

IB112 Základy matematiky

Základy popisné statistiky

Jan Strejček

- Popisuje a sumarizuje informace obsažené ve velkém množství dat pomocí tabulek, grafů a číselných charakteristik.
- Cílem je zpřehlednit informace skryté v datech.
- Některé pojmy popisné statistiky motivují pojmy pravděpodobnosti:
 - relativní četnost motivuje pravděpodobnost,
 - aritmetický průměr motivuje střední hodnotu atd.

- *Základní pojmy*
 - základní soubor a výběr
 - znak
 - datový soubor
- *Jednorozměrný datový soubor*
 - variační obor, rozpětí
 - bodové rozložení četností
 - intervalové rozložení četností
- *Číselné charakteristiky znaků*
 - průměr, modus, medián, kvartily, kvartilový graf
 - rozptyl, směrodatná odchylka
- *Dvourozměrný datový soubor*
 - korelace

Základní pojmy

Definice

*Základní soubor je neprázdná konečná množina E . Prvky množiny E nazýváme **objekty**. Libovolnou neprázdnou podmnožinou množiny E nazýváme **výběr**. Počet prvků výběru nazýváme **rozsah** výběru.*

- Základní soubor se někdy nazývá *populace*.
- Základní soubory bývají pro bližší zkoumání příliš početné (např. všichni občané ČR). Proto se detailně zkumá pouze menší skupina objektů nazývaná výběr.
- Výběr se často získá náhodným postupem. Např. při marketingovém průzkumu se náhodně vyberou telefonní čísla, na která se zavolá.

- Vlastnosti objektů vyjadřujeme číselně pomocí tzv. *znaků*.
- 1 vlastnost = 1 znak

Definice (Znak)

Znakem objektu rozumíme funkci $X : E \rightarrow \mathbb{R}$.

- Objekty mohou mít více znaků.
- Například u lidí můžeme zkoumat znaky jako výška, hmotnost, věk, IQ,...
- Hodnoty znaků pro prvky výběru můžeme reprezentovat datovým souborem.

Definice

(k -rozměrným) *datovým souborem* s rozsahem n rozumíme matici

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}.$$

Řádky matice odpovídají objektům výběru, sloupce jednotlivým znakům.

- Je-li matice jednosloupcová, hovoříme o *jednorozměrném datovém souboru*. Je-li matice dvousloupcová, hovoříme o *dvourozměrném datovém souboru*.
- Hodnoty, kterých znak nabývá, se také nazývají *varianty* nebo *úrovně*.

Příklad datových souborů

Vlevo je trojrozměrný datový soubor s rozsahem 12 popisující výšku a váhu objektů a jejich známku z matematiky. Vpravo pak je jednorozměrný datový soubor popisující pouze výšku objektů.

$$\begin{pmatrix} 161 & 51 & 1 \\ 188 & 82 & 3 \\ 170 & 70 & 2 \\ 174 & 59 & 4 \\ 182 & 95 & 2 \\ 152 & 44 & 3 \\ 193 & 102 & 4 \\ 177 & 73 & 2 \\ 174 & 63 & 1 \\ 188 & 74 & 3 \\ 167 & 61 & 2 \\ 173 & 63 & 2 \end{pmatrix}$$
$$\begin{pmatrix} 161 \\ 188 \\ 170 \\ 174 \\ 182 \\ 152 \\ 193 \\ 177 \\ 174 \\ 188 \\ 167 \\ 173 \end{pmatrix}$$

Jednorozměrný datový soubor

Definice

Jestliže uspořádáme hodnoty jednorozměrného datového souboru do neklesající posloupnosti, získáme jednorozměrný **uspořádaný datový soubor** obvykle značený

$$\begin{pmatrix} x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(n)} \end{pmatrix}, \text{ kde } x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Interval $\langle x_{(1)}, x_{(n)} \rangle$ pak nazveme **variační obor**.

Délka variačního oboru $x_{(n)} - x_{(1)}$ se nazývá **rozpětí** datového souboru.

