Primijenjena i inženjerska matematika

Statistika - predavanja 1 i 2

Tipovi i opisivanje podataka

13.10.2023.

Statistika

- **statistika** - riječ koja je u svakodnevnom životu vezana uz brojčane vrijednosti kojima pokušavamo opisati bitne karakteristike nekog skupa podataka
- Državni zavod za statistiku Republike Hrvatske (`http://www.dzs.hr/`):
  - *prosječna mjesečna isplaćena neto plaća po zaposlenome u pravnim osobama Republike Hrvatske za srpanj 2016. iznosila je* 5594 *kune*
  - *minimalna plaća za razdoblje od 1. lipnja 2012. do 30. lipnja 2012. u Republici Hrvatskoj iznosila je* 2814 *kuna*
  - *stopa registrirane nezaposlenosti za kolovoz 2016. iznosila je* 13.1%

- **statistika** - znanstvena disciplina koja se bavi razvojem metoda **prikupljanja**, **opisivanja** i **analiziranja** podataka te primjenom tih metoda u procesu **donošenja zaključaka** na temelju prikupljenih podataka
- **statističko istraživanje** - fokusirano na skup **objekata**, tj. **jedinki** (ljudi, životinja, biljaka, stvari, država, gradova, poduzeća, itd.) i skup odabranih veličina koje se na njima promatraju (skup **varijabli** ili **obilježja**)
- **populacija** - skup svih jedinki koje se žele obuhvatiti istraživanjem (o kojima se želi zaključivati)
- primjeri 1.1 i 1.2

Primjena statistike u istraživanju podrazumijeva:

- populaciju koja je predmet istraživanja potrebno je detaljno proučiti, zabilježiti njene osnovne karakteristike i ciljeve istraživanja, kreirati kvalitetan uzorak i odabrati metodu za prikupljanje podataka
- izabrati prikladne metode za opis skupa prikupljenih podataka (deskriptivna statistika)
- izabrati prikladne statističke metode za zaključivanje o populaciji na temelju prikupljenih podataka na uzorku (statističko zaključivanje)

U ovom ćemo se kolegiju baviti:

- metodama prikupljanja podataka i kreiranja uzorka
- metodama deskriptivne statistike
- metodama statističkog zaključivanja

Populacija i uzorak

- **populacija** - skup svih jedinki koje su predmet istraživanja, tj. zadovoljavaju neka svojstva bitna za obilježje koje se istražuje
- **uzorak** - populacija može sadržavati vrlo velik broj jedinki i stoga je teško ili čak nemoguće istraživanje provesti na svim jedinkama populacije - tada odabiremo jedan podskup populacije na kojemu je osigurano kvalitetno provođenje istraživanja, a kojeg nazivamo uzorak
- **reprezentativan uzorak** - uzorak u kojem su zastupljene sve tipične karakteristike populacije bitne za istraživanje
- primjeri 2.1 i 2.2
- **slučajan uzorak** - uzorak konstruiran u skladu sa zahtjevom da svaka jedinka populacije ima jednaku vjerojatnost (šansu) biti izabrana u uzorak

Izvori podataka

Način prikupljanja podataka ovisi o karakteristikama obilježja koje je predmet proučavanja:

- podaci iz javnih izvora (knjige, časopisi, novine, Internet)
- podaci iz dizajniranog eksperimenta (istraživač raspoređuje eksperimentalne jedinke u skupine nad kojima vrši eksperimente te bilježi podatke za varijable koje ga zanimaju)
- podaci iz ankete (istraživač sastavlja anketni upitnik, izabire skupinu ljudi koju anketira i na osnovu njihovih odgovora prikuplja podatke)
- podaci prikupljeni promatranjem (istraživač promatra eksperimentalne jedinke u njihovom prirodnom okruženju i bilježi podatke za varijable od interesa)
- primjer 2.3

Organizacija baze podataka

- baza podataka najčešće će biti organizirana u obliku tablice u kojoj svaki redak predstavlja jednu jedinku, a svaki stupac promatrane varijable
- glukoza.sta

↭

Tipovi varijabli - kvalitativne varijable

Kvalitativne (kategorijalne) varijable - varijable čije vrijednosti po svojim svojstvima nisu realni brojevi nego njihove vrijednosti svrstavamo u kategorije. Razlikujemo:

- **Kvalitativne nominalne varijable** - među kategorijama nema prirodnog poretka, npr.
  - boja očiju (plava, smeđa, zelena)
  - krvne grupe (A, B, AB, 0)
  - spol (m ili ž)
  - radna mjesta u školi (spremačica, domar, tajnik, nastavnik, pedagog, ravnatelj)
- **Kvalitativne ordinalne varijable** - među kategorijama postoji poredak, npr.
  - stručna sprema (SSS, VŠS, VSS)
  - opisne ocjene (ništa, malo, srednje, puno)

Tipovi varijabli - numeričke varijable

Numeričke varijable - varijable prirodno primaju vrijednosti iz skupa realnih brojeva, npr.

- postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine
- broj bodova na državnoj maturi iz matematike
- broj ulovljenih komaraca u klopku
- temperatura mora
- koncentracija soli u morskoj vodi

- kategorije kvalitativnih varijabli mogu se izražavati brojevima, no to ih ne čini numeričkim varijablama
- među numeričkim varijablama razlikujemo **diskretne** i **neprekidne** varijable

Tipovi varijabli - diskretne numeričke varijable

Diskretne numeričke varijable - mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti (vrijednosti možemo ispisati u niz), npr.

- broj djece
- broj bodova na državnoj maturi iz matematike
- broj ulovljenih komaraca u klopku
- broj dana u godini s temperaturom zraka većom od $35°$C

Tipovi varijabli - neprekidne numeričke varijable

Neprekidne numeričke varijable - skup mogućih vrijednosti je cijeli skup realnih brojeva ili neki interval, npr.

- postotak prolaznosti na pojedinim ispitima u jednoj akademskoj godini
- temperatura nekog mora
- vodostaj neke rijeke

- u svrhu prikaza podataka i nekih statističkih analiza, vrijednosti numeričke varijable se također mogu svrstati u **kategorije**
- primjer 2.8
- primjer 2.9

↭

Metode opisivanja kvalitativnih podataka

- kvalitativne varijable - primaju vrijednosti koje su razvrstane u kategorije
- pri proučavanju kvalitativnih varijabli pažnju usmjeravamo na zastupljenost pojedine kategorije u uzorku na kojem provodimo istraživanje
- **frekvencija kategorije**, **relativna frekvencija kategorije** - mjere kojima opisujemo zastupljenost jedne kategorije u uzorku

Primjer

primjer 3.1

ispitanik	spol	krvna grupa
1	Ž	A
2	Ž	B
3	M	0
4	Ž	0
5	M	AB
6	M	B
7	Ž	B
8	M	A
9	Ž	AB
10	Ž	A

Metode opisivanja kvalitativnih podataka

- neka varijabla, koju ćemo označiti X, ima k kategorija (recimo $k = 4$ znači da varijabla ima 4 kategorije - npr. krvne grupe)
- označimo pojedine kategorije kao $x_1, x_2, \ldots, x_k$, odnosno, u drugom zapisu $\{x_i : i = 1, \ldots, k\}$
- **frekvencija kategorije** x_i je broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i, $i = 1, \ldots, k$
- frekvenciju kategorije x_i označavamo f_i
- frekvencija pojedine kategorije ovisi o broju izvršenih mjerenja odnosno veličini uzorka

Metode opisivanja kvalitativnih podataka

- **relativna frekvencija (udio) kategorije** x_i je broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i podijeljen s ukupnim brojem izmjerenih vrijednosti za promatranu varijablu, $i = 1, \ldots, k$
- ako je n veličina uzorka (broj svih izmjerenih vrijednosti promatrane varijable), relativnu frekvenciju kategorije x_i računamo kao

$$\frac{f_i}{n}$$

- relativna frekvencija kategorije je mjera zastupljenosti koja daje informaciju o udjelu kategorije u uzorku poznate dimenzije i često se izražava kao postotak

Metode opisivanja kvalitativnih podataka

- **tablično** - tablicom frekvencija i relativnih frekvencija (primjeri 3.2 - 3.4)
- **grafički** - stupčastim dijagramom i kružnim dijagramom frekvencija i relativnih frekvencija (primjeri 3.5 - 3.7)
- u tabličnom i grafičkom prikazu frekvencija i relativnih frekvencija trebaju biti zastupljene sve kategorije promatrane varijable

$\leftrightsquigarrow$

Metode opisivanja numeričkih podataka

Numeričke varijable mogu biti:

- **diskretne** - primaju konačno ili prebrojivo mnogo vrijednosti,
- **neprekidne** - vrijednost može biti bilo koji broj iz nekog intervala realnih brojeva ili bilo koji realan broj

Metode opisivanja numeričkih i ordinalnih podataka - diskretne numeričke i ordinalne varijable

Ako su varijable ordinalne ili ako su numeričke diskretne s malo mogućih vrijednosti, za opis izmjerenih vrijednosti tih varijabli možemo koristiti **iste metode kao pri opisivanju kvalitativnih podataka:**

- tablice frekvencija i relativnih frekvencija
- grafičke prikaze frekvencija i relativnih frekvencija - stupčasti i kružni dijagram
- primjer 3.8

Metode opisivanja numeričkih podataka - neprekidne numeričke varijable

- **neprekidne numeričke varijable** - za prikazivanje podataka iz takvih varijabli su frekvencije, stupčasti i kružni dijagrami napravljeni na osnovu svake pojedine izmjerene vrijednosti nepraktični
- primjer 3.9

