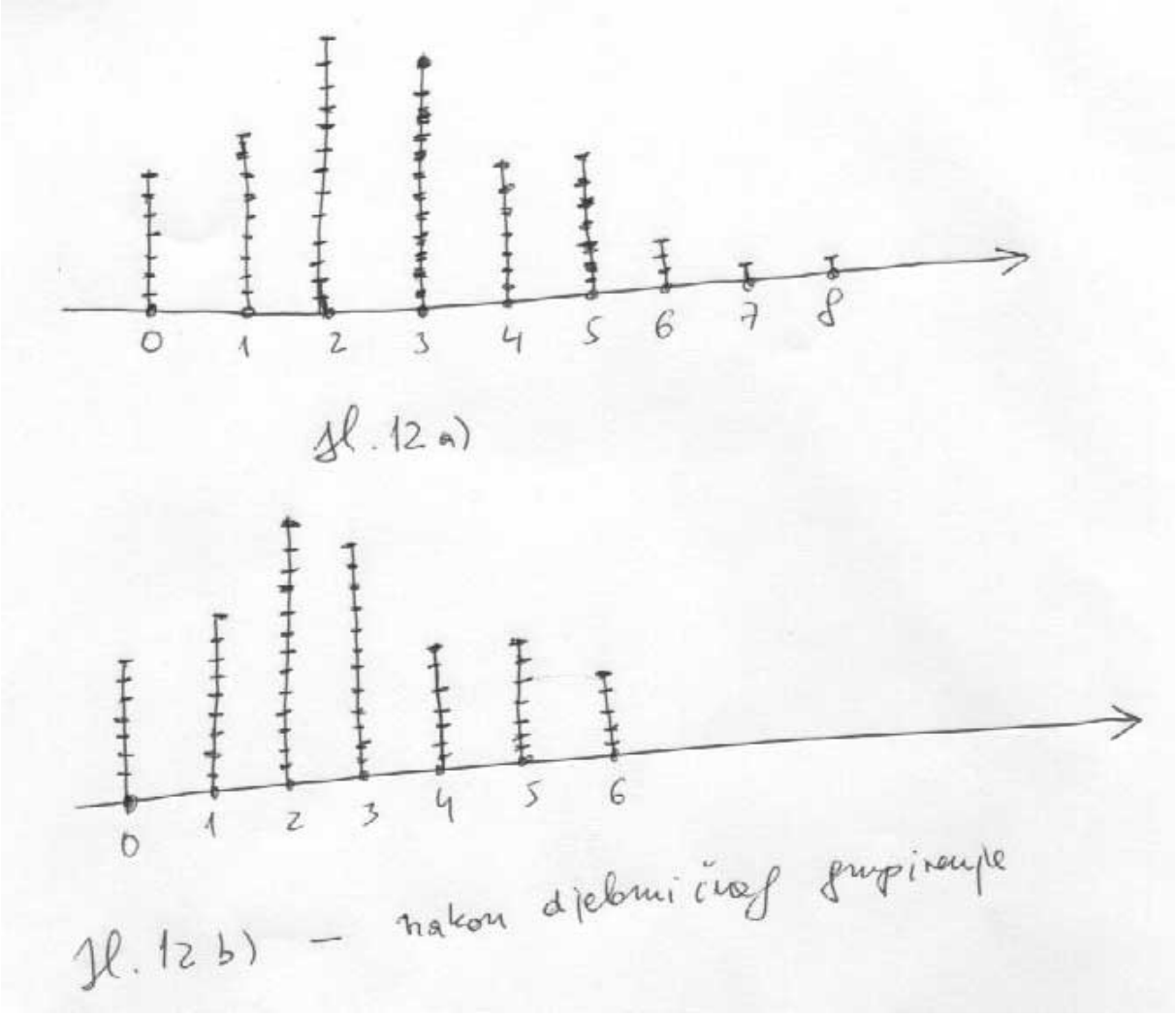
Te podatke možemo predočiti sljedećom tablicom frekvencija.

0	1	2	3	4	5	6	7	8
7	9	13	12	7	7	3	1	1

Umjesto ove stvarne tablice frekvencija, obično se koristi sljedeća, modificirana (nakon djelomičnog grupiranja).