Definice

Nechť je dán jednorozměrný datový soubor. **Vektor variant** je rostoucí posloupnost všech variant vyskytujících se v souboru, obvykle značená

$$\begin{pmatrix} x_{[1]} \\ x_{[2]} \\ \vdots \\ x_{[r]} \end{pmatrix}, \text{ kde } x_{[1]} < x_{[2]} < \dots < x_{[r]}.$$

Příklad

Z uvedeného datového souboru vytvořte uspořádaný datový soubor, určete variační obor, rozpětí datového souboru a vektor variant.

$$\begin{pmatrix} 161 \\ 188 \\ 170 \\ 174 \\ 182 \\ 152 \\ 193 \\ 177 \\ 174 \\ 188 \\ 167 \\ 173 \end{pmatrix}$$

Příklad

Z uvedeného datového souboru vytvořte uspořádaný datový soubor, určete variační obor, rozpětí datového souboru a vektor variant.

161	152
188	161
170	167
174	170
182	173
152	174
193	174
177	177
174	182
188	188
167	188
173	193

← uspořádaný datový soubor

Příklad

Z uvedeného datového souboru vytvořte uspořádaný datový soubor, určete variační obor, rozpětí datového souboru a vektor variant.

161	152
188	161
170	167
174	170
182	173
152	174
193	174
177	177
174	182
188	188
167	188
173	193

← uspořádaný datový soubor

- Variační obor je $\langle 152, 193 \rangle$.
- Rozpětí je 41.

Příklad

Z uvedeného datového souboru vytvořte uspořádaný datový soubor, určete variační obor, rozpětí datového souboru a vektor variant.

$$\begin{pmatrix} 161 \\ 188 \\ 170 \\ 174 \\ 182 \\ 152 \\ 193 \\ 177 \\ 174 \\ 188 \\ 167 \\ 173 \end{pmatrix}$$
$$\begin{pmatrix} 152 \\ 161 \\ 167 \\ 170 \\ 173 \\ 174 \\ 174 \\ 177 \\ 182 \\ 188 \\ 188 \\ 193 \end{pmatrix}$$

← uspořádaný datový soubor

- Variační obor je $\langle 152, 193 \rangle$.
- Rozpětí je 41.

vektor variant →

$$\begin{pmatrix} 152 \\ 161 \\ 167 \\ 170 \\ 173 \\ 174 \\ 177 \\ 182 \\ 188 \\ 188 \\ 193 \end{pmatrix}$$

Bodové rozložení četností

Jestliže je počet variant v jednorozměrném datovém souboru malý, přiřazujeme četnosti jednotlivým variantám. Hovoříme o *bodovém rozložení četností*.

Definice

Nechť je dán jednorozměrný datový soubor $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, v kterém znak X

nabývá r variant. Pak pro každé $j \in \{1, \dots, r\}$ definujeme

n_j - *absolutní četnost* j -té varianty jako počet výskytů $x_{[j]}$ v datovém souboru,

$p_j = \frac{n_j}{n}$ - *relativní četnost* j -té varianty,

$N_j = n_1 + \dots + n_j$ - *absolutní kumulativní četnost* prvních j variant,

$F_j = p_1 + \dots + p_j$ - *relativní kumulativní četnost* prvních j variant.

Tabulka rozložení četností

Definice (Tabulka rozložení četností)

Nechť je dán jednorozměrný datový soubor s r variantami. **Tabulka rozložení četností** nebo též **variační řada** je tabulka následujícího tvaru:

$x_{[j]}$	n_j	p_j	N_j	F_j
$x_{[1]}$	n_1	p_1	N_1	F_1
$x_{[2]}$	n_2	p_2	N_2	F_2
\vdots	\vdots	\vdots	\vdots	\vdots
$x_{[r]}$	n_r	p_r	N_r	F_r

- První dva sloupce tabulky tvoří tzv. **četnostní tabulku**, kterou lze použít ke stručnějšímu zadání jednorozměrného datového souboru.

