Deskriptivna statistika

- Izvodimo eksperiment i bilježimo realizacije varijable X kojom modeliramo neko promatrano obilježje na uzorku (iz neke populacije od interesa).
- Rezultat opažanja varijable X na elementu populacije označavamo s x.
- Opažene vrijednosti od X na uzorku veličine n označavamo s $x_1, \ldots, x_n$.

Primjer (7.1.)

Ocjene iz matematike jednog razreda od 30 učenika na kraju školske godine su:

$$1, 4, 2, 3, 1, 1, 2, 4, 3, 4, 5, 3, 2, 2, 3, 2, 5, 3, 2, 3, 3, 4, 2, 3, 2, 3, 3, 2, 2, 2.$$

Zanima nas ocjena iz matematike kao obilježje točno tog razreda.

- populacija i uzorak su učenici danog razreda, element je pojedini učenik
- $X =$ ocjena iz matematike pojedinog učenika na kraju školske godine
- X je diskretna kvantitativna (ordinalna) varijabla i slika od X je

$$R(X) = \{1, 2, 3, 4, 5\}$$

Osnovna zadaća deskriptivne statistike jest opisivati opažene vrijednosti (na uzorku ili populaciji) od X.

Raspodjela opaženih vrijednosti $x_1, \ldots, x_n$ od X (na uzorku ili populaciji veličine n) opisuje se

(1) frekvencijama

(2) relativnim frekvencijama

(3) kumulativnim frekvencijama

(4) kumulativnim relativnim frekvencijama (u slučaju kvantitativnih varijabli).

Frekvencije i relativne frekvencije

(1) **Frekvencija** je broj pojavljivanja pojedinog x_i u nizu $x_1, \ldots, x_n$.

(2) **Relativna frekvencija** od x_i je omjer frekvencije te vrijednosti i veličine niza $x_1, \ldots, x_n$, tj. n.

(3) **Kumulativna frekvencija** u vrijednosti x_i (pri čemu je niz $x_1, \ldots, x_n$ poredan po veličini od najmanje do najveće vrijednosti) je suma frekvencija vrijednosti manjih ili jednakih od x_i.

(4) **Kumulativna relativna frekvencija** u vrijednosti x_i je omjer kumulativne frekvencije u x_i i veličine niza $x_1, \ldots, x_n$.

a_i	f_i	$\frac{f_i}{30}$	K_{f_i}	$\frac{K_{f_i}}{30}$
1	3	0.1	3	0.1
2	11	0.37	14	0.47
3	10	0.33	24	0.8
4	4	0.13	28	0.93
5	2	0.07	30	1
$\sum$	30	1		

Tablica: Frekvencije, relativne frekvencije, kumulativne frekvencije i kumulativne relativne frekvencije ocjena u razredu iz Primjera 7.1.

- Neprekidne varijable, odnosno raspodjelu njihovih opaženih vrijednosti, diskutiramo na analogan način.
- Vrijednosti neprekidnih varijabli, za razliku od diskretnih, dane su terminima intervala.

Primjer (7.3.)

Mjesečna zarada (u kunama) zaposlenih u nekom poduzeću je dana u tablici.

Mjesečna zarada	Broj radnika - f_i	K_{f_i}	$\frac{f_i}{250}$	$\frac{K_{f_i}}{250}$
$[1000, 2000)$	16	16	0.064	0.064
$[2000, 3000)$	38	54	0.152	0.216
$[3000, 4000)$	66	120	0.264	0.48
$[4000, 5000)$	70	190	0.28	0.76
$[5000, 6000)$	41	231	0.164	0.924
$[6000, 7000)$	18	249	0.072	0.996
$[7000, 8000)$	1	250	0.04	1
$\sum$	250		1	

Grupiranje po razredima

- Opažene vrijednosti diskretnih (kvantitativnih) varijabli mogu biti i grupirane (ako to ima smisla).
- Na primjer, broj automobila po kućanstvu u pravilu ne grupiramo (jer ne baratamo velikim brojevima) dok starost stanovništva ponekad ima smisla grupirati jer, primjerice, nema velike razlike između 44 i 45 godina starosti.
- U slučaju da želimo razrede jednake duljine, veličinu razreda ili broj razreda biramo sami i određujemo iz formule

$$\text{širina razreda} = \frac{\max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\}}{\text{broj razreda}}.$$

- Pojmovi frekvencije, relativne frekvencije, kumulativne frekvencije i kumulativne relativne frekvencije definiraju se analogno kao i kod neprekidnih varijabli.

