6 Opisna statistika

Kocka je bačena 1000 puta, pri čemu se jedinica pojavila 350 puta. Da li je kocka simetrična te da li je način bacanja bio ispravan? Tijekom radnog tjedna student bilježi vrijeme koje provede na putu od kuće do fakulteta pri čemu dobiva sljedeće podatke: 45, 49, 53, 41, 50 minuta. Ako je poznato da je to vrijeme distribuirano po normalnom zakonu, kako možemo odrediti parametre te razdiobe?

Na ova i slična pitanja, odgovor daje **matematička statistika**. U ovom poglavlju upoznat ćemo se s osnovnim pojmovima **opisne** ili **deskriptivne statistike**. Opisna ili deskriptivna statistika je grana statistike koja se bavi predočavanjem i opisivanjem glavnih karakteristika sakupljenih podataka (tablice, grafikoni, histogrami, srednje vrijednosti).

Prilikom opažanja ili eksperimentiranja, pažnja istraživača redovito je usmjerena na jednu ili više veličina. Promatramo li samo jednu veličinu, u oznaci X, onda je rezultat jednog mjerenja realni broj x. Višestrukim ponavljanjem mjerenja veličine X dobivamo konačan niz brojeva $x_1, x_2, \ldots, x_n$ kao rezultat od n ponovljenih mjerenja veličine X. Taj niz je **realizacija** veličine X. Veličina X obično se naziva **statističko obilježje**, a dobiveni niz $x_1, x_2, \ldots, x_n$ predstavlja **statističke podatke** o promatranom statističkom obilježju X.

6.1 Grafički prikaz podataka

U ovoj točki upoznat ćemo se s osnovnim grafičkim prikazima statističkih podataka. Promotrimo ponajprije sljedeći primjer.

Primjer 6.1. *U nekoj srednjoj školi ima* 20 *razrednih odjela. Na kraju polugodišta zabilježeni su sljedeći podatci o broju negativnih ocjena iz matematike u pojedinom razrednom odjelu:*

$$\begin{array}{cccccccccc} 4 & 5 & 3 & 5 & 1 & 3 & 1 & 6 & 6 & 2 \\ 3 & 3 & 4 & 6 & 4 & 3 & 1 & 4 & 1 & 4 \end{array}.$$

Na ovom primjeru objasnit ćemo pojmove koje smo definirali u uvodu ovog poglavlja te ćemo definirati još neke pojmove koji će nam biti važni prilikom grafičkog predočavanja podataka.

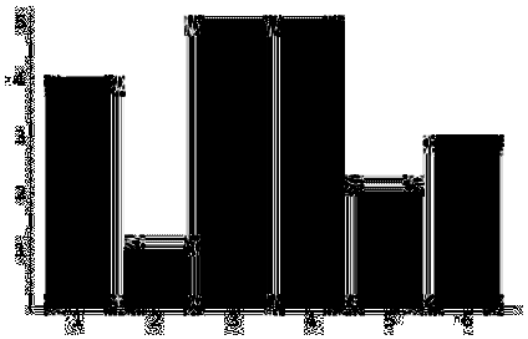
- statističko obilježje X predstavlja broj negativnih ocjena iz matematike u pojedinom razrednom odjelu
- X poprima vrijednosti iz skupa $\{1, 2, 3, 4, 5, 6\}$. Taj skup obično označavamo s $\operatorname{Im} X$, tj. $\operatorname{Im} X = \{1, 2, 3, 4, 5, 6\}$
- u ovom primjeru je $\operatorname{Im} X$ diskretan, tj. konačan skup, pa kažemo da je X **diskretno obilježje**
- obilježje može biti **numeričko** ili **nenumeričko**
- nenumeričko obilježje nazivamo i **kategorija**, npr. spol, boja, vrsta ...

- svakom elementu $a_i \in \operatorname{Im} X$ možemo pridružiti broj f_i tj. **frekvenciju pojavljivanja elementa** a_i u nizu podataka
- broj $f_{r_i} = \frac{f_i}{n}$ naziva se relativna frekvencija elementa a_i. Broj n predstavlja broj ponavljanja pokusa, odnosno broj mjerenja podataka. U ovom primjeru je $n = 20$.

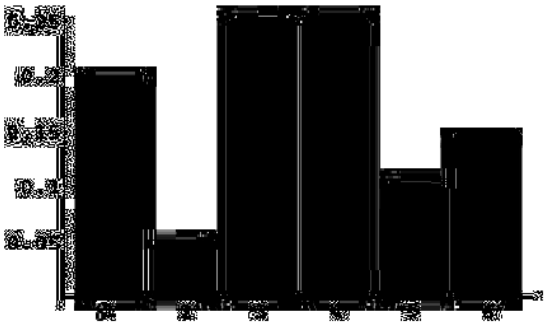
Prikažimo podatke iz Primjera 6.1 u **tablici frekvencija**.

a_i	f_i	f_{r_i}
1	4	$4/20 = 0.2$
2	1	$1/20 = 0.05$
3	5	$5/20 = 0.25$
4	5	$5/20 = 0.25$
5	2	$2/20 = 0.1$
6	3	$3/20 = 0.15$
Σ	20	1.00

Podatke također možemo prikazivati i pomoću **stupčastog dijagrama** (**bar chart**).