Příklad

Je dán jednorozměrný datový soubor s rozsahem 12 obsahující známky z matematiky. Sestavte tabulku rozložení četností.

(
1
3
2
4
2
3
4
2
1
3
2
2
)

Četnostní tabulka zadávající stejný soubor:

$x_{[j]}$	n_j
1	2
2	5
3	3
4	2

Příklad

Je dán jednorozměrný datový soubor s rozsahem 12 obsahující známky z matematiky. Sestavte tabulku rozložení četností.

(1
3
2
4
2
3
4
2
1
3
2
2)

Četnostní tabulka zadávající stejný soubor:

$x_{[j]}$	n_j
1	2
2	5
3	3
4	2

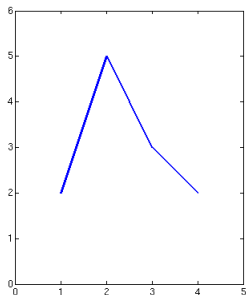
Řešení:

$x_{[j]}$	n_j	p_j	N_j	F_j
1	2	$\frac{2}{12}$	2	$\frac{2}{12}$
2	5	$\frac{5}{12}$	7	$\frac{7}{12}$
3	3	$\frac{3}{12}$	10	$\frac{10}{12}$
4	2	$\frac{2}{12}$	12	$\frac{12}{12} = 1$

Grafická znázornění jednorozměrného bodového rozdělení četností

- *Polygon četností* neboli *spojnicový graf* je lomená čára spojující body se souřadnicemi $(x_{[j]}, n_j)$ pro všechna j .

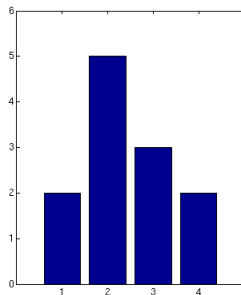
$x_{[j]}$	n_j
1	2
2	5
3	3
4	2



Grafická znázornění jednorozměrného bodového rozdělení četností

- *Sloupcový diagram* je soustava na sebe nenavazujících obdélníků posazených na x -ovou osu, jejichž svislá osa je na nějaké variantě $x_{[j]}$ a výška odpovídá absolutní četnosti n_j .

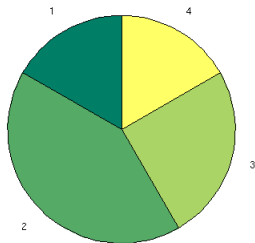
$x_{[j]}$	n_j
1	2
2	5
3	3
4	2



Grafická znázornění jednorozměrného bodového rozdělení četností

- **Výsečový graf** je kruh rozdělený na výseče tak, že poměr obvodu výseče pro variantu $x_{[j]}$ k obvodu kruhu je roven relativní četnosti p_j .

$x_{[j]}$	n_j
1	2
2	5
3	3
4	2

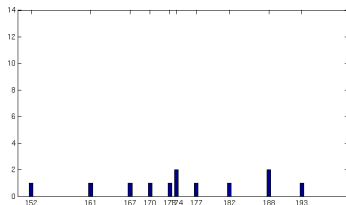


Grafická znázornění jednorozměrného bodového rozdělení četností

- Polygon četností a sloupcový diagram se používají i pro znázornění absolutních kumulativních četností.
- V některých případech může být takový graf názornější.

Příklad:

Sloupcový diagram zobrazující bodové rozdělení četností našeho souboru s výškami jedinců příliš informací nepřináší.

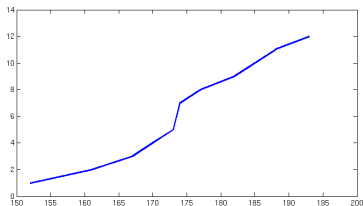


Grafická znázornění jednorozměrného bodového rozdělení četností

- Polygon četností a sloupcový diagram se používají i pro znázornění absolutních kumulativních četností.
- V některých případech může být takový graf názornější.

Příklad:

Polygon absolutních kumulativních četností je v tomto případě přínosnější.