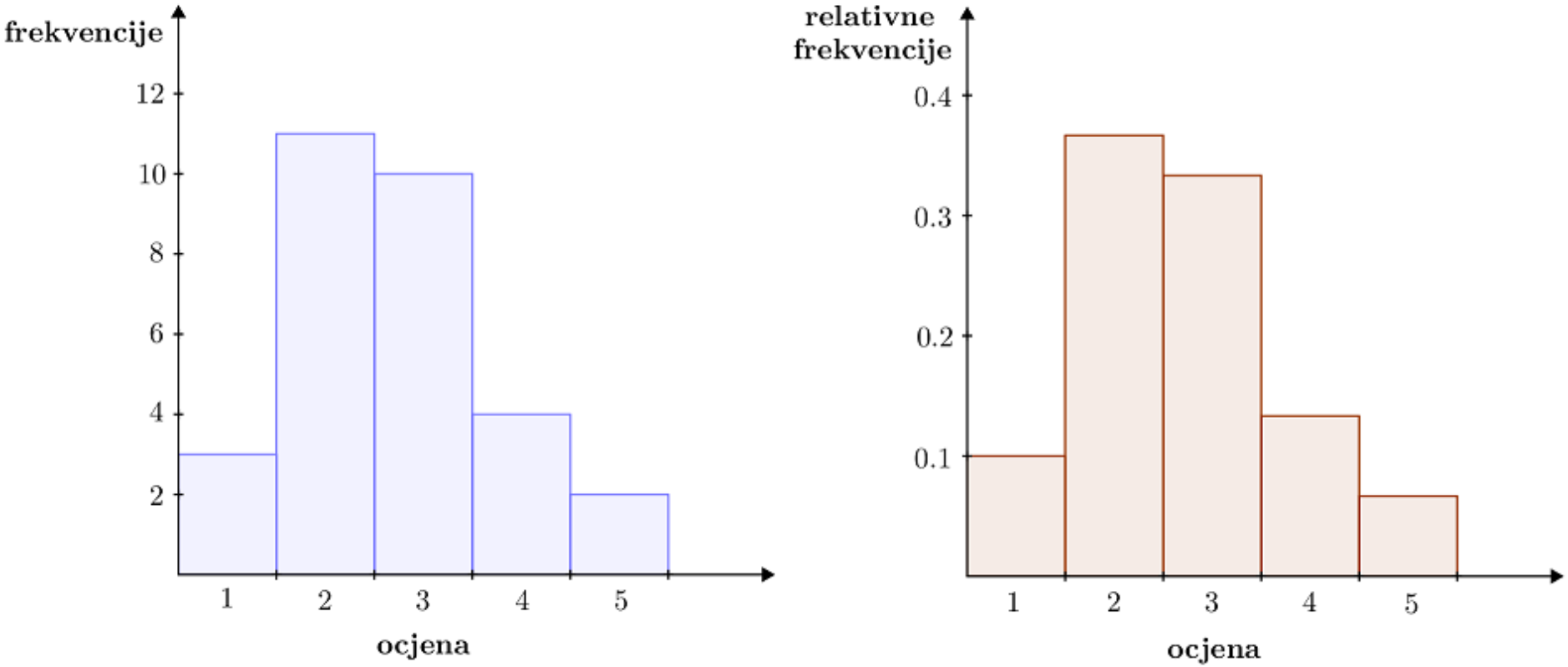
Grafički prikaz

Raspodjelu opaženih vrijednosti od X prikazujemo grafički preko

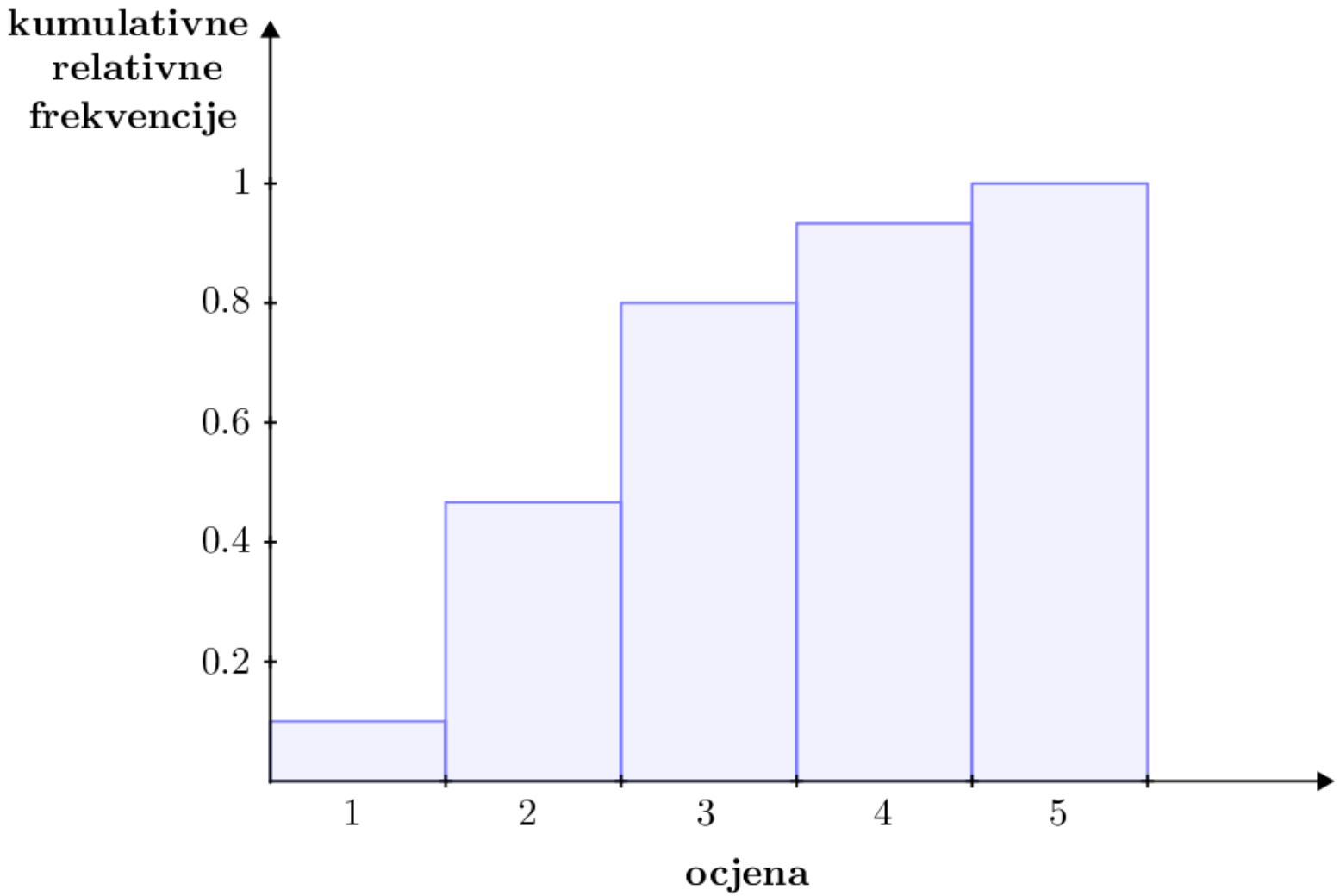
(1) stupčastih dijagrama

(2) histograma.