Metode opisivanja numeričkih podataka - postupak razvrstavanja neprekidnih numeričkih podataka u kategorije

Razvrstavanje izmjerenih vrijednosti neprekidne numeričke varijable u **kategorije**:

- skup svih mjerenih vrijednosti (ili nešto veći skup koji sadrži skup svih mjerenih vrijednosti ali kojega je jednostavnije podijeliti na jednake dijelove) podijeliti na disjunktne intervale jednake duljine
- nema točno definiranog pravila po kojemu bi trebalo definirati duljine intervala niti nihov broj, ali ne smije ih biti niti previše niti premalo da bi cijeli postupak imao smisla i služio svrsi
- kriterij za kategorizaciju vrijednosti kontinuirane numeričke varijable treba biti temeljen na razumijevanju problema koji proučavamo

Tako dobiveni stupčasti dijagram naziva se **histogram** - stupci su postavljeni u koordinatni sustav nad odgovarajućim intervalima

- primjer 3.10

Metode opisivanja numeričkih podataka - mjere deskriptivne statistike

Karakteristika numeričkih varijabli - među njihovim vrijednostima postoji prirodan uređaj i možemo definirati **numeričke karakteristike** tih varijabli koje imaju logičnu interpretaciju i mogu se iskoristiti u cilju prikazivanja skupa mjerenih vrijednosti:

- mjere centra podataka – opisuju centar podataka (tipičnu vrijednost)
- mjere raspršenosti podataka – opisuju raspršenost podataka

Mjere centra podataka
aritmetička sredina podataka

- **aritmetička sredina** (eng. mean) niza izmjerenih vrijednosti (podataka) $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom

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

- npr. neka su $1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8$ izmjerene vrijednosti jedne varijable
- obzirom da ih ima ukupno devet, aritmetička sredina ovog skupa podataka je

$$\frac{1.2 + 2.1 + 3.2 + 4.3 + 5.4 + 6.5 + 7.6 + 8.7 + 9.8}{9} \approx 5.42$$

Mjere centra podataka
medijan podataka

- **medijan** ima značenje izmjerene vrijednosti koja se nalazi na sredini niza podataka kada je on uređen po veličini - barem pola podataka je manje ili jednako medijanu, a istovremeno je barem pola podataka veće ili jednako od medijana
- način njegovog određivanja ovisi o tome imamo li **neparan** ili **paran** broj podataka

Mjere centra podataka
medijan podataka - neparan broj podataka

- ukoliko imamo **neparan broj** izmjerenih vrijednosti, onda postoji podatak koja je na srednjoj poziciji u uređenom skupu izmjerenih vrijednosti, pa njega definiramo kao medijan
- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3$ izmjerene vrijednosti jedne varijable
- prvo ove vrijednosti poredamo po veličini: $1, 1, 2, 2, 2, \mathbf{2}, 3, 5, 5, 6, 7$
- obzirom da ih ima ukupno jedanaest, medijan je vrijednost koja je na šestoj poziciji u tako dobivenom nizu, tj. broj 2

Mjere centra podataka
medijan podataka - paran broj podataka

- ukoliko imamo **paran broj** izmjerenih vrijednosti, onda ne postoji podatak koji je na srednjoj poziciji jer srednju poziciju "zauzimaju" dva podatka - medijan se tada definira kao polovina između ta dva podatka (tj. aritmetička sredina tih dvaju podataka)
- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti jedne varijable
- prvo ove vrijednosti poredamo po veličini:
  $1, 1, 2, 2, 2, \mathbf{2}, \mathbf{3}, 3, 5, 5, 6, 7$
- obzirom da ih ima dvanaest, "sredinu" čine šesti i sedmi podatak, tj. brojevi 2 i 3 - medijan ovog skupa podataka je sredina ta dva broja, tj. medijan je $(2 + 3)/2 = 2.5$

Mjere centra podataka
mod podataka

- **mod** podataka je vrijednost iz niza izmjerenih vrijednosti varijable X kojoj pripada najveća frekvencija (izmjerena je najviše puta)
- mod ne mora biti jedinstven
- npr. neka su

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$$

izmjerene vrijednosti jedne varijable - vrijednost 2 je izmjerena najviše puta (četiri puta) pa je 2 mod ovog skupa podataka

- npr. neka su

$$1, 2, 5, 6, 3, 3, 1, 2, 7, 2, 2, 3, 3$$

izmjerene vrijednosti jedne varijable - najviše puta izmjerene su dvije vrijednosti, tj. 2 i 3 su obje izmjerene točno četiri puta pa mod ovog skupa podataka nije jedinstven nego su mod i 2 i 3

Odabir mjere centra podataka

- Primjer: U jednoj tvrtki s devet zaposlenika zabilježeni su sljedeći iznosi plaća zaposlenika u kunama:

$$4200,\ 3100,\ 3200,\ 3150,\ 3850,\ 14000,\ 15000,\ 3700,\ 3800.$$

Izračunajte aritmetičku sredinu i medijan.