Uočimo kako je na prethodnoj slici visina svakog stupca jednaka odgovarajućoj frekvenciji. Stupčasti dijagram može se crtati i tako da visina svakog stupca bude jednaka odgovarajućoj relativnoj frekvenciji. Tada je ukupna površina svih stupaca jednaka 1, što je često bolje zbog usporedbe za različite vrijednosti n:



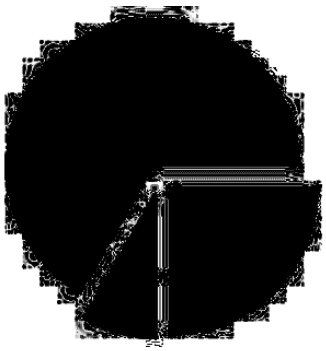
Ukoliko statističko obilježje može poprimiti malo različitih vrijednosti, onda je zgodno nacrtati **strukturni krug** (**pie chart**). Pogledajmo sljedeći primjer.

Primjer 6.2. *Ribar Roko ulovio je* 288 *riba od kojih je* 71 *lokarda,* 24 *zubatca i* 193 *srdele. Nacrtajte strukturni krug za dane podatke.*

Rješenje: Odredimo prvo tablicu frekvencija:

vrsta ribe	f_i	f_{r_i}	%
lokarda	71	$71/288 = 0.2465$	24.65%
zubatac	24	$24/288 = 0.0833$	8.33%
srdela	193	193/288=0.6702	67.02%
Σ	288	1.00	100%

Sada, strukturni krug izgleda ovako:

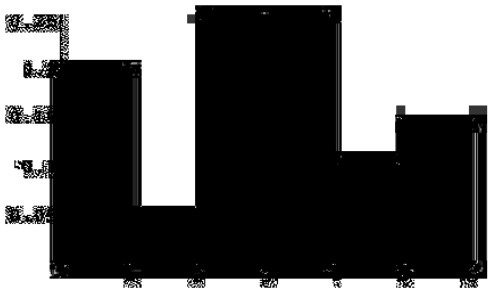


□

Slično kao i stupčasti dijagram, možemo nacrtati i **histogram**. Kod histograma, svaki stupac ima širinu 1 te se svaka dva stupca dodiruju. Visine stupaca jednake su odgovarajućim frekvencijama, odnosno relativnim frekvencijama. Ukoliko je visina svakog stupca jednaka odgovarajućim relativnim frekvencijama, površina svih stupaca je očito jednaka 1. Dakako, crtanje histograma nema smisla za nenumeričke vrijednosti.

Primjer 6.3. *Nacrtajte histogram za podatke iz Primjera* 6.1.

Rješenje:



□

U sljedećem primjeru promatrat ćemo neprekinuto statističko obilježje, odnosno obilježje koje može poprimiti vrijednosti iz nekog intervala realnih brojeva.

Primjer 6.4. *Na sistematskom pregledu studenata Visoke škole za primijenjeno računarstvo mjerena je visina (u metrima)* 30 *studenata. Dobiveni su sljedeći podatci:*

1.86	1.74	1.68	1.59	1.76	1.65	1.79	1.83	1.79	1.69
1.75	1.72	1.86	1.70	1.74	1.76	1.78	1.60	1.76	1.69
1.82	1.79	1.72	1.75	1.72	1.85	1.88	1.78	1.72	1.80

Nacrtajte histogram za dane podatke.

Rješenje: Kao što smo već napomenuli, mjereno statističko obilježje X je neprekinuto, odnosno poprima vrijednosti iz nekog intervala.

Dobivene podatke trebamo rasporediti u razrede. Prvo trebamo odrediti $x_{\max}$ i $x_{\min}$. Iz tablice vidimo da je $x_{\min} = 1.59$ i $x_{\max} = 1.88$. Sada trebamo odabrati adekvatan broj razreda, on je okvirno jednak vrijednosti $\sqrt{n}$. Kako je u našem slučaju $n = 30$, odabrat ćemo prvi veći cijeli broj, a to je $k = 6$.

Nadalje, određujemo zajedničku širinu razreda c. Imamo da je

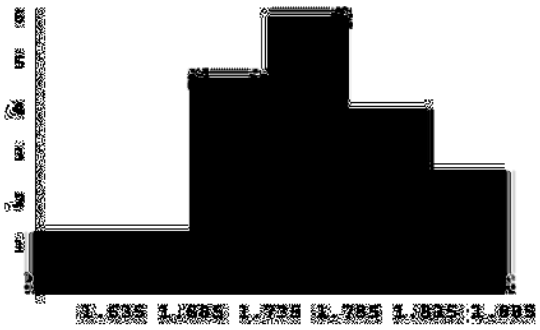
$$\frac{x_{\max} - x_{\min}}{k} = \frac{1.88 - 1.59}{6} = 0.0483.$$

Kako smo podatke mjerili na dvije decimale, tako ćemo i širinu razreda zaokružiti (uvijek na više) na dvije decimale. Prema tome, širina pojedinog razreda je $c = 0.05$.