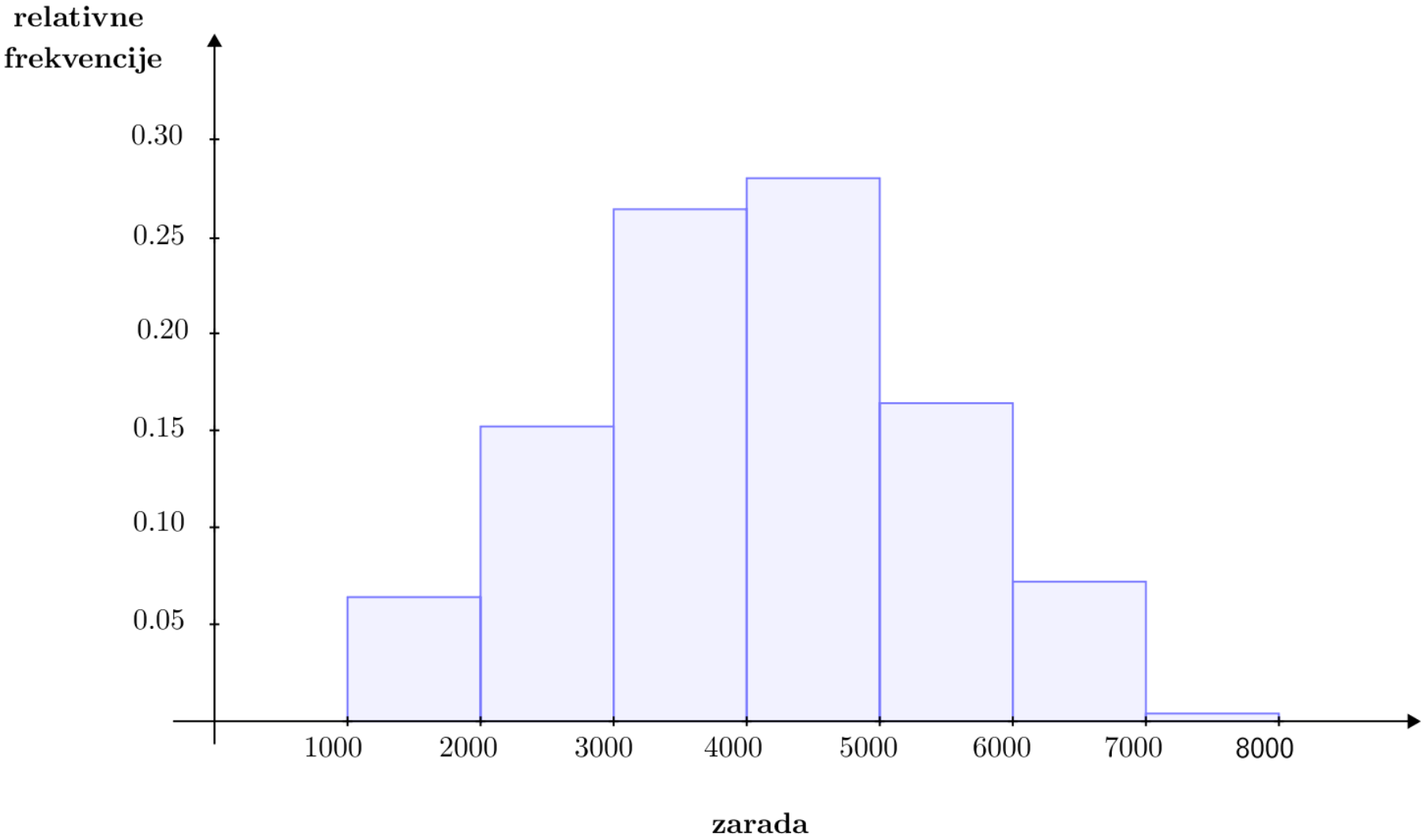
- **Stupčasti dijagram** je grafički prikaz frekvencija (relativnih frekvencija, kumulativnih frekvencija, kumulativnih relativnim frekvencija) opaženih vrijednosti kvalitativnih varijabli i negrupiranih opaženih vrijednosti diskretnih kvantitativnih varijabli.
- **Histogram** je prikaz frekvencija (relativnih frekvencija, kumulativnih frekvencija, kumulativnih relativnih frekvencija) opaženih vrijednosti neprekidnih kvantitativnih varijabli i grupiranih opaženih vrijednosti diskretnih kvantitativnih varijabli.



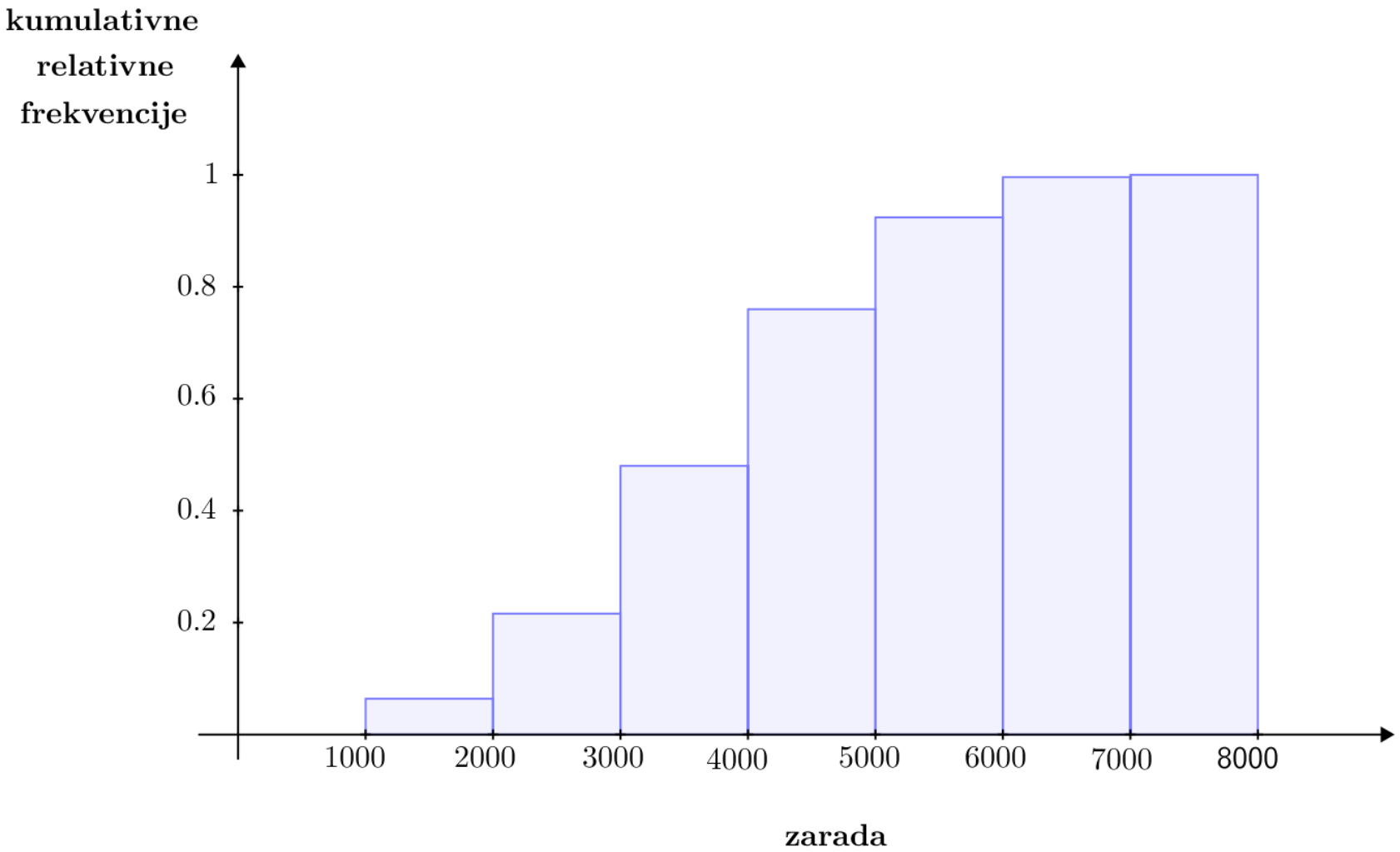
Slika: Stupčasti dijagrami frekvencija i relativnih frekvencija ocjena iz matematike iz Primjera 7.1.



