2 Deskriptivna statistika

Prilikom opažanja ili eksperimentiranja, pažnja istraživača redovito je usmjerena na jednu ili više veličina. Ako se promatra samo jedna veličina, označimo ju s X, onda je rezultat jednog mjerenja jedan realan broj x. Višestrukim ponavljanjem mjerenja veličine X dobiva se konačni niz brojeva $x_1, x_2, \ldots, x_n$ kao rezultat n ponovljenih mjerenja koji nazivamo **realizacija od** X. Veličina X obično se naziva **statističko obilježje**, a dobiveni niz brojeva $x_1, x_2, \ldots, x_n$ **statistički podaci** o promatranom statističkom obilježju X.

2.1 Grafički prikaz podataka

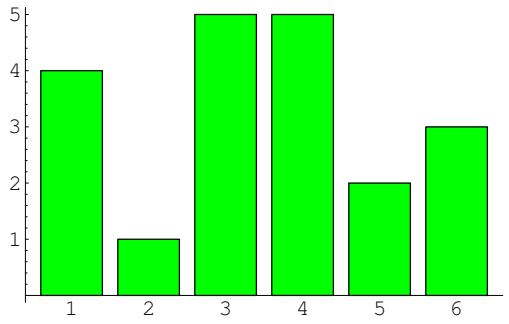
Primjer 8 *Neka X označava broj dobiven bacanjem igrače kocke. Kocku smo bacali 20 puta i dobiveni su sljedeći podaci:* $1, 3, 1, 6, 2, 6, 4, 6, 3, 3, 4, 3, 1, 4, 4, 1, 4, 5, 3, 5$.

- statističko obilježje $X =$ "broj na kocki"
- $\mathrm{Im}X = \{1, 2, 3, 4, 5, 6\} \Rightarrow$ skup svih vrijednosti koje X može poprimiti
- u našem primjeru, $\mathrm{Im}X$ je 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 i slično; možemo mu pridijeliti neku numeričku vrijednost, ali tada nema smisla računati npr. aritmetičku sredinu podataka!
- svakom elementu $a_i \in \mathrm{Im}X$ možemo pridružiti broj $f_i \Rightarrow$ **frekvencija (učestalost) pojavljivanja elementa** a_i u nizu podataka
- broj $f_{r_i} = \dfrac{f_i}{n}$: **relativna frekvencija** od a_i
  (n je broj ponavljanja pokusa, u ovom primjeru $n = 20$)

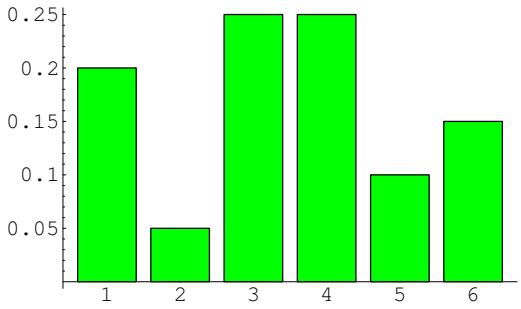
Prikažimo podatke u **TABLICI FREKVENCIJA**

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

GRAFIČKI PRIKAZ PODATAKA POMOĆU **STUPČASTOG DIJAGRAMA (BAR - CHART)**



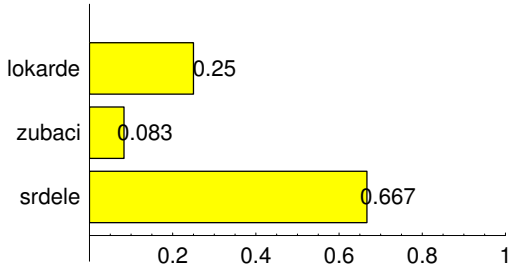
Stupčasti dijagram može se crtati i tako da ukupna površina stupića bude jednaka 1, što je bolje zbog usporedbe, npr. za različite n:



Primjer 9 *U uzorku od 144 ribe ulovljene u Bračkom kanalu, bilo je 36 lokardi, 12 zubataca i 96 srdela.*

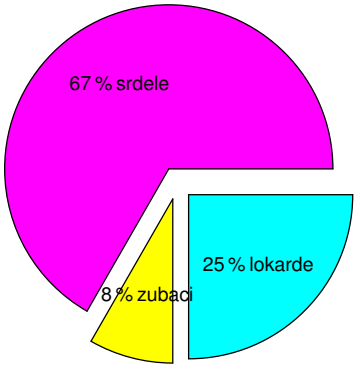
vrsta ribe	frekvencija	relativna frekvencija	
lokarde	36	$\frac{36}{144} = 0.25$	25%
zubaci	12	$\frac{12}{144} = 0.083$	8.3%
srdele	96	$\frac{96}{144} = 0.667$	66.7%
Σ	144	1.00	100%