Konačno, odredimo razrede $I_1, I_2, I_3, I_4, I_5, I_6$. Pri tome $I_1 \cup I_2 \cup I_3 \cup I_4 \cup I_5 \cup I_6$ treba obuhvaćati sve podatke, te je kraj jednog intervala jednak početku sljedećeg intervala. Prema tome imamo ovakvu situaciju:

razredi	f_i	$f_{r_i} = f_i/n$	f_{r_i}/c
$I_1 = [1.585, 1.635]$	2	0.067	1.34
$I_2 = [1.635, 1.685]$	2	0.067	1.34
$I_3 = [1.685, 1.735]$	7	0.233	4.66
$I_4 = [1.735, 1.785]$	9	0.3	6
$I_5 = [1.785, 1.835]$	6	0.2	4
$I_6 = [1.835, 1.885]$	4	0.133	2.66
Σ	30	1	20

Sada možemo nacrtati traženi histogram podataka. Sada širina stupca više nije proizvoljna, ona je jednaka širini razreda, tj. $c = 0.05$. Prema tome, da bi površina svih stupaca bila jednaka 1, na ordinati odabiremo vrijednosti $\frac{f_{r_i}}{c}$, a ne f_{r_i}, zato jer je $20 \cdot c = 20 \cdot 0.05 = 1$. Na kraju, histogram izgleda ovako:



□

6.2 Srednje vrijednosti uzorka

U ovoj točki upoznat ćemo neke pojmove kojima opisujemo srednje vrijednosti nekog uzorka. Preciznije, upoznat ćemo veličine kao što su **aritmetička sredina**, **medijan** i **mod**.

Aritmetička sredina

Neka je X numeričko statističko obilježje, te neka su $x_1, x_2, \ldots, x_n$ realizacije varijable X, odnosno statistički podatci. Aritmetička sredina podataka $x_1, x_2, \ldots, x_n$ je broj

$$\overline{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Ako se u nizu $x_1, x_2, \ldots, x_n$ pojavljuju samo brojevi $a_1, a_2, \ldots, a_k$ s frekvencijama $f_1, f_2, \ldots, f_k$, onda je

$$\overline{x} = \frac{1}{n}\left(f_1 a_1 + f_2 a_2 + \ldots + f_k a_k\right) = \frac{1}{n}\sum_{j=1}^{k} f_j a_j.$$

Uočimo još da je $\{a_1, a_2, \ldots, a_k\} \subseteq \operatorname{Im} X$.

Primjer 6.5. *Kontrolor uzima uzorke od po* 30 *proizvoda i svaki put zapiše broj defektnih proizvoda u uzorku. Nakon* 20 *pregledanih takvih uzoraka dobiveni su sljedeći podatci:*

$$\begin{array}{cccccccccc} 0 & 0 & 1 & 1 & 0 & 0 & 3 & 1 & 0 & 2 \\ 0 & 1 & 0 & 0 & 4 & 0 & 0 & 3 & 2 & 0 \end{array}.$$

Odredite aritmetičku sredinu broja defektnih proizvoda s obzirom na zadane podatke.

Rješenje: Ovdje je $X =$"broj defektnih proizvoda u uzorku od 30 proizvoda". Stoga je $\operatorname{Im} X = \{0, 1, 2, 3, \ldots, 30\}$. Frekvencije pojavljivanja defektnih proizvoda u kontroliranim uzorcima prikazane su u sljedećoj tablici:

i	f_i
0	11
1	4
2	2
3	2
4	1
$5 - 30$	0
Σ	20

Prema tome, tražena aritmetička sredina jednaka je

$$\overline{x} = \frac{11 \cdot 0 + 4 \cdot 1 + 2 \cdot 2 + 2 \cdot 3 + 1 \cdot 4}{20} = \frac{18}{20} = 0.9.$$

□

Medijan

Pojam medijana također promatramo samo za numeričke varijable. Neka je X numerička varijabla. Medijan je vrijednost od X za koju vrijedi da je 50% podataka manje od ili jednako toj vrijednosti i 50% podataka je veće od ili jednako njoj.

Kako bismo odredili medijan m niza $x_1, x_2, \ldots, x_n$, podatke je potrebno poredati po veličini:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Sada, u ovisnosti o tome da li je broj podataka n paran ili neparan, slijedi da je medijan jednak

$$\begin{aligned} m &= x_{(k)}, && n = 2k-1, k \in \mathbf{N} \\ m &= \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), && n = 2k, k \in \mathbf{N}. \end{aligned}$$

Primjer 6.6. *Mjeri se vrijeme izvođenja neke radne operacije. Podatci dobiveni u* 20 *nezavisnih mjerenja su (u sekundama):*

24	28	22	26	24	27	26	25	26	23
30	26	29	25	27	24	26	25	24	27

.

Odredite medijan za dana mjerenja.

Rješenje: Sortirajmo prvo podatke po veličini:

$$22, 23, 24, 24, 24, 24, 25, 25, 25, 26, 26, 26, 26, 26, 27, 27, 27, 28, 29, 30.$$

Sada, kako je $n = 20 = 2 \cdot 10$, slijedi da je

$$m = \frac{1}{2}\left(x_{(10)} + x_{(11)}\right) = \frac{1}{2}\left(26 + 26\right) = 26.$$

□

Formulu za medijan možemo napisati u sažetijem obliku, odnosno da ne razmatramo posebno parni i neparni slučaj. Pri tome ćemo iskoristiti formulu

$$x_{\left(\frac{p}{q}\right)} = x_{(k)} + \frac{r}{q}\left(x_{(k+1)} - x_{(k)}\right),$$

gdje su p i q cijeli brojevi i $p = k \cdot q + r$, $0 \leq r < q$, odnosno $\frac{p}{q} = k + \frac{r}{q}$. Drugim riječima, k je cjelobrojni količnik, a r je ostatak pri dijeljenju broja p s brojem q.