Slika: Stupčasti dijagram kumulativnih relativnih frekvencija ocjena iz matematike iz Primjera 7.1.



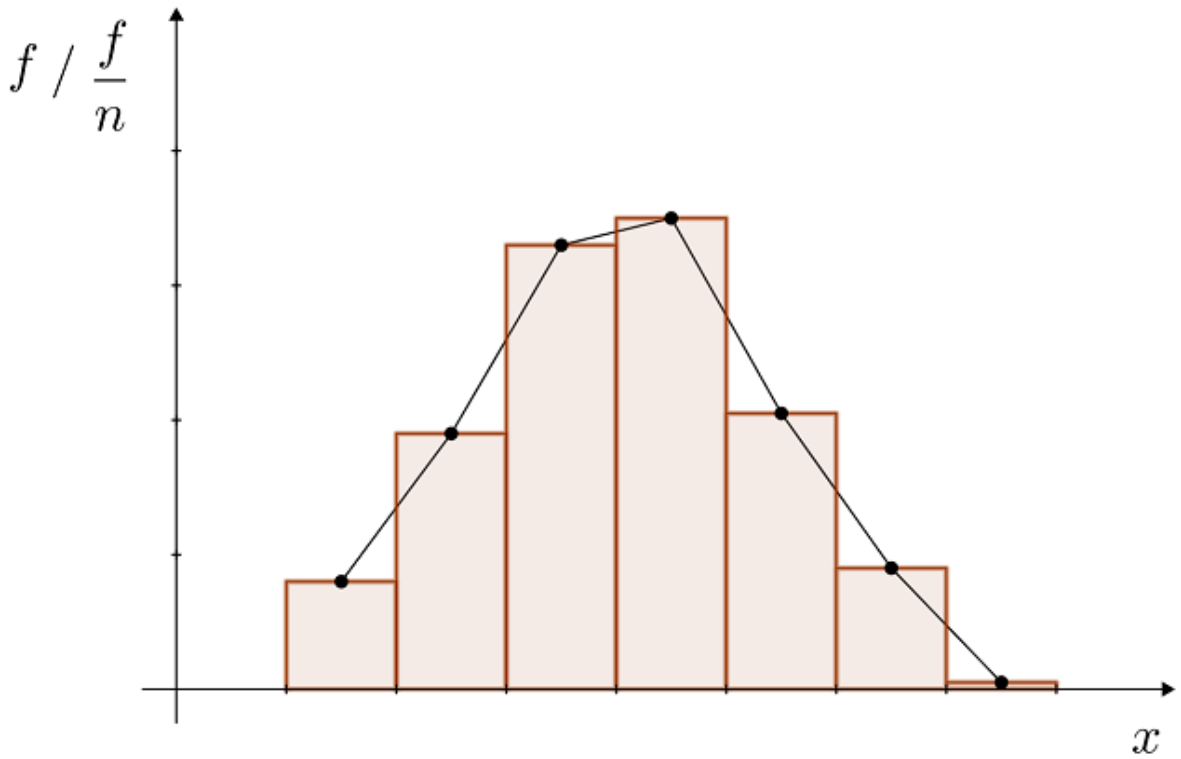
Slika: Histogram relativnih frekvencija plaća po razredima iz Primjera 7.3.



Slika: Histogram kumulativnih relativnih frekvencija plaća po razredima iz Primjera 7.3.

Poligon/krivulja frekvencija

- Ako spojimo sredine gornjih stranica uzastopnih pravokutnika u histogramima ravnim crtama, dobivamo **poligon**.
- Dobivena krivulja naziva se još i **frekvencijskom krivuljom**, **krivuljom relativnih frekvencija**, **krivuljom kumulativnih frekvencija** i **krivuljom kumulativnih relativnih frekvencija**.



U nastavku usredotočit ćemo se na deskriptivnu statistiku kvantitativnih varijabli. Analiziramo numeričke deskriptivne mjere:

(i) **mjere centralne tendencije**

(ii) **mjere raspršenja**

(iii) **mjere oblika**.

- X je kvantitativna varijabla s vrijednostima opažanja $x_1, \ldots, x_n$ na nekom uzorku veličine n.
- Pretpostavljamo da su vrijednosti $x_1, \ldots, x_n$ poredane po veličini od najmanje do najveće.
- U slučaju da je X diskretna kvantitativna varijabla čije su vrijednosti grupirane u n razreda ili neprekidna varijabla dana s n razreda (intervala), onda za vrijednosti $x_1, \ldots, x_n$ uzimamo sredine razreda.

Mjere centralne tendencije

Aritmetička sredina

- **Aritmetičku sredinu** vrijednosti $x_1, \ldots, x_n$ računamo kao

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

- Ako se među brojevima $x_1, \ldots, x_n$ pojavljuju brojevi $a_1, \ldots, a_k$, $k < n$, s frekvencijama $f_1, \ldots, f_k$, onda je

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{k} a_i f_i.$$

- Aritmetička sredina je osjetljiva na stršeće vrijednosti, tzv. **izdvojenice** (engl. outliere).
  Izdvojenice se najčešće javljaju kao posljedica krivog mjerenja ili je podatak točno izmjeren ali predstavlja rijetku pojavu ili dolazi iz neke druge populacije.

Medijan

- **Medijan** vrijednosti $x_1, \ldots, x_n$ je vrijednost za koju vrijedi da je 50% podataka manje ili jednako toj vrijednosti, a 50% podataka je veće ili jednako navedenoj vrijednosti.
- Određujemo ga kao

$$\mathrm{M_e} = \begin{cases} x_{\lfloor \frac{n}{2}+1 \rfloor}, & \frac{n}{2} \text{ nije prirodan broj} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \frac{n}{2} \text{ je prirodan broj.} \end{cases}$$

Primjer (7.4.)

(1) U nizu $1, 7, 9, 14, 18$ medijan je $\mathrm{M_e} = 9$.
(2) U nizu $1, 9, 12, 15, 22, 23$ medijan je

$$\mathrm{M_e} = (12 + 15)/2 = 13.5.$$

Mod

- **Mod** vrijednosti $x_1, \ldots, x_n$ je vrijednost s najvećom frekvencijom.
- Varijabla može imati i više modova.

Primjer (7.5.)