HORIZONTALNI STUPČASTI DIJAGRAM



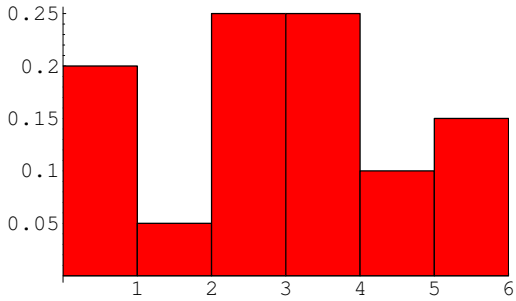
STRUKTURNI KRUG (PIE CHART)

-ako imamo relativno malo različitih vrijednosti koje statističko obilježje može poprimiti



Primjer 10 *Nacrtajmo* ***histogram*** *za podatke iz Primjera 8.*

- svaka 2 susjedna stupića se dodiruju i svaki ima težište u vrijednosti visina f_i ili f_{r_i}
- površina svakog stupića jednaka je relativnoj frekvenciji pa je površina ispod cijelog grafa jednaka je 1
- nema smisla za nenumeričke vrijednosti



Primjer 11 *Mjerena je visina (u metrima) 30 20-ogodišnjaka. Dobiveni su podaci: 1.85, 1.88, 1.78, 1.72, 1.80, 1.72, 1.75, 1.72, 1.79, 1.82, 1.69, 1.76, 1.60, 1.78, 1.76, 1.74, 1.70, 1.86, 1.72, 1.75, 1.69, 1.79, 1.83, 1.79, 1.65, 1.76, 1.59, 1.68, 1.74, 1.86.*

- statističko obilježje $X =$ visina $\Rightarrow$ neprekidno statističko obilježje (poprima vrijednosti iz nekog intervala)
- podatke najprije moramo svrstati u razrede:
  1. odredimo x_{min} i x_{max} : $x_{min} = 1.59,\ x_{max} = 1.88$
  2. izaberemo adekvatan broj razreda (okvirno: $\sqrt{n}$) $\Rightarrow$ k = 6
  3. odredimo zajedničku širinu razreda:

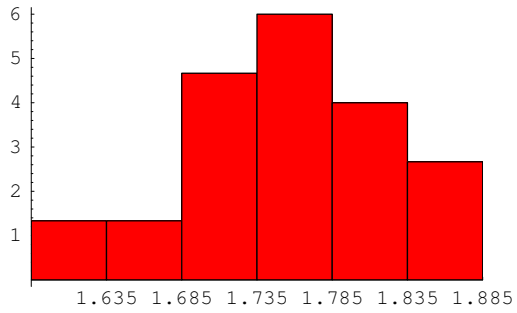
$$c = \frac{x_{max} - x_{min}}{k} = \frac{1.88 - 1.59}{6} = 0.0483 \ \Rightarrow \ \mathbf{c{=}0.05}$$

(uvijek zaokružujemo na više!)

4. odredimo razrede (tj. lijevi prag razreda): $I_1, ..., I_k$
   pritom $I_1 \cup I_2 \cup ... \cup I_k$ mora obuhvaćati sve podatke
   $I_i = [a_{i,1}, a_{i,2}], \quad a_{i,2} = a_{i+1,1}$
   $I_{i+1} = [a_{i+1,1}, a_{i+1,2}]$

RAZREDI	f_i	$f_{r_i} = \frac{f_i}{n}$	$\frac{f_{r_i}}{c}$ = relativna frekv.razreda
$I_1 = [1.585, 1.635]$	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

Nacrtajmo histogram za ove podatke. Širina stupića više nije proizvoljna (sada je jednaka širini razreda, tj. c=0.05), pa da bi suma površina svih pravokutnika (odnosno površina ispod grafa) bila jednaka 1, na ordinatu ucrtavamo $\frac{f_{r_i}}{c}$ a ne f_{r_i}. Naime, $20 \cdot c = 20 \cdot 0.05 = 1$.



STEM AND LEAF DIJAGRAM

stem	leaf
1.5	9
1.6	90958
1.7	82252968640259964
1.8	5802636

stem	leaf
1.5	9
1.6_*	0
1.6^*	5899
1.7_*	2224024
1.7^*	8596865996
1.8_*	023
1.8^*	5866

2.2 Srednje vrijednosti uzorka

2.2.1 Aritmetička sredina uzorka

Aritmetička sredina uzorka je broj

$$\bar{x} := \frac{1}{n}(x_1 + x_2 + \ldots + x_n).$$

Ako je $\mathrm{Im}X = \{a_1, a_2, \ldots, a_k\}$ i pritom se a_i u uzroku ponavlja f_i puta, tada

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot a_i, \quad n = \sum_{i=1}^{k} f_i.$$

- ima smisla samo za numeričke podatke

Primjer 12 *Izračunajte $\bar{x}$ za podatke iz Primjera 11.*

Rješenje:

$$\begin{aligned}\bar{x} &= \frac{1}{30}(1.59 + 1.60 + 1.65 + 1.68 + 2 \cdot 1.69 + 4 \cdot 1.72 + 1.70 + 2 \cdot 1.74 + 2 \cdot 1.75 \\ &\quad +3 \cdot 1.76 + 2 \cdot 1.78 + 3 \cdot 1.79 + 1.80 + 1.82 + 1.83 + 1.85 + 2 \cdot 1.86 + 1.88) \\ &= \frac{52.57}{30} \approx 1.75\end{aligned}$$

□

2.2.2 Medijan uzorka

- uredimo podatke (sortiramo ih po veličini): $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$
- ima smisla samo za numeričke podatke

Medijan uzorka je broj za koji vrijedi da je 50% svih podataka manje od ili jednako njemu i 50% svih podataka veće od ili jednako njemu.

