Dvourozměrný statistický soubor
Korelační a regresní analýza

P1ZST-11a-2026

Ukázka dvourozměrného statistického souboru

Statistická jednotka	Znak X (Výška v cm)	Znak Y (Hmotnost v kg)
1	170	65
2	165	70
3	180	80
4	175	75
5	160	60

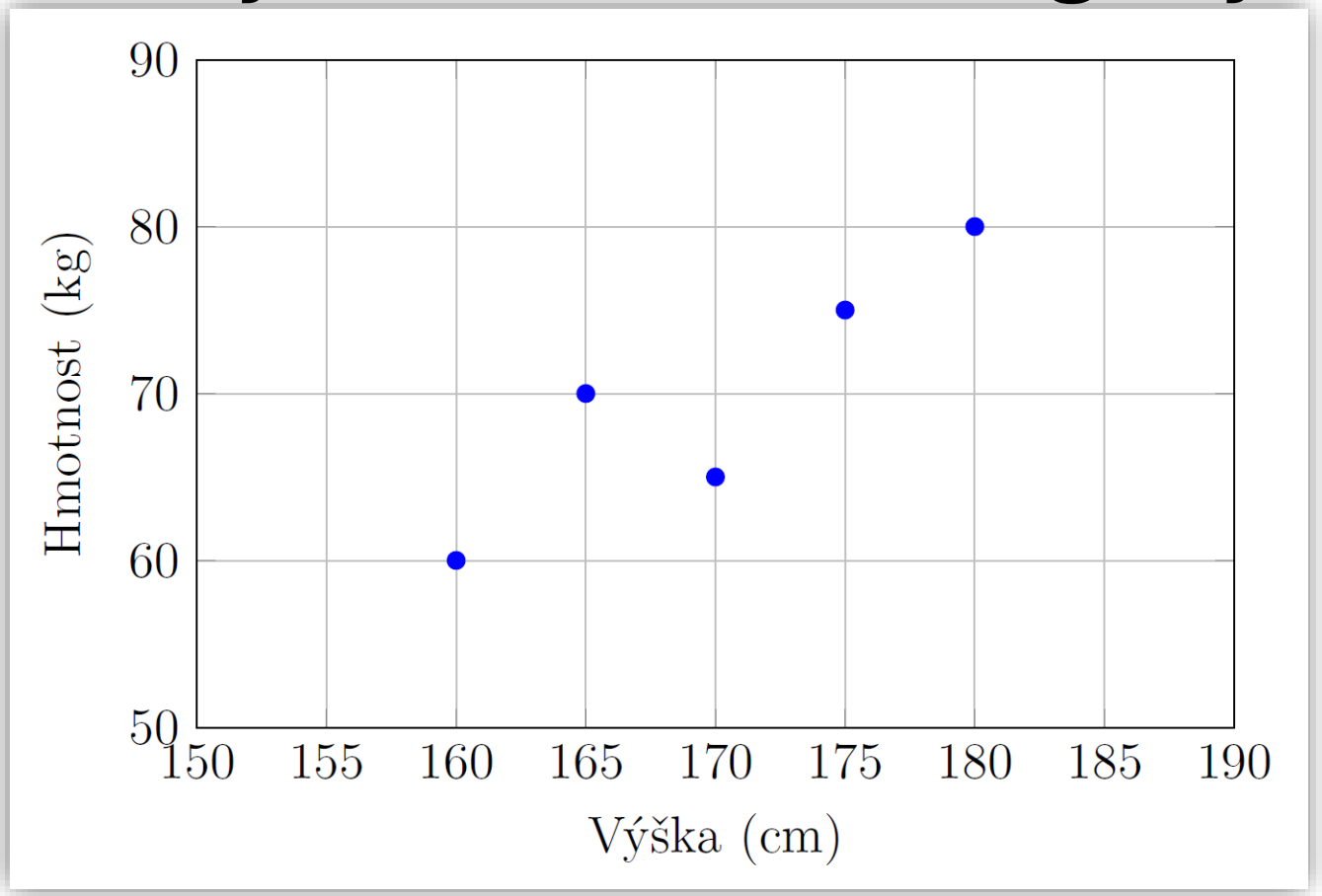
- **Dvojice hodnot**: Každá statistická jednotka má přiřazenou dvojici hodnot (x_i, y_i), kde x_i je hodnota znaku X a y_i je hodnota znaku Y pro i-tou statistickou jednotku.

Tabulkové a grafické zobrazení dvourozměrných dat – Kontingenční tabulky

	Y_1	Y_2	Y_3
X_1	5	7	3
X_2	8	12	4
X_3	6	2	9

- **Řádky tabulky** představují jednotlivé kategorie znaku X.
- **Sloupce tabulky** představují jednotlivé kategorie znaku Y.
- **Buňky tabulky** obsahují absolutní četnosti kombinací hodnot X a Y.

Tabulkové a grafické zobrazení dvourozměrných dat – Bodové grafy



Každý bod v grafu představuje jednu statistickou jednotku a její hodnoty znaků X a Y .

Míry polohy – Aritmetický průměr

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Míry variability a kovariance

- Rozptyl $$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

- Směrodatná odchylka $$s_X = \sqrt{s_X^2}, \quad s_Y = \sqrt{s_Y^2}.$$

- Kovariance

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Vypočítejte základní číselné charakteristiky dvourozměrného statistického souboru.

x	27	31	87	93	114	124	190	193	250	254	264	272	308	324
y	28	21	71	36	30	43	54	54	59	25	82	22	38	22

371	372	440	442	502	503	506	522	556	620	624
56	63	46	24	33	40	41	28	53	38	66

Vypočítejte základní číselné charakteristiky dvourozměrného statistického souboru.