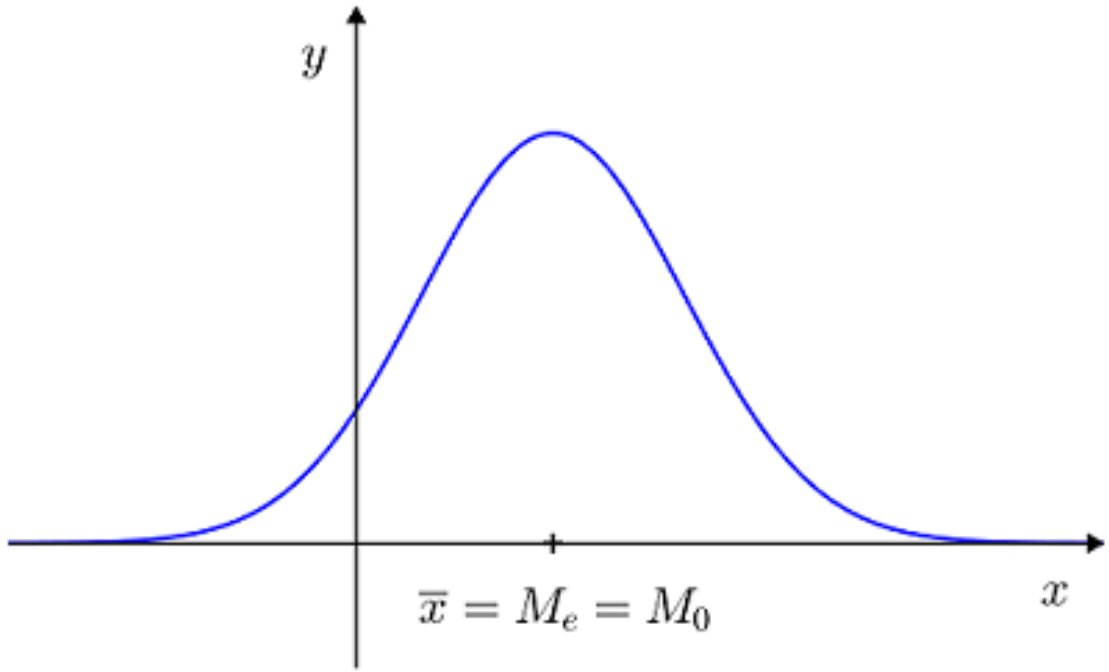
(1) U nizu $1, 1, 2, 3, 4$ mod je

$$\mathrm{M}_0 = 1.$$

(2) U nizu $1, 1, 2, 3, 3, 4$ modovi su

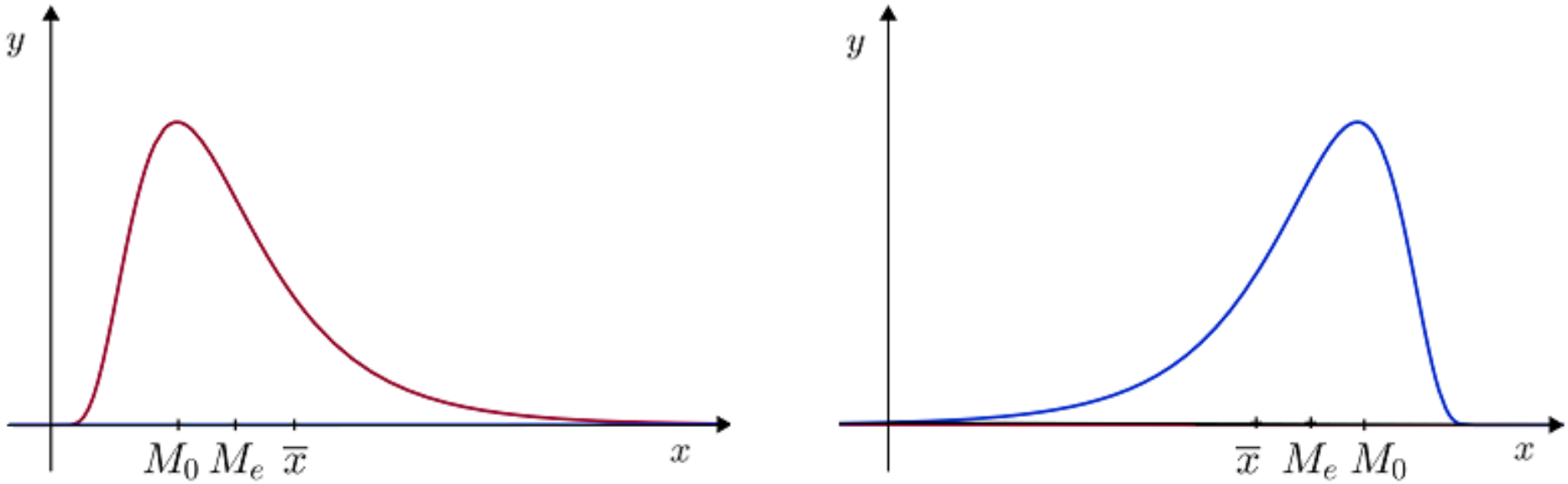
$$\mathrm{M}_0 = 1, 3.$$

- Uočimo da mod i medijan nisu osjetljivi na izdvojenice, za razliku od aritmetičke sredine.
- U slučaju simetričnih ($x_i = -x_{n-i+1}$ za $i = 1, \ldots, \lfloor n/2 \rfloor$) i unimodalnih ($x_1, \ldots, x_n$ imaju samo jedan mod) vrijednosti, aritmetička sredina, mod i medijan se podudaraju.



Slika: Odnos aritmetičke sredine, medijana i moda u slučaju simetričnih unimodalnih vrijednosti

- U slučaju asimetričnih unimodalnih vrijednosti, medijan je uvijek između aritmetičke sredine i moda. Mod se nalazi na mjestu gdje je frekvencijska krivulja najviša, a aritmetička sredina je uvijek na strani na kojoj se nalazi dulji rep krivulje.

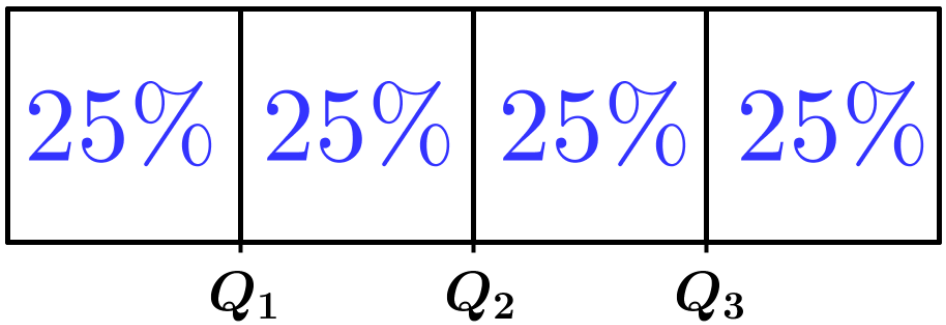


Slika: Odnos aritmetičke sredine, medijana i moda u slučaju asimetričnih unimodalnih vrijednosti

Mjere raspršenja

Kvartili

- **Kvartili** vrijednosti $x_1, \ldots, x_n$ su vrijednosti koje dijele podatke u četiri jednakobrojna dijela.