Ako je broj podataka neparan, tj $n = 2k - 1,\ k \in \mathbf{N}$, tada je $m = x_{(k)}$. Za paran n ($n = 2k$), vrijedi

$$m = \frac{x_{(k)} + x_{(k+1)}}{2}.$$

Općenito, $m = x_{(\frac{n+1}{2})}$. Vrijedi

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

Primjer 13 *Nađite medijan uzorka za podatke iz Primjera 8.*

Rješenje: Sortiramo podatke po veličini:

$$1 \leq 1 \leq 1 \leq 1 \leq 2 \leq 3 \leq 3 \leq 3 \leq 3 \leq \mathbf{3} \leq \mathbf{4} \leq 4 \leq 4 \leq 4 \leq 4 \leq 5 \leq 5 \leq 6 \leq 6 \leq 6$$
$$n = 20 = 2 \cdot 10$$
$$m = \frac{x_{(10)} + x_{(11)}}{2} = \frac{3+4}{2} = 3.5$$

□

2.2.3 Uzorački mod

Mod je ona vrijednost statističkog obilježja koja se u uzorku javlja s najvećom frekvencijom.

- koristan kod statističkih obilježja koja nisu numerička, pa nema aritmetičke sredine

- BIMODALNI UZORAK: uzorak u kojem postoje 2 vrijednosti s jednakom frekvencijom
- UNIMODALNI UZORAK: uzorak u kojem postoji samo jedan mod
- Ako svi podaci imaju istu frekvenciju pojavljivanja u uzorku, tada uzorak nema mod.

Primjer 14 *Nađite mod za podatke iz Primjera 8 i 11.*

Rješenje:

- u Primjeru 8: mod = 3 & mod=4 $\Rightarrow$ bimodalan uzorak
- u Primjeru 11: mod = 1.72

□

2.3 Mjere disperzije ili varijabiliteta

2.3.1 Raspon uzorka

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

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

naziva se **raspon uzorka.**

Primjer 15 *Odredite raspon uzorka iz Primjera 11.*

Rješenje:

$$d = 1.88 - 1.59 = 0.29$$

□

Primjer 16 *Podaci* $1, 2, 3, 10$ *i* $1, 3, 7, 10$ *imaju isti raspon uzorka, ali je njihova "raspršenost" ipak bitno različita. Kažemo da je prvom slučaju broj* 10 *"outlier" jer "strši" u odnosu na ostale podatke* $(1, 2, 3)$ *koji su grupirani.*

2.3.2 Interkvartil

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

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

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

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

Interkvartil: $d_q = q_U - q_L$

Primjer 17 *Nađite formule za donji i gornji kvartil uzorka od* $n = 5, 6$ *i* 7 *podataka.*

Rješenje: Za $n = 5$

$$q_L = x_{\left(\frac{6}{4}\right)} = x_{(1)} + \frac{1}{2}(x_{(2)} - x_{(1)}) = \frac{1}{2}(x_{(1)} + x_{(2)}),$$

$$q_U = x_{\left(\frac{18}{4}\right)} = x_{(4)} + \frac{1}{2}(x_{(5)} - x_{(4)}) = \frac{1}{2}(x_{(4)} + x_{(5)}).$$

Za $n = 6$

$$q_L = x_{\left(\frac{7}{4}\right)} = x_{(1)} + \frac{3}{4}(x_{(2)} - x_{(1)}) = \frac{1}{4}x_{(1)} + \frac{3}{4}x_{(2)},$$

$$q_U = x_{\left(\frac{21}{4}\right)} = x_{(5)} + \frac{1}{4}(x_{(6)} - x_{(5)}) = \frac{3}{4}x_{(5)} + \frac{1}{4}x_{(6)}$$

i za $n = 7$

$$q_L = x_{\left(\frac{8}{4}\right)} = x_{(2)}, \quad q_U = x_{\left(\frac{24}{4}\right)} = x_{(6)}.$$

□

Primjer 18 *Odredite interkvartil za podatke iz Primjera 11.*

Rješenje:

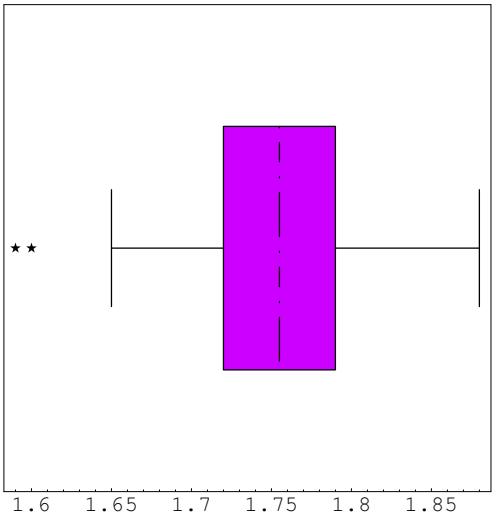
$$\begin{aligned} q_L &= x_{(\frac{n+1}{4})} = x_{(\frac{30+1}{4})} = x_{(7+\frac{3}{4})} = x_{(7)} + \frac{3}{4}\left(x_{(8)} - x_{(7)}\right) \\ &= 1.70 + \frac{3}{4}(1.72 - 1.70) = 1.715 \approx 1.72 \\ q_U &= x_{(\frac{3(n+1)}{4})} = x_{(\frac{93}{4})} = x_{(23+\frac{1}{4})} = x_{(23)} + \frac{1}{4}\left(x_{(24)} - x_{(23)}\right) \\ &= 1.79 + \frac{1}{4}(1.80 - 1.79) = 1.7925 \approx 1.79 \\ d_q &= q_U - q_L = 1.79 - 1.72 = 0.07 \end{aligned}$$

□

Uređenu petorku $(x_{(1)}, q_L, m, q_U, x_{(n)})$ zovemo **karakteristična petorka uzorka.** Pomoću nje crtamo tzv. **"box and whisker" dijagram**, odnosno dijagram pravokutnika.

Primjer 19 *Nacrtajte box and whisker dijagram za podatke iz Primjera 11.*

$x_{(1)} = 1.59,\ q_L = 1.72,\ m = 1.75,\ q_U = 1.79,\ x_{(30)} = 1.88,\ d_q = 0.07$

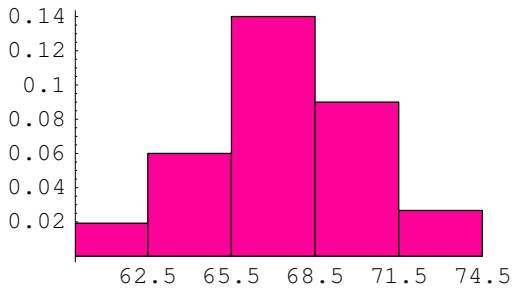


Sve što je izvan $3d_q$ označava se točkom i smatra se "ekstrmnom vrijednošću", tj. "outlier".

Zadatak 1 *U tablici su dane težine 100 studenata PBF-a. Nacrtajte histogram, nađite aritmetičku sredinu, medijan te interkvartil ovog uzorka.*

težina (kg)	*broj studenata*	*sredina razreda*	f_{r_i}	f_{r_i}/c
$60-62$	5	61	0.05	0.017
$63-65$	18	64	0.18	0.06
$66-68$	42	67	0.42	0.14
$69-71$	27	70	0.27	0.09
$72-74$	8	73	0.08	0.027
Σ	100		1	0.334

Rješenje:



Aritmetička sredina:

$$\bar{x} = \frac{1}{100}(61 \cdot 5 + 64 \cdot 18 + 67 \cdot 42 + 70 \cdot 27 + 73 \cdot 8) = 67.45$$

Medijan: U prva 2 razreda upada 5+18=23 podataka, a u prva 3 razreda 5+18+42=65 podataka, što znači da se medijan nalazi negdje unutar 3.razreda, tj. $65.5 \leq m \leq 68.5$. Medijan dobivamo interpolacijom:

$$m = 65.5 + \frac{50-23}{42}(68.5 - 65.5) = 65.5 + \frac{27}{42} \cdot 3 = 67.43$$

Vrijednost medijana može se očitati i sa histograma - medijan je apscisa koja odgovara liniji koja dijeli histogram na 2 dijela jednake površine.

Interkvartil: Najprije moramo odrediti donji i gornji kvartil. Postupak je sličan kao kod određivanja medijana - donji kvartil nalazi se negdje unutar

3.razreda tj. $65.5 \leq q_L \leq 68.5$, dok se gornji kvartil nalazi unutar 4.razreda (budući prva 3 razreda sadrže 65, a prva 4: 5+18+42+27=92 podatka), tj. $68.5 \leq q_U \leq 71.5$. Imamo:

$$q_L = 65.5 + \frac{25-23}{42}(68.5-65.5) = 65.5 + \frac{2}{42}\cdot 3 = 65.643$$
$$q_U = 68.5 + \frac{75-65}{27}(71.5-68.5) = 68.5 + \frac{10}{27}\cdot 3 = 69.61$$
$$d_q = q_U - q_L = 69.61 - 65.643 = 3.967$$

□

Definirajmo još i **koeficijent kvartilne varijacije**:

$$v_q = \frac{d_q}{q_L + q_U}$$

-koristan kada varijabilitet želimo izraziti pomoću RELATIVNE veličine (neovisne o mjernim jedinicama)

2.3.3 Uzoračka varijanca i uzoračka standardna devijacija

Uzoračka varijanca:

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

Uzoračka standardna devijacija:

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

Vrijedi:

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

Ovaj oblik formule je puno praktičniji za računanje.