Odabir mjere centra podataka

- Primjer: Prema izvješću Državnog zavoda za statistiku prosječna neto plaća po zaposlenom u Hrvatskoj za svibanj 2019. iznosila je 6476kn.
  Medijan svih isplaćenih neto plaća iznosi 5590kn prema istom izvješću (informativniji pokazatelj i bolje odražava tipičnu plaću).
  Listopad 2018.: 6281kn i 5172kn.

Odabir mjere centra podataka

Odabir mjere centralne tendencije podataka:

- mod treba koristiti kada se želi naglasiti najčešći podatak, a to je obično slučaj kod kvalitativnih obilježja
- ako se radi o ordinalnim varijablama obično je medijan najpogodniji izbor
- ako uzorak sadrži stršeće vrijednosti, medijan će u pravilu biti bolji pokazatelj tipične vrijednosti od aritmetičke sredine
- kada je obilježje simetrično raspodijeljeno, tada je opravdano koristiti i aritmetičku sredinu i medijan kao mjere centra i vrijednosti će biti vrlo bliske
- kada je raspodjela asimetrična, medijan je bolji reprezentant centra

Odabir mjere centra podataka

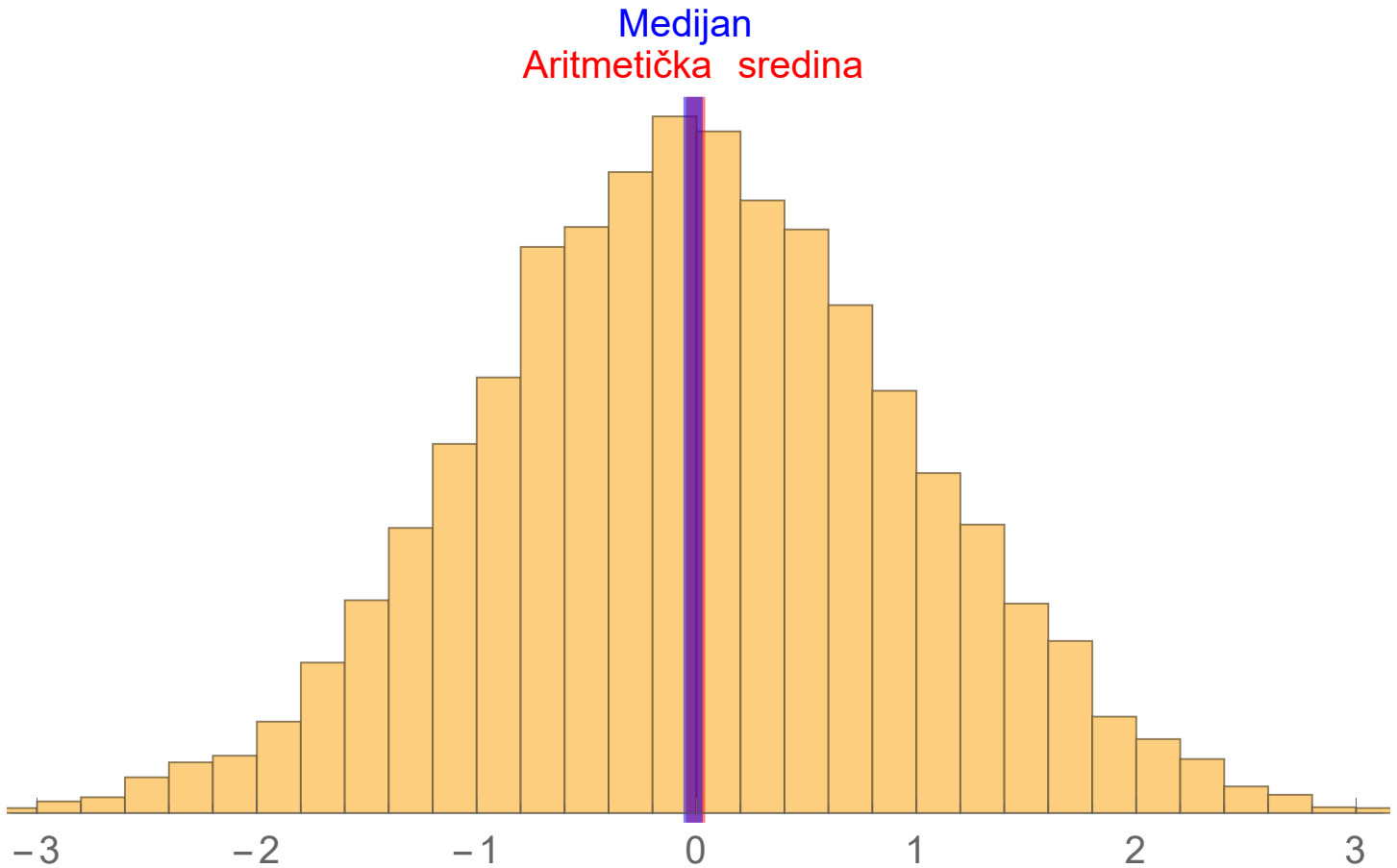

Figure 1: Primjer uzorka iz simetrične distribucije

Odabir mjere centra podataka

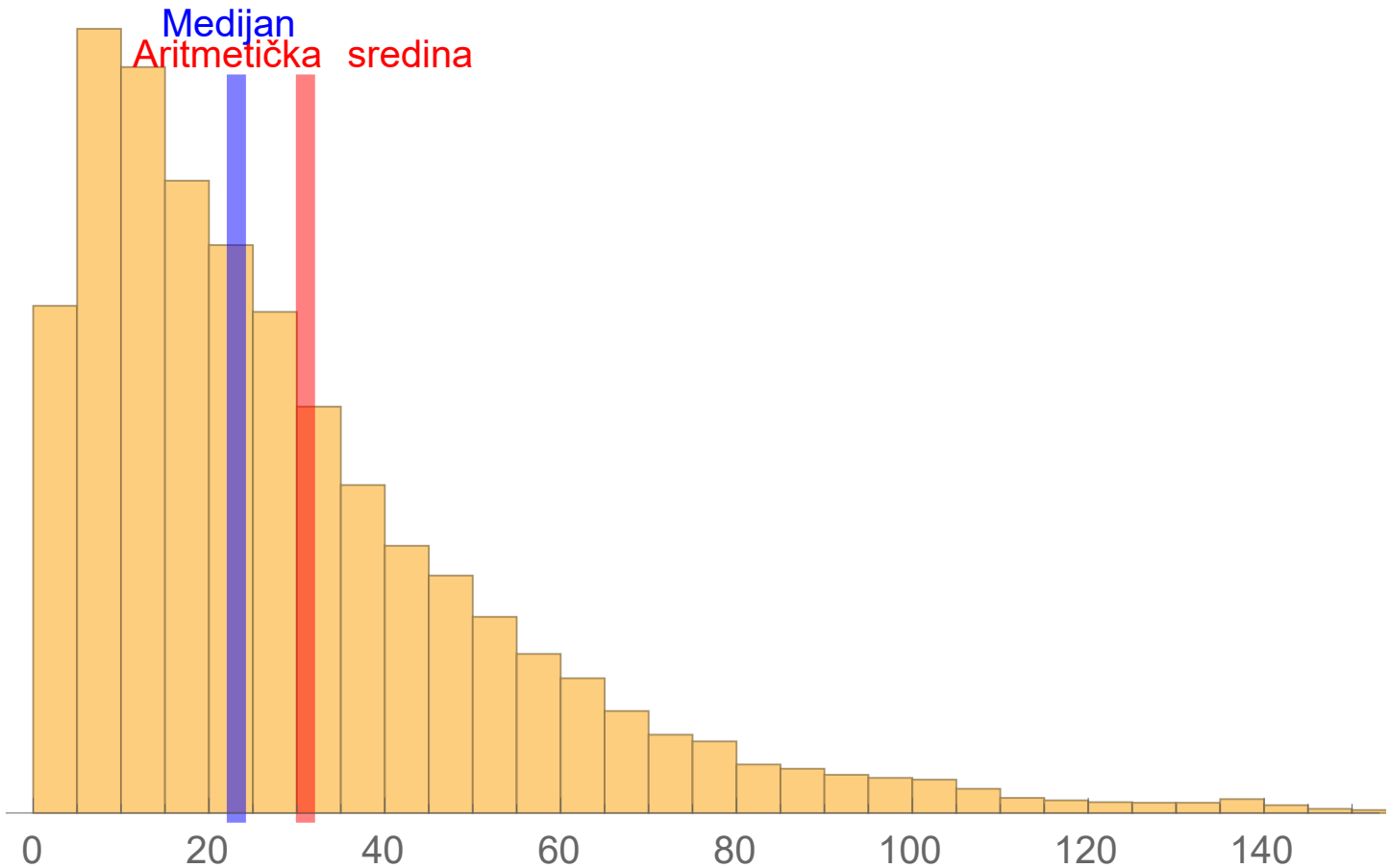

Figure 2: Primjer uzorka iz asimetrične distribucije

Mjere raspršenosti podataka

- **raspršenost** (**varijabilnost**) podataka
- Primjer: neka su zadana tri uzroka
  - $0, 0, -1, 1$
  - $-200, 100, 200, -100$
  - $0, 0, -1, 1, 0, 0$

  Izračunajte aritmetičku sredinu i medijan za svaki uzorak. Je li mjera centra jednako kvalitetna u sva tri uzorka?