Uz takvu notaciju medijan m niza $x_1, x_2, \ldots, x_n$ jednak je

$$m = x_{\left(\frac{n+1}{2}\right)},$$

pa prethodni primjer možemo računati i ovako:

$$m = x_{\left(\frac{21}{2}\right)} = x_{\left(10+\frac{1}{2}\right)} = x_{(10)} + \frac{1}{2}\left(x_{(11)} - x_{(10)}\right) = 26 + \frac{1}{2}(26 - 26) = 26.$$

Navedenu formulu često ćemo koristiti u idućoj točki.

Mod

Mod je ona vrijednost statističkog obilježja koja se u uzorku javlja s najvećom frekvencijom. Koristan je kod statističkih obilježja koja nisu numerička, odnosno gdje ne možemo računati aritmetičku sredinu. Uzorak u kojemu postoje dvije vrijednosti s najvećom frekvencijom naziva se bimodalni uzorak, dok se uzorak sa samo jednim modom naziva unimodalni uzorak. Ako svi podatci imaju istu frekvenciju pojavljivanja u uzorku, tada uzorak nema mod.

Primjer 6.7. *Odredite mod za podatke iz Primjera* 6.1 *i Primjera* 6.6.

Rješenje: U Primjeru 6.1 imamo dvije vrijednosti: mod = 3 i mod = 4. Prema tome, to je bimodalan uzorak. Nadalje, u Primjeru 6.6 je mod = 26, tj. imamo unimodalan uzorak. □

6.3 Mjere raspršenja

U prošloj točki smo definirali veličine kojima opisujemo srednje vrijednosti nekog uzorka. Ovdje ćemo definirati veličine pomoću kojih opisujemo koliko je neki uzorak raspršen, odnosno disperziran. To su **raspon**, **interkvartil**, **varijanca** i **standardna devijacija**.

Raspon podataka

Neka je $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ uređeni niz podataka. Broj

$$d = x_{(n)} - x_{(1)}$$

naziva se raspon uzorka.

Primjer 6.8. *Odredite raspon uzorka iz Primjera* 6.6.

Rješenje: Određivanjem najveće i najmanje vrijednosti dobivamo da je raspon uzorka jednak

$$d = 30 - 22 = 8.$$

□

Interkvartil

Kako bismo definirali interkvartil, potrebne su nam definicije gornjeg i donjeg kvartila.

Donji kvartil q_L je ona vrijednost uzorka za koju vrijedi da je 25% svih podataka manje ili jednako od nje i 75% svih podataka veće ili jednako od nje. Donji kvartil računamo po formuli

$$q_L = x_{\left(\frac{n+1}{4}\right)}.$$

Gornji kvartil q_U je ona vrijednost uzorka za koju vrijedi da je 75% svih podataka manje ili jednako od nje i 25% svih podataka veće ili jednako od nje. Gornji kvartil računamo po formuli

$$q_U = x_{\left(\frac{3(n+1)}{4}\right)}.$$

Interkvartil d_q jednak je razlici gornjeg i donjeg kvartila:

$$d_q = q_U - q_L.$$

Primjer 6.9. *Odredite interkvartil za podatke iz Primjera* 6.4.

Rješenje: Podatke prvo trebamo poredati po veličini:

$$1.59, 1.60, 1.65, 1.68, 1.69, 1.69, 1.70, 1.72, 1.72, 1.72, 1.72, 1.74, 1.74, 1.75, 1.75,$$
$$1.76, 1.76, 1.76, 1.78, 1.78, 1.79, 1.79, 1.79, 1.80, 1.82, 1.83, 1.85, 1.86, 1.86, 1.88.$$

Sada računamo donji i gornji kvartil. Imamo da je

$$\begin{aligned} q_L &= x_{\left(\frac{n+1}{4}\right)} = x_{\left(\frac{30+1}{4}\right)} = x_{\left(7+\frac{3}{4}\right)} \\ &= x_{(7)} + \frac{3}{4}\left(x_{(8)} - x_{(7)}\right) \\ &= 1.70 + \frac{3}{4}\left(1.72 - 1.70\right) = 1.715 \approx 1.72 \end{aligned}$$

i

$$\begin{aligned} q_U &= x_{\left(\frac{3(n+1)}{4}\right)} = x_{\left(\frac{93}{4}\right)} = x_{\left(23+\frac{1}{4}\right)} \\ &= x_{(23)} + \frac{1}{4}\left(x_{(24)} - x_{(23)}\right) \\ &= 1.79 + \frac{1}{4}(1.80 - 1.79) = 1.7925 \approx 1.79. \end{aligned}$$

Prema tome, interkvartil je jednak

$$d_q = q_U - q_L = 1.79 - 1.72 = 0.07.$$

□

Uređenu petorku $\left(x_{(1)}, q_L, m, q_U, x_{(n)}\right)$ nazivamo **karakteristična petorka uzorka**. Pomoću nje crtamo tzv. "box and whisker" dijagram, odnosno **dijagram pravokutnika**. Pri formiranju dijagrama pravokutnika "outlieri" su sve vrijednosti koje su od donjeg ili gornjeg kvartila udaljene za više od $\frac{3}{2}d_q$. "Brkovi" su najmanja i najveća vrijednost koje nisu outlieri. Outlieri se posebno naznačavaju na dijagramu pravokutnika.

Primjer 6.10. *Na nekom fakultetu je odabran uzorak od* 40 *studenata i izmjerene su im visine:*

140	188	175	176	177	168	162	181
183	187	187	162	184	161	180	169
195	171	170	199	181	169	189	191
172	182	183	178	180	165	185	205
183	187	188	182	163	179	178	188

(a) Odredite karakterističnu petorku uzorka.