Ako se u uzroku $x_1, x_2, \ldots, x_n$ vrijednosti $a_1, a_2, \ldots, a_k$ pojavljuju s frekvencijom $f_1, f_2, \ldots, f_k$, onda vrijedi:

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

Primjer 20 *Izračunajte uzoračku varijancu s^2 i uzoračku standardnu devijaciju s za podatke iz Primjera 11.*

Rješenje:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{k} f_i \cdot a_i^2 - n\bar{x}^2\right) = \frac{1}{29}\left[(1.59^2 + 1.60^2 + 1.65^2 + 1.68^2 + 2 \cdot 1.69^2\right. \\ &+ 1.70^2 + 4 \cdot 1.72^2 + 2 \cdot 1.74^2 + 2 \cdot 1.75^2 + 3 \cdot 1.76^2 + 2 \cdot 1.78^2 + 3 \cdot 1.79^2 \\ &+ \left. 1.80^2 + 1.82^2 + 1.83^2 + 1.85^2 + 2 \cdot 1.86^2 + 1.88^2) - 30 \cdot 1.75^2\right] \approx 0.0051 \\ s &= +\sqrt{s^2} = 0.071 \end{aligned}$$

□

Kakav je značaj standardne devijacije?

Izračunajmo koliki se postotak podataka nalazi u intervalu $[\bar{x} - ks, \bar{x} + ks]$ gdje je $k \in \{1, 2, 3, 4\}$.

Označimo s l_k broj podataka koji se nalzi izvan tog intervala tj.,

$$l_k := \#\{x_i : |x_i - \bar{x}| > ks\}.$$

$$\begin{aligned} (n-1)s^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{\{i:|x_i-\bar{x}|\le ks\}}(x_i - \bar{x})^2 + \sum_{\{i:|x_i-\bar{x}|>ks\}}(x_i - \bar{x})^2 \\ &\ge \sum_{\{i:|x_i-\bar{x}|>ks\}}(x_i - \bar{x})^2 > k^2 s^2 l_k \end{aligned}$$

$$\Rightarrow s^2 > k^2 s^2 \frac{l_k}{n-1} > k^2 s^2 \frac{l_k}{n} \Leftrightarrow \frac{1}{k^2} > \frac{l_k}{n} \Leftrightarrow 1 - \frac{l_k}{n} = \frac{n - l_k}{n} > 1 - \frac{1}{k^2}.$$

k	%
1	$-$
2	$\frac{3}{4} \cdot 100\% = 75\%$
3	$\frac{8}{9} \cdot 100\% = 89\%$
4	$\frac{15}{16} \cdot 100\% = 93.75\% \approx 94\%$

2.4 Mjere lokacije

Medijan, te gornji i donji kvartil spadaju u mjere lokacije. Tu su još i:

- **DECILI**: k-ti uzorački decil je broj $D_k = x_{(\frac{k(n+1)}{10})}, \quad k = 1, 2, \ldots, 9$
  (k/10 podataka je manje ili jednako njemu)
- **PERCENTILI:** k-ti uzorački percentil je broj $P_k = x_{(\frac{k(n+1)}{100})}, \quad k = 1, 2, \ldots, 99$
  (k% podataka je manje ili jednako njemu)
- decili su specijalni slučaj percentila: $D_1 = P_{10},\ D_2 = P_{20}, \ldots, D_9 = P_{90}$

Zadatak 2 *Izmjeren je kapacitet na 485 istovrsnih kondenzatora. Rezultati mjerenja su dani sljedećom tablicom frekvencija (podaci su u μF zaokruženi na dvije decimale)*

i	*razred*	f_i	$\bar{a}_i$	d_i	$f_i d_i$	$f_i d_i^2$	f_{r_i}	$F_i = \sum_{k=1}^{i} f_{r_k}$
1	$19.58 - 19.62$	3	19.60	-6	-18	108	0.006	0.006
2	$19.63 - 19.67$	5	19.65	-5	-25	125	0.010	0.016
3	$19.68 - 19.72$	5	19.70	-4	-20	80	0.010	0.026
4	$19.73 - 19.77$	20	19.75	-3	-60	180	0.041	0.067
5	$19.78 - 19.82$	35	19.80	-2	-70	140	0.072	0.139
6	$19.83 - 19.87$	74	19.85	-1	-74	74	0.153	0.292
7	$19.88 - 19.92$	***92***	19.90	0	0	0	0.190	0.482
8	$19.93 - 19.97$	83	19.95	1	83	83	0.171	0.653
9	$19.98 - 20.02$	70	20.00	2	140	280	0.144	0.797
10	$20.03 - 20.07$	54	20.05	3	162	486	0.111	0.908
11	$20.08 - 20.12$	27	20.10	4	108	432	0.056	0.964
12	$20.13 - 20.17$	12	20.15	5	60	300	0.025	0.989
13	$20.18 - 20.22$	2	20.20	6	12	72	0.004	0.993
14	$20.23 - 20.27$	3	20.25	7	21	147	0.006	0.999
	Σ	485			319	2507		