- Računamo ih po principu sličnom računanju medijana. Primijetimo odmah da je drugi kvartil vrijednost za koju vrijedi da je 50% podataka manje ili jednako toj vrijednosti, a 50% podataka veće ili jednako navedenoj vrijednosti, odnosno $Q_2 = \mathrm{M_e}$.

Kvartili

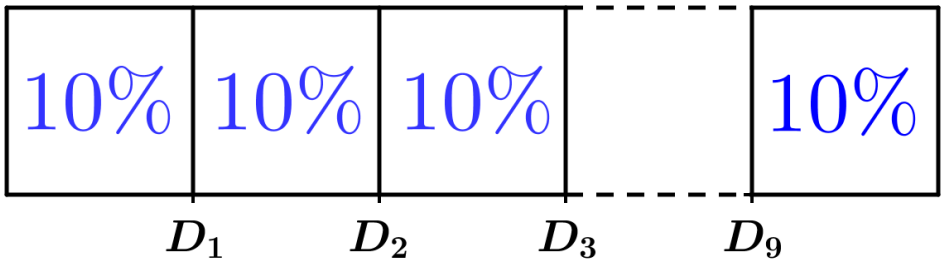
Prvi i treći kvartil dobivamo na sljedeći način:

$$Q_1 = \begin{cases} x_{\lfloor \frac{n}{4}+1 \rfloor}, & \frac{n}{4} \text{ nije prirodan broj} \\ \dfrac{x_{\frac{n}{4}} + x_{\frac{n}{4}+1}}{2}, & \frac{n}{4} \text{ je prirodan broj.} \end{cases}$$

$$Q_3 = \begin{cases} x_{\lfloor \frac{3n}{4}+1 \rfloor}, & \frac{3n}{4} \text{ nije prirodan broj} \\ \dfrac{x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1}}{2}, & \frac{3n}{4} \text{ je prirodan broj.} \end{cases}$$

Decili

- Analogno se definiraju **decili**.

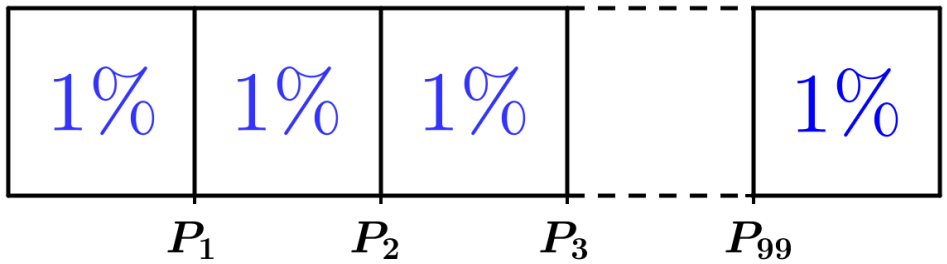


- Za $i = 1, 2, \ldots, 9$ vrijedi:

$$D_i = \begin{cases} x_{\lfloor \frac{in}{10}+1 \rfloor}, & \frac{in}{10} \text{ nije prirodan broj} \\ \dfrac{x_{\frac{in}{10}} + x_{\frac{in}{10}+1}}{2}, & \frac{in}{10} \text{ je prirodan broj.} \end{cases}$$

Percentili

- Na jednak način definiramo i **percentile**, vrijednosti koje dijele podatke na sto jednakobrojnih dijelova.



- Za $i = 1, 2, \ldots, 99$ vrijedi:

$$P_i = \begin{cases} x_{\lfloor \frac{in}{100}+1 \rfloor}, & \frac{in}{100} \text{ nije prirodan broj} \\ \frac{x_{\frac{in}{100}} + x_{\frac{in}{100}+1}}{2}, & \frac{in}{100} \text{ je prirodan broj.} \end{cases}$$

Primjer (7.6.)

Na prvom zimskom ispitnom roku, pismenom ispitu iz kolegija Vjerojatnost i statistika pristupilo je 24 studenata. Broj bodova (od maksimalno mogućih 100) koje su studenti ostvarili na ispitu je sljedeći:

$$53, 72, 82, 45, 36, 88, 20, 31, 81, 68, 75, 58,$$
$$24, 67, 54, 93, 98, 70, 30, 5, 34, 7, 2, 61.$$

Prvo poredamo podatke po veličini (od najmanjeg do najvećeg):

$$2, 5, 7, 20, 24, 30, 31, 34, 36, 45, 53, 54, 58,$$
$$61, 67, 68, 70, 72, 75, 81, 82, 88, 93, 98.$$

(1) **Prvi i treći kvartil:**

$n/4 = 6$ i $3n/4 = 18$ su prirodni brojevi, pa prvi i treći kvartil računamo na sljedeći način:

$$Q_1 = \frac{x_{\frac{24}{4}} + x_{\frac{24}{4}+1}}{2} = \frac{x_6 + x_7}{2} = \frac{30 + 31}{2} = 30.5,$$

$$Q_3 = \frac{x_{\frac{3\cdot 24}{4}} + x_{\frac{3\cdot 24}{4}+1}}{2} = \frac{x_{18} + x_{19}}{2} = \frac{72 + 75}{2} = 73.5.$$

Vrijednost prvog kvartila nam govori da je 25% studenata na završnom ispitu iz Statistike ostvarilo 30.5 bodova ili manje, dok je 75% studenata ostvarilo 30.5 bodova ili više.

(2) **Prvi i deveti decil:**

Da bismo izračunali prvi i deveti decil podataka potrebno je provjeriti jesu li $n/10$ i $9n/10$ prirodni brojevi. Budući da je $n/10 = 2.4$ i $9n/10 = 21.6$, decile računamo na sljedeći način:

$$D_1 = x_{\lfloor \frac{24}{10}+1 \rfloor} = x_{\lfloor 3.4 \rfloor} = x_3 = 7,$$

$$D_9 = x_{\lfloor \frac{9\cdot 24}{10}+1 \rfloor} = x_{\lfloor 22.6 \rfloor} = x_{22} = 88.$$

Vrijednost devetog decila nam govori da je 90% studenata na završnom ispitu dobilo 88 bodova ili manje, dok je najboljih 10% studenata ostvarilo 88 bodova ili više.

Raspon i interkvartilni raspon

- **Raspon** vrijednosti $x_1, \ldots, x_n$ je razlika najveće i najmanje vrijednosti:

$$R = x_n - x_1.$$

- **Interkvartilni raspon** vrijednosti $x_1, \ldots, x_n$ je razlika između trećeg i prvog kvartila:

$$I_Q = Q_3 - Q_1.$$

Interkvartilni raspon nam daje raspon u kojem se nalazi srednjih 50% podataka.