(b) Nacrtajte dijagram pravokutnika.

Rješenje:
(a) Uredimo podatke:

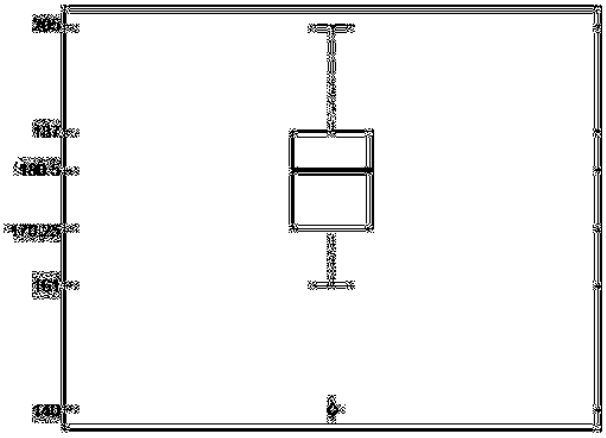
140	161	162	162	163	165	168	169
169	170	171	172	175	176	177	178
178	179	180	180	181	181	182	182
183	183	183	184	185	187	187	187
188	188	188	189	191	195	199	205

Tada je
$n = 40$,
$x_{(1)} = 140$,
$q_L = x_{\left(\frac{40+1}{4}\right)} = x_{\left(10+\frac{1}{4}\right)} = x_{(10)} + \frac{1}{4}\left(x_{(11)} - x_{(10)}\right) = 170 + \frac{1}{4}(171 - 170) = 170.25$,
$m = x_{\left(\frac{40+1}{2}\right)} = x_{\left(20+\frac{1}{2}\right)} = x_{(20)} + \frac{1}{2}\left(x_{(21)} - x_{(20)}\right) = 180 + \frac{1}{2}(181 - 180) = 180.5$,
$q_U = x_{\left(\frac{3(40+1)}{4}\right)} = x_{\left(30+\frac{3}{4}\right)} = x_{(30)} + \frac{3}{4}\left(x_{(31)} - x_{(30)}\right) = 187 + \frac{3}{4}(187 - 187) = 187$,
$x_{(40)} = 205$.
Prema tome, karakteristična petorka je $(140, 170.25, 180.5, 187, 205)$.

(b) Interkvartil iznosi $d_q = q_U - q_L = 187 - 170.25 = 16.75$ pa je

$$q_L - \frac{3}{2}d_q = 145.125 \quad \text{i} \quad q_U + \frac{3}{2}d_q = 212.125.$$

Sada lagano crtamo dijagram pravokutnika.



□

Napomenimo još da medijan, te gornji i donji kvartil spadaju u mjere lokacije. Tu još spadaju decili, percentili i općenito kvantili, ali ovdje nećemo uvoditi te veličine.

Uzoračka disperzija i standardno odstupanje

Uzoračka disperzija (varijanca) s^2 niza $x_1, x_2, \ldots, x_n$ dana je formulom

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\,,$$

pri čemu je $\overline{x}$ aritmetička sredina uzorka $x_1, x_2, \ldots, x_n$.

Uzoračko standardno odstupanje s (devijacija) jednako je drugom korijenu uzoračke disperzije, odnosno

$$s = +\sqrt{s^2}.$$

Formula za uzoračku disperziju je nepraktična za računanje pa ćemo je prikazati u pogodnijem

obliku. Naime, vrijedi

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

Ovaj oblik formule je puno praktičniji za računanje.

Nadalje, ako se u uzorku $x_1, x_2, \ldots, x_n$ vrijednosti $a_1, a_2, \ldots, a_k$ pojavljuju s frekvencijama $f_1, f_2, \ldots, f_k$, onda vrijedi

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(a_i-\overline{x})^2\cdot f_i = \frac{1}{n-1}\left(\sum_{i=1}^{k}f_i\cdot a_i^2-n\overline{x}^2\right).$$

Primjer 6.11. *Izračunajte disperziju i standardno odstupanje za podatke iz Primjera* 6.6.

Rješenje: Napravimo prvo frekvencijsku tablicu u koju ćemo unijeti vrijednosti f_ia_i i $f_ia_i^2$.

a_i	f_i	f_ia_i	$f_ia_i^2$
22	1	22	484
23	1	23	529
24	4	96	2304
25	3	75	1875
26	5	130	3380
27	3	81	2187
28	1	28	784
29	1	29	841
30	1	30	900
Σ	20	514	13284

Sada je

$$\overline{x} = \frac{514}{20} = 25.7,$$

$$s^2 = \frac{1}{19}\left(13284 - 20\cdot 25.7^2\right) = 3.91$$

i

$$s = \sqrt{3.91} = 1.98.$$

□

6.4 Mjere oblika

Slično kao što se definira uzoračka disperzija, može se definirati i uzorački centralni moment reda k, $k \in \mathbf{N}$:

$$\mu_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^k .$$

Uočimo da je centralni moment reda 1 uvijek jednak nuli. Naime, vrijedi

$$\mu_1 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i - n\overline{x} \right) = \frac{1}{n-1} (n\overline{x} - n\overline{x}) = 0.$$

Centralni moment reda 2 je upravo uzoračka disperzija. Značajan je i centralni moment reda 3, odnosno

$$\mu_3 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^3 .$$

U tu svrhu pogledajmo sljedeći primjer.