(1) Nacrtajte histogram. (DZ)

(2) Kako bi procijenili aritmetičku sredinu i varijancu uzroka?

(3) Kako bi procijenili medijan te gornji i donji kvartil?

Rješenje:

$$n = \sum_{i=1}^{14} f_i = f_1 + \ldots + f_{14} = 485$$

Budući je $n = 485$ vrlo velik broj, $\frac{1}{n-1}$ u formuli za s^2 približno je jednak $\frac{1}{n}$. Dovoljno je, dakle, uzeti:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (\bar{a}_i - \bar{x})^2 \text{ gdje je } \bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot \bar{a}_i$$

Nadalje, širina razreda je $c = 0.05$. Definirajmo:

$$d_i := \frac{\bar{a}_i - \bar{a}_0}{c} \Leftrightarrow \bar{a}_i = \bar{a}_0 + c \cdot d_i,$$

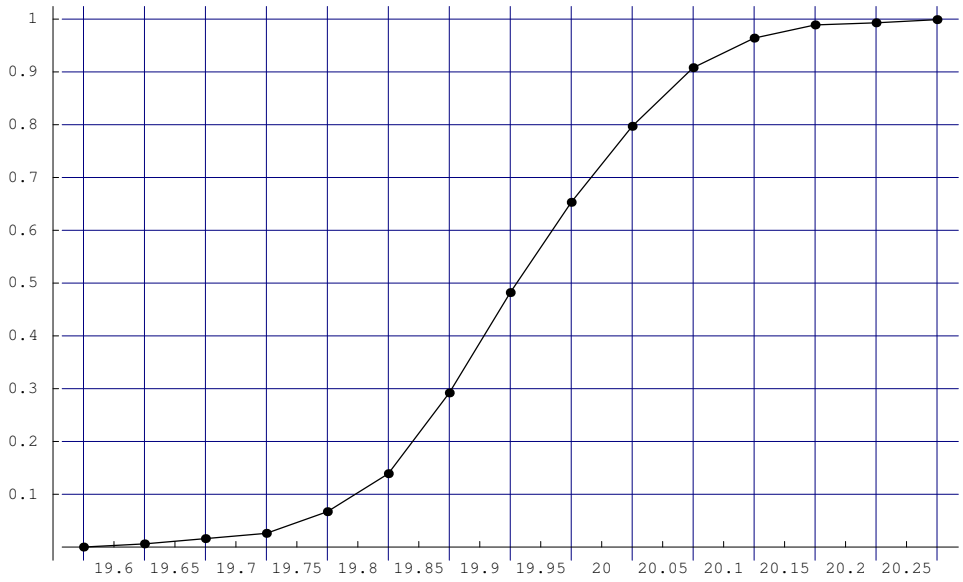
gdje je $\bar{a}_0$ referentna vrijednost aritmetičkog niza $\bar{a}_1, \ldots, \bar{a}_k$. Za $\bar{a}_0$ se obično uzima vrijednost s najvećom frekvencijom. Dakle, $\bar{a}_0$ je mod (ili jedan od). U ovom zadatku $\bar{a}_0 = 19.90$. Imamo:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{k} f_i \cdot \bar{a}_i = \frac{1}{n}\sum_{i=1}^{k} f_i(\bar{a}_0 + c \cdot d_i) = \frac{1}{n}\left(\bar{a}_0 \sum_{i=1}^{k} f_i + c\sum_{i=1}^{k} f_i \cdot d_i\right) \\
&= \bar{a}_0 + c \cdot \bar{d}, \text{ gdje je } \bar{d} = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot d_i, \\
s^2 &= \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (\bar{a}_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (\bar{a}_0 + c \cdot d_i - \bar{a}_0 - c \cdot \bar{d})^2 \\
&= \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (c(d_i - \bar{d}))^2 = c^2 \cdot \frac{1}{n}\sum_{i=1}^{k} f_i \cdot (d_i - \bar{d})^2 = \ldots = c^2\left[\frac{1}{n}\sum_{i=1}^{k} f_i d_i^2 - \bar{d}^2\right]
\end{aligned}$$

Iz podataka dobivamo da je

$$\bar{d} = \frac{319}{485} = 0.658 \Rightarrow \bar{x} = 19.90 + 0.05 \cdot 0.658 = 19.93\ \mu F$$
$$s^2 = 0.05^2\left(\frac{1}{485} \cdot 2507 - 0.658^2\right) = 0.012 \Rightarrow s = 0.11\ \mu F$$

Kod određivanja medijana, te donjeg i gornjeg kvartila pomoći će nam **graf kumulativnih relativnih frekvencija** koji je prikazan na donjoj slici.