Roztříděný datový soubor

Jestliže je počet variant v jednorozměrném datovém souboru blízký rozsahu souboru, pak variační obor pokryjeme systémem disjunktních intervalů. Hodnoty znaku pak nahradíme příslušností do intervalu. Hovoříme o *roztříděném datovém souboru*.

- Číselnou osu rozdělíme na intervaly

$$(-\infty, u_1), (u_1, u_2), (u_2, u_3), \dots, (u_r, u_{r+1}), (u_{r+1}, \infty)$$

tak, aby krajní intervaly neobsahovaly žádnou pozorovanou hodnotu (s těmito intervaly dále nepracujeme).

- (u_j, u_{j+1}) nazveme *j-tý třídící interval*, kde $j \in \{1, \dots, r\}$.
- $d_j = u_{j+1} - u_j$ je *délka j-tého intervalu*.
- $x_{[j]} = \frac{u_j + u_{j+1}}{2}$ je *střed j-tého intervalu*.
- Počet tříd volíme podle *Sturgesova pravidla* $r = 1 + 3,3 \cdot \log_{10} n$ (je to pouze doporučení, pro $n < 500$ často volíme vyšší r)
- Intervaly obvykle volíme tak, aby měly stejnou délku.

Intervalové rozložení četností

Hodnoty z jednorozměrného datového souboru pak rozdělíme do r zvolených intervalů. *Intervalové rozložení četností* pak definujeme podobně jako bodové rozložení četností.

Definice

Nechť je dán jednorozměrný datový soubor a r třídících intervalů. Pak pro každé $j \in \{1, \dots, r\}$ definujeme

n_j - *absolutní četnost* j -tého třídícího intervalu počet prvků datového souboru padajících do intervalu (u_j, u_{j+1}) ,

$p_j = \frac{n_j}{n}$ - *relativní četnost* j -tého třídícího intervalu,

$f_j = \frac{p_j}{d_j}$ - *četnostní hustota* j -tého třídícího intervalu,

$N_j = n_1 + \dots + n_j$ - *absolutní kumulativní četnost* prvních j třídících intervalů,

$F_j = p_1 + \dots + p_j$ - *relativní kumulativní četnost* prvních j třídících intervalů.

Tabulka rozložení četností

Definice (Tabulka rozložení četností)

Nechť je dán jednorozměrný datový soubor s r variantami. **Tabulka rozložení četností** je tabulka následujícího tvaru:

(u_j, u_{j+1})	d_j	n_j	p_j	f_j	N_j	F_j
(u_1, u_2)	d_1	n_1	p_1	f_1	N_1	F_1
(u_2, u_3)	d_2	n_2	p_2	f_2	N_2	F_2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
(u_r, u_{r+1})	d_r	n_r	p_r	f_r	N_r	F_r

- První a třetí sloupec tabulky tvoří tzv. **četnostní tabulku**, která popisuje roztríděný datový soubor.

Příklad

Je dán uspořádaný datový soubor popisující výšku osob. Četností tabulkou zapište roztříděný datový soubor (počet tříd určete Sturgesovým pravidlem) a sestavte tabulku rozložení četností.

(
152
161
167
170
173
174
174
177
182
188
188
193
)

Příklad

Je dán uspořádaný datový soubor popisující výšku osob. Četností tabulkou zapište roztříděný datový soubor (počet tříd určete Sturgesovým pravidlem) a sestavte tabulku rozložení četností.

(152)
161
167
170
173
174
174
177
182
188
188
193)

Řešení:

- Rozsah je $n = 12$. Počet tříd tedy bude $r = 5$.

Příklad

Je dán uspořádaný datový soubor popisující výšku osob. Četností tabulkou zapište roztříděný datový soubor (počet tříd určete Sturgesovým pravidlem) a sestavte tabulku rozložení četností.