Uzoračka varijanca i standardna devijacija

- **Uzoračka varijanca** vrijednosti $x_1, \ldots, x_n$ daje nam približno srednje kvadratno odstupanje podataka od aritmetičke sredine:

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

  ili, kada su podaci dani svojim frekvencijama,

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(a_i - \overline{x}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i a_i^2 - \frac{n}{n-1}\overline{x}_n^2.$$

- **Uzoračka standardna devijacija** vrijednosti $x_1, \ldots, x_n$, koja daje približno srednje odstupanje podataka od $\overline{x}_n$, dana je sa

$$\boxed{s_n = \sqrt{s_n^2}}.$$

Koeficijent varijacije

- **Koeficijent varijacije** računamo kao

$$\boxed{V = \frac{s_n}{\overline{x}_n}}.$$

- Koristimo ga ako imamo dva uzorka ili čak dva eksperimenta i ako želimo usporediti njihove aritmetičke sredine i varijance (standardne devijacije).

Mjere oblika

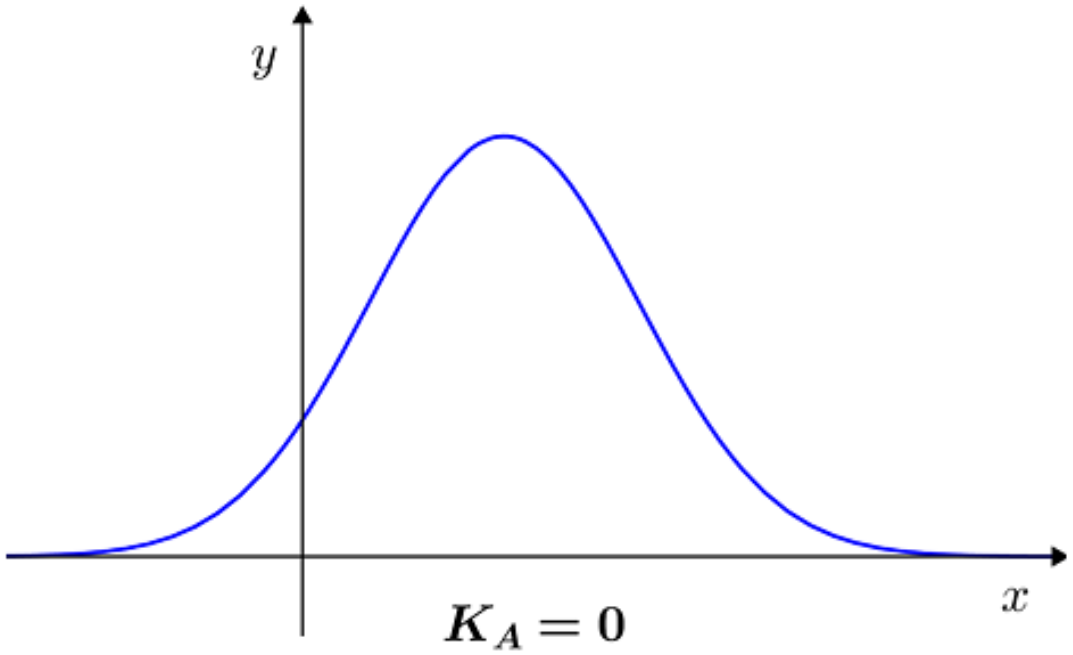
Koeficijent asimetrije

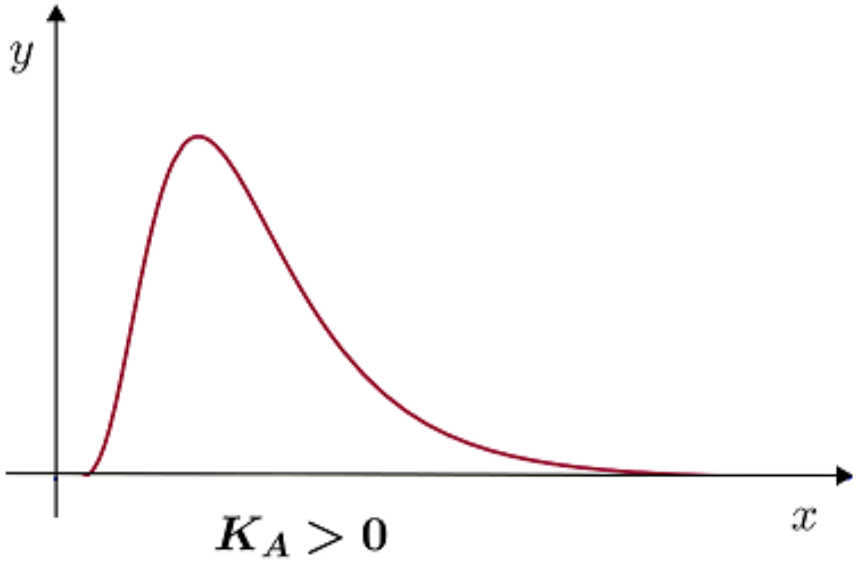
Koeficijent asimetrije vrijednosti $x_1, \ldots, x_n$ daje nam podatak o simetričnosti pripadne frekvencijske krivulje.

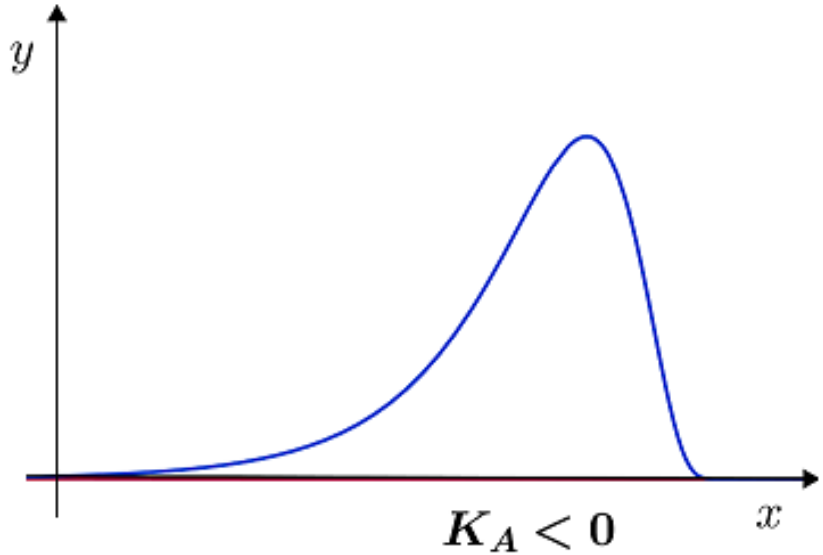
Računa se po sljedećoj formuli:

$$\mathrm{K_A} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^3}{s_n^3} = \frac{\frac{1}{n}\sum_{i=1}^{k} f_i(a_i - \overline{x}_n)^3}{s_n^3}.$$