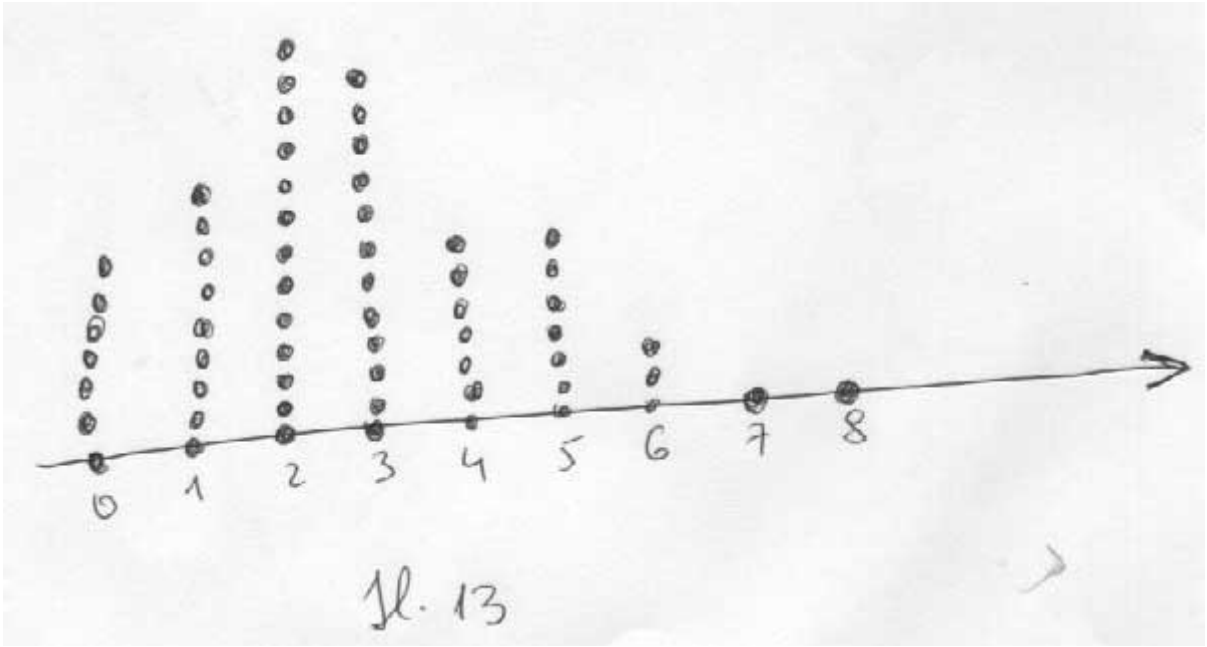
x_i |	0	1	2	3	4	5	6 ili više
f_i |	7	9	13	12	7	7	5

Tu smo tri šestice, jednu sedmicu i jednu osmicu stavili skupa u razred koji smo nazvali *šest ili više*.
U ovom je primjeru riječ o diskretnom obilježju pa za grafičko predočavanje podataka nije pogodan histogram, već **dijagram frekvencija** kao na slici 12.



Tu smo iznad svake vrijednosti x_i na koordinatnom pravcu postavili dužinu kojoj je duljina jednaka pripadnoj frekvenciji f_i. Ukupna duljina ovih dužina jednaka je ukupnom broju podataka n (u ovom je primjeru n = 60).

Još je uobičajenije grafičko predočavanje pomoću **dijagrama relativnih frekvencija**, u kojemu je duljina dužine iznad pojedinog podatka x_i jednaka pripadnoj relativnoj frekvenciji f_i/n. Ukupna duljina dužina sad je jednaka 1.
Dijagram frekvencija mogli smo predočiti i točkicama kao na slici 13.



Sl. 13

Aritmetička sredina, standardna devijacija, korigirana standardna devijacija, raspon i kvartili dobiju se kao i kod kontinuiranih obilježja. Dakle, na tri decimale dobijemo:

$$\bar{x} = \frac{0\cdot 7 + 1\cdot 9 + 2\cdot 13 + 3\cdot 12 + 4\cdot 7 + 5\cdot 7 + 6\cdot 3 + 7\cdot 1 + 8\cdot 1}{60} = 2.783$$

s' = 1.881
s = 1.896
Raspon = 8-0 = 8
Prvi kvartil $q_1 = 1$ (objasnite)
Drugi kvartil – medijan = 3 (objasnite)
Treći kvartila $q_3 = 4$ (objasnite).

Napomena. Statističke veličine u primjerima poput Primjera 13. često se ne računaju za sve podatke, već približno, tj. koristeći se podatcima iz modificirane tablice (nakon djelomičnog grupiranja). Tada se dobije:

$$\bar{x} = \frac{0\cdot 7 + 1\cdot 9 + 2\cdot 13 + 3\cdot 12 + 4\cdot 7 + 5\cdot 7 + 6\cdot 5}{60} = 2.733$$

s' = 1.769
s = 1.784
Raspon = 6 - 0 = 0
Prvi kvartil $q_1 = 1$ (objasnite)
Drugi kvartil – medijan = 3 (objasnite)
Treći kvartila $q_3 = 4$ (objasnite).