Mjere raspršenosti podataka
varijanca i standardna devijacija podataka

- **varijanca** i **standardna devijacija** karakteriziraju raspršenost podataka oko aritmetičke sredine
- varijanca niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2,$$

a standardna devijacija je kvadratni korijen varijance, tj.

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}.$$

Mjere raspršenosti podataka
varijanca i standardna devijacija podataka

- npr. neka su izmjerene vrijednosti jedne varijable

$$1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8$$

- iz primjera znamo da je aritmetička sredina ovog skupa podataka približno jednaka 5.42, pa su varijanca i standardna devijacija ovog skupa podataka

$$s_n^2 \approx \frac{1}{8}\sum_{i=1}^{9}(x_i - 5.42)^2 \approx 8.86, \quad s_n \approx \sqrt{\frac{1}{8}\sum_{i=1}^{9}(x_i - 5.42)^2} \approx 2.98$$

Mjere raspršenosti podataka
varijanca i standardna devijacija podataka

- Primjer: neka su zadana tri skupa podataka
  - $0, 0, -1, 1$
  - $-200, 100, 200, -100$
  - $0, 0, -1, 1, 0, 0$

  Izračunajte varijancu i standardnu devijaciju za sva tri uzorka.

Mjere raspršenosti podataka
kvantili

- p-**kvantil** ili **kvantil reda** p (postotna vrijednost) za neki izabrani broj $p \in (0, 100)$, označimo je x'_p, definira se poštujući zahtjev da je barem $p\%$ izmjerenih vrijednosti varijable manje ili jednako x'_p, dok je barem $(100 - p)\%$ vrijednosti veće ili jednako x'_p
- medijan je kvantil reda 0.5 (50%-tna vrijednost)
- kvantil reda 0.25 (25%-tna vrijednost) zove se **donji kvartil**
- kvantil reda 0.75 (75%-tna vrijednost) zove se **gornji kvartil**
- kao i kod računanja medijana, ako se na traženoj poziciji za računanje postotne vrijednosti nalaze dva podatka u uređenom skupu izmjerenih vrijednosti, postotnu vrijednost određujemo kao njihovu aritmetičku sredinu

Mjere raspršenosti podataka
kvantili

- na osnovu n podataka želimo odrediti p-kvantil, $p \in (0, 100)$
- izračunamo vrijednost $j = np/100$
- ako j nije prirodan broj onda je p-kvantil podatak na poziciji $\lceil j \rceil$ - najmanji prirodni broj veći od j (prvi sljedeći)
- ako je j prirodan broj onda je p-kvantil aritmetička sredina podataka na pozicijima j i $j + 1$

Mjere raspršenosti podataka
postotna vrijednost, donji i gornji kvartil

- npr. neka su $1, 2, 5, 6, 6, 1, 3, 7, 3, 3, 3, 3$ izmjerene vrijednosti jedne varijable
- prvo ove vrijednosti poredamo po veličini:
  $1, 1, 2, 3, 3, 3, 3, 3, 5, 6, 6, 7$
- želimo li odrediti donji kvartil, potrebno je prvo odrediti četvrtinu podataka (25%) - obzirom da imamo 12 podataka, četvrtinu (25%) čine tri podatka
- treći podatak u gornjem skupu je broj 2, a četvrti 3 - donji kvartil je 2.5
- deveti broj u gornjem skupu podataka je broj 5, a deseti 6 - gornji kvartil je 5.5

Mjere raspršenosti podataka
najmanja i najveća vrijednost, raspon podataka

- ako su $x_1, x_2, \ldots, x_n$ izmjerene vrijednosti varijable X, označimo najmanju od njih **(minimum)** $x_{\min}$, a najveću od njih **(maksimum)** $x_{\max}$
- **raspon** (eng. range) podataka - razlika najveće i najmanje vrijednosti u skupu izmjerenih vrijednosti varijable $(x_{\max} - x_{\min})$
- npr. neka su izmjerene vrijednosti jedne varijable $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ - 1 je najmanja izmjerena vrijednost, a 7 najveća, pa je raspon ovog skupa izmjerenih vrijednosti $7 - 1 = 6$

Mjere raspršenosti podataka
maksimalno odstupanje od "prosjeka"

- **maksimalno odstupanje izmjerenih vrijednosti varijable od "prosjeka"** - veći od brojeva $(\bar{x}_n - x_{\min})$ i $(x_{\max} - \bar{x}_n)$, tj. broj

$$\max\{(\bar{x}_n - x_{\min}), (x_{\max} - \bar{x}_n)\}.$$

- npr. neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti neke varijable X:

$$x_{\min} = 1, \quad x_{\max} = 7,$$

$$\bar{x}_n = \frac{1+2+5+6+5+1+2+7+2+2+3+3}{12} = 3.25$$

- maksimalno odstupanje izmjerenih vrijednosti ove varijable od prosjeka:

$$\max\{3.25 - 1, 7 - 3.25\} = \max\{2.25, 3.75\} = 3.75$$

Grafička metoda opisivanja numeričkih podataka - kutijasti dijagram

- korištenjem numeričkih karakteristika numeričkih varijabli skup mjerenih vrijednosti može se prikazati grafički pomoću **kutijastog dijagrama** (eng. box plot, boxplot ili box-and-whisker plot)
- kutijastim dijagramom prikazujemo odnos pet numeričkih karakteristika skupa izmjerenih vrijednosti: minimalnu vrijednost, donji kvartil, medijan, gornji kvartil i maksimalnu vrijednost
- na kutijastom dijagramu se također označavaju takozvane **stršeće vrijednosti** skupa podataka, ako postoje
- primjer 3.20