- $\mathrm{K_A} = 0$ - podaci su simetrični
- $\mathrm{K_A} < 0$ - podaci su negativno asimetrični
- $\mathrm{K_A} > 0$ - podaci su pozitivno asimetrični


y
x
$K_A = 0$


y
x
$K_A > 0$


y
x
$K_A < 0$

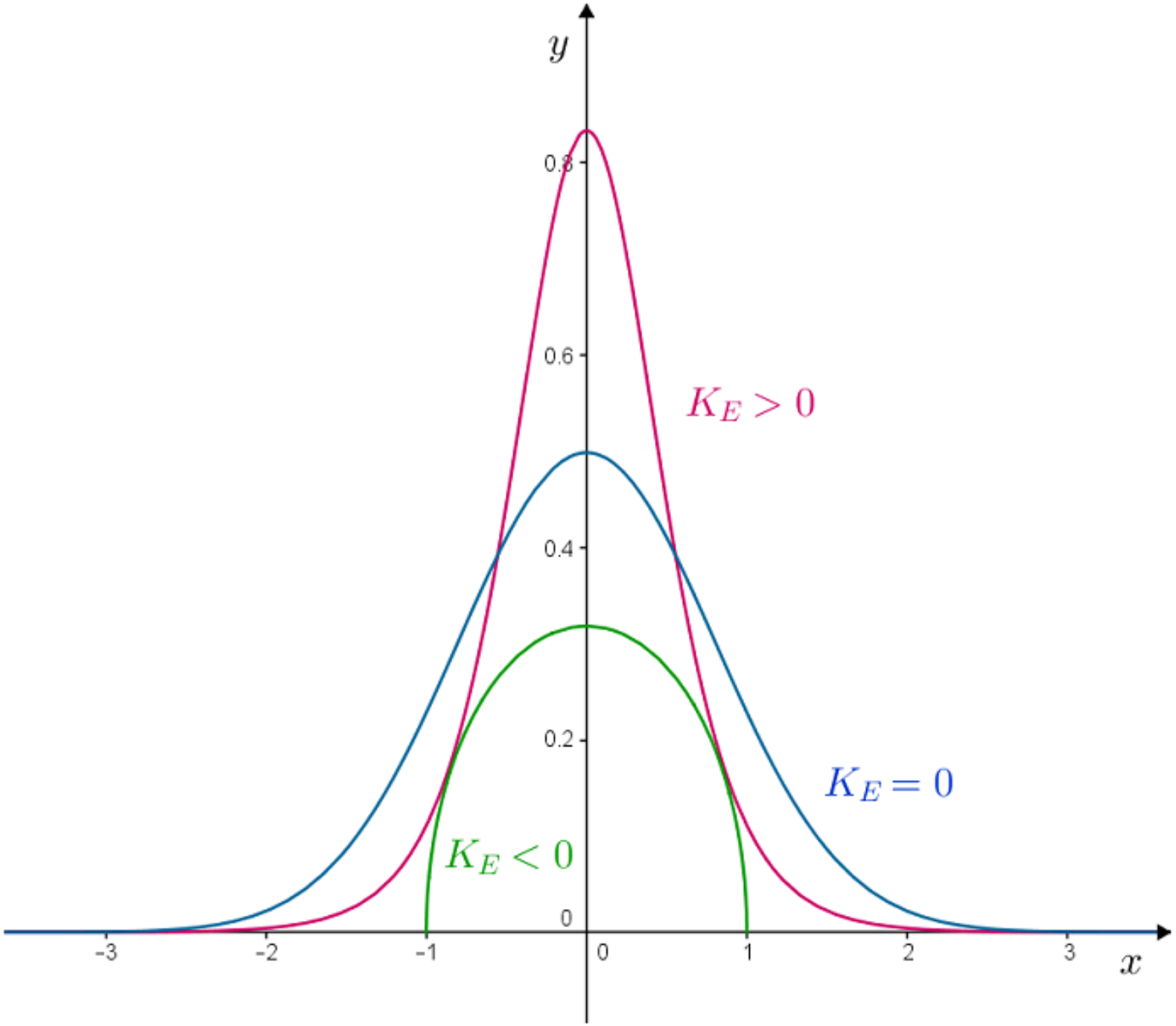
Koeficijent zaobljenosti

Koeficijent zaobljenosti vrijednosti $x_1, \ldots, x_n$ daje nam podatak o "zaobljenosti" (spljoštenosti) pripadne frekvencijske krivulje oko aritmetičke sredine.

Računa se po formuli:

$$\mathrm{K_E} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^4}{s_n^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{k} f_i(a_i - \overline{x}_n)^4}{s_n^4} - 3.$$

- vrijednosti $x_1, \ldots, x_n$ su dobivene opažanjem normalne slučajne varijable $\implies \mathrm{K_E} \approx 0$
- uspoređuje razdiobu (frekvencijsku krivulju) podataka $x_1, \ldots, x_n$ s normalnom razdiobom $N(\overline{x}_n, s_n^2)$


y
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
x
$K_E > 0$
$K_E = 0$
$K_E < 0$

Primjer (7.7.)

U tablici je dana raspodjela broja kvarova nekog uređaja dobivena na uzorku veličine 114.

Broj kvarova - a_i	Broj uređaja - f_i	$f_i/114$
0	3	0.026
1	9	0.078
2	15	0.131
3	26	0.228
4	38	0.333
5	18	0.158
6	5	0.044
$\sum$	114	1

Izračunajte koeficijent asimetrije i zaobljenosti.

Prvo izračunamo

$$\overline{x}_{114} = 3.412 \qquad \text{i} \qquad s_{114} = 0.997.$$

Izračunajmo sada koeficijent asimetrije i koeficijent zaobljenosti danih podataka:

$$\begin{aligned} \mathrm{K_A} &= \frac{\frac{1}{114}\sum_{i=1}^{7} f_i(a_i - \overline{x}_{114})^3}{s_{114}^3} \\ &= \frac{\frac{1}{114}(3(0-3.412)^3 + \cdots + 5(6-3.412)^3)}{0.997^3} \\ &= -1.091, \end{aligned}$$

$$\begin{aligned} \mathrm{K_E} &= \frac{\frac{1}{114}\sum_{i=1}^{7} f_i(a_i - \overline{x}_{114})^4}{s_{114}^4} - 3 \\ &= \frac{\frac{1}{114}(3(0-3.412)^4 + \cdots + 5(6-3.412)^4)}{0.997^4} - 3 \\ &= 6.909. \end{aligned}$$

Po negativnom predznaku koeficijenta asimetrije zaključujemo da su podaci negativno asimetrični.