Primjer 6.12. *Zadan je uzorak* $0.2, 1.5, 2.5, 4.5, 5.5, 6.8$. *Odredite centralni moment reda* 3 *za taj uzorak.*

Rješenje: Aritmetička sredina zadanog uzorka jednaka je

$$\overline{x} = \frac{0.2 + 1.5 + 2.5 + 4.5 + 5.5 + 6.8}{6} = \frac{21}{6} = 3.5.$$

Sada je centralni moment reda 3 jednak

$$\begin{aligned} \mu_3 &= \frac{1}{5} \left((0.2 - 3.5)^3 + (1.5 - 3.5)^3 + (2.5 - 3.5)^3 + (4.5 - 3.5)^3 + (5.5 - 3.5)^3 + (6.8 - 3.5)^3 \right) \\ &= \frac{1}{5} (-35.937 - 8 - 1 + 1 + 8 + 35.937) = 0. \end{aligned}$$

□

Iz prethodnog primjera zaključujemo kako je kod uzorka simetričnog s obzirom na njegovu aritmetičku sredinu, centralni moment reda 3 uvijek jednak nuli. Stoga se pomoću te veličine definira takozvani **koeficijent asimetrije** uzorka.

Koeficijent asimetrije uzorka definiran je formulom

$$\alpha_3 = \frac{\mu_3}{s^3} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3 .$$

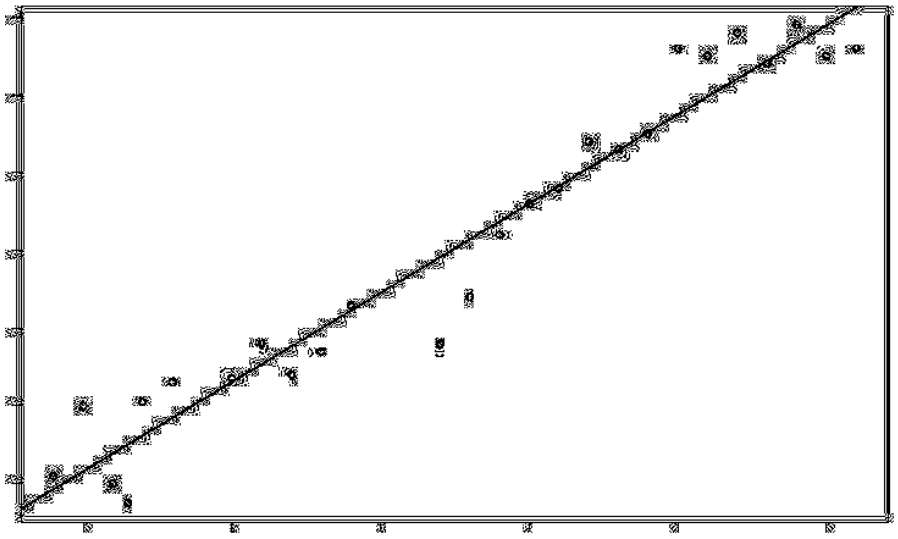
Dakako, ako se u uzorku $x_1, x_2, \ldots, x_n$ vrijednosti $a_1, a_2, \ldots, a_k$ pojavljuju s frekvencijama $f_1, f_2, \ldots, f_k$, onda je

$$\alpha_3 = \frac{1}{n-1} \sum_{i=1}^{k} f_i \cdot \left(\frac{a_i - \overline{x}}{s} \right)^3$$

Ako je $\alpha_3 = 0$ kažemo da je uzorak simetričan. Nadalje, ako je $\alpha_3 > 0$ kažemo da je uzorak pozitivno asimetričan, dok je za $\alpha_3 < 0$ uzorak negativno asimetričan.

6.5 Linearna regresija

Pretpostavimo da imamo n parova podataka (x_i, y_i), $i = 1, 2, \ldots, n$. Želimo odrediti vezu između nezavisne varijable x i zavisne varijable y. Nadalje, pretpostavimo da je ta veza linearna, odnosno da je graf pripadajuće funkcije pravac $y = \alpha x + \beta$. Drugim riječima, mi tražimo pravac koji najbolje aproksimira dane parove točaka, kao na slici.



Pomoću takozvane metode najmanjih kvadrata dobiva se da je procjenitelj pravac

$$y = \hat{\alpha} x + \hat{\beta},$$

gdje je

$$\hat{\alpha} = \frac{s_{xy}}{s_x^2}, \quad \hat{\beta} = \overline{y} - \hat{\alpha}\overline{x},$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right),$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \right),$$

i

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \right).$$

Primjer 6.13. *Šestero studenata upitano je koliko su se sati pripremali za završni ispit. Njihovi odgovori na to pitanje uspoređeni su s bodovima na ispitu (maksimalni broj bodova je* 30*). Dobiveni su ovi rezultati:*

sati učenja	0.50	1.25	1.50	2.00	2.25	3.50
bodovi	19	23	25	26	28	30

(a) Procijenite pravac linearne regresije za ove podatke.

(b) Procijenite koliko student ima bodova na ispitu ako se za njega pripremao jedan sat.

(c) Procijenite koliko se za ispit pripremao student koji ima 24 boda na ispitu.