(152)
161
167
170
173
174
174
177
182
188
188
193)

Řešení:

- Rozsah je $n = 12$. Počet tříd tedy bude $r = 5$.
- Volíme stejnou délku tříd $d_j = 10$. Zvolíme tedy třídy $(150, 160)$, $(160, 170)$, $(170, 180)$, $(180, 190)$, $(190, 200)$.
- Četnostní tabulka vypadá následovně:

(u_j, u_{j+1})	n_j
$(150, 160)$	1
$(160, 170)$	3
$(170, 180)$	4
$(180, 190)$	3
$(190, 200)$	1

Z četnostní tabulky už snadno spočítáme tabulku rozložení četností:

(u_j, u_{j+1})	n_j
(150, 160)	1
(160, 170)	3
(170, 180)	4
(180, 190)	3
(190, 200)	1

Z četnostní tabulky už snadno spočítáme tabulku rozložení četností:

(u_j, u_{j+1})	n_j
$(150, 160)$	1
$(160, 170)$	3
$(170, 180)$	4
$(180, 190)$	3
$(190, 200)$	1

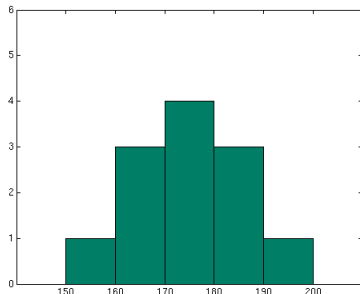
(u_j, u_{j+1})	d_j	n_j	p_j	f_j	N_j	F_j
$(150, 160)$	10	1	$\frac{1}{12}$	$\frac{1}{120}$	1	$\frac{1}{12}$
$(160, 170)$	10	3	$\frac{3}{12}$	$\frac{3}{120}$	4	$\frac{4}{12}$
$(170, 180)$	10	4	$\frac{4}{12}$	$\frac{4}{120}$	8	$\frac{8}{12}$
$(180, 190)$	10	3	$\frac{3}{12}$	$\frac{3}{120}$	11	$\frac{11}{12}$
$(190, 200)$	10	1	$\frac{1}{12}$	$\frac{1}{120}$	12	$\frac{12}{12}$

Grafická znázornění jednorozměrného intervalového rozdělení četností

- Používáme podobné grafy jako u bodového rozdělení četností.
- V případě polygonu četností místo konkrétních hodnot použijeme středy intervalů.
- Místo sloupcového grafu používáme *histogram*, což je v vlastně totéž, akorát bez mezer mezi sloupci (šířka sloupce odpovídá délce intervalu).
- Lze použít i výsečový graf.
- Polygon četností a histogram lze použít i pro znázornění absolutních kumulativních četností.

Histogram pro roztríděný datový soubor z předchozího příkladu.

(u_j, u_{j+1})	n_j
$(150, 160)$	1
$(160, 170)$	3
$(170, 180)$	4
$(180, 190)$	3
$(190, 200)$	1



Číselné charakteristiky znaků

Všechny charakteristiky definujeme pro neroztříděné datové soubory. Varianty pro roztříděné soubory lze dohledat.

Aritmetický průměr

$$\blacksquare m = \frac{1}{n} \sum_{i=1}^n x_i$$

- Aritmetický průměr je citlivý na extrémně odchýlené hodnoty.

Modus

- Je to hodnota, která se v datovém souboru vyskytuje nejčastěji.
- Není určeno jednoznačně (více hodnot může mít maximální absolutní četnost).

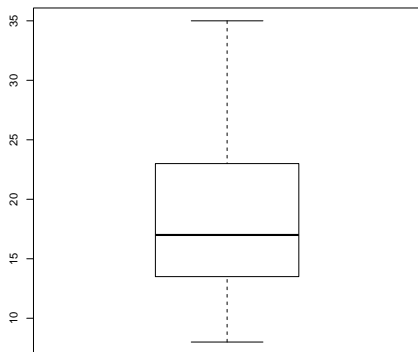
α -kvantil, medián, horní a dolní kvartil