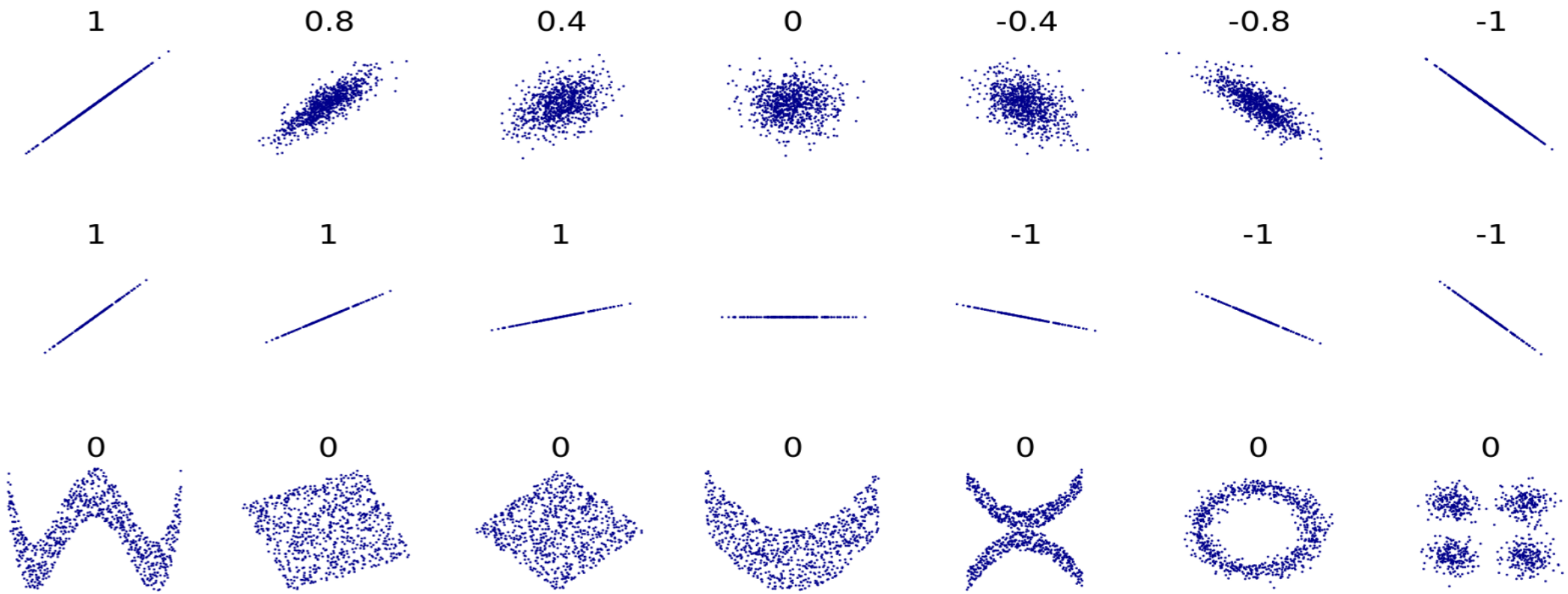
$x \backslash y$	20	30	40	50	60	70	80
250	19	5					
350	23	116	11				
450	1	41	98	9			
550		4	32	65	7		
650		1	4	21	46	3	
750			1	2	11	13	1
850					1	3	2

Výpočet (Pearsonova) korelačního koeficientu

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y} = \frac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \cdot \sum(y_i - \overline{y})^2}}$$

Excel: Pomocí funkce `CORREL(array1, array2)` lze získat stejný výsledek.

Vizualizace (ne)lineární závislosti



https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/1920px-Correlation_examples2.svg.png

- Interpretace hodnot Pearsonova korelačního koeficientu
- 0-0,19: mezi znaky X a Y není lineární vztah
- 0,20-0,39: mezi X a Y je slabý pozitivní lineární vztah
- 0,40-0,59: mezi X a Y je středně silný pozitivní lineární vztah
- 0,60-0,79: mezi X a Y je silný pozitivní lineární vztah
- 0,80-1: mezi X a Y je velmi silný pozitivní lineární vztah
- analogicky pro záporné hodnoty

Zde jsou data pro prodej dvou produktů v různých týdnech. Určete, zda mezi prodejem těchto produktů existuje lineární vztah.

Prodeje produktu A	100	105	110	95	115	90	120	85	125	80
Prodeje produktu B	200	180	205	185	190	185	190	195	200	190

- Znázorněte data pomocí bodového grafu
- Pokud graf naznačuje lineární vztah, vypočtěte korelační koeficient.

Nástrahy korelačního koeficientu

- Prozkoumejte listy Korelace1 – Korelace3

Nástrahy korelačního koeficientu

- Prozkoumejte listy Korelace1 – Korelace3

Data o erupcích gejzíru Old Faithful v Yellowstonském národním parku.

- Soubor má 272 pozorování a pouze dvě číselné proměnné:
  - eruptions (Délka erupce): Jak dlouho gejzír stříkal vodu (v minutách).
  - waiting (Čekání): Jak dlouho se čekalo na tuto erupci od té předchozí (také v minutách).

Klíčové vlastnosti:

- Bimodální rozdělení:
  - Vytvořte bodový graf a zkuste odhadnout, co by to mělo znamenat.
  - Vypočtěte průměrnou dobu čekání. Co nám říká?
- Vysoký korelační koeficient:
  - Vypočtěte jej a srovnejte s bodovým grafem.

Data o erupcích gejzíru Old Faithful v Yellowstonském národním parku.