Rješenje: (a) Ovdje je $n = 6$, pa za pripadne aritmetičke sredine $\overline{x}$ i $\overline{y}$ imamo da je

$$\overline{x} = \frac{1}{6}\left(0.50 + 1.25 + 1.50 + 2.00 + 2.25 + 3.50\right) = 1.833$$

i

$$\overline{y} = \frac{1}{6}\left(19 + 23 + 25 + 26 + 28 + 30\right) = 25.167.$$

Nadalje,

$$\sum_{i=1}^{6} x_i^2 = 0.50^2 + 1.25^2 + 1.50^2 + 2.00^2 + 2.25^2 + 3.50^2 = 25.375,$$

pa je

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{1}{5}\left(25.375 - 6 \cdot 1.833^2\right) = 1.043.$$

Dalje imamo

$$\begin{aligned}\sum_{i=1}^{6} x_i y_i &= 0.50 \cdot 19 + 1.25 \cdot 23 + 1.50 \cdot 25 + 2.00 \cdot 26 + 2.25 \cdot 28 + 3.50 \cdot 30 \\ &= 295.750,\end{aligned}$$

odakle je

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}\right) = \frac{1}{5}\left(295.750 - 6 \cdot 1.833 \cdot 25.167\right) = 3.793.$$

Konačno, koeficijenti $\hat{\alpha}$ i $\hat{\beta}$ su jednaki

$$\hat{\alpha} = \frac{s_{xy}}{s_x^2} = \frac{3.793}{1.043} = 3.637$$

i

$$\hat{\beta} = \overline{y} - \hat{\alpha}\overline{x} = 25.167 - 3.637 \cdot 1.833 = 18.5,$$

pa je traženi pravac linearne regresije

$$y = \hat{\alpha}x + \hat{\beta} = 3.637x + 18.5.$$

(b) Student koji se pripremao 1 sat, po ovoj metodi ima približno

$$y(1) = 3.637 \cdot 1 + 18.5 = 22.137 \approx 22$$

boda na ispitu.
(c) Ovdje trebamo naći x takav da je

$$3.637x + 18.5 = 24.$$

Rješenje te jednadžbe je $x \approx 1.50$, pa se student za ispit pripremao oko sat i pol. □

Usko u vezi s linearnom regresijom je **Pearsonov koeficijent korelacije**. Pearsonov koeficijent korelacije r varijabli x_i i y_i, $i = 1, 2, \ldots, n$, definiran je formulom

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \cdot \frac{y_i - \overline{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y}.$$

Može se pokazati da Pearsonov koeficijent korelacije uvijek poprima vrijednosti unutar intervala $[-1, 1]$, odnosno uvijek je $-1 \leq r \leq 1$.

Ako je $r = 0$ onda nema korelacije između varijabli x_i i y_i, $i = 1, 2, \ldots, n$. Ako je $r > 0$ onda je korelacija pozitivna. To znači da ako varijabla x raste, onda u pravilu raste i varijabla y. S druge strane, ako je $r < 0$ korelacija je negativna pa ako x raste, onda u pravilu varijabla y pada.

Primjer 6.14. *Odredite Pearsonov koeficijent korelacije za podatke iz Primjera* 6.13.

Rješenje: U prethodnom primjeru smo dobili da je $s_{xy} = 3.793$ i $s_x^2 = 1.043$, odakle je $s_x = 1.021$. Trebamo još izračunati s_y. Slično kao i u prethodnom primjeru dobivamo

$$\sum_{i=1}^{6} y_i^2 = 19^2 + 23^2 + 25^2 + 26^2 + 28^2 + 30^2 = 3875,$$

odakle je

$$s_y^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \right) = \frac{1}{5} \left(3875 - 6 \cdot 25.167^2 \right) = 14.947,$$

odnosno $s_y = 3.866$. Konačno, uvrštavanjem dobivenih podataka dobivamo da je

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{3.793}{1.021 \cdot 3.866} = 0.96,$$

što je vrlo visoka korelacija. □

6.6 Zadatci za ponavljanje

Zadatak 6.1. Tijekom 20 dana zabilježeni su sljedeći podatci o broju prometnih nesreća na širem području nekog grada:

3	0	3	5	0	4	0	6	1	3
3	4	0	3	1	4	5	0	10	1

.

(a) Prikažite zadane podatke pomoću frekvencijske tablice.

(b) Odredite aritmetičku sredinu uzorka.

(c) Odredite disperziju i standardno odstupanje uzorka.

(d) Odredite mod uzorka.

(e) Odredite medijan uzorka.

(f) Odredite raspon uzorka.

(g) Odredite donji i gornji kvartil, te interkvartil uzorka.

(h) Odredite koeficijent asimetrije uzorka.

Rješenje:

(a) Tablica frekvencija izgleda ovako:

a_i	f_i	f_{r_i}
0	5	$5/20 = 0.25$
1	3	$3/20 = 0.15$
3	5	$5/20 = 0.25$
4	3	$3/20 = 0.15$
5	2	$2/20 = 0.1$
6	1	$1/20 = 0.05$
10	1	$1/20 = 0.05$
Σ	20	1.00

(b) Kako je $n = 20$, aritmetička sredina iznosi

$$\overline{x} = \frac{5 \cdot 0 + 3 \cdot 1 + 5 \cdot 3 + 3 \cdot 4 + 2 \cdot 5 + 1 \cdot 6 + 1 \cdot 10}{20} = \frac{56}{20} = 2.8.$$

(c) Izračunajmo prvo disperziju:

$$s^2 = \frac{1}{19}\left(5 \cdot 0^2 + 3 \cdot 1^2 + 5 \cdot 3^2 + 3 \cdot 4^2 + 2 \cdot 5^2 + 1 \cdot 6^2 + 1 \cdot 10^2 - 20 \cdot 2.8^2\right) = 6.59.$$

Zbog toga je standardno odstupanje jednako $s = \sqrt{s^2} = \sqrt{6.59} = 2.567$.