Detekcija stršećih vrijednosti

Stršeća vrijednost - podatak koji je značajno veći ili manji u odnosu na druge izmjerene vrijednosti jedne varijable i čije je pojavljivanje najčešće vezano uz jedan od sljedećih razloga:

- podatak je ili netočno izmjeren ili krivo unesen u bazu podataka
- podatak dolazi iz druge populacije (ne iz populacije koju promatramo u kontekstu problema kojeg proučavamo)
- podatak je točno izmjeren i unesen u bazu, ali predstavlja rijetku pojavu u populaciji
- primjer 3.21

Oblik distribucije uzorka

- prema obliku histograma možemo nešto reći o razdiobi (raspodjeli, distribuciji) uzorka
- lijevi rep distribucije je teži ako su podaci manji od prosjeka udaljeniji od njega nego podaci s desne strane prosjeka (oni koji su veći od njega) - i obratno
- simetrične distribucije - repovi su podjednaki
- asimetrične distribucije - jedan rep je teži od drugoga (lijevo nagnute ili desno nagnute)

Oblik raspodjele uzorka

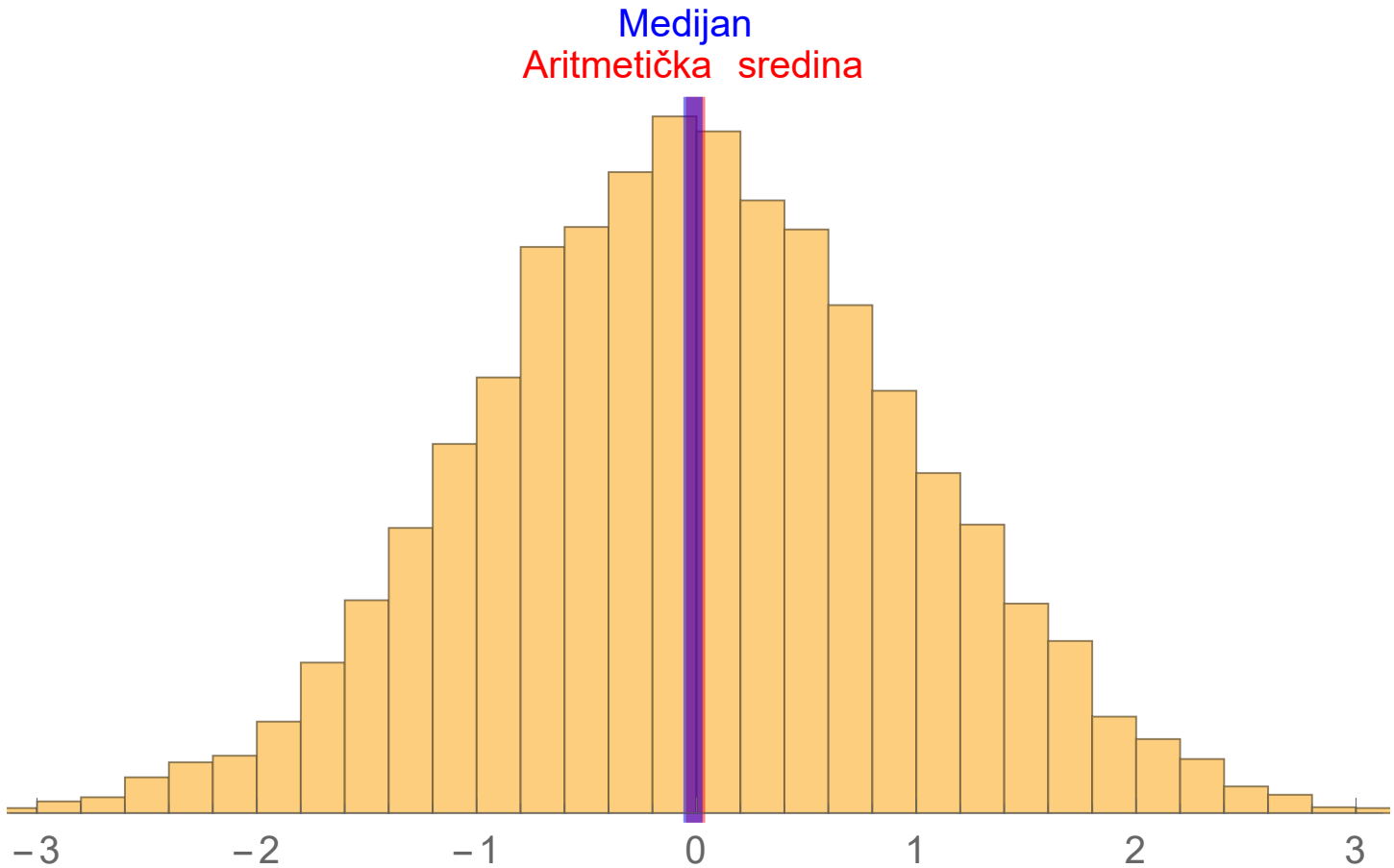

Figure 3: Primjer uzorka iz simetrične distribucije

Oblik raspodjele uzorka

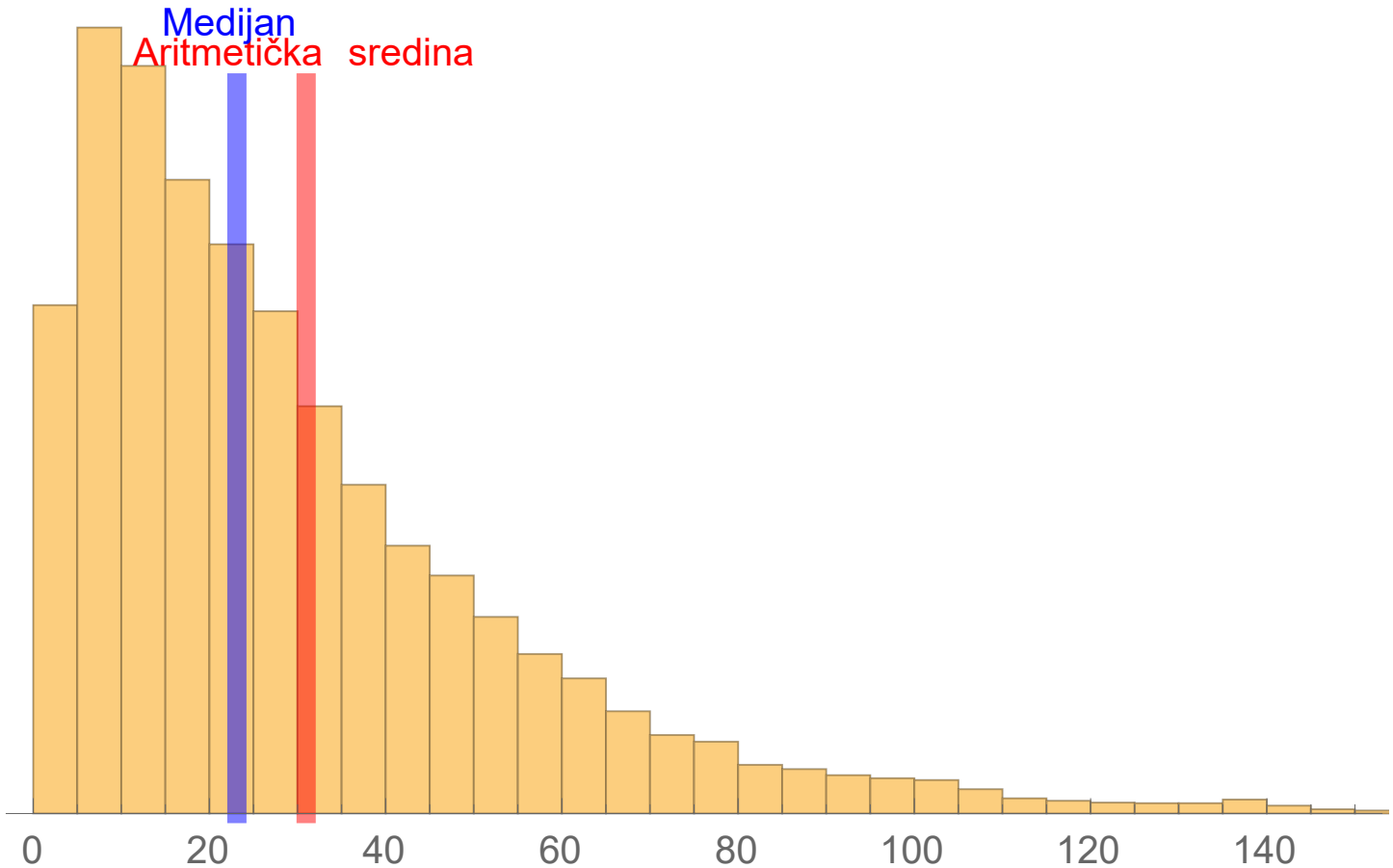

Figure 4: Primjer uzorka iz asimetrične distribucije - desno nagnute (desni rep je duži - teži)