Medijan m je x-koordinata točke $(m, 0.5)$ na grafu kumulativnih relativnih frekvencija. Ta točka leži na pravcu određenom točkama $(\bar{a}_7, F_7)$ i $(\bar{a}_8, F_8)$ pa medijan možemo izračunati linearnom interpolacijom:

$$\frac{1}{2} - F_7 = \frac{F_8 - F_7}{\bar{a}_8 - \bar{a}_7}(m - \bar{a}_7)$$
$$\frac{1}{2} - 0.482 = \frac{0.654 - 0.482}{0.05}(m - 19.925) \;\Leftrightarrow\; m = 19.93\ \mu F$$

Slično se mogu izračunati donji q_L i gornji kvartil q_U. Njima su na grafu pridružene, redom, točke $(q_L, 0.25)$ i $(q_U, 0.75)$:

$$\frac{1}{4} - F_5 = \frac{F_6 - F_5}{\bar{a}_6 - \bar{a}_5}(q_L - \bar{a}_5) \;\Leftrightarrow\; q_L = 19.84\ \mu F$$
$$\frac{3}{4} - F_8 = \frac{F_9 - F_8}{\bar{a}_9 - \bar{a}_8}(q_U - \bar{a}_8) \;\Leftrightarrow\; q_U = 19.98\ \mu F$$

□

2.5 Mjere oblika

Slično kao što se definira uzoračka varijanca, može se definirati **uzorački k-ti centralni moment**, $k \in \mathbb{N}$:

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

Specijalno,

$$\mu_1 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n-1}\sum_{n=1}^{n} x_i - \frac{n\bar{x}}{n-1} = \frac{n\bar{x}}{n-1} - \frac{n\bar{x}}{n-1} = 0$$
$$\mu_2 = s^2$$
$$\mu_3 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Primjer 21 *Promatrajmo uzorak:* $1, 2, 4, 5$. *Srednja vrijednost tog uzorka je* $\bar{x} = \frac{1}{4}(1+2+4+5) = 3$.
S druge strane, 3.centalni moment tog uzorka je

$$\mu_3 = \frac{1}{3}\left((1-3)^3 + (2-3)^3 + (4-3)^3 + (5-3)^3\right) = 0$$

Odavde možemo zaključiti da kada je uzorak simetričan s obzirom na aritmetičku sredinu, 3.centalni moment $\mu_3 = 0$.

Koeficijent asimetrije uzorka (skewness) definiran je s:

$$\alpha_3 = \frac{\mu_3}{s^3} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{1}{n-1}\sum_{i=1}^{k} f_i \cdot \left(\frac{a_i - \bar{x}}{s}\right)^3$$

Vrijedi:

(i) $\alpha_3 = 0 \Rightarrow$ uzorak je SIMETRIČAN

(ii) $\alpha_3 > 0 \Rightarrow$ uzorak je POZITIVNO ASIMETRIČAN

(iii) $\alpha_3 < 0 \Rightarrow$ uzorak je NEGATIVNO ASIMETRIČAN

2.6 Linearna regresija

Imamo n parova podataka $(x_i, y_i),\ i = 1, \ldots, n$. Želimo odrediti vezu između nezavisne varijable x (nju možemo kontrolirati) i zavisne varijable y. Pretpostavimo da je veza **linearna**, tj. da je graf pripadajuće funkcije **pravac** $y = \beta x + \alpha$. Želimo odrediti procjenitelj za taj pravac oblika

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

Pravac određujemo **metodom najmanjih kvadrata**, tj. želimo minimizirati funkciju

$$L(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

po nepoznatim parametrima pravca α i β. Tražene vrijednosti su rješenja sustava jednadžbi:

$$\frac{\partial L}{\partial \alpha} = 0, \frac{\partial L}{\partial \beta} = 0$$

$$\Leftrightarrow -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0, \quad -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0$$

$$n \cdot \alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Uvodimo oznake

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

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

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

Uz ove oznake sustav jednadžbi se može napisati na sljedeći način:

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i / : n \Rightarrow \alpha + \beta\bar{x} = \bar{y}$$

$$\alpha \cdot n\bar{x} + \beta((n-1)s_x^2 + n\bar{x}^2) = (n-1)s_{xy} + n\bar{x}\bar{y}$$

$$\Rightarrow (n-1)(\beta s_x^2 - s_{xy}) = n\bar{x}(\bar{y} - \alpha\bar{x} - \beta) = 0 \quad (\text{ zbog } \alpha + \beta\bar{x} = \bar{y})$$