(d) U ovom uzorku imamo dvije vrijednosti s najvećom frekvencijom, tj. $\text{mod} = 0$ i $\text{mod} = 3$.

(e) Da bismo odredili medijan, prvo ćemo podatke poredati po veličini:

$$0, 0, 0, 0, 0, 1, 1, 1, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 10.$$

Sada je medijan jednak

$$m = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{21}{2}\right)} = x_{\left(10+\frac{1}{2}\right)} = x_{(10)} + \frac{1}{2}\left(x_{(11)} - x_{(10)}\right) = 3 + \frac{1}{2}(3-3) = 3.$$

(f) Raspon uzorka je $d = x_{(20)} - x_{(1)} = 10 - 0 = 10$.

(g) Donji i gornji kvartil su redom jednaki

$$q_L = x_{\left(\frac{n+1}{4}\right)} = x_{\left(\frac{20+1}{4}\right)} = x_{\left(5+\frac{1}{4}\right)} = x_{(5)} + \frac{1}{4}\left(x_{(6)} - x_{(5)}\right) = 0 + \frac{1}{4}(1-0) = 0.25,$$

$$q_U = x_{\left(\frac{3(n+1)}{4}\right)} = x_{\left(\frac{3(20+1)}{4}\right)} = x_{\left(15+\frac{3}{4}\right)} = x_{(15)} + \frac{3}{4}\left(x_{(16)} - x_{(15)}\right) = 4 + \frac{3}{4}(4-4) = 4,$$

pa je interkvartil jednak

$$d_q = q_U - q_L = 4 - 0.25 = 3.75.$$

(h) Formulu za koeficijent asimetrije možemo zapisati u obliku

$$\alpha_3 = \frac{1}{(n-1)s^3}\sum_{i=1}^{n} f_i(a_i - \overline{x})^3,$$

pa je

$$\begin{aligned}\alpha_3 = \frac{1}{19 \cdot 2.567^3}\Big[&5 \cdot (0-2.8)^3 + 3 \cdot (1-2.8)^3 + 5 \cdot (3-2.8)^3 + 3 \cdot (4-2.8)^3 \\ &+2 \cdot (5-2.8)^3 + 1 \cdot (6-2.8)^3 + 1 \cdot (10-2.8)^3\Big] = 0.95.\end{aligned}$$

□

Zadatak 6.2. Profesorica povijesti je tijekom školskog sata ispitala 6 učenika. Nakon što ih je ocijenila, upitala je svakog od učenika koliko je sati jučer proveo pred računalom igrajući igrice. U priloženoj tablici x_i označava dobivenu ocjenu, a y_i broj sati provedenih pred računalom:

x_i	5	5	4	4	3	2
y_i	3	1	2	3	4	5

(a) Odredite pravac linearne regresije za dane podatke.

(b) Odredite Pearsonov koeficijent korelacije za dane podatke. Da li je korelacija između varijabli pozitivna ili negativna?

Rješenje: (a) Imamo 6 parova podataka, pa za pripadne aritmetičke sredine $\overline{x}$ i $\overline{y}$ imamo da je

$$\overline{x} = \frac{1}{6}(5+5+4+4+3+2) = 3.833$$

i

$$\overline{y} = \frac{1}{6}(3+1+2+3+4+5) = 3.$$

Nadalje,

$$\sum_{i=1}^{6} x_i^2 = 5^2+5^2+4^2+4^2+3^2+2^2 = 95,$$

pa je

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{1}{5}\left(95 - 6\cdot 3.833^2\right) = 1.37.$$

Dalje imamo

$$\sum_{i=1}^{6} x_i y_i = 5\cdot 3+5\cdot 1+4\cdot 2+4\cdot 3+3\cdot 4+2\cdot 5 = 62,$$

odakle je

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}\right) = \frac{1}{5}(62 - 6\cdot 3.833\cdot 3) = -1.399.$$

Konačno, koeficijenti $\hat{\alpha}$ i $\hat{\beta}$ su jednaki

$$\hat{\alpha} = \frac{s_{xy}}{s_x^2} = -\frac{1.399}{1.37} = -1.021$$

i

$$\hat{\beta} = \overline{y} - \hat{\alpha}\overline{x} = 3 + 1.021\cdot 3.833 = 6.913,$$

pa je traženi pravac linearne regresije

$$y = \hat{\alpha}x + \hat{\beta} = -1.021x + 6.913.$$

(b) U prvom dijelu zadatka dobili smo da je $s_{xy} = -1.399$ i $s_x^2 = 1.37$, odakle je $s_x = 1.17$. Trebamo još izračunati s_y. Slično kao i u prethodnom primjeru imamo

$$\sum_{i=1}^{6} y_i^2 = 3^2+1^2+2^2+3^2+4^2+5^2 = 64,$$

odakle je

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) = \frac{1}{5}\left(64 - 6 \cdot 3^2\right) = 2,$$

odnosno $s_y = 1.414$. Konačno, uvrštavanjem dobivenih podataka dobivamo da je

$$r = \frac{s_{xy}}{s_x \cdot s_y} = -\frac{1.399}{1.17 \cdot 1.414} = -0.846,$$

pa su varijable negativno korelirane. □