- Nechť $\alpha \in (0, 1)$. α -kvantil je číslo x_α , které rozděljuje uspořádaný datový soubor na dolní úsek obsahující podíl α ze všech dat a na horní úsek obsahující podíl $1 - \alpha$ ze všech dat.
- Pokud $n \cdot \alpha$ je celé číslo c , pak $x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2}$.
- Pokud $n \cdot \alpha$ není celé, vezmeme nejbližší větší celé číslo c a $x_\alpha = x_{(c)}$.
- **Medián** je $x_{0,5}$, tedy rozděljuje soubor na horní a dolní polovinu. Oproti aritmetickému průměru není citlivý na extrémně odchýlené hodnoty.
- **Dolní kvartil** je $x_{0,25}$, tedy rozděljuje soubor na dolní čtvrtinu a horní tři čtvrtiny.
- **Horní kvartil** je $x_{0,75}$, tedy rozděljuje soubor na dolní tři čtvrtiny a horní čtvrtinu.

Kvartilový graf (boxplot)

Grafické znázornění pěti charakteristik polohy:

- největší hodnota
- horní kvartil
- medián
- dolní kvartil
- nejmenší hodnota



Rozptyl

- Čím větší rozptyl, tím větší proměnlivost souboru.
- Průměrná kvadratická odchylka od aritmetického průměru.

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

Směrodatná odchylka

- Čím větší směrodatná odchylka, tím větší proměnlivost souboru.
- $s = \sqrt{s^2}$

Příklad

Je dán uspořádaný datový soubor popisující výšku osob. Spočítejte zmíněné číselné charakteristiky a nakreslete kvartilový graf.

(
152
161
167
170
173
174
174
177
182
188
188
193
)

Je dán uspořádaný datový soubor popisující výšku osob. Spočítejte zmíněné číselné charakteristiky a nakreslete kvartilový graf.

(152)
161
167
170
173
174
174
177
182
188
188
193)

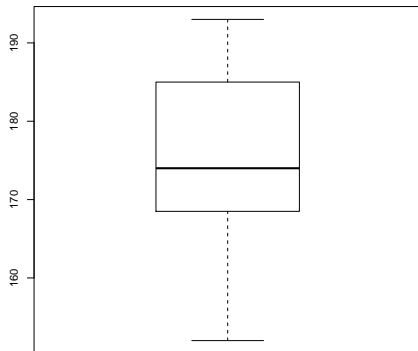
Řešení:

- Aritmetický průměr je $m = \frac{2099}{12} = 174,916\bar{6}$.
- Modus může být 188 nebo 174.
- Medián je $x_{0,5} = \frac{x_{(6)}+x_{(7)}}{2} = 174$.
- Dolní kvartil je $x_{0,25} = \frac{x_{(3)}+x_{(4)}}{2} = 168,5$.
- Horní kvartil je $x_{0,75} = \frac{x_{(9)}+x_{(10)}}{2} = 185$.
- Rozptyl s^2 je přibližně 127,9.
- Směrodatná odchylka je s je přibližně 11,3.

Příklad

Je dán uspořádaný datový soubor popisující výšku osob. Spočítejte zmíněné číselné charakteristiky a nakreslete kvartilový graf.

(152)
161
167
170
173
174
174
177
182
188
188
193)



Příklad

Je dán uspořádaný datový soubor popisující výšku osob. Spočítejte zmíněné číselné charakteristiky a nakreslete kvartilový graf.

102

161

167

170

173

174

174

177

182

188

188

193

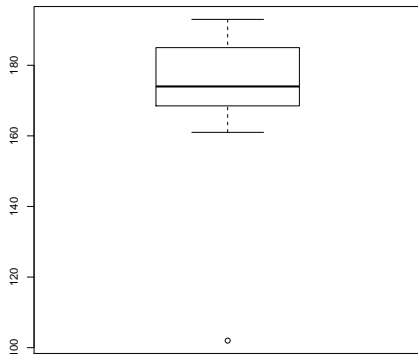
Řešení:

- Aritmetický průměr je $m = \frac{2049}{12} = 170,75$.
- Modus může být 188 nebo 174.
- Medián je $x_{0,5} = \frac{x_{(6)} + x_{(7)}}{2} = 174$.
- Dolní kvartil je $x_{0,25} = \frac{x_{(3)} + x_{(4)}}{2} = 168,5$.
- Horní kvartil je $x_{0,75} = \frac{x_{(9)} + x_{(10)}}{2} = 185$.
- Rozptyl s^2 je přibližně 509,9.
- Směrodatná odchylka je s je přibližně 22,6.