Dakle,

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

Zadatak 3 *U tablici su dani podaci o broju emitiranih reklama tijekom mjesec dana za neki proizvod i ostvarenoj zaradi na tom proizvodu. Procijenite pravac regresije za ove podatke.*

broj reklama	16	59	65	43	82	90	31	22
promet (u tisućama kuna)	18	63	28	71	85	98	20	25

Rješenje:

$$n = 8$$

$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{1}{8}(16 + 59 + 65 + 43 + 82 + 90 + 31 + 22) = 51$$

$$\bar{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{1}{8}(18 + 63 + 28 + 71 + 85 + 98 + 20 + 25) = 51$$

$$\sum_{i=1}^{8} x_i^2 = 16^2 + 59^2 + 65^2 + 43^2 + 82^2 + 90^2 + 31^2 + 22^2 = 26080$$

$$\Rightarrow s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{7}\left(26080 - 8 \cdot 51^2\right) = 753.143$$

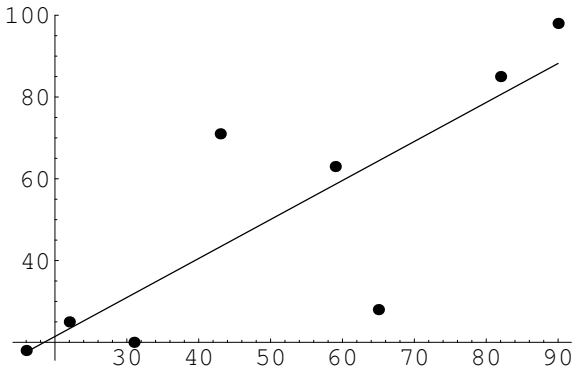
$$\sum_{i=1}^{8} x_i y_i = 16 \cdot 18 + 59 \cdot 63 + 65 \cdot 28 + 43 \cdot 71 + 82 \cdot 85 + 90 \cdot 98 + 31 \cdot 20 + 22 \cdot 25 = 25838$$

$$\Rightarrow s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) = \frac{1}{7}\left(25838 - 8 \cdot 51 \cdot 51\right) = 718.571$$

$$\Longrightarrow \hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{718.571}{753.143} = 0.954$$

$$\Longrightarrow \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 51 - 0.954 \cdot 51 = 2.346$$

$$\Longrightarrow y = 0.954 \cdot x + 2.346$$



□

2.7 (Pearsonov) koeficijent korelacije

Isto kao i u prethodnom poglavlju, pretpostavljamo da imamo n parova podataka $(x_i, y_i),\ i = 1, \ldots, n$. Cilj nam je izmjeriti "povezanost" veličine x i y pomoću tih podataka.

Standardizirajmo:

$$\frac{x_1 - \bar{x}}{s_x}, \frac{x_2 - \bar{x}}{s_x}, \ldots, \frac{x_n - \bar{x}}{s_x}, \quad \frac{y_1 - \bar{y}}{s_y}, \frac{y_2 - \bar{y}}{s_y}, \ldots, \frac{y_n - \bar{y}}{s_y}.$$

Na taj način početni niz dvodimenzionalnih podataka prelazi u dvodimenzionalni niz podataka

$$\left(\frac{x_i - \bar{x}}{s_x}, \frac{y_i - \bar{y}}{s_y} \right), i = 1, \ldots, n.$$

Standardiziranjem smo postigli centriranje niza (eliminacija utjecaja srednje vrijednosti) i normalizaciju niza (eliminacija mjernih jedinica i svođenja raspršenja na jediničnu standradnu devijaciju). Tada povezanost od x i y mjerimo pomoću **Pearsonovog koeficijenta korelacije**:

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

Primjetimo da ako je $y_i = \alpha + \beta x_i$ za sve $i = 1, \ldots, n$ (tj. ako su x i y u

egzaktnoj linearnoj vezi), da je

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\cdot\frac{\alpha+\beta x_i-\alpha-\beta\bar{x}}{|\beta|s_x}\right) = \frac{\beta}{|\beta|}\frac{1}{s_x^2}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{\beta}{|\beta|}.$$

Dakle,

1. $r = 0$ nema korelacije
2. $r = 1(> 0)$ pozitivna korelacija (ako x raste, i y u pravilu raste)
3. $r = -1(< 0)$ negativna korelacija (ako x raste, y u pravilu pada)

Primjer 22 *Nađite koeficijent korelacije za podatke iz Zadatka 3.*

Rješenje:

$$s_{xy} = 718.571$$
$$s_x^2 = 753.143 \;\Rightarrow\; s_x = 27.4435$$
$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) = \frac{1}{7}\left(27972 - 8\cdot 51^2\right) = 1023.43 \;\Rightarrow\; s_y = 31.9911$$
$$\Longrightarrow\; r = \frac{s_{xy}}{s_x\cdot s_y} = \frac{718.571}{27.4435\cdot 31.9911} = 0.818467$$

što je razmjerno visoka korelacija. □