Příklad

Je dán uspořádaný datový soubor popisující výšku osob. Spočítejte zmíněné číselné charakteristiky a nakreslete kvartilový graf.

(102
161
167
170
173
174
174
177
182
188
188
193)



Dvourozměrný datový soubor

- Připomeňme, že dvourozměřný datový soubor je matice tvaru

$$\begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}.$$

- Pro dvourozměřný soubor můžeme definovat analogické pojmy k většině pojmů zavedených pro jednorozměřný prostor, namátkou:
 - rozříděný dvourozměřný datový soubor
 - středy tříd, absolutní (kumulativní) četnost, relativní (kumulativní) četnost
 - četnostní tabulky, ...
- Ke grafickému znázornění dvourozměřných datových souborů lze použít *rozptylový graf* (nerozříděný soubor), případně *stereogram* (dvourozměřný histogram, pro rozříděné soubory).

- Zajímavá číselná charakteristika dvourozměrného souboru.
- Udává míru lineární závislosti znaků X a Y .
- Předpokládáme, že $s(x)$, $s(y)$ jsou nenulové směrodatné odchylky znaků X a Y a m_x , m_y jejich aritmetické průměry.
- Pak koeficient korelace je

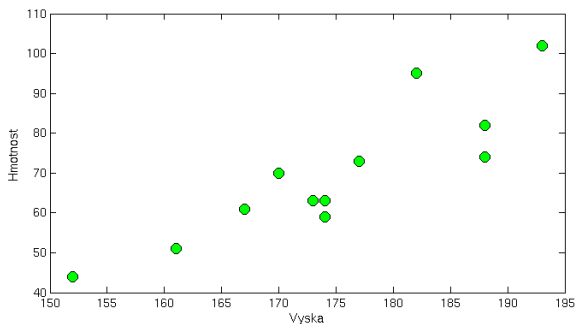
$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - m_x)(y_i - m_y)}{s(x)s(y)}.$$

- Vždy platí $-1 \leq r \leq 1$, přičemž r nabývá krajních hodnot, pokud jsou znaky zcela lineárně závislé, tj. pokud $y_i = ax_i + b$. Je-li $r = 1$, tak body (x_i, y_i) leží na rostoucí přímce, pro $r = -1$ leží na klesající přímce. Hodnoty r blízké 0 vyjadřují, že závislost X , Y není lineární, případně jsou X , Y nezávislé.

Příklad

Uvažme dvojrozměrný datový soubor výšek a hmotností.
Rozptylový graf vypadá následovně.

161	51
188	82
170	70
174	59
182	95
152	44
193	102
177	73
174	63
188	74
167	61
173	63

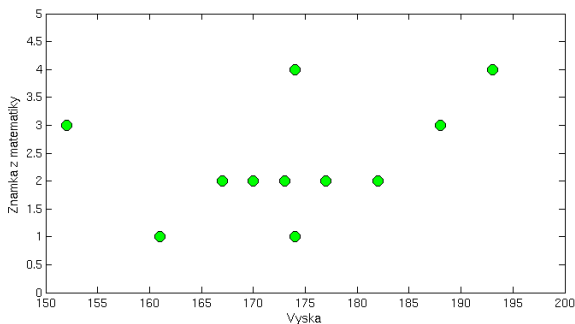


Koeficient korelace je 0,8784.

Příklad

Uvažme dvojrozměrný datový soubor výšek a známek z matematiky. Rozptylový graf vypadá následovně.

161	1
188	3
170	2
174	4
182	2
152	3
193	4
177	2
174	1
188	3
167	2
173	2



Koeficient korelace je 0,4049.